DOCTOR OF PHILOSOPHY

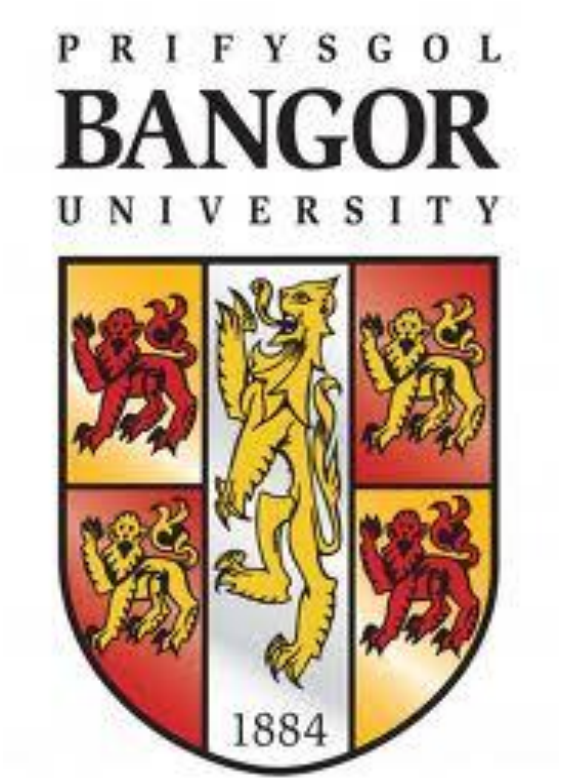
Defining marine meiofaunal genetic diversity using 454 sequencing

Garcia da Fonseca Batista, Vera

Award date:
2011

Awarding institution:
Bangor University

Prifysgol Bangor



Bangor University

Defining marine meiofaunal genetic diversity using 454 sequencing

A thesis submitted for the degree of Doctor of Philosophy in the School of Biological Sciences, Bangor University

by

Vera Alexandra Garcia da Fonseca Batista

October 2011

Summary

Biodiversity is the product of millions of years of evolution and forms the basis of earth's life support system, but the magnitude and relative diversity of global species richness remains unknown. On earth there may be over 100 million species but fewer than 2 million have been formally described. Coverage across different biological groups is very uneven with a known taxonomic deficit especially marked for microfauna and meiofauna due to problematic identification and hyper abundant representatives. Soft-bottom benthic meiofauna are ubiquitous, highly abundant organisms that play a crucial role in marine ecosystem functioning. Nevertheless, quantifying community structure using standard morphological approaches requires highly skilled taxonomists that are in short supply, and is very time consuming. The development of massively parallel sequencing has paved the way to explore microbial and meiofaunal diversity in time and space. Several studies have used pyrosequencing to assess the diversity of bacteria and archaea in the marine environment but there has been comparatively limited focus on eukaryotes. Moreover, diversity estimates derived from earlier second-generation sequencing studies are now known to be artificially inflated and skewed, due to several problems ranging from DNA manipulation and PCR amplifications to bioinformatic analyses. Here, I initially provide an overview of the emerging field of meiofaunal biodiversity assessment, using 454 Roche sequencing. The field differs substantially from environmental sequencing of prokaryotes and even some protists and so warrants separate attention. Empirically, meiobenthic richness of numerous phyla was estimated at alpha (local) and beta (European) scales, illustrating the extensive, but also spatial nature of meiofaunal richness and putative distribution patterns. Further to this, I sequenced carefully constructed artificial nematode control communities to assess the drivers (richness and genetic diversity) of DNA recombinant (or chimera) formation in environmental DNA PCR reactions. These advances provide a fast, objective and cost-effective way of accurately exploring and elucidating biodiversity in environmental samples.

Contents

List of Tables

List of Figures

List of Abbreviations

bp	Base pairs
18S, rRNA	18S ribosomal RNA
A+T	Adenine + Thymine
ARB	(*arbor* = tree) Software for environment sequence data
BLAST	Basic Local Alignment Search Tool
CO_2	Carbon dioxide
COI	Cytochrome oxidase subunit I
CTAB	Cetyltrimethylammonium Bromide
DESS	dimethyl sulphoxide, disodium EDTA and saturated NaCl
DMSO	dimethyl sulphoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
DOTUR	Distance-Based OTU and Richness
EDTA	Ethylenediaminetetraacetic Acid
emPCR	Emulsion PCR
ESPRIT	Estimating species Richness
hr, min, sec	Hour, minute, second
IPTG	Isopropyl-beta-D-thiogalactopyranoside
ITS	Internal transcribed spacer
k-mer	number of bases in DNA molecules
Kb	Kilo bases
kg, g, mg, µg, ng, pg	Kilogram, gram, milligram, microgram, nanogram, picogram
L, ml, µl	Litre, millilitre, microlitre
M, µM	Molar, micromolar
MgCl2	Magnesium chloride
MID	Molecular Identifier tags
MOTU	Molecular Operational Taxonomic Unit
MPS	Massively Parallel Sequencing
MUSCLE	Multiple Sequence Comparison by Log-Expectation
N, W	North, West
NaCl	Sodium Chloride
nSSU, nLSU	nuclear small (18S) and large (28S) subunit ribosomal RNA

OCTU	Operational Clustered Taxonomic Unit
OCTUPUS	Operationally Clustered Taxonomic Units for Parallel-tagged Ultra Sequencing
OTU	Operational Taxonomic Unit
PCR	Polymerase chain reaction
Pfu	Plaque forming unit
QC	Quality control
rDNA	Ribosomal DNA
RNA	Ribonucleic acid
SDS	Sodium dodecyl sulphate
SILVA	(*silva*=forest) Comprehensive ribosomal RNA database
Tris-HCl	Tris (hydroxymethyl) aminomethane-hydrochloric acid
UK	United Kingdom
UV	Ultraviolet
VAMPS	Visualization and Analysis of Microbial Population Structures

List of publications, posters and communications

Publications

VG Fonseca, B Nichols, D Lallias, GR Carvalho, DM Power, C Quince, and S Creer (2011). Sample richness and diversity as a driver of chimera formation in nSSU metagenetic data. (*submitted to Nucleic Acid Research*).

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

S Creer*, VG Fonseca*, DL Porazinska, RM Giblin-Davis, W Sung, DM Power, GR Carvalho, ML Blaxter, PJD Lambshead and WK Thomas (2010). Ultrasequencing of the meiofaunal biosphere: practice, promises and pitfalls. *Molecular Ecology* 19 (1): 4-20. (*joint first authors)

Posters

VG Fonseca, G Carvalho, W Sung, HF Johnson, DM Power, SP Neill, M Packer, ML Blaxter, PJD Lambshead, K Thomas and S Creer (2008). Second-generation environmental sequencing unmasks marine metazoan biodiversity. 2nd Conference of the European Consortium for the Barcode of Life, 2-4 June 2010 Braga, Portugal.

VG Fonseca, PJD Lambshead, T Ferrero, D Lunt, K Thomas, W Sung, DM Power, G Carvalho & S Creer (2008). Defining marine meiobenthic diversity using 454 ultrasequencing. World Conference on Marine Biodiversity, 11-15 November 2008, Valencia, Spain.

VG Fonseca, J Lambshead, D Lunt, K Thomas, DM Power, G Carvalho & S Creer (2008). Metagenetic Diversity and Environmental Genomics of Marine Meiofaunal Communities. PhD Conference Day. Bangor University, School of Biological Sciences. Brambell Building 16June. (Prize for best Poster).

Communications

VG Fonseca, GR Carvalho, W Sung, HF Johnson, DM Power, SP Neill, M Packer, ML Blaxter, PJD Lambshead, WK Thomas, and S Creer. Defining marine meiofaunal genetic diversity at fine geographic scales using deep sequencing techniques. July 17, 2009. Molecular Ecology and Fisheries Genetics Laboratory. School of Biological Sciences, Environment Centre Wales, Deiniol Road, College of Natural Sciences, Bangor University, Gwynedd LL57 2UW, UK.

VG Fonseca, GR Carvalho, W Sung, HF Johnson, DM Power, SP Neill, M Packer, ML Blaxter, PJD Lambshead, WK Thomas, and S Creer. Second-generation environmental sequencing unmasks marine metazoan biodiversity. February 12, 2010. Comparative Molecular Endocrinology Group, Centre for Marine Studies, University of Algarve, Gambelas, Portugal.

Acknowledgements

First, I would like to acknowledge the Portuguese Foundation for Science and Technology (FCT) for financing my PhD grant (SFRH/BD/27413/2006) that made possible the concretization of this thesis. I would also like to acknowledge Bangor University for accepting me as a PhD student. Support to this thesis by a NERC New Investigator Grant NE/E001505/1 and MGF Grant 167 to Simon Creer is acknowledged.

Simon Creer, my main supervisor, I am very grateful for all your help and for making sure I had all the conditions that made this work possible. Thank you for sharing your experience, always reassuring me that everything was going to be fine along this bumpy road called PhD! Thank you for all the guidance, feedback and mentoring during these four years. Thanks for your willingness to discuss every single thing and keen availability.

Gary Carvalho, my co-supervisor, thanks for the warm welcome in the MFEGL lab in Bangor. Thanks for encouraging me and for your guidance, through all steps of the thesis. Most of all thank you very much for sharing your knowledge, advice and your positive disposition that always made things look less impossible!

Deborah Power, my co-supervisor, thank you so much for always encouraging me to proceed. Thank you very much for welcoming me back to Portugal and making sure I had the conditions to continue the work. Thank you for trusting me, for the endless support, motivation and fruitful discussions. Thanks for always being so optimistic it is very motivating!

I would like to thank the members of my supervisory committee, Prof Roger Hughes and Dr Wolfgang Wuster for their guidance during the three years I have spent in Bangor.

During these four years I have learned a lot but most of all that scientific research is teamwork. This thesis represents the contribution of several persons, all my co-authors, that I had the privilege to meet and who shared their knowledge and made this thesis better and more interesting. Thanks Si, Gary, Debbie, Way, Chris, Ben, Delphine, Margaret, Mark, John, Kelley, Harriet and Kevin.

I am very grateful to Way Sung for all his efforts to develop and improve a pipeline that made possible the sorting of massive number of sequences (Chapter2)!

I would like to thank Chris Quince and Ben Nichols who welcomed me in Glasgow and made possible the sorting of the second round of massive

numbers of sequences (Chapter 4 and Chapter 5)!! Thank you both for your help with the data analysis and scientific discussions that substantially improved this thesis.

Thank you Delphine Lallias, my lab mate and friend for all your help, especially during this last year that I was in Portugal. You have made possible the DNA quantification of critical samples (Chapter 5).

I would also like to thank Alex Dudarenko, Lucy Price and Harriet Johnson for their help with the video captured nematodes (not shown in thesis) and sediment analysis (Chapter 3), while conducting their undergraduate studies.

Thank you Dorota Porazinska, for sharing some of your tropical rain forest data (Chapter 2)! Thank you very much for your help and skype scientific discussions.

Thank you Simeon Walker for solving any computer problem, for your prompt availability and willingness to help me.

Thank you very much Wendy Grail, "**diolch yn fawr iawn" for** all the help in the lab and advice. Thanks Jan Albin and Jane Smith for always being so helpful and available for any administrative matter.

Thank you very much Karen Goodwin and Gillian Clayton for always being so kind and for the immense help to sort out any administrative matter.

Thank you also Kevin Ashelford and Margaret Hughes for making sure sequencing was performed on time and for you kind welcome in Liverpool.

Special and huge thanks to my friends Delphine and Sonia for the three excellent years spent in Bangor! Thank you Delphine it's been a genuine pleasure getting to know you better and I hope we stay friends for many years to come, despite the distance. Thank you Sonia for making me feel at home with the long Portuguese chats and for sharing Portuguese treats. I am immensely grateful to both of you for all your help and support but most of all for your friendship. I miss lunch times, BlueSky, the weekends and the long conversations of about everything, right? I miss you guys, you've been amazing and I will always remember you.

I would like to thank everyone from the MEFGL group that made everything go by so smoothly with a great scientific and friendly atmosphere. Thank you for all the outdoors activities, barbecues on the beach, Xmas dinners, enthusiastic scientific meetings but mainly for welcoming and helping me in the lab. Thanks Si, Gary, Delphine, Sonia, Wendy, Jan, Jane, Joana, Carla, Nicklas, Lindsay, Axel, Markos, Sarah, Wol, Serinde, Andrew, Jenny, Mark dB, Martin, Marj, Mark C, Filipe, Aitor, Roger H, Wolfgang, Cathy, Safi and Darren.

I would also like to thank my friends in the lab in Portugal for their support and friendship. Thank you Joao, it's been great working with you and thanks for all the help during this final stage! Thank you Bela for the support, friendship and decaff coffee-breaks to release the pressure! Thanks to Alexandra, Elsa, Maria, Ricardo, Patricia, Liliana, Pedro and Rita T. for all the support and cheering up.

A special thank to my neighbours in Felinheli Sylvie, Emyr, Elen and Lois for making my stay in Wales so pleasant. Thank you for the warm welcome and for always being so thoughtful.

I would like to thank my friends Matt and Delphine who always welcomed me, and my family in their home, as if it was ours. Thank you for the lovely time and for showing us the beautiful Welsh gardens. Thank you for having me at your home in Cowny when I needed and to make sure I had a nice stay.... Despite the Volcano ash!!!

Thanks to my long, long journey friends Vera and Carla who know me better than anyone and despite the distance I know I can always count on them. Thank you for all the support.

I am very grateful to Irineu, my father-in-law, for always showing concern and giving me advice through all these years. Thank you very much for your friendship and immeasurable help.

Thank you Alexandre, my dear friend and brother-in-law, for sharing your great sense of humour, your attention and most of all for all your help when we needed.

A special thanks to my parents who made literally everything possible! Thank you for your endless help, strength, love and support. I love you.

Last but most important, thank you Fred for everything! For your strength, persistence, patience and love through all these years. Thank you for your immeasurable support during this final stage of the thesis. Thank you and Daniel for being so special and making my life more meaningful and happy.

CHAPTER 1 - General Introduction

Classification of biodiversity

Biodiversity or biological diversity is a term used to describe the numerous life forms found on Earth. One of the characteristics of the living world is its complex structure and hierarchy and the need of classification as a requirement to understand the diversity of living organisms and their biological and evolutionary relationships. Ultimately classification provides the unifying scaffold upon which biological knowledge is assembled. Researchers in many areas of biology use classification (taxonomy) as a structure for comparative genetics/genomics and also as the basis for drawing broad conclusions about the diversity of living organisms (Parfrey *et al.* 2006).

The diversity of life forms have been classified into four to six primary "kingdoms", mainly assembling organisms that share common features and evolutionary traits into taxonomic groups. The most significant system is the Whittaker five kingdom structure, recognizing Monera (prokaryotes) and four eukaryotic kingdoms: Animalia (Metazoa), Plantae, Fungi and Protista (Simpson & Roger 2004). Nowadays the existence of six-kingdoms of life it is widely accepted and includes the Bacteria (prokaryotes) and five eukaryotic kingdoms: the Protozoa, Animalia, Fungi, Plantae and Chromista (Cavalier-Smith 2004, 2010). Perspectives on the classification of the eukaryotic diversity have also become a daunting task in recent years. Eukaryotes (organisms containing nuclei) are highly distinct but also a highly variable group (Baldauf 2003) that encompass an incredible morphological diversity (Parfrey *et al.* 2006) and are very complex at the cellular and genetic level. Nonetheless phylogenomic analysis has helped to clarify the evolutionary links between major groups of eukaryotes allowing the delineation of five (Koonin 2010) or six supergroups (Roger & Simpson 2009). The main eukaryotic groups proposed are the Opisthokonta, Amoebozoa, Plantae, Chromalveolata, Rhizaria and Excavata (Figure 1.1). In brief the Opisthokonts contain animals, true fungi and some unicellular groups, including the free-living choanoflagellates and also some free-living amoeba (Simpson & Roger 2004). The Amoebozoa include a diversity of predominantly amoeboid members that are mainly free-living. The Chromalveolata comprise unicellular and multicellular forms with four major groups of eukaryotic algae: dinoflagellates, cryptophytes, haptophytes and stramenopiles but also include

non-photosynthetic forms (Simpson & Roger 2004; Parfrey *et al.* 2006). The Rhizaria unites a heterogeneous group of free-living unicellular flagellates and amoebae, namely the Foraminifera and Radiolaria (Burki *et al.* 2010; Cavalier-Smith 2010) and the Excavata are unicellular eukaryotes, most of which are heterotrophic flagellates (Simpson & Roger 2004).

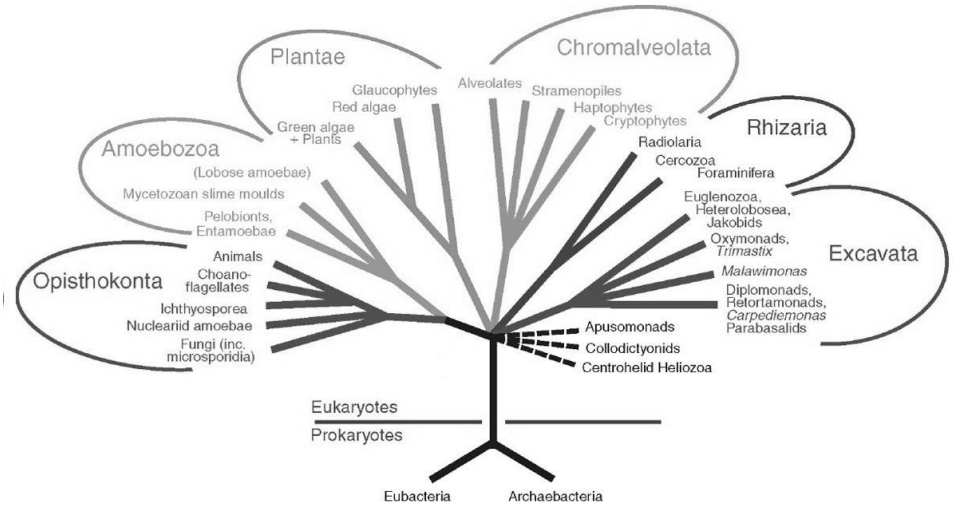


Figure 1.1- A diagrammatic tree depicting the organization of eukaryotes into six major groups (from (Simpson & Roger 2004)).

The ultimate goal in taxonomy and biology in general is to assess species identification so that almost all studies in biology, whether at the level of molecules, cells, individuals or populations, are typically referenced to the level of the species. The process of allocating individuals to a given species clearly depends on the criteria by which species are defined and delimited (Balakrishnan 2005), which are in turn connected by the concept of what is a species. Although no concept is universally accepted, one of the most documented and applied is the 'biological species concept' (Mayr 1963; Avise & Wollenberg 1997; Hendry 2009), where species are groups of actually or potentially interbreeding individuals that are reproductively isolated (Velasco 2008). This concept is sufficiently ambiguous to allow for some genetic exchange between species, since some estimates hold that 25% of all plant species and 10% of all animal species hybridize successfully with at

least one other species, (Velasco 2008; Hendry 2009). The "biological species concept" does not take in to account groups that are geographically separated (allopatric) and it is hard to predict if these groups would collapse into a single species if they did become sympatric. In addition, it does not englobe organisms without sexual reproduction, such as viruses and most eukaryotic microbes. For many reasons the species concept continues to be the subject of much debate (Jensen 2010) and unconventional species concepts revolve around the magnitude of morphological or genetic differences between groups (Hendry 2009). The difficulty of a species concept is in deciding just how big a difference is to delimit or identify a species within a particular group.

Identifying a certain group or level by taxonomic classification can be supported by molecular/ biochemical data and by morphological characteristics (Avise & Wollenberg 1997; Blaxter 2003b; Cavalier-Smith 2006) which can be used solely or in combination. Establishing whether a certain morphologically or genetically distinguishable group of organisms represents a species has proved to be a challenging task (Avise & Wollenberg 1997; De Queiroz 2007; Vogler & Monaghan 2007; Jensen 2010). Difficulties vary from intraspecific genetic difference from the level of a few nucleotides to whole chromosomes (Parfrey *et al.* 2006), particularly conservative genes such as those coding for rRNA (Fenchel & Finlay 2006; Parfrey *et al.* 2006) and also accuracy of molecular *versus* nuclear markers (Shaw 2002) to the existence of "cryptic" diversity (Blouin 2002; Bickford *et al.* 2007).

Components of biodiversity

Biodiversity covers a range of different levels of organisation from the genetic variation between individuals and populations, to species diversity, assemblages, habitats, landscapes and biogeographical provinces (Gray 1997; Sala *et al.* 2000; Loreau *et al.* 2001; Hooper *et al.* 2005b). There are, in general, three main categories of biodiversity: genetic diversity, species diversity and habitat diversity (Gray 2000; Feral 2002).

Genetic diversity characterizes the amount of genetic information within and among individuals of a population, a species, an assemblage, or a community (Feral 2002). It is extremely important because it represents a requisite for evolutionary

adaptation to a changing environment (Gray 1997; Sala & Knowlton 2006). To understand many ecological and evolutionary processes, it is crucial to document genetic diversity of species, populations, or individuals of interest (Müller *et al.* 2003). In addition, it is necessary to understand the mechanisms for creating and maintaining the observed patterns of diversity. Thus, studies of genetic diversity have the potential to provide insights into many fields including conservation biology, population and community ecology, and evolutionary biology (Feral 2002; Hooper *et al.* 2005b). On the other hand, species diversity is the most commonly used synonym for biodiversity, and it mainly represents the species richness/ number of a species in an area (Whittaker 1972; Gray 1997; Hooper *et al.* 2005b). Important conceptual components of species diversity include richness, evenness, dominance, and rarity of species (Wilsey *et al.* 2005). These components are characterized by the way in which the presence of each species is weighted by factors such as abundance or biomass (Magurran 1998; Magurran 2004). With species richness, each species contributes to diversity in the same manner regardless of its abundance or biomass. Some measures, such as species evenness, weight each species by its relative abundance or biomass. Species diversity indices (e.g., Simpson's 1/D or Shannon's H) represent composite measures, and are sometimes calculated so that richness and evenness are mathematically independent (Smith & Wilson 1996). Other measures centre on a restricted subset of species. Species dominance (e.g., Berger-Parker index) is the relative importance of the one species contributing the most to total abundance or biomass. In contrast, species rarity is a measure of the proportion of species that meet the restriction that their relative abundance or biomass is below some threshold (Wilsey *et al.* 2005; Chao *et al.* 2006). The number of species alone may not be the best predictor of ecosystem properties and biodiversity (Purvis & Hector 2000; Hooper *et al.* 2002) because the number of individuals per species varies and complete inventories of all the species in the world is practically impossible. Finally, habitat diversity represents the variation in the collection of assemblages, communities, and habitats within a region (Whittaker 1972).

Although these are the three main components of biodiversity, there are others such as taxonomic and functional diversity (Norse 1993). Taxonomic or phyletic diversity is highest in the sea (Gray 1997; Bouchet 2006) and it involves the variation and variability of phyla of organisms (Brunel 2006). For example, in the

marine environment, considered to have high phyletic diversity, 32 out of the 33 animal phyla are present (Norse 1993). Functional diversity is the range of functions that are performed by organisms in a system (Gray 1997). Functional diversity is usually associated with phenotypic diversity since phenotypic diversity between individuals, populations, and species is usually described in terms of the variation in external morphology or traits of individuals (Hooper *et al.* 2002; Hooper *et al.* 2005b; Carvalho *et al.* 2010). Thus functional diversity, which is also tightly associated with genetic diversity, represents an important measure of the adaptation of the organism to its environment because it interacts with biotic and abiotic factors of the environment. To understand how changes in biodiversity influence species loss, and ultimately ecosystem properties, it is fundamental that the functional traits of a species be recognized (Chapin *et al.* 2000; Sala *et al.* 2000; Hooper *et al.* 2005b; Sala & Knowlton 2006).

Biodiversity is very dynamic in nature as populations are constantly evolving and all of its constituent components overlap and interconnect in a complex set of relationships (Feral 2002). Biodiversity results from combined effects of speciation, extinction and climatic changes at all levels (Sala *et al.* 2000; Loreau *et al.* 2001; Hooper *et al.* 2005b; Sala & Knowlton 2006). Because biodiversity loss causes alteration or destruction of habitats (Loreau *et al.* 2001; Hooper *et al.* 2005b) and ecosystem processes (Naeem & Wright 2003; Covich *et al.* 2004; Loreau *et al.* 2006) the study and understanding of biodiversity is of major importance in understanding biotic/abiotic interactions in ecosystem function.

Measuring biodiversity

There is an increasing interest in quantifying the patterns in species diversity by comparing components of diversity that occur within (alpha-diversity) and among samples (beta-diversity) at multiple sampling scales (Ugland *et al.* 2005; Thrush *et al.* 2010). Within-habitat diversity or alpha-diversity represents a group of organisms that interact and compete for the same resources or share the same environment and it represents species richness within a given area. Between-habitat diversity or beta-diversity is the degree of change in species composition between locations or communities (Gray 2000). High beta-diversity implies low similarity between

species composition of different habitats. It is usually expressed in terms of a similarity index between communities (or species turnover rate) between different habitats in the same geographical area (Gray 2000; Ricotta 2010). The interplay of the mean within-group diversity and between-group diversity is named geographical diversity or gamma-diversity (Gray 2000). Nonetheless, the complexity of scales and units makes it impossible to assess community biodiversity using a single measure (Sala & Knowlton 2006).

The measure of biological diversity most commonly used in ecology is species richness (number of species) (Magurran 1998; Magurran 2004), as it not only allows basic comparisons among sites but is also a fundamental measure of community status (Gotelli & Colwell 2001). However, all biologists who sample natural communities are overwhelmed with the problem of how well a sample reflects a community's true diversity (Hughes & Hellmann 2005). This is mainly because observed species richness depends strongly on sample size; thus it is not reasonable to directly compare observed species richness in samples of unequal size. The techniques of species accumulation curves (Colwell & Coddington 1994) and rarefaction (Sanders 1968), plotting species richness versus sample size, are often used to compare observed species richness in two communities at a common sample size (Hughes & Hellmann 2005).

Species–accumulation curves are used to evaluate the effectiveness of sampling or to compare species richness among habitats using rarefaction (Gotelli & Colwell 2001; Magurran 2004; Shaw *et al.* 2008). This approach does not need to be area-based because rarefaction explicitly controls for differences in the numbers of individuals among samples (Gotelli & Colwell 2001; Hughes *et al.* 2001; Hughes & Hellmann 2005). Individual-based rarefaction provides the expected numbers of species with increasing numbers of individuals sampled, assuming a random sample of individuals in the community (Crist & Veech 2006). Sample-based rarefaction describes the average number of accumulated species as the number of samples increases (Gotelli & Colwell 2001; Gotelli *et al.* 2010). Such analysis provides an important step in quantifying the contributions of different habitats to broad-scale patterns of species richness or temporal patterns of richness within habitats (Crist & Veech 2006). A crude indication of species richness are the rarefaction curves, where cumulative species richness represent a function of the numbers of individuals sampled, assuming that the ranking on sample diversity will not change with further

sampling (Hughes & Hellmann 2005). Rarefaction curves are often constructed from samples taken within habitats to determine the efficacy of sampling the true species richness of a given habitat, or to compare species richness among habitats on an equal-effort basis (Gotelli & Colwell 2001). Rarefaction accounts for the fact that large samples have more species (or any taxonomic unit) than small samples even if they are drawn from the same community (Hughes & Hellmann 2005). As ecological samples are not usually large enough to include more than half the potential species, the rest of the distribution is hidden in the non-observed species (Wilsey *et al.* 2005). Although rare species are arguably the most important components of diversity, their abundances are inevitably poorly typified (Chao *et al.* 2009). The difficulty becomes even more acute as species richness increases. The more speciose a sample or taxa is the harder it will be to access true levels of diversity. Nonetheless, rarefaction methods either sample-based or individual-based allow a meaningful standardization or sampling effort and permits comparison of datasets (Hughes *et al.* 2001).

Patterns in biodiversity

Biological diversity is not evenly distributed over the surface of the planet and several global patterns of spatial variation in biodiversity have been explored. These patterns can be different in hotspots and coldspots (highs and lows) of diversity, change with spatial scale (e.g. species–area relationships and relationships between local and regional richness) and along gradients across space or environmental conditions (e.g. latitude, longitude, altitude, depth, peninsulas, bays, isolation, productivity/energy and aridity) (Sala *et al.* 2000; Gaston 2007, 2009). The most famous large-scale biodiversity pattern in ecology is the latitudinal gradient (Lambshead *et al.* 2000) where there is a tendency of lower latitudes to have more species than higher latitudes (Fuhrman 2009; Gaston 2009). There is probably higher productivity in lower latitudes providing more resources that can generate more niches but also higher temperatures in low latitudes increase the metabolic rate and make biological processes, including speciation, occur faster (Fuhrman 2009). Nevertheless, no consensus has been reached about the mechanisms behind this phenomenon (Willig *et al.* 2003; Hillebrand 2004).

There are at least two rival interpretations to explain biodiversity patterns and community assembly. The first is that there is limited dispersal of species combined with unrestricted entry into communities (dispersal assembly models), such theory was proposed by Hubbell (2001) as "the unified neutral theory of biodiversity", where *neutrality* assumes that all individuals in a community are strictly equivalent regarding their prospects of reproduction and death (Chave 2004). The second interpretation is that species' dispersal is combined with environmental filtering (niche-assembly models) (Leibold *et al.* 2004). Dispersal assembly models predict a progressive decay of community similarity in space and through time, reflecting the effect of dispersal limitation and the stochastic replacement of individuals from the community. In contrast, niche-assembly models are expected to have more predictable community composition among sites and/or time periods characterized by similar environmental conditions (Chave 2004). In between these hypotheses Lourens Baas-Becking (Baas-Becking 1934) hypothesised that "everything is everywhere but the environment selects" (Finlay 2002). The 'everything is everywhere' hypothesis implies a lack of biogeographic patterns and provides evidence of the high dispersal potential of microorganisms (Finlay 2002; Foissner 2006; Martiny *et al.* 2006; Ramette & Tiedje 2007; Fuhrman 2009; Cermeno *et al.* 2010; Schauer *et al.* 2010). This theory supports the notion that current distributions of organisms are the result of ecological and historical factors, including dispersion by wind, water and animals, and adaptations to local conditions that change over space and time (Fuhrman 2009). However the ubiquity model might be bias because of under-sampling and misidentification of the samples (Mitchell & Meisterfeld 2005). Furthermore, currently there are several microorganims with known restricted distributions, such as protists (Foissner 2006; Foissner 2007) and flagellates (Bass *et al.* 2007) and also amoeba (Smith & Wilkinson 2007).

Marine biodiversity: meiofauna

The term meiofauna was introduced by Mare (1942) to describe benthic metazoan fauna 'of intermediate size'. Twenty-five years ago, scientists believed that the ca. 1.6 million species described represented approximately 50% of plant and animal species on the planet; his estimate has now increased to 1.7-1.8 million described

species with estimates of between 10-100 million species probably still remaining to be discovered (Bouchet 2006). In the marine environment, the total number of marine species is still unknown, and estimates range from 500,000 (May 1992; Gray 1996), 5 million (Poore & Wilson 1993), to more than 10 million species (Lambshead 1993; Gray 1997). These numbers are biased as there is a serious deficit in taxonomic expertise (Buyck 1999; Tautz *et al.* 2003) so that many highly diverse groups of organisms are disregarded (Fig. 1.2). Also, the fact that many taxa considered to be the same may actually be different (Knowlton 1986, 1993; Hebert *et al.* 2004) may lead to a 10-fold underestimation of marine biodiversity (Sala & Knowlton 2006). In addition, many studies rely only on macrofauna data and do not take into consideration meiofaunal taxa (Hebert *et al.* 2003b) and this presumably significantly biases the number of described marine species. Studies on meiofauna suggest that these organisms are of pivotal ecological importance and comprise the major part of ocean biodiversity (Rysgaard *et al.* 2000; Vanaverbeke *et al.* 2000; Gheskiere *et al.* 2005b; Rundell & Leander 2010).

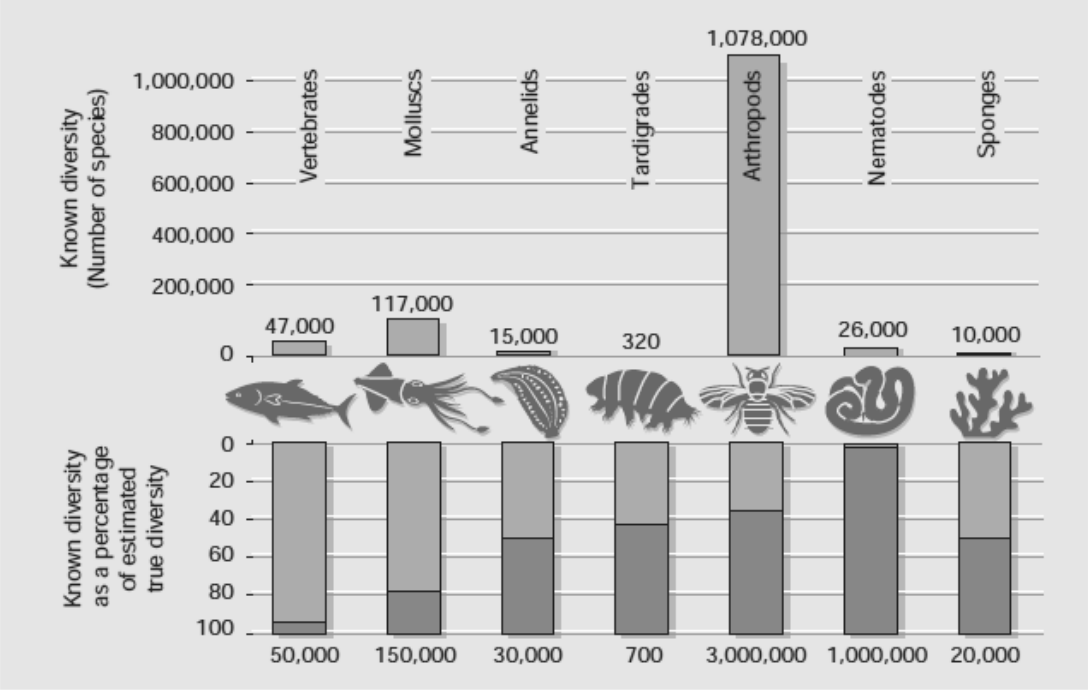


Figure 1.2- Known and estimated diversity of a selection of animal phyla (from Blaxter (2003). Known species numbers for larger taxa, like vertebrates and some molluscs are closer to reality than for smaller taxa, like bacteria, arthropods and nematodes where estimates are far from reality.

Most quantitative studies on meiofauna communities have established that the minimum size of meiofauna retained on the sieves used to process samples, is between 63 μm to 500 μm (Gheskiere *et al.* 2006; Rundell & Leander 2010). Meiofaunal abundance per unit area is much higher than for macrofauna, leading to higher meiofauna richness within a particular patch (Snelgrove & Smith 2002) though this is not very informative on a broader scale.

Nematodes are often the most abundant metazoans in the soft sediments of estuaries and intertidal sediments (Rysgaard *et al.* 2000). They are characterised by direct benthic development, with the whole of their life-cycle closely coupled to the sediment and with no specific dispersal phase (Nybakken 1997). In estuarine sediments, meiofaunal organisms facilitate biomineralization of organic matter and augment nutrient regeneration (Aller & Aller 1992; Berg *et al.* 1998; Rysgaard *et al.* 2000) and can also be a major dietary component of commercial species (Carlson *et al.* 1997). Meiofauna exhibit high sensitivity to anthropogenic inputs, making them excellent indicators of estuarine pollution (Kennedy & Jacoby 1999; Mirto & Danovaro 2004; Gheskiere *et al.* 2005b). Distribution and abundance of the meiofauna are mainly controlled by physio-chemical factors such as grain size, redox potential and tidal exposure (Nybakken 1997; Gheskiere *et al.* 2005a). Nonetheless, the presence of the biogenic structure (e.g., animal tubes or burrow, system root) is also known to affect meiofaunal distribution and abundance (Bell *et al.* 1978).

The most abundant and ubiquitous phylum of the meiofauna on earth (Lambshead 2004) and in nearshore marine and estuarine waters is the Nematoda (Vincx *et al.* 1994; Lambshead & Boucher 2003; Urban-Malinga *et al.* 2006). Nematodes are so cosmopolitan that recently were also found to inhabit hypoxic conditions in a deep mine in South Africa (Borgonie *et al.* 2011) and are also known to inhabit Antartic regions (Andrassy & Gibson 2007; Barnes *et al.* 2008) . They often outnumber those of all other meiofaunal taxa collectively, as they constitute 80-90% of metazoans on average (Lambshead 1993; Soltwedel 2000). Because Nematoda represent the most abundant and rich phyla in the meiobenthos brief ecological aspects will be addressed.

Nematode biology, biodiversity and ecological importance

Nematodes, or roundworms, are an ancient and diverse group of organisms (Platt *et al.* 1984). Although many taxonomic discrepancies are found within the phylum Nematoda (Meldal *et al.* 2007), small subunit ribosomal DNA sequences support three major clades: Chromadoria, Enoplia and Dorylaimia (Blaxter *et al.* 1998; De Ley & Blaxter 2004). The first two include various groups of marine, estuarine and freshwater nematodes, while the Dorylaimia are mainly fresh water (Abebe *et al.* 2008). Nematodes are unsegmented pseudocoelomates, generally translucent and with a simple body plan (Heip *et al.* 1985; Hajibabaei *et al.* 2005; Abebe *et al.* 2008). The body is essentially an elongated cylinder consisting of a body wall with cuticle, epidermis, somatic musculature, digestive and nervous system (Figure 1.3) (Heip *et al.* 1985; Hajibabaei *et al.* 2005). Development normally includes an egg stage and three to four juvenile stages with a moult at the end of each stage. The generation time is species dependent and varies from a few days to a year or more (Abebe *et al.* 2008).

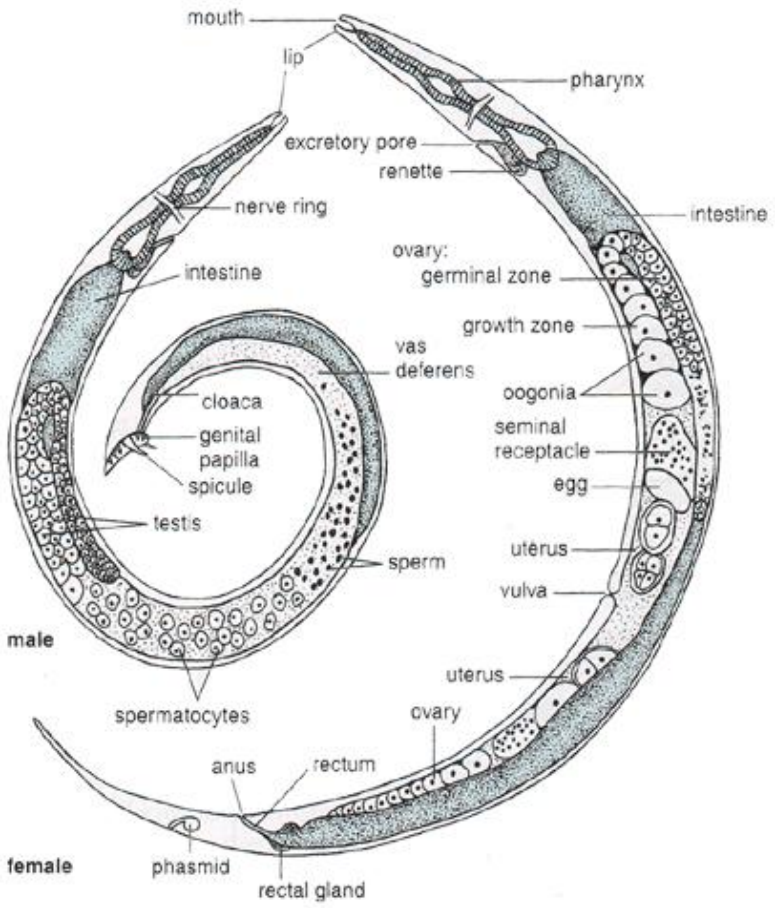


Figure 1.3- General features of male and female nematode (from http://classes.seattleu.edu/biology/biol235/hodin/nematodePriapulidGroup/nematodes/formAndFunction.htm).

Nematodes exploit multiple ecological niches and include free-living terrestrial and marine microbivores, meiofaunal predators, herbivores, and animal and plant parasites (Heip *et al.* 1985; Coghlan 2005). Parasitic nematodes are important pathogens of humans, infecting several thousand people, and causing damage to domesticated animals and crops (Heip *et al.* 1985; Mitreva *et al.* 2005). Whereas the importance of parasitic nematodes has been recognized for decades, the same is not true for free-living species, especially those inhabiting freshwater and marine environments. Free-living nematodes remain relatively unstudied, despite their abundance (millions per m^2) in soils and sediments and their occurrence in a broad range of habitats (Heip *et al.* 1985). Although free-living nematodes are also found in terrestrial habitats (Freckman *et al.* 1997) the majority are marine species

(Lambshead & Boucher 2003). They inhabit littoral (Heip *et al.* 1985; Lambshead 1986), freshwater (Andrassy & Gibson 2007) and abyssal sediments (Copley *et al.* 2007), where they frequently dominate the meiofauna size fraction.

The density, diversity and/or composition of nematode assemblages has been related to differences in sediment composition, hydrodynamic conditions, salinity, organic content and food resource availability (Heip *et al.* 1985; Soetaert *et al.* 1994; Soetaert *et al.* 1995; Li *et al.* 1997; Moens *et al.* 1999b; Steyaert *et al.* 1999; Tita *et al.* 2002; Somerfield *et al.* 2003; Steyaert *et al.* 2003). Nematode abundance also seems to decline with increasing water depth (Soltwedel 2000; Vanreusel *et al.* 2000) and distance from continents (Cook *et al.* 2000). Hence, the highest nematode abundance tends to be found in rich lowlands, marshes and marine mud around the coastline except for some tropical areas (Alongi 1987; Boucher & Clavier 1990).

Global marine nematode species richness may exceed 1 million (Lambshead 2004) but only a few thousand are described (De Ley *et al.* 2005). Since marine nematodes are so diverse and abundant they are believed to be of major ecological importance playing an important role in decomposition processes (Snelgrove *et al.* 1997; Austen 2004) and recycling of nutrients (Austen 2004). They have also proved to be highly sensitive indicators of anthropogenic stress in a range of situations (Lambshead 1986; Lambshead & Paterson 1986; Schratzberger *et al.* 2000; Austen *et al.* 2002; Steyaert *et al.* 2007). The inherent ecological importance of the Nematoda, high species diversity, abundance and ubiquity, coupled with high reproductive rates with no dispersal phase (Lambshead & Paterson 1986; Boucher & Lambshead 1995), a short life span and a sedentary life inhabiting mostly sediments, make nematodes ideal candidates for bio-monitoring studies (Platt *et al.* 1984; Schratzberger *et al.* 2000; Gheskiere *et al.* 2005a).

Traditional nematode taxonomy is supported by morphological traits only. These include differences in head (size, shape and orientation of sensilla, feeding apparatus and sensory amphids) and tail structure but also in the reproductive system shape (Heip *et al.* 1985). Some of the morphological differences in nematodes can be associated with different feeding mechanisms (predators, omnivores, deposit feeders, epigrowth feeders) (Moens *et al.* 1999b; Macas *et al.* 2007). Despite their apparent similarity and simple morphology nematodes occupy very different roles and trophic positions in sediments (Heip *et al.* 1985) making the identification of functional guilds ecologically very important and informative (Yeates & Bongers 1999;

Bremner *et al.* 2006). Moreover, studies on macrobenthic invertebrates have shown that combining functional and taxonomic diversity can reveal different relationships between assemblages (Bremner *et al.* 2003; Bostrom *et al.* 2006; Bremner *et al.* 2006). Nevertheless, studying the ecological importance of nematode worms is limited mostly because identification at the species level is problematic in many cases (Snelgrove *et al.* 1997; Lambshead & Boucher 2003). This is mainly due to the subjectivity of individual taxonomists, impasse in nomenclature and the existence of cryptic species (Cook *et al.* 2005; De Ley *et al.* 2005). In addition, the massive number of nematodes to be identified in a single sample, preparing slides for microscope observation, the need for high resolution microscope and the fact that most species can only be identified from adult characters (Litvaitis *et al.* 1994 ; Warwick & Clarke 1998; Floyd *et al.* 2002; Powers 2004; De Ley *et al.* 2005) makes identification of nematodes using external morphology problematic and very time consuming.

Molecular identification strategies

Recently there has been an escalation in the range of DNA-based technologies used in biodiversity studies, by establishing methodologies for species identification either through DNA barcoding (Hebert *et al.* 2004; Lambert *et al.* 2005; Ward *et al.* 2005; Hajibabaei *et al.* 2006; Pfenninger *et al.* 2007), DNA arrays or phylochips (Loy *et al.* 2002; Pfunder *et al.* 2004; Garaizar *et al.* 2006; Peplies *et al.* 2006; Tobler *et al.* 2006; Allen *et al.* 2007), and/ or through cutting edge sequencing techniques like Massively Parallel Sequencing (MPS) (Sogin *et al.* 2006; Huber *et al.* 2007; Porazinska *et al.* 2009; Medinger *et al.* 2010). Such principles were built on much earlier applications of molecular methods to describe and investigate taxonomic diversity and relationships (Ferguson 1980; Hillis & Moritz 1990) providing the necessary technological and conceptual platform to advance biodiversity assessment.

A DNA-based taxonomy system should be straightforward. It should encompass sample collection and sample character (e.g. tissue from a given individual); a validated DNA extraction method; PCR amplification of one or more regions of a chosen gene and sequencing of the target region. The resulting sequences should be made available through a public database, linked to a species description, including a

definition of its taxonomic status (Tautz *et al.* 2003; Wilson *et al.* 2005). The aim of such a system of classification is to establish defined molecular operational taxonomic units (MOTU; Floyd *et al.* 2002; Blaxter *et al.* 2005) on the basis of sequence differences in short, orthologous marker gene sequences (Tautz *et al.* 2003; Blaxter *et al.* 2004). Floyd *et al.* (2002) initially defined a MOTU as a group of sequences that differed from one another by three or four nucleotides in a 500-bp gene region. Although it is unclear how well MOTU diversity corresponds to species or ecological diversity (Vogler & Monaghan 2007; Abebe *et al.* 2011), it certainly has an important role in evaluating genetic diversity within defined taxa and/or communities. This follows the general definition of operational taxonomic units (OTU) as groups of organisms used in a taxonomic study without designation of taxonomic rank (Floyd *et al.* 2002). A variety of molecular markers have been used to identify or delimit species or species-like units (Shinn *et al.* 2000; Bucklin *et al.* 2007; Bellemain *et al.* 2010; Nassonova *et al.* 2010). The most widely used markers to study environmental samples are the nSSU and nLSU genes (nuclear small subunit and large subunit ribosomal of 18S rDNA) (De Ley *et al.* 2005; Bhadury *et al.* 2006), mainly because of their highly divergent regions flanked by conserved regions (Figure 1.4).

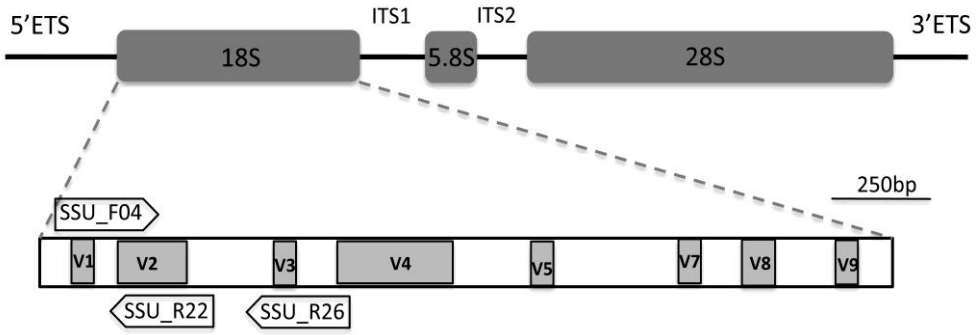


Figure 1.4- Diagram of the ribosomal DNA (rDNA) gene and location of the primers used in this study (SSU_F04; SSU_R22; SSU_R26) within the small subunit gene (SSU or 18S). The internal (ITS1 and ITS2) and external (5' ETS and 3' ETS) transcribed spacers, the 5.8S gene as well as the large subunit gene (LSU or 28S) are also illustrated. Within the 18S the V1–V9 boxes denote the variable regions of the 18S rDNA subunit (V6 exists in prokaryotes only). Position of the primers and sizes are approximate.

Conservation within sequences facilitates alignment and primer design for different OTUs while divergent regions are suitable for species discrimination (Floyd *et al.* 2002). In addition, the nSSU genes are multicopy, making them relatively easy to amplify (Abebe *et al.* 2011). In contrast, molecular markers like the internal transcribed spacer (ITS) and cytochrome c oxidase subunit 1 (COI) are not very commonly used in meiobenthic DNA-based studies. This is predominantly because in nematodes ITS regions are very difficult to align (Floyd *et al.* 2002) and the COI gene is characterised by unusual molecular evolutionary rates and processes. Nevertheless, COI combined with other markers could be used at the species levels to differentiate cryptic species (Blouin *et al.* 1998; Smith *et al.* 2006; Derycke *et al.* 2007).

Next Generation Sequencing

The automated Sanger method is considered as a "first-generation" technology, and newer methods are referred to as next-generation sequencing (NGS) (Metzker 2010). NGS is being developed for single individuals but also for whole-populations (metagenomics) (Noonan *et al.* 2006; Bohannon 2007; Frias-Lopez *et al.* 2008; Lazarevic *et al.* 2009; Nowrousian 2010; Quaiser *et al.* 2011; Unterseher *et al.* 2011) revolutionizing the study of ecology and evolution. Metagenomics and metagenetics

both apply to analysis of a multi- genome unit, or community, where the latter is widely applied by the use of homologous genes whereas metagenomics represents a more functional approach usually encompassing different genes (Hugenholtz & Tyson 2008; Mou et al. 2008). With the advances of next generation sequencing a metagenetics approach might help to dissect ecological questions (Porazinska *et al.* 2010) at a new level of precision probably advancing the entire field of ecology by answering questions not possible before in macroecological studies.

Massively Parallel Pyrosequencing (MPS) is one next generation sequencing method (Hudson 2008) developed by 454 Life Sciences that has dramatically reduced the time and cost constraints of DNA sequencing (Margulies *et al.* 2005). It uses pyrophosphate release as a method for detection of base incorporation, and is capable of sequencing hundreds of thousands of DNA molecules in parallel on a picotitre plate, also named as pyrosequences (Hall 2007; Meyer *et al.* 2007; Meyer *et al.* 2008b). Furthermore, it not only produces large amounts of data at a low cost, but also allows sequencing of environmental DNA without a prior cloning step (Ronaghi *et al.* 1998; Ronaghi 2001; Edwards *et al.* 2006; Rabouille *et al.* 2006; Turnbaugh *et al.* 2006; Meyer *et al.* 2007). It is now possible to identify multiple discrete samples on a single sequencing plate using sequence information (Binladen *et al.* 2007; Meyer *et al.* 2007) and the full power of this novel sequencing technology is and can be applied to numerous, diverse samples simultaneously. The first groundbreaking publication detailing this method described the use of MPS for shotgun sequencing and *de novo* assembly of a bacterial genome (Margulies *et al.* 2005). The approach is now routinely used for metagenetic analyses to characterize natural assemblages of microorganisms, mostly prokaryotes, from a variety of ecosystems (Sogin *et al.* 2006; Huber *et al.* 2007; Stoeck *et al.* 2010; Huse *et al.* 2008; Amend *et al.* 2010; Roossinck *et al.* 2010; Youssef *et al.* 2010; Orsi *et al.* 2011; Pawlowski *et al.* 2011; Unterseher *et al.* 2011), shotgun transcriptome analyses (Hughes & Vogler 2006; Gracey 2007; Weber *et al.* 2007; Coppe *et al.* 2010; Blanca *et al.* 2011) and a range of other applications.

In the last decade, the use of next-generation sequencing has allowed a much deeper sampling of environmental biodiversity by producing many orders of magnitude more sequence information than the Sanger-sequencing approach (Haas *et al.* 2011), even so an accurate assessment of biodiversity is central to any biological study. The identification of taxa present in a given community is likely to be crucial

for understanding what that community does; nonetheless different methodological approaches can result in discrepant estimates from the same sample (Quince *et al.* 2009). The unveiling of highly diverse and uncharacterized environmental samples has hypothesized the existence of a the "rare biosphere" (Sogin *et al.* 2006). Nonetheless, this has proved to be a controversial matter because pyrosequencing generates a substantial amount of sequences with intrinsic error rates that could lead to overestimation of the number of rare OTUs (Kunin *et al.* 2009; Quince *et al.* 2011). Currently the method widely used to measure biodiversity in environmental samples is to PCR amplify the region of interest followed by pyrosequencing (Huse *et al.* 2008; Haas *et al.* 2011). Nonetheless, amplification from a multi-template population can be problematic since artificial chimeras can be generated (Gonzalez *et al.* 2005; Smyth *et al.* 2010). Chimeras between two different DNA molecules with high sequence similarity (homologue genes) take place during PCR when incomplete extension occurs in one round of PCR and thus the resulting fragment acts as a primer for a different sequence in the next step (Wang & Wang 1997). This has become a common a problem in environmental studies as chimeras produce diversity that is not present in the original sample (von Wintzingerode *et al.* 1997; Hamp *et al.* 2009; Reeder & Knight 2009; Stoeck *et al.* 2009; Huse *et al.* 2010) and thus levels of biodiversity become inflated and unrealistic. In fact, it has been estimated that up to 30% of PCR products from a standard PCR are artificial chimeras (Wang & Wang 1997; Cronn *et al.* 2002). Multiple factors seem to influence the formation of PCR-induced chimeras and include; PCR cycles, Taq polymerase, extension time, the quality of DNA and shorter amplicons (Wang & Wang 1997; Shafikhani 2002; Gonzalez *et al.* 2005; Lahr & Katz 2009; Smyth *et al.* 2010). In response to the problem, several bioinformatic approaches have been developed and optimized in order to identify these multiple errors in pyro-sequenced environmental datasets (Zhang *et al.* 2000; Cole *et al.* 2003; Huber *et al.* 2004; Ashelford *et al.* 2005; Gonzalez *et al.* 2005; Ashelford *et al.* 2006; Haas *et al.* 2011; Quince *et al.* 2011).

Advances in nematode molecular identification

Due to the high level of technical expertise required to accurately identify marine nematode species and a potential deficit of 960,000 unnamed species, it is unlikely that identification of all marine nematode species will be achieved only by

morphological-based taxonomy (Lambshead & Boucher 2003). However, advances have been made with new molecular techniques that provide high throughput methods for the identification and classification of micro and meiofaunal taxa (De Ley *et al.* 2005). It has also been suggested that combining both taxonomic and molecular techniques may help clarify the reliability of nematode species biodiversity estimates (Bahr *et al.* 2005; Blaxter *et al.* 2005; Bhadury *et al.* 2006, 2008). Nevertheless, much of the genetic variation in marine habitats has been uncovered by the advent of molecular techniques alone (Knowlton 2000; Feral 2002). In fact, molecular appraisals of currently accepted marine nematode morphospecies are now uncovering genetic structuring indicative of extensive cryptic speciation (Derycke *et al.* 2005). Currently, amplification and sequencing of diagnostic regions of nematode DNA have become the major source of new information for advancing our understanding of evolutionary and genetic relationships (Hajibabaei *et al.* 2007a; Meldal *et al.* 2007). Combining MOTUs with improvements in environmental DNA extraction promises a rapid method for assessing the diversity of marine nematodes, *en mass*, at any given location (Porteous *et al.* 1997; Floyd *et al.* 2002; Blaxter 2004; Blaxter *et al.* 2005). The 18S rRNA gene has proven particularly useful in constructing viable nematode DNA barcodes and has been used in a number of studies to identify assortments of marine (Bhadury *et al.* 2006), and terrestrial free living nematodes (Donn *et al.* 2008), and both plant and animal parasitic species (Casiraghi *et al.* 2006; Riga *et al.* 2007). Molecular taxonomy could denote clear advantages when comparing to traditional taxonomy such as (i) applicability to a wide range of taxa, and different life stages, including those possessing few distinctive morphological features; (iii) a standardized approach for sample processing, interpretation, and comparison across different studies and (v) the ability to taxonomically characterize large numbers of samples that are typical of most ecological studies (Caron *et al.* 2009; Abebe *et al.* 2011). Regardless, sequences alone cannot say much about a species ecological role (Tautz *et al.* 2002). Information about the interaction of organisms with their environment will be incomplete because they lack morphological information that relates to function (Abebe *et al.* 2011). Open relational databases such as Nematol (http://nematol.unh.edu/) serve as portals for the collation of morphological, molecular and ecological data pertaining to nematode phylogeny and biodiversity. Also, by combining bioinformatic databases and high throughput methodologies with

videocaptured digital images of known specimens may allow the interpretation of both morphological and molecular data of nematode taxonomy, thus facilitating a less subjective approach to biodiversity appraisals.

Aim and outline of the thesis

The aim of this thesis was to use massively parallel sequencing to estimate molecular diversity of littoral meiofauna populations around the UK and mainland Europe. Advances in sequencing technologies are known to allow the characterization of environmental communities from different ecosystems, yet some flaws are currently known to exist. This thesis is divided in chapters: CHAPTER 2 is an overview of the 454 meiofaunal metagenetics in biodiversity assessment, in addition to presenting and discussing novel datasets from marine and tropical rain forest habitats. CHAPTER 3 describes the relative richness of multiple metazoan meiofaunal phyla inhabiting the marine benthos on two sandy beaches in the UK, using second generation sequencing approaches. The data quantify previously unknown and substantial levels of relative richness that refute currently accepted paradigms of phylum rank abundance. In CHAPTER 4 levels of alpha and beta diversities and community composition for meiobenthic communities are described using next-generation sequencing technique. This will clarify our understanding of the scale and extent of community variation across marine ecosystems. CHAPTER 5 describes how chimera formation during PCR can be influenced by species richness, sample taxonomic similarity and nSSU molecule secondary structure. In CHAPTER 6 a final discussion and concluding remarks can be found.

CHAPTER 2 - Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises

Introduction

Robust, quantified biodiversity assessment is key to deep understanding of the relationship between biodiversity and ecosystem functioning. The effects of major anthropogenic stressors on global ecosystems, including elevated CO_2, pollution, habitat loss and fragmentation, add urgency to this field, demanding an increasing focus on mechanistic and predictive studies. However, investigating the role of biodiversity in maintaining ecosystem function, resilience and recovery (Sutherland *et al.* 2006) can be meaningfully addressed only if biodiversity can first be identified. The identity of macrofaunal and floral communities can be ascertained by teams of trained taxonomists/ecologists with their skills being augmented by globally integrated molecular barcoding approaches (Hebert *et al.* 2003a; Hajibabaei *et al.* 2007b). Similarly, recent advances in sequencing power and the molecular identification of microbes are facilitating the more realistic characterization of the phylogenetic affinities, identity (DNA sequences), composition (Sogin *et al.* 2006; Huber *et al.* 2007) dynamics and even functional capacity (Edwards *et al.* 2006; Mou *et al.* 2008) of prokaryotic communities. There remains, however, a well-acknowledged biodiversity identification gap related to eukaryotic meiofaunal organisms (Blaxter 2003b; Blaxter & Floyd 2003; Tautz *et al.* 2003; Blaxter *et al.* 2005).

Meiofaunal taxa are a paraphyletic assemblage, grouped on the basis of size (i.e., organisms that pass through a 0.5 mm sieve but are retained on 25 to 65 μm sieves). Approximately 60% of animal phyla have meiofaunal representatives and meiofaunal Platyheminthes, Nemertea, Nematoda, Rotifera, Annelida, Arthropoda, Tardigrada, Mollusca and Chordata have taxa that occupy key roles in marine, freshwater and terrestrial habitats (Higgins & Thiel 1988; Giere 2009). Meiofaunal assemblages are dominated by nematodes and are characterized by high abundances (up to 10^8 individuals per 1 m^2) and diversity (up to 60 morphological species per 75 cm^3 of sediment) (Lambshead 2004). Thus, although meiofaunal organisms are conceptually and demonstrably ecologically important (Snelgrove *et al.* 1997; Danovaro *et al.* 2008a), current estimates of global species richness remain a matter of conjecture (Lambshead & Boucher 2003). For nematodes, global estimates of species richness range from 100,000 to 1,000,000, but only ca. 23,000 species have been described

(Platt & Warwick 1983; Coomans 2000), and contemporary studies routinely recover between 30-40% of sampled taxa that are new to science (Lambshead 2004). Meiofaunal taxon diversity and abundance is so great that effectively studying communities requires a huge investment in resources and labour. The effort expended in assigning only 10% of nematodes to known species was 120-fold that required to successfully assign all vertebrate morphospecies to known taxa (Lawton *et al.* 1998) in tropical forest habitats.

The identification bottleneck associated with meiofaunal taxonomy is confounded by a range of taxonomic hurdles: the small size and fragility of organisms, convergent evolution, morphological conservatism (Derycke *et al.* 2005; Bhadury *et al.* 2008; Derycke *et al.* 2008b; Fontaneto *et al.* 2009) and developmental and sexual variation in morphology (Tautz *et al.* 2003; Lambshead 2004; Blaxter *et al.* 2005). Perhaps the most restricting factor in meiofaunal research is the mismatch between the diversity and abundance of multiple phyla occupying a range of ecological niches and habitats and the number of taxonomists that are able to simultaneously identify and catalogue meiofaunal diversity. In order to address this impediment, it has been suggested that *en mass* molecular identification of meiofaunal communities may significantly advance knowledge and progress in meiofaunal research (Blaxter & Floyd 2003; Markmann & Tautz 2005). Whilst the molecular identification of meiofaunal communities shares similarities with current molecular barcoding (Hebert *et al.* 2003b) and microbial phylotype approaches (Kemp & Aller 2004; Shaw *et al.* 2008), there remains a difference in methodology and taxonomic richness and diversity.

Phylotypes, molecular operational taxonomic units (MOTUs) and barcoding for the identification of biodiversity

With a molecular barcoding approach, a standardized homologous region of the genome (e.g. the mitochondrial cytochrome oxidase subunit I gene [COI] for animals) is used for species identification, and is linked to a virtual or actual physical molecular voucher specimen (Hebert *et al.* 2003a; Ratnasingham & Hebert 2007). However, when dealing with individuals or communities of microscopic organisms, the whole voucher specimens are sacrificed usually in order to extract genomic DNA

(Blaxter *et al.* 2005; De Ley *et al.* 2005). Advances in videocapture technology of microscopic organisms (De Ley *et al.* 2005) and individual organismal PCRs (Floyd *et al.* 2002; De Ley *et al.* 2005; Floyd *et al.* 2005; Bhadury *et al.* 2006; Meldal *et al.* 2007) can overcome this problem and forge a link between taxon ecology/morphology and community-based DNA analyses. Such research provides potential for linking taxonomy, phylogeny (Forest *et al.* 2007; Warwick & Somerfield 2008) functional (Petchey & Gaston 2006) and molecular ecology. It also effectively engages and links morphological taxonomists with molecular ecologists, a connection that will be vital for a holistic approach towards ecosystem-based research. However, standard barcoding approaches are not appropriate for large-scale environmental analyses mainly because of extensive abundances and putative hyperdiversity of some taxa (e.g. nematodes, Lambshead (2004); Lambshead & Boucher (2003)). Further to this, the extent of taxonomic coverage and lack of taxonomic expertise, manpower and resources makes the task of barcoding environmental samples inefficient.

Instead, the proposed identification of operational taxonomic units (OTUs) in eukaryotic metagenetic analyses has more in common with prokaryotic phylotype (Kemp & Aller 2004) delineation than with species identification using standardized barcoding approaches. The term metagenomics is sometimes used to consider the analysis of any environmentally-derived genomic DNA (Hugenholtz & Tyson 2008). Here though, there is a distinction between metagenetics, the large-scale analysis of taxon richness via the analysis of homologous genes, and metagenomics, the functional analysis of environmentally derived DNA from unculturable organisms (Edwards *et al.* 2006; Rodriguez-Brito *et al.* 2006; Blow 2008; Hugenholtz & Tyson 2008; Mou *et al.* 2008).

Bacterial phylotypes are groups of sequences that are created by subjecting a larger community of sample-derived shotgun sequences, to a user-defined base pair cut-off algorithm. In most cases, phylotypes of a particular grouping (e.g. 97% for bacteria, Venter *et al.* 2004; Shaw *et al.* 2004) are used as a proxy for "species". Although microbial communities can be orders of magnitude more diverse than micro-eukaryotic communities, the similarities of their intractable community compositions have led to similar approaches in studying eukaryotic protists (Moon-van der Staay *et al.* 2001; Moreira & Lopez-Garcia 2002) and meiofaunal organisms (Floyd *et al.* 2002; Blaxter & Floyd 2003). For meiofaunal organisms, Floyd *et al.*

(2002) formally defined the molecular operational taxonomic unit (MOTU) concept whereby sequences derived from individual specimens are defined as belonging to the same MOTU, based on a user-defined cut-off. The term was later extended to community DNA extractions in Blaxter *et al.* (2005). Normally, the MOTUs do not have any formal correlation with published species descriptions. However, correlations can be achieved by *de novo* elucidation of cryptic species (Abebe & Blaxter 2003) bioinformatic sequence comparisons to existing databases (with both molecular and morphological data), further sequencing, or future classifications, termed "reverse taxonomy" (Markmann & Tautz 2005).

Environmental metagenetics

Until recently, most molecular identification was achieved using Sanger chain-termination sequencing (Kemp & Aller 2004; Venter *et al.* 2004). However, there has recently been a rise in the use of ultrasequencing platforms (Margulies *et al.* 2005) for metagenetic identification of microbial phylotypes using homologous gene region (Sogin *et al.* 2006; Hall 2007; Huber *et al.* 2007) derived from environmental DNA. The recent increases in sequencing throughput represent a significant shift in our ability to disentangle the biotic complexity of ecosystems. From sample collection to data analysis, there are numerous steps, questions and an exponentially large number of hypotheses that could be tested in order to optimally analyse environmental meiofaunal diversity.

Here, an overview of the relevant focal areas is provided in an attempt to highlight potential approaches and pitfalls in meiofaunal metagenetics. Secondly, two datasets derived from independent ultrasequencing experiments of marine benthic and tropical rain forest habitats are presented. The aim is to illustrate the advantages and limitations of ultrasequencing approaches in addressing large-scale identification of complex eukaryotic communities. The tropical rain forest case study predominantly targeted nematodes, whereas the marine example targeted collective meiofauna (extended to include organisms ranging from 45 µm to 1000 µm in size). The approaches and data presented do not test specific hypotheses regarding metagenetic analyses, but provide a resource that will be useful to researchers wishing to pursue

similar research. Although meiofaunal organisms are the primary focus, the general principles are easily transferrable to other eukaryotic as well as prokaryotic taxa.

Methodological overview and rationale

Sample preservation and extraction

Once an ecologically suitable sampling strategy has been designed, an appropriate decision needs to be made regarding sample processing. Given the diverse and dynamic nature of the micro- and meiofauna, it is predicted that after removing a small subsample of the community, a natural progression of ecological interactions will change the population composition. It is therefore important to either preserve or process samples shortly after collection. Some experiments (e.g. those with small sample sizes or local collection regimes) may lend themselves to field processing. Others will necessitate sample, and more importantly, DNA preservation. Such decisions are based predominantly on logistics. Moreover, if sampling regimes are extensive and geographically diverse, it is preferable to deal with a large number of small samples, rather than vice versa.

Formalin is the preferred fixative for morphological analyses of the meiofauna (Giere 2009), but specimens fixed in formalin yield low quality and degraded DNA (but see Thomas *et* al. 1997 and Bhadury *et al.* 2005). Conversely, samples fixed in ethanol yield DNA optimal for downstream molecular manipulations, but corrupted morphological features due to osmotically driven shrinkage (Bhadury *et al.* 2006). In an attempt to overcome these constraints, samples are often split between formalin and ethanol preservation to yield distinct samples for morphological and molecular genetic analysis, respectively. Adopting a split sampling approach, however, not only creates a problem of potentially unequal community composition between samples, but also precludes obtaining both morphological and molecular data from the same individual (Yoder *et al.* 2006). An answer to the preservation issue is the use of a solution of 20% DMSO, 0.25 M disodium EDTA, saturated with NaCl, pH 8.0, known as DESS by (Yoder *et al.* 2006). Originally proposed for the preservation of avian blood samples (Seutin *et al.* 1991), DESS has yielded PCR-ready DNA from individual nematodes and communities of entire soil/sediment samples for up to 1

year at room temperature. DESS works by inactivating naturally occurring nuclease activities by a combination of a severe osmotic shock, followed by rapid transportation of disodium EDTA and NaCl into tissues enabled by DMSO (Yoder *et al.* 2006). As with all DNA preservation approaches, it is important to optimize the DNA to buffer ratio to achieve effective preservation. Such is particularly the case for wet soil/sediment samples where the inclusion of significant volumes of sample water may dilute either the concentration of ethanol and/or DESS, preventing complete inhibition of nuclease activities.

Meiofaunal organisms must always be extracted from the substrate because the biomass is orders of magnitude lower than the actual sample volume of soil, sediment, or water. Separation can be achieved by employing several approaches (reviewed in Somerfield *et al.* 2005)), including those that rely on agitation of the sample in large volumes of water followed by retention of the community on sieves. Such mass decantation approaches rely on the different settling speeds of abiotic particles compared to the biotic fraction. Medium to coarse grain sediments can often be decanted successfully by mass decantation alone, but muddy or high in clay aggregate samples may require prior rinsing or sonication (Murrell & Fleeger 1989; Giere 2009). Following mass decantation, samples are frequently cleaned using flotation/centrifugation approaches using either sugar solution (Jenkins 1964; Esteves & Silva 1998), or Ludox®, a colloidal silica solution with a specific gravity tailored to user specifications (Markmann & Tautz 2005; Giere 2009). While passive methods recover both living and dead components of the community, active methods (e.g. Bearmann Funnel and its modifications) recover only living components because they depend on organismal locomotion (Baermann 1917; Whitehead & Hemming 1965). During all of these procedures, it is important to note that communities are continually manipulated via the use of measuring cylinders, funnels and stainless steel sieves. From a DNA-based perspective therefore, the potential for cross contamination of a minor fraction of biodiversity between samples is a concern. Cleaning of apparatus should therefore be rigorous and standardized, with pressurized water augmented by autoclaving and UV treated where possible. Cross contamination can be tested by performing intermittent negative control experiments, involving no samples.

Experimental Design

Case Studies:

1. Marine littoral benthos

Three 44 mm x 1000 mm sediment cores were taken ca. 10 m apart from the low water intertidal zone from the beach at Littlehampton on the south coast of England, UK, during July, 2007. Samples were stored for approximately 6 months at room temperature in DESS solution (ratio of 1:3, volume drained sediment to DESS volume respectively). A meiofaunal fraction designed to include the larger nematodes (45 µm-1000 µm) was isolated by mass decantation, followed by Ludox® (specific gravity 1.16) centrifugation, utilizing combinations of stainless steel and Millipore disposable nylon net filters (Millipore Corporation).

2. Tropical Rainforest

In March 2007, soil, litter and understory habitats were sampled at La Selva Biological Station, Costa Rica following the protocol described by Powers *et al.* 2009). Briefly, four locations (at 200 m, 300 m, 400 m and 500 m markers) along the Sendero Suroeste trail were selected. Within each location, a sampling plot (22 m radius circle, 1520 m^2) was divided into 4 quadrants. Within each quadrant, one random canopy tree and one random understory tree were selected as sampling points. Two soil (15 cm depth) and two litter (overlying soil) samples were collected from 15 cm x 15 cm areas (within 1 m to 2 m away from the canopy and understory trees). A total of 8 subsamples (2 trees X 4 quadrants) were pooled to make up one composite soil sample and one composite litter sample per plot. The epiphytic material (e.g. lichen, moss, algae) present on the surface of stems of canopy and understory trees was collected to represent canopy sample. Each tree was sampled at three equidistant (between 2.5 cm and 2.5 m from the soil surface) vertical strata. A 15 by 15-cm area was sampled in each of the strata for a total of 24 subsamples (3 strata X 2 trees X 4 quadrants) pooled to form one composite canopy sample per plot. Samples were stored in a cooler and transported to Universidad Nacional for immediate processing. Litter and canopy samples were cut into smaller pieces, mixed thoroughly, and 15-30 g of subsamples were used for nematode extraction. Litter and canopy subsamples were further chopped in a blender in 150 ml of deionized water

for 10 s and set onto cotton wool filters (s'Jacob & van Bezooijen 1984) placed in extraction trays. Nematodes were collected at 24 and 48 hr intervals and immediately counted for total abundance under an inverted microscope. Nematodes from soils were extracted from ~100 g of subsamples using sugar flotation and centrifugation (Jenkins 1964) and counted immediately.

Community DNA extraction

Following sample manipulation for traditional meiofaunal ecology studies, the community is retained on 25-45 µm stainless steel sieves or filters. Samples are then rinsed from sieves using approximately 40-60 mL of water, ethanol or DESS, depending on experimental design. However, for the purposes of DNA extraction, all target organisms have to be removed from solution and placed in a suitable cell lysis buffer for DNA extraction. Specimen retrieval can be achieved by removal from Ludox-water interfaces (Markmann & Tautz 2005), centrifugation (but note, the specific gravity of DESS is unsuitable for centrifugation separation), successive subtraction and examination of aliquots of water, or using disposable sieves and 45 µm meshes. The aim is to reduce the community into a volume from which genomic DNA can be effectively liberated.

Once removed from the sample, DNA can be extracted from taxa, but before proceeding, two issues should be considered. First, although extraction methods target organisms of a desired size range, the sample is likely to contain additional taxa such as bacteria, Archaea, Fungi, Plantae etc., present in the environment, adsorbed on the surface of, and present in the guts of targeted organisms. Second, decaying organic matter, containing humic substances and secondary metabolites (e.g. polyphenols, tannins and polysaccharides (Zhou *et al.* 1996; Porebski *et al.* 1997) can potentially inhibit PCR and sequencing reactions.

To achieve effective DNA extraction and overcome the problem of environmentally derived inhibitors, several approaches have been developed to obtain PCR-ready genomic templates from environmental samples. Sample cell disruption can be more effective using bead beating, though there is a risk of shearing DNA into smaller fragments (Picard *et al.* 1992). Conversely, using longer, more gentle treatments, such as spinning wheels, sodium dodecyl sulphate (SDS)

(Huber *et al.* 2002; Sogin *et al.* 2006), enzymes, heat, or freeze thaw processes, generally yield higher molecular weight genomic DNA extracts (Zhou *et al.* 1996; Porteus *et al.* 1997; von Wintzingerode *et al.* 1997). Environmental DNA extraction protocols either use a combination of CTAB, phenol, chloroform, caesium chloride etc. (Sambrook *et al.* 1989; Porteus *et al.* 1997) or proprietary chemicals to clean DNA extracts (e.g. Epicentre SoilMaster™, ZR Soil Microbe™ and Mobio PowerSoil™ DNA extraction kits) in association with various column formats. However, given that most proprietary environmental kits are designed for extracting DNA from microbes, they usually have a maximum capacity of ca. 250 µg of DNA, or eluates of ca. 200 µL, and it is widely acknowledged that overloading results in poor DNA yields. Consequently, for eukaryotic environmental work, either the community has to be partitioned into 250 mg of DNA aliquots, or somehow digested in a very low volume lysis buffer prior to kit usage. An alternative may be to use a combination of traditional lysis, followed by a large capacity DNA extraction kit, as used in the marine case study here.

Case Studies:

1. Marine littoral benthos

Out of seven different DNA extraction methods tested on whole community marine environmental samples, the QIAMP DNA Blood Maxi Kit (Qiagen) yielded the most consistent PCR results (**Figure 2.1**) and was used for subsequent DNA extractions.

Meiofaunal Metazoan		Nematoda
Primer	Sequence	
SSU_F04	5'- G C T T G T C T C A A A G A T T A A G C C -3'	100
% identity	99 96 96 95 98 99 97 97 95 98 99 99 99 99 99 99 99 99 98 98 98	
SSU_R22	5'- G C C T G C T G C C T T C C T T G G A -3'	100
% identity	100 100 99 99 100 100 100 100 100 100 98 99 88 100 97 100 83 99 100	
NF1	5'- G G T G G T G C A T G G C C G T T C T T A G T T -3'	100
% identity	99 100 100 100 100 99 100 100 100 100 100 100 98 100 99 99 100 100 98 100 100 99 99 100	
18Sr2b	5'- T A C A A A G G G C A G G G A C G T A A T -3'	100
% identity	100 99 100 99 100 100 100 100 98 100 100 100 100 100 100 100 92 97 98 95 100	

Figure 2.1 - The 18S rDNA primer sets used in the marine littoral meiobenthos and tropical rainforest analyses. The number beneath each nucleotide base highlights the conservation of the priming site (calculated by visual inspection) derived from an alignment of ~170 sequences derived from NCBI representing each of the phyla containing meiofaunal representatives (supplied in the Supporting Information). Sequences representing Nematoda were subsampled throughout the currently accepted phylogenetic range presented in Meldal et al. (2007) and all base pair positions were 100% conserved in all primer pairs but see Porazinska et al. (2009) for further primer bioinformatic comparisons.

2. Tropical Rainforest

Samples were transferred into ZR BashingBead Lysis Tubes (Zymo Research Corp, Santa Ana, CA) and disrupted using a Mini-BeadBeater (BioSpec Products, Inc. Bartlesville, OK) at maximum speed for 2 min. Genomic DNA was extracted using a ZR Soil Microbe DNA kit according to the manufacturer's protocol.

Choice of genomic loci for delineation of meiofaunal molecular operational taxonomic units (MOTUs)

There are clear conceptual differences between metagenetic and specimen-based barcoding analyses. Hebert *et al.* (2003a) chose the mitochondrial COI gene as the standardized barcoding gene for animals for a number of well-established reasons. Mitochondrial DNA has a haploid mode of inheritance, elevated rate of molecular evolution, lacks introns and has limited recombination (Clayton 1984; Wilson *et al.* 1985; Avise 1994; Piganeau *et al.* 2004; Tsaousis *et al.* 2005). Moreover, indels are

rare in mtDNA protein coding genes, universal primers for the COI gene are fairly robust (Folmer *et al.* 1994; Zhang & Hewitt 1997) and the mode of molecular evolution of COI usually facilitates species discrimination whilst also retaining phylogenetic information for the majority of animal taxa (Hebert *et al.* 2003a; Hebert *et al.* 2003b). Unfortunately though, the COI gene is not optimal for molecular taxonomic identification purposes for nematodes because nematode mitochondria have high mutational rates, display excessive saturation, biased substitution patterns and are very A+T rich (Blouin *et al.* 1998; Blouin 2000). Furthermore, primers used for most barcoding studies (Folmer *et al.* 1994) are poorly conserved across nematode diversity, and alternate conserved regions for primer design are not evident (Blouin *et al.* 1998). Given that meiofaunal communities can comprise between 50-90% nematodes (Lambshead & Boucher 2003), other nuclear markers may be more appropriate for meiofaunal metagenetic studies (Blaxter 2003a; Blaxter *et al.* 2003).

It is widely acknowledged that alternative markers are required for certain taxa, and attempts are being made to include suites of markers in DNA barcoding. Examples of alternative markers include nuclear ribosomal RNA genes that have been used for decades to identify phyla of microscopic eukaryotes. It was first demonstrated in the 1960's that ribosomal RNA genes (rDNA) and their gene products (rRNA) could be used for the taxonomic classification of microbial species (Doi & Igarashi 1965; Dubnau *et al.* 1965; Pace & Campbell 1971). The genes coding for rRNA are particularly well suited for molecular taxonomy, because they are universally found in all cellular organisms and are of relatively large size. They also contain both highly conserved and variable regions that facilitate the design of very conserved primers that amplify diagnostic regions (Woese 1987; Floyd *et al.* 2002; Markmann & Tautz 2005; Carvalho *et al.* 2009).

For eukaryotes, both the nuclear 18S small subunit (nSSU) and 28S large subunit (nLSU) rDNA genomic regions are excellent candidate genes for molecular identification as they are present in tandemly repeated, multiple copies (50-150 copies per cell), and undergo concerted evolution (Markmann & Tautz 2005). The latter two attributes facilitate their amplification from microscopic organisms that are highly conserved within a species, and divergent among species. DNA barcoding studies utilizing rDNA genes have focused on the more variable portions of the genes: the D2-D3 'diversity loop' regions of the 28S and the 5' region of 18S. While

both 18S and 28S are probably equally useful for molecular identification, both cases presented here utilize 18S, predominantly because for free-living nematodes at least, universal 18S primer sets are more consistent for PCR amplification than 28S primer sets (Bhadury *et al.* 2006; Porazinska *et al.* 2009). There are also significantly more 18S than 28S sequences in public repositories (e.g. recent SILVA databases contain 868,390 18S vs. 143,653 28S entries) (Pruesse *et al.* 2007), enabling more accurate and comprehensive taxonomic assignment to query sequences (Blaxter 2003a). What remains less clear is the extent to which variation in 18S or 28S genes follows the division of individual organisms into biological species. Within Nematoda, some species have identical 18S sequences, while other congeneric species differ by over 2% (Blaxter *et al.* 1998). 18S (and 28S) are good markers for deep phylogeny (Blaxter *et al.* 1998) but may be less suited to distinguishing between closely related taxa: available data suggest that the D2-D3 loop of 28S may be the better marker in this respect (Ye *et al.* 2007; Subbotin *et al.* 2008).

Marine littoral benthos and tropical rainforest case studies

To select optimal 18SrDNA primer pair combinations, the genomic location of available primers (http://nematol.unh.edu/, http://www.nematodes.org/) and numbers of segregating sites spanning primer pairs ca. 400 bases apart (recommended for Roche 454 GSFLX sequencing) were investigated. Furthermore, the resolving power of target regions (Porazinska *et al.* 2009) and level of primer sequence conservation across meiofaunal metazoans was also considered. Consequently, two candidate regions, defined by primers SSU_F04 and SSU_R22 towards the 5' end and NF1 and 18Sr2b towards the 3' end of the 18SrDNA gene (**Figure 2.2**) were used independently in the marine and rainforest samples.

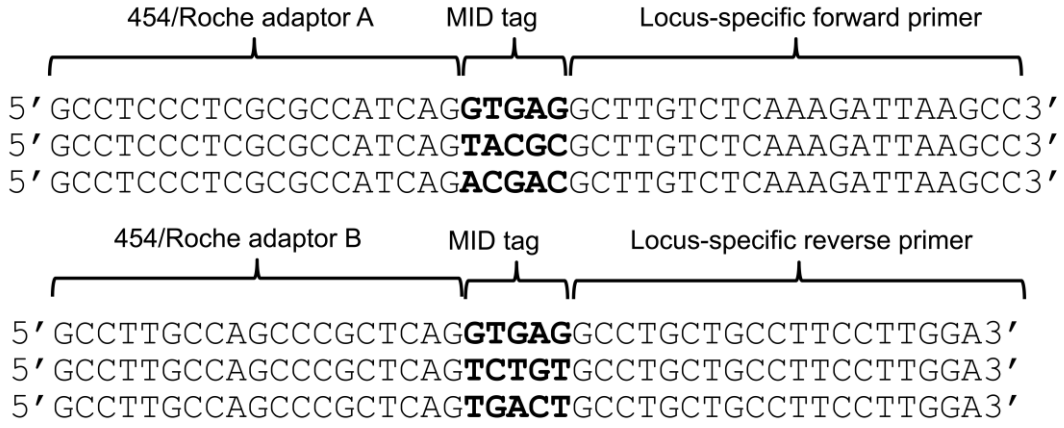


Figure 2.2 - Examples of three independent primer sets for the identification of mixed metagenetic amplicon pools. The Roche 454 adaptor precedes a five base molecularly identified (MID) tag immediately prior to locus-specific forward and reverse primers respectively. Bidirectional and unidirectional sequencing can be achieved by combining A and B adaptors and appropriate Roche 454 emulsion PCR kits (II or III). In the current example, sequencing from the 5' end of the forward primer could be performed by hybridizing the Roche 454 adaptor B onto the beads during emulsion PCR and sequencing with the A sequencing adaptor.

PCR and sequencing strategies

At the time of writing, three ultrasequencing platforms (the Roche 454 GS Titanium Series, the Illumina SOLEXA Genome Analyzer and the Applied Biosystems SOLiD™ System) were readily accessible by the research community. Presently, the Roche 454 system is the intuitive choice for any form of metagenetic, or metagenomic analysis, simply because of greater read lengths and subsequent clarity of annotation of individual reads (Blow 2008; Hugenholtz & Tyson 2008). Whereas the Illumina Genome Analyzer and ABI SOLID™ generate many gigabases of sequence data partitioned into 150bp base reads per instrument run, the Roche 454 Titanium platform generates about 400Mb of data from 0.8 million 400-450 base reads.

Consensus sequence accuracies of the Roche 454 sequencers range from 99.97% to 99.9984%, with individual per-base error rates of between 0.6% and 0.49%. The large majority of per-base errors (between 39%-98%) are derived from misreading of the lengths of nucleotide homopolymers effects, including extensions (insertions), incomplete extensions (deletions) and carry forward errors (insertions and

substitutions) (Huse *et al.* (2007); Margulies *et al.* (2005); http://454.com). Nucleotide homopolymers are relatively rare in coding genes and in 18S and 28S in particular but quite common in nSSU variable region four (V4).

Following the publication of Margulies *et al.* (2005) *de novo* sequencing of *Mycoplasma genitalium*, it was clear that parallel pyrosequencing represented a paradigm shift in the cost and volume of sequencing compared to chain-termination (Sanger *et al.* 1977) approaches. However, in order to utilize such sequencing power for multi-sample metagenetic investigations, methods had to be devised to pool and then recover amplicons on single, or multiple Roche 454 picotitre plates. Physical gaskets can be used to partition Roche 454 picotitre plates from between 2 and 16 samples (as used in Sogin *et al.* 2006), but this sample multiplicity is inevitably associated with lower overall sequence throughput per picotitre plate. A number of ways have been suggested to separate samples post-run *in silico.* These range from pooling of different easily identifiable loci (Thomas *et al.* 2006), to use of individually molecularly identified (MID) linkers to independent samples (Meyer *et al.* 2008, 2007; Parameswaran *et al.* 2007; Roche 454). However, for metagenetic samples, incorporating MID linkers and universal Roche 454 adaptors into fusion primer sets (Binladen *et al.* 2007) is probably the easiest and most cost-effective way of tagging amplicons. Following Binladen *et al.* (2007) study-specific forward primers can be synthesized preceded by a sample-specific MID tag and either Roche 454's A or B universal adaptor sequences **(Figure 2.2)**. Thus, each experimental sequence will begin with the MID tag and the PCR primer, and these can be recognized via pattern-matching algorithms to sort individual reads into sample sets. Binladen *et al.* (2007) initially proposed the use of 2 base tags yielding 16 (4^2) different MID combinations. However, Huse *et al.* (2007) strongly recommends the use of MID adaptors that differ by at least two bases to limit the potential of misallocation due to errors in the MID sequence itself. Eighty two of the possible 1024 (4^5) five base MID tags that can be combined to fulfil these criteria are currently listed at the Josephine Bay Paul Center's Visualization and Analysis of Microbial Population Structures (VAMPS) website http://vamps.mbl.edu/resources/keys.php, and are supplied here in the supplementary review file. For even higher stringency, Hamady *et al.* 2008) and Hamady & Knight 2009) constructed 1,544 optimal eight base error-correcting barcodes based on Hamming codes, which minimise redundancy. Given the nature of Roche 454

sequencing, homopolymers are also best avoided and it is optimal (although probably not essential with contemporary read lengths) to utilise primers that require the least number of parallel sequencing nucleotide flows (Huse *et al.* 2007; Meyer *et al.* 2007; Meyer *et al.* 2008a) to maximise sequencing efficiency through the adapters and primers.

Case Studies:

1. Marine littoral benthos

The 18S rDNA fragment spanning the primers SSUF04 (5'-GCTTGTAAAGATTAAGCC-3') and SSUR22 (5'- GCCTGCTGCCTTCCTTGGA-3') (Blaxter *et al.* 1998) was amplified using MID-tagged fusion primers using 1 µl of genomic DNA template (1:500 dilutions) in a 40 µL reaction using *Pfu* DNA polymerase (Promega), according to manufacturers' recommendations. Sample-specific PCR reactions involved a 2 min denaturation at 95 ºC, then 35 cycles of 1 min at 95 ºC, 45s at 57 ºC, 3 min at 72 ºC and final extension of 10 min at 72ºC. Negative controls were included for all amplification reactions. Electrophoresis of PCR products was carried out on a 2% Top Vision^{TM} LM GQ Agarose (Fermentas) gel and the expected 450bp fragment was purified using the QIAquick Gel Extraction Kit (Qiagen), following the manufacturer's protocol. All purified PCR products were then quantified with an Agilent Bioanalyser 2100, diluted to the same concentration, pooled and sequenced (A-Amplicon, alongside 10 additional unrelated experimental samples) on a half-plate of a Roche 454 GSFLX sequencer at Liverpool University's Advanced Molecular Genetics Facility, UK.

2. Tropical Rainforest

Individual PCR amplifications were performed following protocols described in detail elsewhere (Porazinska *et al.* 2009) using tagged fusion primers and 1 µL of DNA template. A total of 12 metagenetic samples (all derived from single PCR replicates) were pooled and sequenced together on a single GSFLX half-plate at the Interdisciplinary Center for Biotechnology Research at the University of Florida, Gainesville, FL. Through earlier experiments with artificially assembled nematode communities, it was determined that the use of a single in-house PCR replicate as well as a single emulsion PCR and pyrosequencing run is sufficient for both qualitative and quantitative nematode community analysis (Porazinska *et al.* In

Review). To illustrate potential information that can be drawn from metagenetic studies, data from a single location within the transect (i.e. one soil, one litter, and one canopy) is presented here.

Bioinformatic analyses of metagenetic datasets

The increase in read number and read length generated through contemporary ultrasequencing platforms requires novel sequence analysis packages that reduce computational runtime and increase OTU clustering efficiency. Read scaling has made algorithms using direct pairwise comparisons of all available sequences (N*(N-1)) computationally intractable (Yu *et al.* 2006; Huson *et al.* 2007) and has reduced the efficiency of programs using distance matrices methods (N*(N-1)/2) (Schloss & Handelsman 2005). Distance matrix methods have been used previously for bacterial metagenetic analyses (Sogin *et al.* 2006) . However, the time required for the generation of distance matrices can increase exponentially with an increase in sequence number or metagenetic diversity, and the derivation of a distance matrix inevitably includes estimation of pairwise alignments. Rapid processing of large read numbers requires either reduction of this search space by heuristic avoidance of irrelevant comparisons or implementation of approaches less bound by problems of pairwise comparison. The latter refers to k-mer algorithms, which can cluster sequences based upon the probability of matching a particular word between sequences (Sun *et al.* 2009).

Sequence entry order is a primary concern for developing OTU clustering algorithms. Available programs assign sequences to OTUs based upon fixed distances from an initial seed (Blaxter *et al.* 2005; Sun *et al.* 2009). Consequently, an outlier seed can heavily influence generation of OTUs. While it is possible to randomize the sequence entry order, this process becomes increasingly inefficient as progressively more reads are used. Furthermore, randomization of sequence order can lead to variations in the final OTUs (Floyd *et al.* 2002; Blaxter *et al.* 2005).

The analyses that have been performed here are based upon **O**perational **C**lustering of **T**axonomic **U**nits from **P**arallel **U**ltra**S**equencing (OCTUPUS, Sung *et al.* in review), a program that attempts to address both seeding and runtime problems by interlacing sequence alignments and pairwise comparison in order to generate

OTUs (beta version available from the Thomas laboratory on request). OCTUPUS takes advantage of k-mer algorithms (Zhang *et al.* 2000) to make pairwise comparisons against consensus sequences, and can be faster than distance matrix methods (Schloss & Handelsman 2005) or k-mer comparisons using unique sequences (Sun *et al.* 2009). The consensus sequences OCTUPUS uses are continually evolving based upon the sequences assigned to the OTU. Once repeated multiple alignments of the OTU sequences result in an unchanging consensus sequence, the OTU is considered a "fixed OCTU". Each OCTU potentially represents one taxonomic group based upon the identity cut-off. By using fixed pairwise comparisons against a variable consensus sequence, seeding error can be reduced.

Data Analysis

Sequences generated from the Roche 454 GSFLX from both the marine littoral benthic and tropical rain forest habitats were first checked for quality using Lucy (Chou & Holmes 2001) at default parameters. The sequences were then trimmed, binned according to MID tags and clustered at 95, 96, 97, 98, and 99% similarity match using the OCTUPUS pipeline (Sung *et al.* in review). Fixed OCTUs were then compared by megablast (Altschul *et al.* 1997) against the NCBI database. A major concern with the analysis of PCR-generated homologous gene regions is the formation of *in vitro* recombinant DNA molecules, or chimaeras, where molecules from two different origins artificially combine during PCR (Meyerhans *et al.* 1990). One quick and objective way of flagging a putative chimera is to use the "greedy" (taking in more sequences than it should) nature of the megablast algorithm and compare the length of matched bases from the top hit in a megablast search to the length of the query sequence. As long as the database sequence is longer than the query sequence, and a portion of the 3' end does not match, it is likely that the query is a recombinant. Given that recombinant molecules can form at any position along a DNA sequence (Qiu *et al.* 2001), and referring to previous analyses including control datasets (Porazinska *et al.* 2009), we applied a strict quality filter that allowed a four base length difference between a query OCTU sequence and the matched database sequence for further analyses.

Data overview and interpretation

Community PCR and sequencing yielded a total of 29756 high quality sequences over 200 bases from the marine samples (core 1: 9893, core 2: 9908 and core 3: 9955, GenBank numbers to be allocated), generating between 246 and 1327 putative non-chimeric OCTUs between the 95%-99% cut-offs (**Figure 2.3a**). For the tropical rain forest, the three samples yielded a total of 40334 high quality sequences of at least 200 bases (soil: 23742, litter: 10854 and canopy: 5738, Genbank numbers to be allocated), generating between 625 and 5671 putative non-chimeric OCTUs between the 95%-99% similarity match (**Figure 2.3b**). Putative chimera detection for the total datasets ranged from 35%-38% and 44%-49% for the F04-R22 (marine) and NF1-18Sr2b (rainforest) data respectively. Many, but not all putative chimeric OCTUs were made up of low copy number reads and accounted for ca. 20% of reads used in generating the OCTUs.

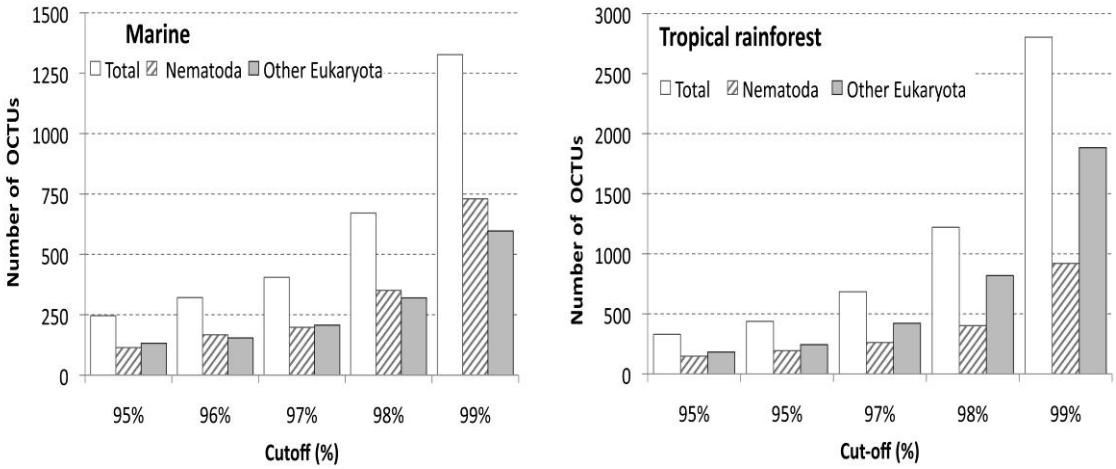


Figure 2.3- Number of operational clustering of taxonomic units (OCTUs) found in (a) the marine littoral benthos and (b) tropical rainforest case studies for each base cut-off. Putative non-chimeric OCTU numbers are presented for the total data, Nematoda and other Eukaryota (including OCTUs with BLAST hits to 'environmental samples' representing unclassified taxa).

PCR recombination and chimera formation

The proportion of putative chimeras representing OCTUs is disconcerting, especially as recent control experiments on artificial nematode communities only identified a level of 0.4% of total reads (Porazinska *et al.* 2009). PCR recombination will (a.) suggest the existence of sequences that do not actually exist in the investigated sample, but also (b.) give a false impression of organism richness (Markmann & Tautz 2005). *In vitro* recombination of homologous DNA leading to chimeric molecules is widely reported in the microbial literature and detected in databases (von Wintzingerode *et al.* 1997; Qiu *et al.* 2001; Ashelford *et al.* 2005), and levels up to 33% have been reported from meiofaunal communities (Markmann & Tautz 2005). In the latter example, as in many chimera detection approaches (Huber *et al.* 2004; Shaw *et al.* 2008), query sequences were split into 50:50 blast query fragments and were identified as putative chimeras if the 5' query and the 3' query had best BLAST matches to different taxa, suggesting that the 5' and 3' ends of the sequence are derived from different species. Upon revisiting the data, it is apparent that the 50:50 blast approach can fail to detect some chimeric formations, especially for recombinants that occur further from the sequence midpoint. Accordingly, a stringent base matching approach at the 3' end of the query sequence against complete reference sequences appears to be the more conservative way of approaching chimera quality control, at least for the taxa involved in this study. Slightly more putative chimeras were generated from the 3' end of the 18S rDNA terrestrial samples, compared to the 5' end of the 18S rDNA marine samples. It is therefore tempting to suggest that the 5' end may be less susceptible to chimera formation, perhaps due to differing secondary structures and levels of polymorphism (von Wintzingerode *et al.* 1997; Qiu *et al.* 2001) though the experiments vary in many aspects. In the marine case study, environmental samples were stored in DESS, whereas the tropical rain forest samples were processed without storage. Additionally, the two studies used different DNA polymerases and extraction methods; Promega *Pfu* and spinning wheel digestion for the marine samples and New England Biolabs DyNAzyme Hot Start polymerase and bead beating disruption for the terrestrial samples, potentially also affecting the incidence of chimera formation (Qiu *et al.* 2001). Furthermore, as is usually the case with ultrasequencing

experiments, there was no replication of the actual 454 step, and so it is difficult to draw accurate conclusions without further experimentation and hypothesis testing.

In many ways, metagenetic ultrasequencing experiments are the ideal "breeding ground" for recombinant DNA molecules. They are based on the amplification of homologous regions from a large number of potentially highly related organisms (von Wintzingerode *et al.* 1997; Qiu *et al.* 2001). The taxonomic composition of the samples may contribute to the level of chimera formation (Qiu *et al.* 2001), but more empirical work needs to be done to assess the level of chimera formation in community-based PCR. The chimera detection approach applied here is particularly aggressive and could also exclude taxa that incorporate five base or more indels in BLAST assignment. Further solutions may therefore be necessary to advance the field of chimera detection, but given that chimeric molecule formation is potentially highly spatially stochastic, a quick and ideal solution may be unattainable without reference to control datasets. It is therefore better to try and reduce the level of DNA recombination within environmental PCRs by adhering to the following procedures (a.) performing "gentle" methods of DNA extraction (enzymatic digestion and using spinning wheels) (Huber *et al.* 2002), rather than bead beating approaches, (b.) increasing polymerase extension times and (c.) where possible, reducing the number of PCR cycles to the minimum (e.g. 20) (Meyerhans *et al.* 1990; von Wintzingerode *et al.* 1997; Qiu *et al.* 2001).

Sample and taxon coverage

Both sampling and sequencing approaches achieved between five and twenty three times deeper sequence coverage per core than is usually revealed with chain termination clone library approaches (Kemp & Aller 2004; Blaxter *et al.* 2005; Markmann & Tautz 2005) at approximately 1% of the cost. The coverage and cost-effectiveness of Roche 454 sequencing therefore brings substantial advantages to studies aiming to elucidate the molecular genetic richness of complex eukaryotic communities.

According to BLAST results, both primer sets (SSU_F04-SSU_R22 and NF1-18Sr2b) amplified homologous 18S gene regions from a substantial proportion of not only meiofaunal, but representatives of the Kingdoms Protista, Plantea and Fungi, in addition to those OCTUs with BLAST hits to "environmental samples". The latter are generally representative of the total data (e.g. comprising ca. 50% nematodes, in addition to further eukaryota) and further manual BLASTing can refine the taxonomic assignment of specific groupings if required. It is clear therefore that both primer sets are very highly conserved in eukaryotes. The primary difference between the two primer sets is that SSU_F04 and SSU_R22 span a more variable region of the 18SrDNA gene (ca. 30% more polymorphic sites) compared to the NF1-18Sr2b region. Of the meiofaunal phyla that may have been expected to be present in both environments, notable exceptions are Cnidaria, Nemertea, Rotifera, Brachiopoda and Echinodermata in the marine habitat, and Platyhelminthes, Annelida and Mollusca in the tropical rain forest habitats. This might have been mainly due to the small scale sampling and the patchy nature of meiobenthic organisms in the sediments. Further to this, seasonal variations in organisms life history strategies and primer bias towards some phyla must also be considered. Visualizing the conservation of the marine primer sets within the small subunit reference database from SILVA using ARB (Ludwig *et al.* 2004; Pruesse *et al.* 2007) suggests that all of the above, with the exception of cnidarians, should have amplified if genomic DNA was available in the PCR reaction. Therefore, these phyla were either not present in these samples (reflecting actual biology, or the result of taxon extraction methods), or competitive PCR (von Wintzingerode *et al.* 1997) may have prevented amplification of the missing phyla. Revisiting the priming sites of SSU_F04 and SSU_R22 however, reveals that ca. 50% of cnidarians have a base pair mismatch at the penultimate 3'position of SSU_R22, suggesting that primer mismatching will reduce the amplification of cnidarians in similar studies. In the tropical rainforest case, although undetected in the samples presented here, both flatworms and annelids were recovered in the remaining replicate samples. Absence of molluscs may be associated with the exclusive nature of the extraction methods.

It is clear from **Figure 2.3** that OCTU generation at multiple different cut-offs provides very different estimates of richness per sample as OCTUs are created at ever deeper levels of phylogenetic resolution. At fine levels, intraspecific variation will be sampled in some taxa, whereas at deeper levels, certain taxa will be grouping

on the basis of genera, order and higher taxonomic levels (Shaw *et al.* 2008). Many nematode morphospecies can be separated on the basis of very low 18S sequence divergence (e.g. 2%) (Blaxter *et al.* 1998), whereas intragenomic and intraspecific sequence variation will invariably be higher in other species and phyla. It is likely that the OCTU cut-off level that broadly correlates with species will occur between 95% and 99% 18S sequence divergence, but there will obviously be exceptions according to the actual species involved in the samples. Without explicitly referring to species, OCTU discrimination does however provide comparative metrics that can appraise relative diversity between samples.

Considering the 97% OCTU cut-off, the marine samples two (174 OCTUs) and three (160 OCTUs) contained more than twice the OCTU richness of sample 1 (71 OCTUs) (**Figure 2.4**). Thus, although all three intertidal cores were collected within 10m of each other minor changes in microhabitat (e.g. sediment grain size, detritus and organic matter and bacterial assemblages) can significantly alter meiofaunal richness between samples, even at microspatial scales (Giere 2009).

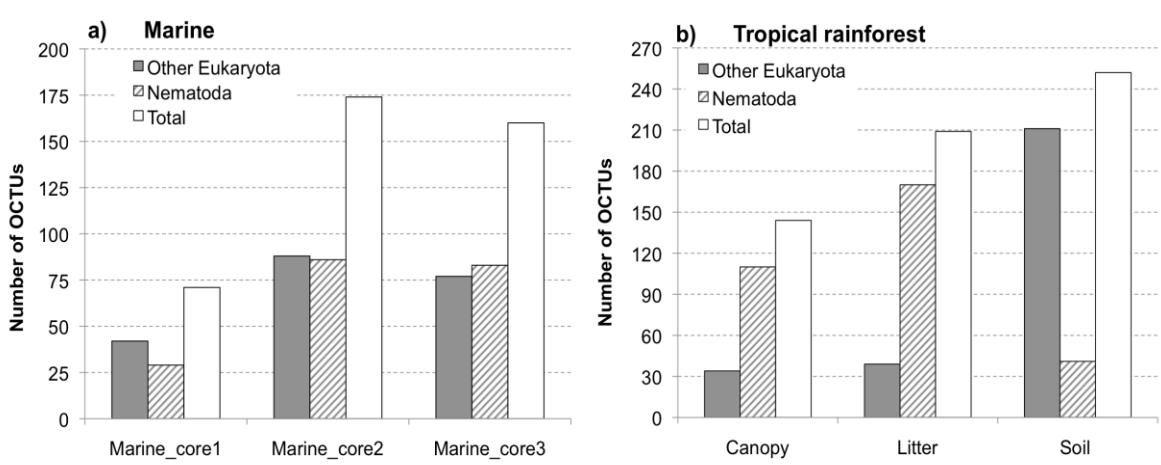


Figure 2.4- Number of putative non-chimeric OCTUs (clustered at 97% similarity) found (a) in the marine littoral benthos and (b) tropical rainforest case studies for sample site. Data are provided for totals, Nematoda and other Eukaryota (including OCTUs with BLAST hits to 'environmental samples' representing unclassified taxa).

In the terrestrial dataset, while the soil habitat had fewer nematode OCTUs (35) than either the litter (149) or canopy (97), the pattern was reversed for other eukaryotes, particularly for mites (soil: 179, litter: 6, canopy: 1) (**Figures 2.4 and**

2.5). As expected, plant-parasitic nematodes were more diverse and abundant in the soil environment, with bacterial- and fungal-feeding nematodes predominating in the litter and canopy. No omnivorous/predatory nematodes were observed in the soil. Recalling that the extraction methods used in the tropical rain forest study were optimized for nematode taxa, the diversity patterns regarding eukaryotes other than nematodes may be inaccurate. For instance, the nearly complete absence of mites and springtails in litter and canopy seems unrealistic. Also, extremely low recovery of fungal sequencing reads is unusual, but appears not be an artefact of primer conservation. It is likely that extraction methods, biology and/or competitive PCR interactions (von Wintzingerode *et al.* 1997) may have been the cause of the lack of fungal sequences in the terrestrial dataset.

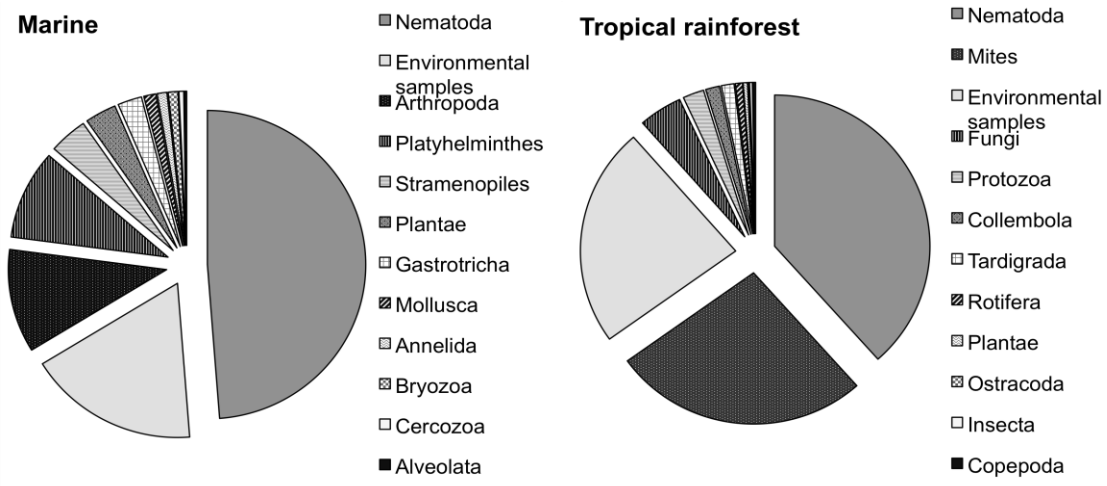


Figure 2.5- Pie chart illustrating the relative proportion of OCTUs (clustered at 97% similarity) belonging to each taxonomic grouping found in (a) the marine littoral benthos and (b) the tropical rainforest case studies. BLAST hits to 'environmental samples' represent unclassified taxa.

Perspective and future directions in eukaryotic environmental metagenetics

Ultrasequencing accompanied by BLAST annotation can clearly assist in the assessment of relative MOTU richness from large numbers of ecological samples. The increased throughput in sequencing afforded by new generation sequencers enables faster access (weeks rather than years) to larger amounts of data spanning the

breadth of the eukaryotic domain of life, at a fraction of the cost. Environmental metagenetics may therefore widen the identification bottleneck and increase the flow of information that is currently associated with the taxonomy and identification of smaller organisms in biodiversity assessments (Blaxter 2003a; Blaxter & Floyd 2003; Tautz *et al.* 2003). There are however, some fundamental limitations regarding the approach. Relating species to MOTUs will always be a contentious issue, but bioinformatic sequence comparisons and additional molecular assisted taxonomy will inevitably bridge the gap as further voucher specimens are linked to sequences (Blaxter 2003a; Blaxter *et al.* 2003; Markmann & Tautz 2005).

The data here refer explicitly to the relative number or richness (McIntosh 1967) of MOTUs and not diversity, that takes into consideration both richness and evenness (Good 1953; Hurlbert 1971; Magurran 2004). Since prokaryotes are unicellular and accepting the limitations of PCR-based approaches, prokaryote ecologists often make the assumption that numbers of reads reflect phylotype diversity (ie, one 16S sequence per individual prokaryote) in metagenetic datasets (Kemp & Aller 2004). The same assumption however, cannot be readily made for multicellular organisms that comprise different numbers of cells that will change in relation to developmental stage (especially using multicopy 18S markers). Very little work has been performed on the quantitative aspect of environmental metagenetics, but preliminary investigations suggest that number of reads may not correlate with small scale differential amounts of DNA template (Binladen *et al.* 2007), or numbers of individuals (Porazinska *et al.* 2009). Given the unequivocal need for quantitative assessments in biodiversity assessment, investigations that are more comprehensive will be required to test the quantitative nature of metagenetic datasets at a range of taxonomic levels. Further to this, trials of independent PCRs (standard and emPCR for Roche 454 sequencing) of the same samples will enable the assessment of replicability within datasets, regarding both quantification and chimera formation.

The need for homology in MOTU derivation

It is clear that single primer pair combinations will never coamplify all taxa and so it may be necessary to use primer cocktails (Ivanova *et al.* 2007), or more than one diagnostic region per study. The latter solution however, only becomes

comparatively meaningful if the multiple diagnostic regions can be linked to the same genome from which they were derived. Otherwise, the result is two unlinked MOTU derivations and independent BLAST annotations. We feel that the two regions featured here are optimal for eukaryotic metagenetic analyses and perhaps in the future, with ever increasing read lengths and paired-end methods, it may be possible to incorporate both in MOTU derivation. Further data comparisons will elucidate which may be optimal, but at least for the time being other researchers are urged to use the same locus and regions to enable comparative analyses, following a similar ethos to the barcoding movement.

To conclude, perhaps one of the most crucial issues of environmental metagenetics is to maintain a bioinformatic paper trail so that the scientific community will be able to BLAST annotate future queries and link MOTUs to already existing MOTUs. By linking independent datasets, there is the potential to facilitate the integration of all metagenetic datasets opening up the possibility of ecosystem-based approaches to a range of spatially and temporally heterogeneous evolutionary and ecological questions. We are investigating database mechanisms of how to achieve this and look forward to developments in the emerging field of eukaryotic environmental metagenetics.

CHAPTER 3 - Second-generation environmental sequencing unmasks marine metazoan biodiversity

Introduction

Biodiversity is the product of millions of years of evolution and forms the basis of earth's life support system, but the magnitude and relative diversity of global species richness remains unknown. On earth there may be over 100 million species (Blaxter 2003b) but fewer than 2 million have been formally described (May 1988). There is also a pronounced bias towards the study of larger organisms, leaving the most speciose communities that are dominated by microscopic organisms, understudied. In order to study diverse environments dominated by small taxa, second generation sequencing has been employed for the quantification of bacteria, archaea (Sogin *et al.* 2006; Huber *et al.* 2007) and viruses (Angly *et al.* 2006; Desnues *et al.* 2008) but large knowledge gaps still exist regarding the organization of diversity within several eukaryotic kingdoms, including the Metazoa. The Kingdom Metazoa, also known as Animalia, consists of multicellular heterotrophic organisms ranging from Porifera to Chordata. Contemporary phylogenetic studies routinely recover a monophyletic Bilateria – Triploblasta clade (including deep-branching Ecdysozoa, Lophotrochozoa and Deuterostomia) (Dunn *et al.* 2008; Philippe *et al.* 2009), but no consensus view exists of the precise relationships between the Bilateria, and the basal groups of Porifera, Ctenophora, Placozoa, Cnidaria and Coelenterata (Dunn *et al.* 2008; Philippe *et al.* 2009). Marine benthic metazoan communities display some of the highest α-diversity on the planet and occupy one of the largest ecosystems on earth, where only 1% of species are estimated to be known (Snelgrove 1999). Benthic meiofauna (small metazoans between 45 µm and 500 µm in size), comprise members encompassing 60% of animal phyla and represent a major part of marine biodiversity (Snelgrove 1999). Dominated by nematodes and characterized by high abundances (up to 10^8 individuals per $1m^2$) and diversity (Lambshead 2004), meiofaunal assemblages perform essential roles in marine ecosystem processes, namely nutrient cycling, secondary production, sediment transport and mineralisation (Giere 2009).

A metagenetic approach is applied (ie. the large-scale analysis of taxon richness via the analysis of homologous genes) using second-generation sequencing of the 18S nuclear small subunit (nSSU) ribosomal RNA (rRNA) gene to assess simultaneously the relative levels of richness and patterns of diversity of multiple metazoan phyla across an ecological gradient in a temperate benthic ecosystem. The

heterogeneous levels of accumulating taxon richness derived from the benchmarked analyses were broadly congruent with those derived from intensive morphological assessments, but MEGABLAST annotation revealed a previously unidentified phylogenetic breadth of microbial metazoan life. Moreover, the finding that the largely predacious turbellarian Platyhelminthes represent a substantial proportion of benthic diversity quantifies their hitherto unrecognised ecological importance in benthic food chains. Annotated metagenetic analyses enable the objective assessment of microbial biodiversity throughout all ecosystems, facilitating understanding of linkages between microbial biodiversity and ecosystem processes.

Methods

Sample collection

Twenty four benthic samples were collected from the low tide mark along an 800 m transect, using a standard corer methodology (Platt & Warwick 1988) from marine sandy beach substrata in Prestwick (three every 100 m between 55º30.481'N, 4º37.489'W and 55º30.194'N, 4º37.368'W) and a further three from Littlehampton (50º48.021'N, 00º32.530'W), UK, during summer 2007. The latter sampling site was used as a geographically disparate out-group comparison. For the sequencing analysis, each biological sample comprised three pooled 44 mm diameter x 100 mm benthic samples taken approximately 10 m apart. An additional core was taken for sediment analysis. All samples were immediately fixed in 500 ml storage pots containing 300 ml of DESS (20% DMSO and 0.25 M disodium EDTA, saturated with NaCl, pH 8.0) (Yoder *et al.* 2006). The meiofaunal size fraction was mechanically separated from the sand and concentrated by decanting five times with filtered tap water through a 45 µm filter. Subsequent separation from fine silt was achieved by repetitive centrifugation in 1.16 specific gravity (sg) LUDOX-TM solution (de Jonge & Bouwman 1977). Following centrifugation, each sample was retained on a distinct mesh sieve which was then folded, sliced and placed in a 15 ml falcon tube and kept at -80°C until DNA extraction. Samples were lysed overnight at 55°C in lysis buffer (100 mM Tris-HCl, pH7.5; 100 mM NaCl; 100 mM EDTA; 1% SDS, 500 µg/ ml proteinase K), assisted by spinning wheel mixing, and DNA

extracted with the QIAamp DNA Blood Maxi Kit (Qiagen) following the manufacturer's protocol.

Primer design and PCR strategy

The genes coding for ribosomal RNA have been used for decades for the identification of microbial species (Doi & Igarashi 1965; Pace & Campbell 1971) and are well suited for taxonomy. Mainly because they are ubiquitous in cellular organisms; are of relatively large size; contain highly conserved and variable regions that facilitate primer design; and are present in tandemly repeated, multiple copies enabling efficient PCR amplification (Floyd *et al.* 2002; Markmann & Tautz 2005; Carvalho *et al.* 2010). For environmental metagenetic discovery, the choice of the nuclear small subunit (nSSU) makes sense due to the considerable more reference sequences are available in public databases for nSSU (1,246,462) than the large subunit (180,344) (Pruesse *et al.* 2007). Moreover, 'universal' nSSU primers have been shown to amplify more taxa from mock communities than those designed for the LSU (Porazinska *et al.* 2009). The 5' region of the nSSU exhibits more segregating sites than the 3' region in Metazoa (Floyd *et al.* 2002) and so was selected as the target area for OCTU discrimination. MEGA version 4.1 (Tamura *et al.* 2007) was used to align and compare a wide range of metazoan nSSU sequences with respect to existing degenerate primers and putative new priming sites spanning a region between 250-500bp in length (a combination of average mean read length and maximum permitted amplicon size respectively, of a 454 Life Sciences, Roche Applied Science GSFLX amplicon sequencing run). Of all primer permutations, the SSU_FO4 (5'-GCTTGTCTCAAAGATTAAGCC-3') and SSU_R22 (5'-GCCTGCTGCCTTCCTTGGA-3') (Blaxter *et al.* 1998) primers were selected as they exhibited pronounced homology across meiofaunal phyla, but also flanked a highly divergent region of the nSSU and were used for subsequent amplicon generation. Fusion primers were then developed in which a proprietary primer sequence (Adaptor A or B) of the Roche 454 GSFLX sequencing technology and a sample-specific five nucleotide key tag (to differentiate between multiple samples) were included (Binladen *et al.* 2007). All fusion primers were designed using PRIMER PREMIER 5.0 (Premier Biosoft International, Palo Alto, Calif.),

considering physical and structural properties of the oligonucleotides (such as annealing temperature, G+C percentage, hairpins and false priming). The forward fusion primers were ca. 45 bases in length and designed such that the 454 A-adaptor is followed by the tag (each of which differed by at least two bases) (Binladen *et al.* 2007), and then by the experimental forward primer (SSU_FO4). The reverse primers were designed similarly, to optimise thermal compatibility. PCR amplification of the specified nSSU region was performed using 1 µl of genomic DNA template (1:500 dilutions) in 3 x 40 µl independent reactions using *Pfu* DNA polymerase (Promega). PCR conditions involved a 2 min denaturation at 95 °C, then 35 cycles with 1 min 95 °C, 45 s 57 °C, 3 min 72 °C and final extension of 10 min at 72°C. Negative controls (ultrapure water only) were included for all amplification reactions. Electrophoresis of triplicate PCR products was undertaken on a 2% gel with Top Vision™ LM GQ Agarose (Fermentas), and the expected 450 bp fragment was purified using the QIAquick Gel Extraction Kit (Qiagen) following the manufacturer's instructions, prior to pooling identical samples. All purified PCR products were then quantified with an Agilent Bioanalyser 2100, diluted to the same concentration, pooled together to create a single sample and sequenced in one direction (A-Amplicon) on half a plate of a Roche 454 GSFLX (454 Life Sciences, Roche Applied Science) sequencing platform at Liverpool University's Centre for Genomic Research, UK.

Sediment analysis

Particle size analysis was carried out using a Malvern Mastersizer 2000, which uses laser diffraction to calculate the particle size distribution for individual sediment grains in the range 0.02-2000 µm. To prevent flocculation, prior to testing, the samples were immersed for 24 hours in distilled water with added dispersant (sodium hexametaphosphate). The Mastersizer determines particle size distribution by volume and the results of the particle size analysis are reported as the cumulative median grain size.

Data Analysis

Sequences generated from the Roche 454 GSFLX pyrosequencing were analysed using the Operational Clustering of Taxonomic Units from Parallel Ultrasequencing pipeline (OCTUPUS, supplementary software available at http://OCTUPUs.sourceforge.net/). Briefly, OCTUPUS comprises a number of Perl scripts that concatenate quality trimming (Chou & Holmes 2001), tag matching and size culling, prior to the assignation of user defined substitutional difference based cut-off clustering. The clustering module of OCTUPUS involves three steps. Initially, sequences are compared successively to each other by MEGABLAST (Zhang *et al.* 2000) to define different OCTU groups, separated by a user-defined genetic distance. If an unassigned sequence matches an existing OCTU sequence (e.g. 97% similarity or more), the companion sequences are aligned using MUSCLE (Edgar 2004) and if the resulting consensus sequence differs from the original alignment, the consensus OCTU sequence is changed to reflect the diversity of sequences within the OCTU cluster. If the consensus OCTU does not change following a pre-set number of novel comparisons, further consensus OCTU matching sequences will be placed within the OCTU cluster, bypassing a computationally intractable number of multiple sequence alignments that would otherwise be required in the analysis of large metagenetic datasets. Benchmarking trials have shown that ten additions to the OCTU cluster without consensus amendment provides a stable estimate of OCTU numbers and was used throughout here. OCTUs were annotated using MEGABLAST (megablast -d database path -D 2 -p 90 -a 2 -b 1 -v 1 -i infile -F F > outfile) against the downloaded GenBank/EMBL/DDBJ nucleotide database and taxonomic annotation was restricted to matches of 90% and higher.

Acknowledging concerns regarding the misinterpretation of levels of richness due to the formation of recombinant DNA molecules in environmental DNA sequencing (von Wintzingerode *et al.* 1997; Reeder & Knight 2009) an aggressive putative chimera-culling regime embedded within the OCTUPUS pipeline for the primary dataset was adopted. The OCTUPUS chimera screening was followed by manual removal of putative false positive OCTUs (i.e. those that exhibited more than 10 consecutive base pair mismatches with the MEGABLAST reference sequence) and retrieval of clear false negative chimeras (i.e. those that exhibited 100% length matches with already sequenced taxa). In order to compare independent estimates of

taxon richness, the OCTUPUS clustering algorithm was also tested against the HCluster algorithm of ESPRIT (Sun *et al.* 2009) on a subsample of the larger dataset. Finally, OCTU richness generated from the above OCTUPUS algorithm was benchmarked at a range of percentage similarity cut-off against a reference dataset comprising the metagenetic analysis of a phylogenetically diverse combination of 41 nematode (Porazinska *et al.* 2009) species. Scripts for preprocessing the above data are available from the Natural Environment Research Council's Environmental Bioinformatics Centre Script Repository (http://nebc.nerc.ac.uk/tools/scripts/general-bioinformatics). For the 54 OCTUs exhibiting BLAST hits below 90% identity, manual MEGABLAST searches were performed and phylogenetic affinities investigated via the NCBI taxonomy browser. Sequence data and all associated fusion primer codes have been deposited in the GenBank/EMBL/DDBJ short read archive under accession number SRA009394.2/PWick_LHampton_2007. For direct ecological comparisons of between sample OCTU richness, the original dataset was reanalysed using 15,000 randomly picked sequences (over 200 bases in length, n= 135,000) from each sample, prior to OCTUPUS clustering and annotation. The total number of OCTUs generated from the original and standardized datasets was significantly correlated (Spearman's coefficient: $\rho = 0.783$, $p = 0.0132$), nevertheless, interpretations derived from direct comparisons of richness between samples refer to the standardized dataset. Sample cluster analyses (UPGMA) were performed using the Multivariate Statistical Package (Kovach 1999) using Sorensen's Coefficient on a binary (presence/absence of OCTU) data matrix. Phylum-specific rarefaction curves for the Prestwick transect were generated using EstimateS (Version 8.2.0, R. K. Colwell) using a range of estimators (e.g. ACE, Chao1, Jackknife1 and Bootstrap) that yielded very similar results. The ACE abundance-based coverage estimator (Chazdon *et al.* 1998) was used because it represents a consensus view of the analyses and has proven to work well for the analysis of metagenetic datasets (Huber *et al.* 2007).

Results

Taxon richness estimates

Nuclear small subunit rRNA amplicons were generated from eight benthic samples from the low tide zone of an estuarine beach near Prestwick on the West coast of Scotland, and from one sample from a beach in Littlehampton in the South of England. The amplicons were processed for sequencing on the Roche 454 FLX platform generating a total of 353,896 sequences that were quality filtered to 305,702 for downstream analysis (**Table I**).

Table I. Total number of pyrosequencing sequences after quality control and chimera screening. Data are shown for reads passing the initial quality trimming, tag matching and size culling steps (QC tag reads) and reads underpinning OCTUs that were estimated to be non-chimeric (Chimera check reads) from the Prestwick and Littlehampton (LH) sample sites.

Samples	Prestwick								LH	TOTALS
Sample	1	2	3	4	5	6	7	8	1	
QC tag reads	27206	23877	21042	19031	49476	39054	35794	54227	35995	305702
Chimera check reads	13758	14774	13534	12856	42741	23932	29641	40234	25751	217221

When performing metagenetic assessments of taxon richness, it is important to cluster taxonomic units in a biologically meaningful fashion, since even slight differences in similarity cut-offs and using different algorithms can result in significantly different estimates of richness. The taxon clustering comparisons between ESPRIT (Sun *et al.* 2009) and OCTUPUS (the Operational Clustering of Taxonomic Units from Parallel UltraSequencing, supplementary software available at http://octupus.sourceforge.net/) (**Figure 3.1**) on the subsampled beach pyrosequencing data show that ESPRIT had higher-estimates than OCTUPUS (between 1.1x-4.4x, over the 90-99% cut-off range).

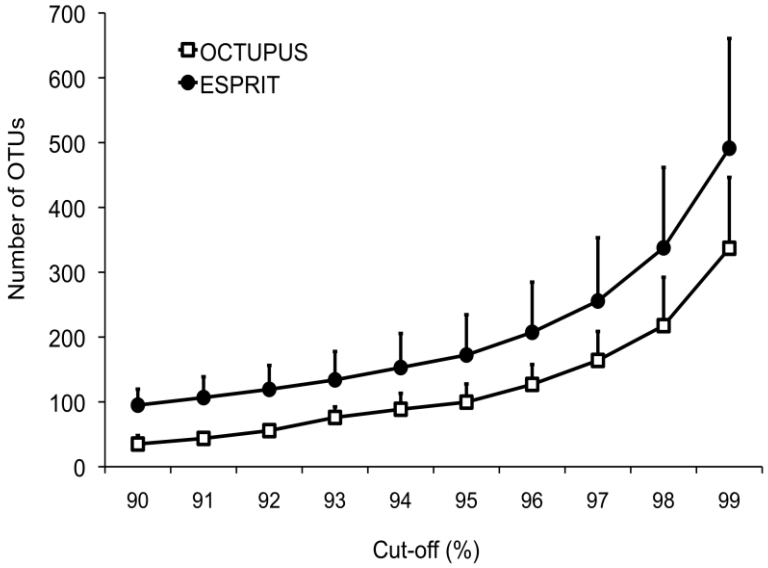


Figure 3.1. Lineage-through-time plots for OCTUPUS and ESPRIT. Mean number of OTUs plotted against each percentage identity cut-off (90-99% similarity) generated using 5000 subsampled sequences (>199 bases in length) from three independent sample sites (Prestwick 2, 7 and Littlehampton 1) using OCTUPUS (squares) and ESPRIT's HCluster (Sun *et al.* 2009) (circles) OTU clustering. Values are given as average ± s.d. (n= 9).

Phylogenetic (Blaxter *et al.* 1998) and barcoding (Bhadury *et al.* 2006) studies based on the analysis of chain-termination nSSU gene sequences suggest that intraspecific divergence at least in nematode species is low (1-2%). However, the true level of intragenomic and intraspecific variation among rRNA gene repeats is largely unknown and genome wide analyses reveal a dominant set of conserved sequences accompanied by rare variant sequences (Stage & Eickbush 2007). Our benchmarking exercise, performed against a reference control dataset comprising 41 species (Porazinska *et al.* 2009) revealed that the 96% similarity OCTUPUS clustering algorithm with accompanying chimera screening estimated taxon richness that was most closely aligned with species richness. At 96% similarity, OCTUPUS resulted in 37 operational clustered taxonomic units (OCTUs), whereas at 97% OCTUPUS resulted in 51 OCTUs from the control metagenetic analysis (Porazinska *et al.* 2009). Thus, although richness may be underestimated by at least 10%, a more conservative approach was adopted by setting a 96% identity OCTU cut-off for all subsequent numerical comparisons. At this cut-off, an OCTU is likely to (at worst) correspond to a group of related species. Following the 96% similarity OCTUPUS

clustering strategy, the total number of putatively non-chimeric tag reads and OCTUs was 217,221 and 428, respectively. Prior to chimera screening, 1013 OCTUs were clustered from the initial quality screened 305,702 reads.

Community richness is closely linked to the environment

The peak of standardized OCTU richness for all phyla was within samples 6, 7 and 8 from Prestwick, and cluster analyses indicated clear and fine scale hierarchical distinctions in OCTU composition within and between the two sites, with clear divergence of samples from Littlehampton (**Figure 3.2**). OCTU richness and sediment grain size were positively correlated (Spearman's correlation coefficient, $n = 9$, $\rho = -0.82$, $p = 0.0108$).

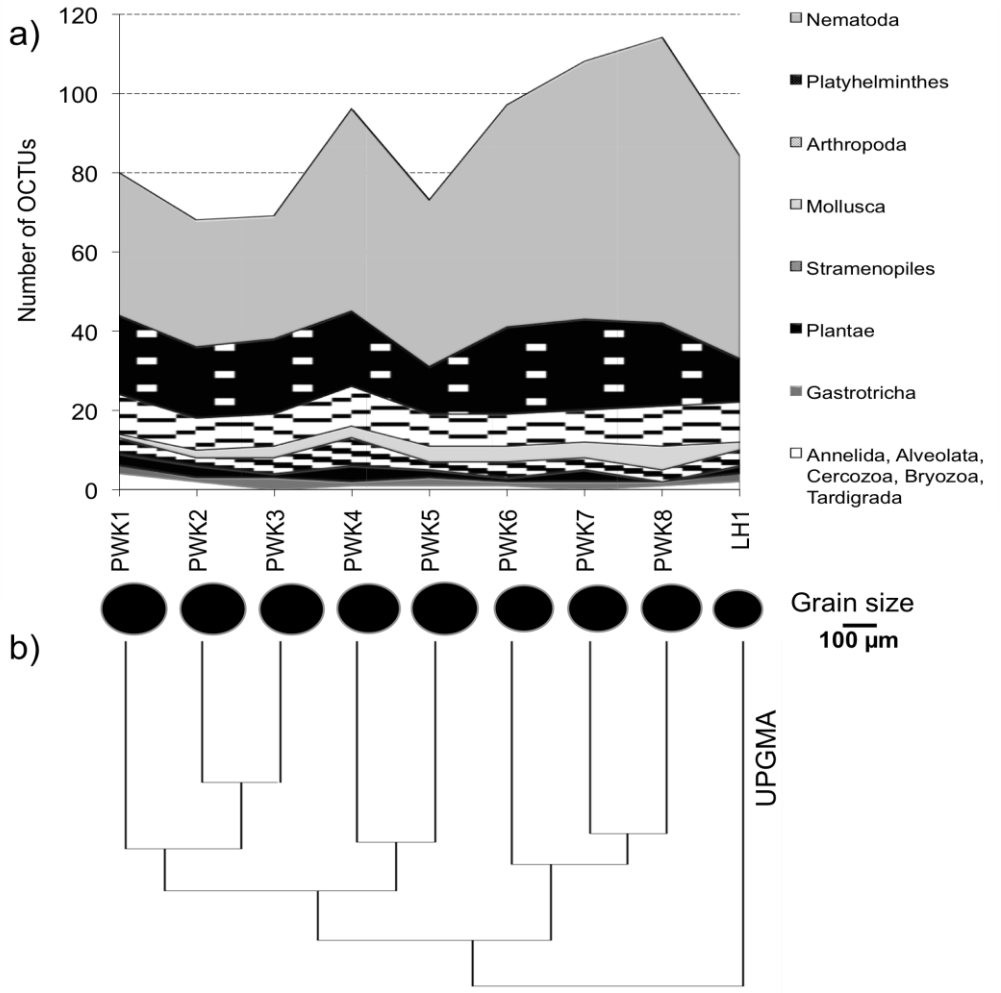


Figure 3.2. Taxon richness and community similarity in relation to ecology and space. (a) Number of different OCTUs per sample for each phylum after data standardization derived from the Prestwick (8 sampling sites) and Littlehampton (1 sampling site) marine littoral benthos; (b) grain size represents the relative 50% cumulative median grain size (µm) per site, and cluster analyses (UPGMA) using Sorensen's Coefficient represent the number of shared OCTUs between the nine independent samples.The positive relationship between grain size and sample richness is highly significant (Spearman's correlation coefficient, n=9, $\rho = -0.83$, $P = 0.0108$).

Dominance of the Nematoda and rise of the Platyhelminthes

Plotting phylum richness rank for all nine independent samples (**Figure 3.3**) shows that the Nematoda are the most OCTU-rich in all nine samples, with Platyhelminthes and Arthropoda ranking second and third in all, respectively.

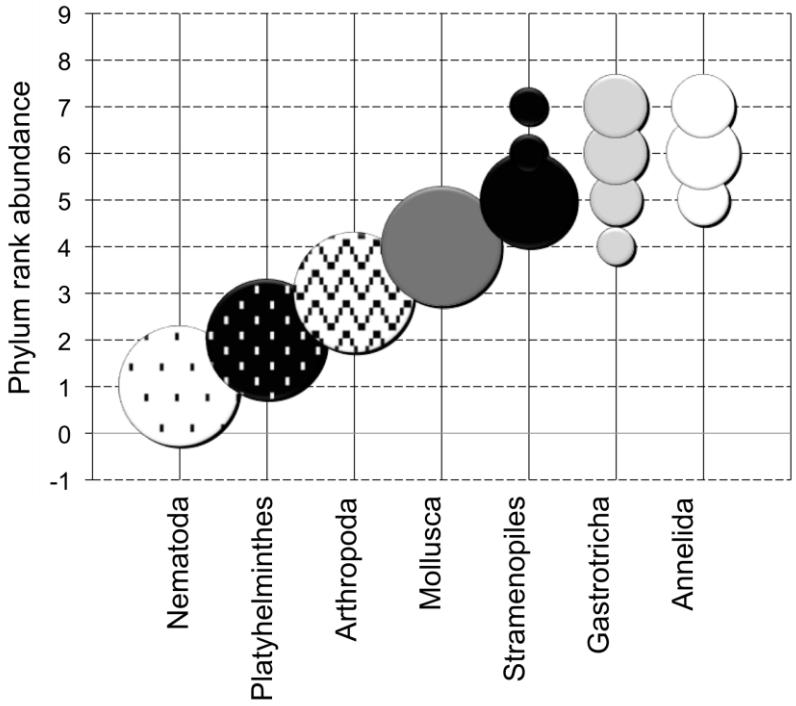


Figure 3.3. Phylum rank abundance plot. Community assemblage OCTU richness rank order for the main phyla recovered from the Prestwick and Littlehampton samples (after data standardization). The frequency of ranking (out of the 9 samples) is represented by the diameter of the symbol at each rank. Single symbols per phylum represent a constant ranking, whereas multiple symbols highlight variance in phylum rank order throughout the samples.

Annotated metagenetic analyses can yield robust relative richness estimates and here it was possible to assign 374 OCTUs to phylum (**Figure 3.4**). The PCR primers used are not fully specific to Metazoa, and thus other nSSU genes were sequenced from protist taxa from the Alveolata (2 distinct OCTUs), Cercozoa (3 OCTUs) and stramenopiles (15 OCTUs). Of the metazoan OCTUs, 182 were from Nematoda, at least three times more than from any other individual meiofaunal taxon (**Figure 3.4**). Platyhelminthes (61 OCTUs) was the second richest phylum, followed by the Arthropoda (29 OCTUs including Copepoda, Ostracoda and Malacostraca), Mollusca (22 OCTUs), Gastrotricha (7 OCTUs), Annelida (6 OCTUs), and five less rich phyla (e.g. Bryozoa, Echinodermata, Cercozoa, Rotifera and Alveolata with between 1-3 OCTUs each).

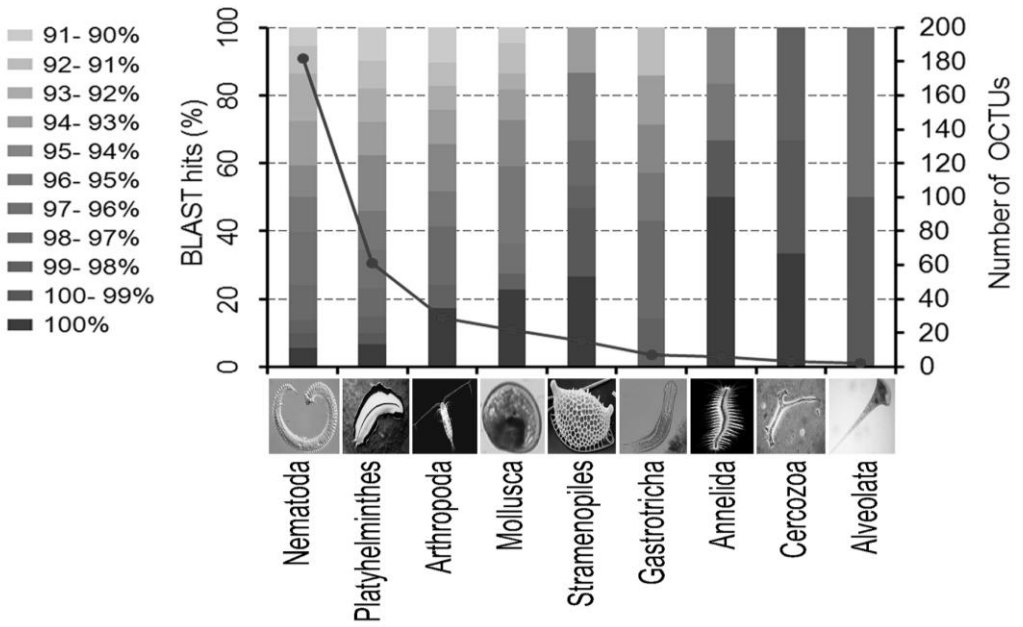


Figure 3.4. Percent identity to known sequences and number of OCTUs found for the main phyla. Number of different OCTUs for the main phyla recovered from the Prestwick and Littlehampton meiofaunal samples and their different levels of identity to nSSU in the GenBank/EMBL/DDBJ nucleotide database.

Metagenetics reveals phylogenetically distinct taxa

According to the comparisons of the OCTU sequences with the NCBI database the majority (95%) of Nematoda OCTUs have never been sequenced before (**Figure 3.4, Table II**).

Table II. OCTUs with 100% identity to nSSU sequences in the GenBank/EMBL/DDBJ database.

OCTU number	Total sequences	Accession Number	Phyla	Class	Species
5	28217	AY854224.1	Nematoda	Chromadorea	*Anaplectus sp.*
43	1773	AY854198.1	Nematoda	Enoplea	*Viscosia viscosa*
153	133	AY854217.1	Nematoda	Chromadorea	*Spirinia parasitifera*
320	2239	AY854224.1	Nematoda	Chromadorea	*Daptonema normandicum*
378	13	AY854209.1	Nematoda	Chromadorea	*Dichromadora sp.*
177	244	AY854226.1	Nematoda	Chromadorea	*Daptonema setosum*
317	115	FJ040466.1	Nematoda	Chromadorea	*Sabatieria pulchra*
381	728	AY284693.1	Nematoda	Chromadorea	*Theristus agilis*
3	66604	GU139755.1	Nematoda	Enoplea	*Mesacanthion sp.*
15	5798	AY284695.1	Nematoda	Chromadorea	*Theristus agilis*
233	543	AY775746.1	Platyhelminthes	Turbellaria	*Schizochilus choriurus*
40	699	AY775766.1	Platyhelminthes	Turbellaria	*Proxenetes quadrispinosus*
18	794	AJ012508.1	Platyhelminthes	Turbellaria	*Diascorhynchus rubrus*
25	15111	AJ012531.1	Platyhelminthes	Turbellaria	*Paromalostomum fusculum*
4	23256	AM774524.1	Mollusca	Bivalvia	*Angulus tenuis*
127	43	DQ279943.1	Mollusca	Bivalvia	*Chamelea striatula*
146	337	DQ640516.1	Mollusca	Bivalvia	*Mytilus trossulus*
109	7	EF526454.1	Mollusca	Bivalvia	*Modiolus modiolus*
658	6	DQ093437.1	Mollusca	Gastropoda	*Littorina littorea*
415	28	AF448164.1	Annelida	Polychaeta	*Scolelepis squamata*
253	12	AF508122.1	Annelida	Polychaeta	*Ophelia bicornis*
176	90	AF508125.1	Annelida	Polychaeta	*Protodriloides symbioticus*
184	81	AY446901.1	Arthropoda	Maxillopoda	*Centropages hamatus*
307	29	EU868740.1	Arthropoda	Malacostraca	*Crangon crangon*
86	27	L81939.1	Arthropoda	Maxillopoda	*Calanus pacificus*
446	13	GU594643.1	Arthropoda	Maxillopoda	*Oithona sp.*
557	5	AB087108.1	Stramenopiles	Phaeophyceae	*Dictyota linearis*
186	24	AB178865.1	Stramenopiles	Oomycetes	*Haliphthoros sp.*
851	4	EU818944.1	Stramenopiles	Bacillariophyta	*Odontella aurita*
515	4	AJ519935.1	Stramenopiles	Labyrinthulida	*Aplanochytrium stocchinoi*
391	4	AY620355.1	Rhizaria	Cercozoa	*Uncultured cercozoan*
964	1	DQ073794.1	Echinodermata	Echinoidea	*Spatangus raschi*

Similarly, only 6.5% of Platyhelminthes and 17.2% of Arthropoda OCTUs had 100% identity to previously sequenced specimens. The Annelida, Mollusca and Stramenopiles, however, had 50%, 23% and 26.6%, respectively, of their OCTUs with a 100% identity to previously sequenced taxa. None of the Gastrotricha OCTUs were identical to previously sequenced taxa. Overall, 54 OCTUs (representing 7247 individual sequences) had identities below 90% to any reference nSSU sequence (OCTUs<90% identity), and may represent previously unsampled diversity. Given the nature of the adopted clustering algorithm, the OCTUs<90% identity do not show any pattern of variation that would be expected of a chimeric assemblage of the already defined 428 OCTUs. Fifteen of the OCTUs<90% identity (300 sequences)

were robustly placed within identified phyla (mainly Nematoda). A further 36 were either sisters to the sequences from known phyla, or represented deep, likely misplaced branches (**Figure 3.5**).

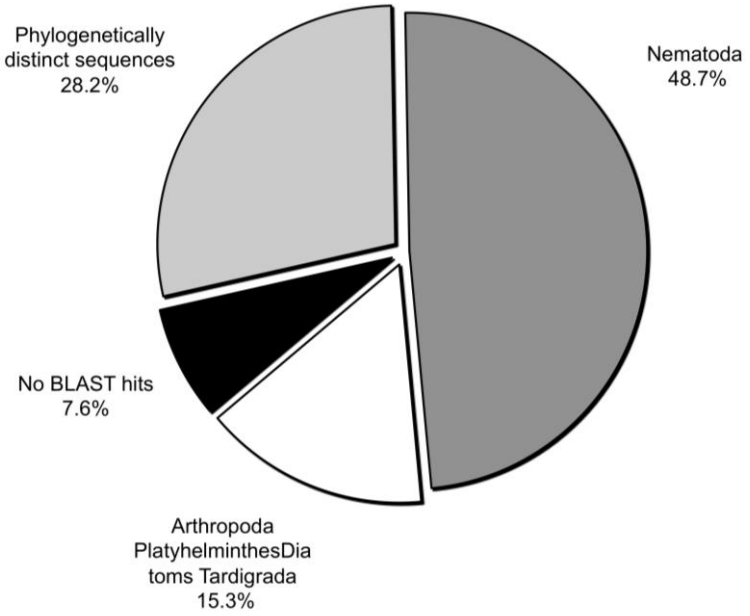


Figure 3.5. Putative taxonomic classification of 39 OCTUs that had less than 90% sequence identity to nSSU in GenBank/EMBL/DDBJ.

Surprisingly 11 OCTUs<90% identity, comprising 107 sequences in total distributed across the sampling area were not affiliated to any phyla. The similarity of the latter sequences ranged between 83-89% to known nSSU genes, suggesting that they may represent unknown, distinct lineages. Furthermore, three OCTUs resulted in no MEGABLAST hit, but two of these exhibited signatures of relaxed selection, indicating that they may represent translocated "orphan" nSSU genes, that are no longer constrained by function. Resampling to generate longer sequences beyond the 200 base pairs presented here will be required to confirm or refute the biological reality of these phylogenetically distinct sequences.

Metazoan richness curves do not approach saturation

Along the Prestwick transect, the slope of OCTU rarefaction curves at 96% cut-off for the Nematoda, Platyhelminthes, Arthropoda, Mollusca, Stramenopiles and Annelida phyla did not reach an asymptote (**Figure 3.6**). Thus, still 217,221 samples and 374 OCTU-defined taxa fail to achieve saturation, even for low abundance phyla where rarefaction curves tend to converge (Tipper 1979).

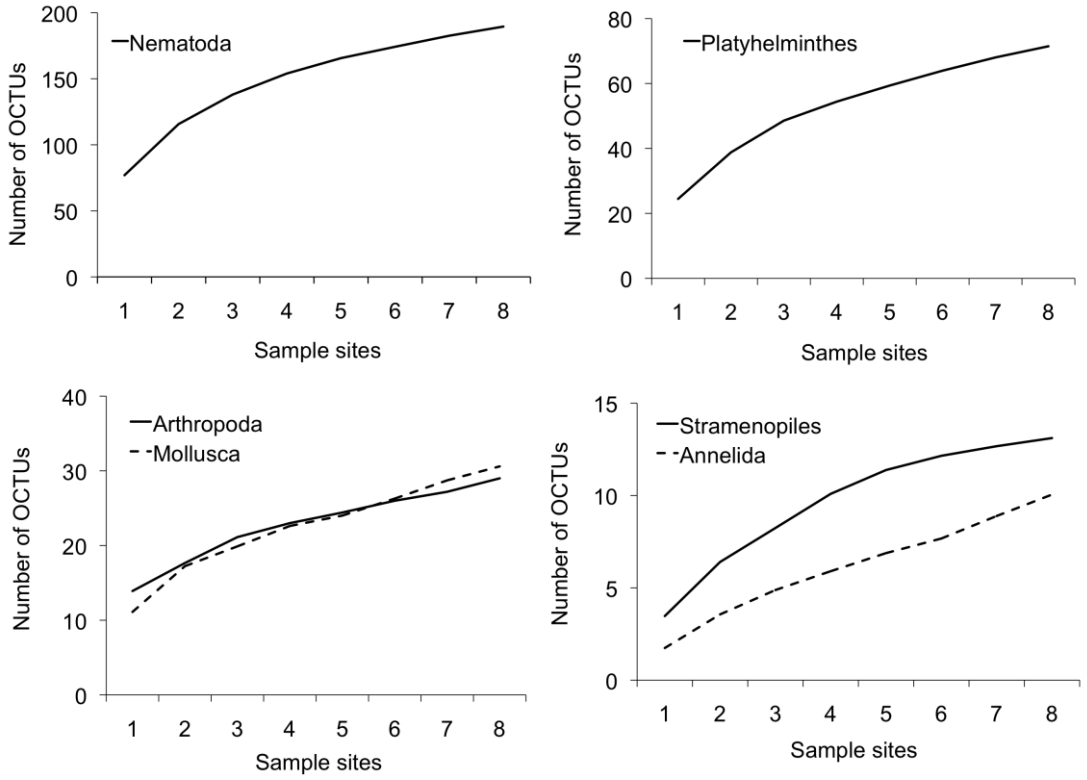


Figure 3.6. Rarefaction curves of the abundance-based coverage estimation (ACE) diversity estimator. Plots are shown for the (a) Nematoda, (b) Platyhelminthes, (c) Arthropoda (blue) and Mollusca (red dashed), (d) Stramenopiles (blue) and Annelida (dashed red) at 96% identity OCTU cut-off for the Prestwick meiobenthic samples 1-8. Curves were estimated from 100 randomizations, without replacement, using EstimateS, version 8.2.0.

Discussion

Both ESPRIT (Sun *et al.* 2009) and OCTUPUS can be used to cluster OTUs from the size of the environmental metagenetic dataset presented here, without the need of a computing cluster. Nonetheless, ESPRIT may overestimate OTU richness, presumably because it counts both substitutional and independent insertion/deletion (indel) events in OTU clustering. One of the most widely publicized artefacts of 454 Roche pyrosequencing is its inherent inaccuracy in calling homopolymer runs, including extensions (insertions), incomplete extensions (deletions) and carry forward errors (insertions and substitutions) (Margulies *et al.* 2005; Huse *et al.* 2007; Kunin *et al.* 2009). Therefore, clustering algorithms that use indel data informatively, especially as independent events at homologous base positions are likely to be susceptible to higher OTU estimates. Analyses herein, and those of Sun *et al.* (Sun *et al.* 2009) suggest that OCTUPUS generates a conservative number of OTUs compared to both ESPRIT and combinations of MUSCLE (Edgar 2004) and DOTUR (Schloss & Handelsman 2005). Nonetheless, it is imperative when possible to benchmark the analytical framework against real or simulated data, in order to estimate taxon richness as realistically as possible (Kunin *et al.* 2009).

The patterns of taxon richness along the Prestwick transect exhibited remarkable differences even at a micro-geographic scale. Although numerous environmental factors influence meiobenthic distribution and assemblage, grain size, examined here, is known to be the predominant driver of meiofaunal community structure and diversity (Giere 2009). Nonetheless, the fine scale community structuring also indicates that there are likely to be a host of additional biotic (e.g. prokaryote communities and organic matter) and abiotic (sediment grain shape, surface composition) micro-geographical factors responsible for community structuring within the benthos (Giere 2009).

Nematodes account for approximately 80% of all individual animals on earth (Bongers 1988), but quantifying the levels of relative richness between major metazoan phyla has not been practical hitherto, due to the taxonomic magnitude of the challenge. Remarkably, along only an 800 m transect, 182 Nematoda OCTUs were found, compared with 450 species of nematode that have been described from around the entire British Isles (Warwick *et al.*, 1998). From a geographical

perspective, these data represent the discovery of 40% of the previously known phylum richness from a transect that represents 0.004% of the length of the British coastline (ca. 17,820km, Ordnance Survey). Molecular definition of substantial meiofaunal taxon richness at micro-geographic scales are in alignment with intensive morphological studies (Warwick *et al.* 2006), yet recent metagenetic studies from the microbial biosphere have suggested that taxon richness may be orders of magnitude more than previously expected (Angly *et al.* 2006; Sogin *et al.* 2006; Huber *et al.* 2007; Desnues *et al.* 2008). Reeder & Knight (2009), Quince *et al.* (2009) and Huse *et al.* (2010) have recently provided alternative interpretations, based on pyrosequencing error, DNA recombination and clustering approaches, for earlier studies reporting extensive richness and rare biospheres derived from Roche 454 pyrosequencing. The data here has been quality/size filtered, only substitution variance has contributed to taxon identification and a particularly aggressive chimera screen has been applied to featured OCTUs. In combination with the benchmarking exercise performed against known taxa, it is therefore highly unlikely that the delineated taxa are based on sequencing error and taxon estimates are likely to underestimate richness by at least 10%. Although global nematode diversity estimates range from 1 to 100 million species, only 24,000 species have been described (Platt & Warwick 1983) with approximately 2000 marine species catalogued in Europe (SMEBD 2009). Such relative paucity of taxonomic knowledge of nematode biodiversity is a product of their numerical dominance, the number of taxa, and the high level of cryptic species: this taxon is thus a prime candidate for taxonomic exploration using metagenetic analyses. Why are Nematoda so dominant, in both abundance and diversity, in meiofaunal communities? Benthic nematodes are small in size and possess rapid generation times, and these attributes may have facilitated rapid adaptation to local conditions and habitats, especially in interstitial niches. The taxonomic richness of nematodes is additionally likely promoted by ancient ancestry (Aguinaldo *et al.* 1997) and the absence of a dispersal phase, promoting speciation across structurally and spatially heterogeneous environments during glacial and interglacial periods. Nematodes also have cryptobiotic adaptations, including resting eggs and highly resistant larvae (Jonsson 2005), and they are considered to be especially resistant to environmental challenges, being among the last taxa to disappear in an environmental catastrophe (Boucher & Lambshead 1995). As no other meiofaunal phylum shares all such characteristics

simultaneously, nematodes may have had the time and the potential to diversify, endowing them with collectively optimal traits for domination of the marine benthos. Considering all metazoan taxa, the detection of 428 OCTUs on only two sites is likely a gross underestimate of actual metazoan richness. 70% of Nematoda OCTUs were unique to Prestwick and 58% to 100% of the OCTUs for the other phyla were only present in Prestwick. In the absence of large-scale dispersal events sustaining ubiquitous meiobenthic communities (Giere 2009), levels of marine meiobenthic α-diversity, at the very least, in other parts of the world are likely to exceed expectations.

Contrary to most morphological assessments of marine meiobenthic richness (Heip *et al.* 1988; Warwick *et al.* 2006), Platyhelminthes were consistently ranked as the second richest phylum in all the investigated samples. Platyhelminthes normally have to be identified from living or well-preserved material and their low ranking in other surveys perhaps arises from a global lack of taxonomic expertise, from extensive crypsis, and from delicate body structures that do not survive harsh extraction methods. Given that the majority of free-living turbellarian Platyhelminthes encountered here occupy a top predacious role in benthic ecosystems (Reise 1988), the empirically demonstrated prominence of this phylum proves that conventional diversity assessments provide a notably distorted perspective of trophic relationships within the benthos (Giere 2009). Only 6 polychaete annelid OCTUs were identified, despite their dominance of benthic infaunal biomass (Warwick *et al.* 2006; Giere 2009), probably because biomass dominance is affected by non-meiofaunal taxa (with only reproductive propagules and larvae being sampled here). Similarly, molluscan OCTUs comprised 21 bivalves and a single gastropod (*Littorina littorea*), likely representing larval and juvenile stages that are temporary members of the benthic meiofauna. Here, a large percentage of OCTUs that have never been sequenced before were also uncovered, despite the nSSU region sampled here being the most frequently used genomic region for meiofaunal barcoding studies (Floyd *et al.* 2002; Blaxter *et al.* 2005). Some of our metagenetic sequences were clearly attributable to phyla, but were not closely related to previously sequenced taxa, possibly uncovering a high level of phylogenetic novelty across a micro-geographic scale that has not been shown from morphological or conventional DNA barcoding studies (Bhadury *et al.* 2006) of the meiobenthos.

Quantifying biodiversity in species-rich environments raises questions of whether the sampling effort was sufficient to fully represent the natural community. Here, despite an extensive sampling strategy the slope for OCTU rarefaction curves remained incomplete and the true standing diversity of an unremarkable Scottish beach is likely to exceed this estimate by an undetermined number of taxa. Such observations illustrate clearly the likely extent of marine meiofaunal diversity, and importantly reveal that extensive sampling combined with ultrasequencing technologies make realistic estimates of genetic diversity tractable in one of the largest ecosystems on earth.

The present data offer novel insights into the organization, magnitude and identity of meiofaunal richness, but the mechanism of taxon delineation here warrants additional focus. In a multi-site terrestrial survey, 60 times more traditional taxonomic scientific effort had to be expended in assigning only 10% of meiofaunal taxa to known species, compared to parallel studies that successfully assigned all vertebrate morphospecies to known taxa (Lawton *et al.* 1998). Lambshead (Lambshead 1986), using morphological taxonomy, spent three years describing 113 nematode morphospecies from the Firth of Clyde region (including the Prestwick sampling site), from approximately one fifth of the volume of sediment analysed here. Our study identified 182 Nematoda OCTUs in approximately 3% of the time. Moreover, whereas the expertise of morphological taxonomists is of necessity restricted to certain groups, high-throughput nSSU sequencing enables the simultaneous OCTU delineation of multiple phyla from large numbers of samples. Importantly, metagenetic analyses can also be linked retrospectively to morphospecies by using reverse taxonomy (Markmann & Tautz 2005). The approaches described provide a rapid, objective and cost-effective framework for exploring links between phyletic diversity, ecosystem structure and function. Such advances promise to impact substantially on our ability to elucidate and predict the relationship between biodiversity and ecosystem function (Solan *et al.* 2004; Worm *et al.* 2007; Danovaro *et al.* 2008a): issues central to notions of resilience, recovery and sustainability (Palumbi *et al.* 2008).

CHAPTER 4 – Marine meiofauna biodiversity distribution patterns and estimates

Introduction

Biodiversity is the main biological focus of habitat quality that represents both intraspecific variation (genetic diversity) and community variation (richness, abundance and evenness of species) at different spatial scales (Whittaker 1972; Jost 2007). Distribution patterns are relatively well understood in macroscopic organisms such as plants and animals, but are strongly debated in small-bodied taxa (Lambshead 1993; Platt 1994; Lambshead & Boucher 2003) such as eukaryotic meiobenthic fauna. The meiofaunal size fraction comprises communities of small sized organisms (50 – 500µm), that are essential to the ecosystem functioning of benthic ecosystems due to their high abundance, diversity, fast turnover rates (Platt & Warwick 1988), role in nutrient cycling (Gee & Warwick 1996; Warwick *et al.* 2006) and mineralization processes (Gheskiere *et al.* 2004).

The description and interpretation of large-scale ecological processes and patterns, defined as macroecology (Brown 1995), of marine biodiversity is still controversial and subject to many discussions (Heip *et al.* 1990; May 1994; Lambshead *et al.* 2000; Allen *et al.* 2002; Thompson & Townsend 2006; Enquist *et al.* 2007; Howeth & Leibold 2010). One of the key goals in macroecology is to identify the processes that underlie the patterns that are found (McGill & Collins 2003). Such patterns are usually assumed to occur due to the differing characteristics of species, which in a given environmental milieu cause one species to be common and another rare, or one species to be a specialist adapted to a narrow range of conditions, whereas another is a generalist that can be found everywhere (Bell 2001). Efforts to develop a global picture of diversity in the sea are hampered by the small number of key studies, the varied sampling protocols applied, the different diversity indices employed to measure diversity, and the varying levels of taxonomic resolution applied within particular studies (Clarke 1992; Clarke & Crame 2010). Gray (1997) stated that marine biodiversity is higher in benthic (bottom-related) rather than in pelagic (in the water column) systems, and on coasts rather than in the open ocean, since there is a greater range of habitats near the coasts. Although molecular data have revealed the vast scope of microbial diversity in virtually all habitats (Green et al. 2004; Slapeta et al. 2005), the fundamental conflict between ubiquitous and endemic distribution

patterns, in relation to the 'everything is everywhere' dispute (Finlay 2002; Foissner 2007; Cermeno & Falkowski 2009; Nolte et al. 2010), is far from being resolved.

The development of massively parallel sequencing has paved the way to explore microbial and meiofaunal diversity in time and space (Sogin *et al.* 2006; Huber *et al.* 2007; Amaral-Zettler *et al.* 2009). Several studies have used pyrosequencing to assess the diversity of bacteria and archaea in the marine environment (Galamba *et al.* 2001; Huber *et al.* 2007), but there has been comparatively limited focus on eukaryotes (Brown *et al.* 2009). Of the studies that are available, research questions have focused almost exclusively on diversity (Amaral-Zettler *et al.* 2009; Stoeck *et al.* 2009), with little data available for understanding macroecological patterns of distribution. Next generation sequencing studies have tried to link genetic diversity directly to species richness by sorting out operational taxonomic units (OTUs; *proxies* for species) from the genetic diversity data. Measures of abundance are more difficult to assess in molecular taxonomic studies of microorganism biodiversity because of underlying PCR bias (Amend *et al.* 2010; Bellemain *et al.* 2010) and poor correlation of the amount of gene copy numbers to biomass or cell frequency (von Wintzingerode *et al.* 1997; Amend *et al.* 2010).

Biodiversity estimates are still far from reality, many habitats have been poorly sampled and several species-rich taxonomic groups, especially smaller organisms, remain poorly studied (Costello *et al.* 2006; Costello *et al.* 2010). About 11 to 33% of European marine metazoan species have been predicted to be undescribed (Costello & Wilson 2011), and outside European seas it appears that the proportion of undescribed species is much higher (Bouchet 2006; Brandt et al. 2007a; Brandt et al. 2007b). In this thesis complete levels of richness were achieved for some European meiobenthic samples, within the sampled scale. Further to this and for the first time are also shown levels of α and β diversities and community composition for meiobenthic communities around the UK coast and mainland European samples derived from next-generation sequencing. This will help clarify the scale and extent of community richness and composition across eukaryotic marine meiobenthic ecosystems.

Material and Methods

Sample collection

Sixty six sandy marine benthic samples were collected using a standard corer methodology (Platt & Warwick 1988) from 23 sampling stations around the UK and France, Spain, Portugal and Gambia (**see Supplementary Figure S1, Supplementary Table SI**) during the summers of 2007-2008. Distances between the 23 sampling stations represent approximate distances between multiple points using Google Earth **(Supplementary Figure S2).** The UK samples comprised 16 sampling sites; France/ Spain/ Portugal comprised 2 sampling sites each and Gambia one sampling site. Each biological sample comprised a single 44 mm diameter x 100 mm benthic core and was taken approximately 10 m apart from companion samples. An additional core was taken for sediment analysis. All samples were immediately fixed in 500 ml storage pots containing 300 ml of DESS (20% DMSO and 0.25 M disodium EDTA, saturated with NaCl, pH 8.0) (Yoder *et al.* 2006). The meiofaunal size fraction and DNA extraction were performed according to Chapter 3.

Primer design and PCR strategy

The primers SSU_FO4 (5'-GCTTGTCTCAAAGATTAAGCC-3') and SSU_R22 (5'- GCCTGCTGCCTTCCTTGGA-3') were used to amplify approximately 450bp of the nuclear small subunit rDNA (18S rDNA). Primer selection was based on nuclear small subunit rDNA (18S rDNA) pronounced homology across meiofaunal phyla, but also highly divergent regions (Blaxter *et al.* 1998) and were used for subsequent amplicon generation. Fusion primers, PCR amplification and 454 sequencing were then developed according to Chapter 2.

Data Analysis and generation of Operational Taxonomic Units

During the course of the PhD program, substantial advances were made in the analysis of 454 Roche amplicon sequencing datasets. Prompted by discussions in Reeder and Knight (2009) and Quince et al. (2009) on the confounding effects of PCR/sequencing errors and chimera formation regarding second generation environmental sequencing datasets, I adopted the AmpliconNoise algorithm for the analysis of the large European dataset. Thus, sequences were analysed using

AmpliconNoise followed by chimera removal using Perseus (Quince *et al.* 2011). The AmpliconNoise algorithm *denoises* pyrosequencing data by reducing per base sequencing errors, comprising filtering, flowgram and sequence clustering steps. Following analyses of control communities, it has been shown to reduce noise by approximately 50% in environmental data sets (Quince *et al.* 2011). Chimeras were identified using Perseus (Quince *et al.* 2011). This algorithm generates a Chimera Index (CI) for each read an index greater than or equal to zero with higher values corresponding to reads that are most likely to be chimeras. *Denoised* and chimera-removed sequences were then used to identify Operational Taxonomic Units (OTUs). OTUs were calculated using a complete linkage-clustering algorithm, measuring the distance between the most distant members in each cluster, at a 99% identity cut-off. Based on previous studies suggesting that different algorithms might influence estimates of taxon richness specially at higher cut-offs (Quince *et al.* 2009; Hao *et al.* 2011; Quince *et al.* 2011), the AmpliconNoise algorithm (Quince *et al.* 2011) was tested against the OCTUPUS clustering algorithm (http://octupus.sourceforge.net/) on the present normalized dataset at different cut-off percentage identities. Taxonomy assignment was performed using MEGABLAST (megablast -d database path -D 2 -p 90 -a 2 -b 1 -v 1 -i infile -F F > outfile) against the downloaded GenBank/EMBL/DDBJ nucleotide database and OTUs annotation was restricted to matches of 90% and higher, using the OCTUPUS annotation and parsing toolkit.

Diversity and community analysis

For direct ecological comparisons of between sample OTU richness, the original dataset was normalized using 9,490 (i.e the lowest coverage achieved in any one sample) randomly picked sequences (over 200 bases in length, n= 218,276) from each sample, prior to de-noising and OTU clustering. The total number of OTUs generated from the original and normalized datasets was significantly correlated (Spearman's coefficient: $\rho = 0.535E\text{-}5$, $p = 0.0001$), nevertheless, interpretations derived from direct comparisons of richness between samples refer to the standardized dataset.

Sample-specific rarefaction curves were performed using the DiversityEstimates software package via AmpliconNoise and phylum-specific rarefaction curves were

generated using EstimateS (Version 8.2.0, R. K. Colwell) using a range of species richness estimators (e.g. ACE, Chao1, Jackknife1 and Bootstrap) that yielded very similar results. The ACE abundance-based coverage estimator (Chazdon *et al.* 1998) was used because it represents a consensus view of the analyses and has proven to work well for the analysis of metagenetic datasets (Huber *et al.* 2007). In order to assess the similarity between samples according to meiofaunal community composition, Bray-Curtis similarity distances were computed between samples based on a presence/absence similarity matrix (data was prior square-root transformed to down-weight the importance of very abundant OTUs without losing the influence of rarer OTUs). Site similarities were then visualized by multidimensional-scaling (MDS) using Primerv6 software (Clarke & Gorley 2006). The ecologically distinct samples from Gambia were removed from the MDS analysis, since their inclusion precluded effective visualization of multivariate ordination of the European samples. In order to test for significant differences of community assemblages between sampling sites a permutational multivariate analysis of variance (PERMANOVA) (Anderson 2001) was used. Analyses were based on Bray-Curtis dissimilarities on untransformed data of OTU presence/absence matrix over the 22 sampled sites with 1000 permutations. To test if there was an association between geographic distances and community composition per phylum a nonparametric Mantel-type test (RELATE) based on distance matrices derived from geographic distance and community composition (presence-absence data) was performed using PRIMERv6. The biodiversity of the different samples was also reported as the number of operational taxonomic units (OTUs) and quantified using Shannon-Wiener biodiversity measure (H', using log base 2) and the equitability of meiofauna assemblages was estimated as Pielou's index (evenness J'). The Shannon–Wiener index (Ricotta & Szeidl 2006) is a nonparametric diversity index that combines estimates of richness (the total number of OTUs) and evenness (the relative abundance of OTUs). For example, communities with one dominant species have a low index, whereas communities with a more even distribution have a higher index. Where appropriate, sequential Bonferroni corrections, which are more sensitive to false positives than the standard Bonferroni technique, were applied (Rice 1988).

Results

Sequence data and sampling efficiency

Amplicons were generated from 23 benthic sampling sites from the low tide zone of marine sandy beaches around the UK, France, Spain, Portugal and Gambia. The total number of reads derived from Roche 454 FLX platform for all sampled sites was 877,423 which, after de-noising and chimera removal, were reduced to 694,802 sequences further used for downstream analysis. Clustering comparisons derived from OCTUPUS and AmpliconNoise show that OCTUPUS generates more Operational Taxonomic Unit (OTU) especially at higher similarity cut-offs, underestimating richness at lower similarity cut-offs (**Figure 4.1**). The OTU richness did not vary substantially throughout the tested cut-offs, and since many of the primary objectives of the current chapter are related to distribution, the 99% cut-off was chosen to avoid the confounding effect of "clumping" species genotypes within a lower level cut-off. Nonetheless both normalized and lower similarity cut-offs were performed on all analyses and exhibited the same ecological patterns (data not shown) and no substantial richness differences (**Supplemental Table SII**).

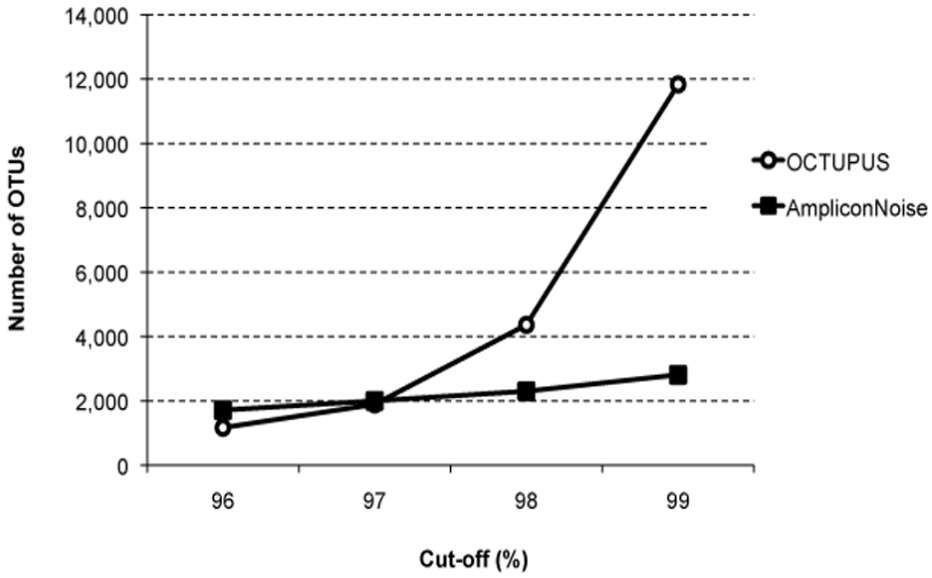


Figure 4.1. Lineage-through-time plots for OCTUPUS and AmpliconNoise. Number of OTUs plotted against each percentage identity cut-off (96-99% similarity) on a normalized dataset (4000 sequences randomly selected) using OCTUPUS (circles) and AmpliconNoise (squares) OTU clustering.

Sampling saturation profiles varied between the sites under study showing that sequencing effort was sufficient for some samples and far from completion for others (**Figure 4.2a, 4.2b**). From a qualitative perspective, sufficient sequencing effort was achieved for samples derived from Seaham, Dunnet Bay, Harwich, and Prestwick (UK) and Gambia (Africa). Samples showing a representative sampling effort and almost showing signs of saturation included Mera and Sada (Spain), Skye Staffin, Littlehampton, Exe, Fraserburgh (UK), St Jean Luz (France) and Vila Nova de Milfontes (Portugal). Sampling sites far form reaching an asymptote included Egremont, Moggs Eye, Porthawan, Sheerness, Firth of Fourth, Freshwater West, Silecroft (UK), Cap Ferret (France) and Praia Limpa (Portugal). It is worth noticing that Freshwater West and Sheerness showed high number of sampled sequences but were still not representative of the true community richness whereas sites with fewer sampling effort are fully representative of local community richness (e.g Prestwick, St Jean Luz and Gambia). Rarefaction curves also give an indication that Sheerness represents one of the most diverse sampled site whereas Vila Nova de Milfontes one of the less diverse.

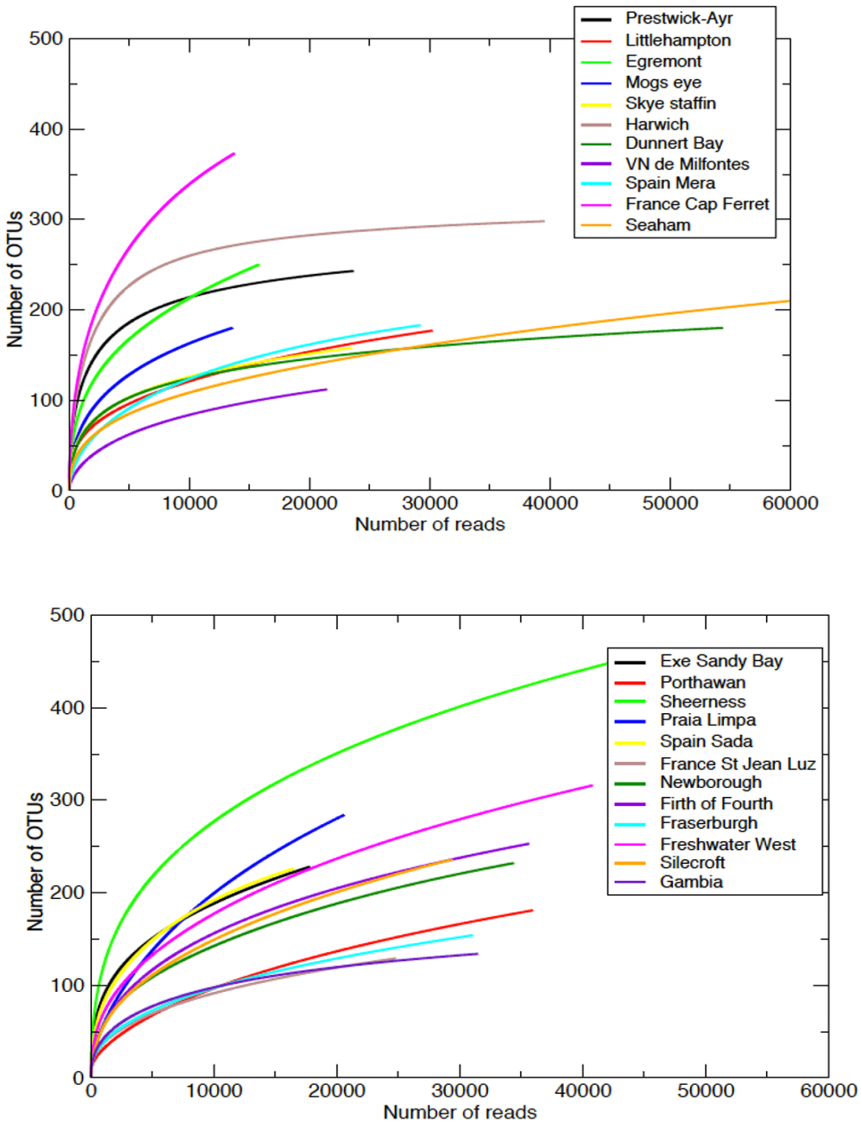


Figure 4.2. Sampling saturation profiles for samples collected in the UK, France, Spain, Portugal and Gambia, at a 99% sequence similarity level for OTU clustering. UK samples: Skye Staffin, Littlehampton, Exe, Fraserburgh, Egremont, Mogs Eye, Porthtawan, Dunnet Bay, Sheerness, Firth of Fourth, Freshwater West, Newborough and Silecroft. France samples: St Jean Luz and Cap Ferret. Spain samples: Mera and Sada. Portugal samples: Praia Limpa and Vila Nova de Milfontes.

Community diversity (α and β) and similarity

Relative OTU numbers at a 99% similarity cut-off indicate that Cap Ferret (France) and Sheerness (UK) were the most species rich samples with 313 and 282 different OTUs, respectively, followed by Harwich (UK) with 247 different OTUs. Egremont, Prestwick and Exe (UK) showed ca. 200 different OTUs. There were two further identifiable groups of samples that showed similar OTU richness. The first,

with higher OTU richness (ca. 160 different OTUs), comprised samples from Freshwater West, Firth of Fourth, Sada, Silecroft, Praia Limpa, Newborough and Mogs Eye. The second group, representing samples with lower levels of richness (ca. 100 different OTUs) included Mera, Sky Staffing, Seaham, Littlehampton, Dunnet Bay, Fraserburgh, Porthawan, St Jean Luz and Vila Nova de Milfontes (**Figure 4.3**). The diversity of the different samples observed by means of the Shannon (H') and Pielou's (J') indexes were highly correlated (ρ =0.9871, P=3.26E-16, p<0.0001) and the higher the OTUs richness the more even the OTUs distribution within a sample (**Supplementary Figure S3**). Furthermore the patterns of OTUs relative numbers throughout the sampled sites are similar to the ones observed for H' and J' indexes.

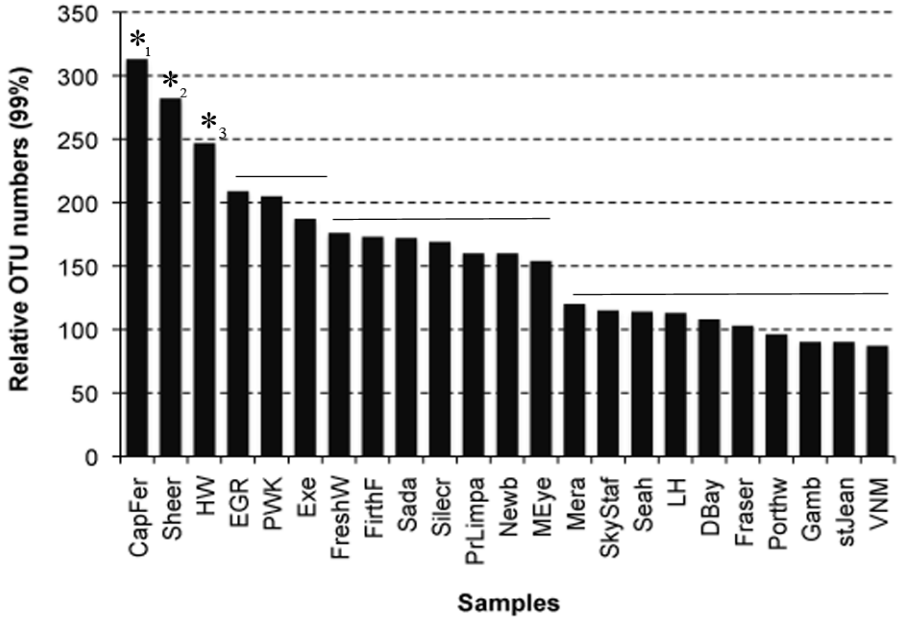


Figure 4.3. Relative OTU numbers for each sample site at a 99% sequence similarity level on normalized dataset. Stars highlight the first three samples with high OTU richness and black bars indicate samples with high, middle and low OTU richness. Skye Staffing (SkyStaf), Littlehampton (LH), Exe (Exe), Fraserburgh (Fraser), Egremont (EGR), Mogs Eye (MEye), Porthawan (Porthw), Sheerness (Sheer), Firth of Fourth (FirthF), Freshwater West (FreshW) and Silecroft (Silecr). France samples: St Jean Luz (stJean) and Cap Ferret (CapFer). Spain samples: Mera (Mera) and Sada (Sada). Portugal samples: Praia Limpa (PrLimpa) and Vila Nova de Milfontes (VNM).

Observations from numbers of shared vs. unique OTUs show that the majority of the sampled sites had 25-40% of unique OTUs present on each site, numbers only greatly surpassed by Gambia samples that had 74% of unique OTUs only sharing 26% of the OTUs with other sites. Further to Gambia, Cap Ferret and Sheerness samples also showed high percentages of unique OTUs, with 60% and 53% unique OTUs, respectively (**Figure 4.4**).

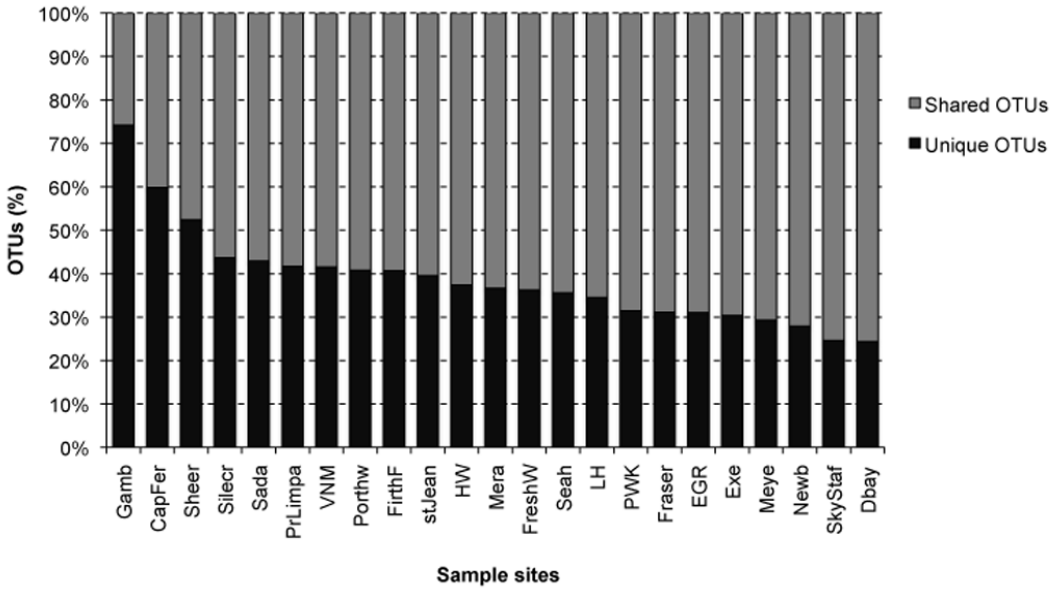


Figure 4.4. Relative shared and unique OTU percentages for each sample site at a 99% sequence similarity level on normalized dataset. Shared OTUs are represented by the grey bars and unique OTUs by the black bars. Skye staffing (SkyStaf), Littlehampton (LH), Exe (Exe), Fraserburgh (Fraser), Egremont (EGR), Mogs Eye (MEye), Porthawan (Porthw), Sheerness (Sheer), Firth of Fourth (FirthF), Freshwater West (FreshW) and Silecroft (Silecr). France samples: St Jean Luz (stJean) and Cap Ferret (CapFer). Spain samples: Mera (Mera) and Sada (Sada). Portugal samples: Praia Limpa (PrLimpa) and Vila Nova de Milfontes (VNM).

The MDS ordination of sites by community composition does not show an obvious pattern of geographic separation between sites (**Figure 4.5**). Nonetheless, MDS-analysis data combined with the PERMANOVA indicates that overall there are significant differences ($p<0.001$) in community composition between sites. The ordination of samples by (MDS) shows that in general regions geographically further apart tend to be less similar, nonetheless exceptions to the rule exist. Samples clustering alone are Sheerness, Freswater West, Skystaffin, Dunnet Bay, Silecroft,

Seaham, Egremont and Newborough. It is worth noticing that samples from Spain, Portugal and France have similar community composition together with Prestwick and Littlehampton. Vila Nova de Milfontes in Portugal seams to represent an outlier.

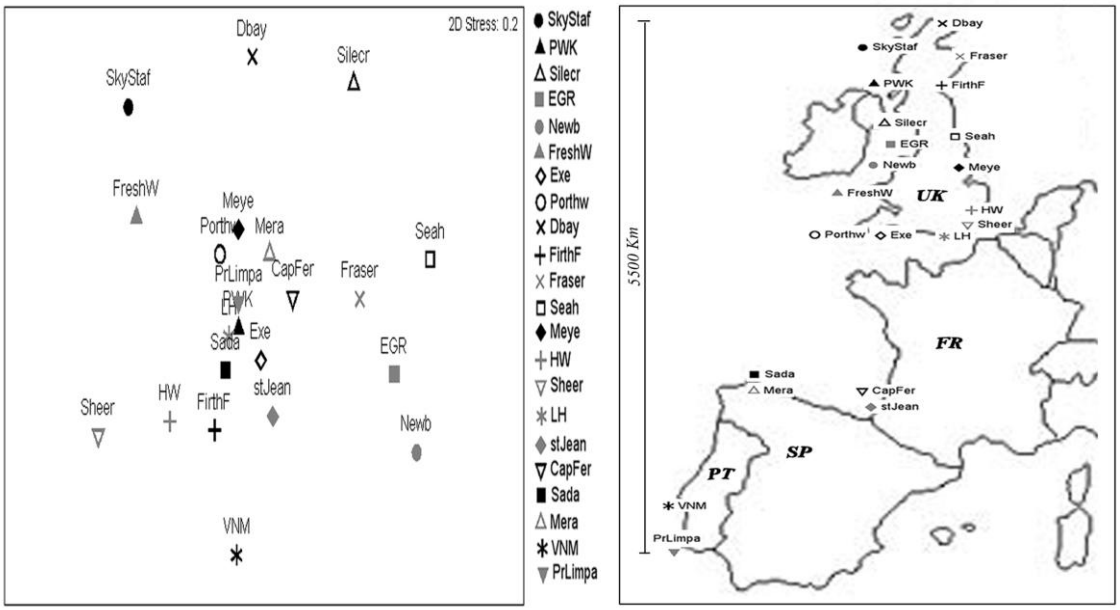


Figure 4.5. Multi-Dimensional scaling (MDS) configuration plot for the sampled sites based on square-root transformed data of OTUs abundance (a) and sample localization map (b). Plotted are sixteen samples around the United Kingdom (UK), two samples from France (FR), Spain (SP) and Portugal (PT). Skye Staffing (SkyStaf), Littlehampton (LH), Exe (Exe), Fraserburgh (Fraser), Egremont (EGR), Mogs Eye (MEye), Porthawan (Porthw), Sheerness (Sheer), Firth of Fourth (FirthF), Freshwater West (FreshW) and Silecroft (Silecr). France samples: St Jean Luz (stJean) and Cap Ferret (CapFer). Spain samples: Mera (Mera) and Sada (Sada). Portugal samples: Praia Limpa (PrLimpa) and Vila Nova de Milfontes (VNM).

Community composition and richness

The taxonomic distribution and abundance of the main meiofauna representatives was explored (**Figure 4.6, Supplementary Table SII**). The total number of OTUs found in all sampled sites was of 2183 and could be assigned to 23 different phyla, but five phyla were represented by no more than 3 OTUs (Orthonectida, Tunicata, Gnathostomulida, Porifera and Excavata). From the 23 different phyla 14 corresponded to marine meiofauna representatives. By far the most abundant phyla throughout all samples were the Nematoda, which formed up to 40% of all meiofauna OTUs for the majority of the samples. Platyhelminthes were the second

most abundant followed by the Arthropoda (mainly represented by Copepoda) with 20% and 10% representation in general, respectively. On average, there seems to be a clear taxonomic ranking pattern that follows Nematoda, Platyhelminthes, Copepoda, Annelida and Gastrotricha, nonetheless exceptions exist in some European samples where Platyhelminthes and even Copepods were ranked first. There was a clear taxonomic distribution pattern, where samples from the UK appeared to have a more homogenous community composition where the other European sampled sites proved to be more heterogeneous. Spearman's correlation (sequential Bonferroni corrected) showed a significant inversely proportional association of OTUs richness between the Nematoda and the Platyhelminthes (ρ =0.0025, α >0,05) but no such association was found between other meiobenthic phyla.

Correlations between OTU richness and grain size were performed using Spearman's correlation coefficient (ρ) nonetheless no significant association was found (ρ =0.377, p =0.325). Further more, Spearman's correlation showed a signification relationship ($p<0.05$) between the main meiofaunal phyla and geographic distance whereas the same correlation was not found for the protists (**Table III**).

Table III. Spearman's correlation (ρ) and significance test (p) between community similarity and geographic distance derived from a Mantel-type test, for the main meiofauna phyla and protists (shaded in grey: Rhizaria, Alveolata and Stramenopiles). $p>0.05$ are non-significant.

Phyla	Spearman's correlation (ρ)	Significance (p)
Nematoda	0.279	0.002
Platyhelminthes	0.354	0.001
Copepod	0.219	0.01
Annelida	0.145	0.065
Mollusca	0.252	0.001
Tardigrada	0.169	0.03
Rhizaria	-0.007	0.518
Alveolata	0.072	0.178
Stramenopiles	0.013	0.41

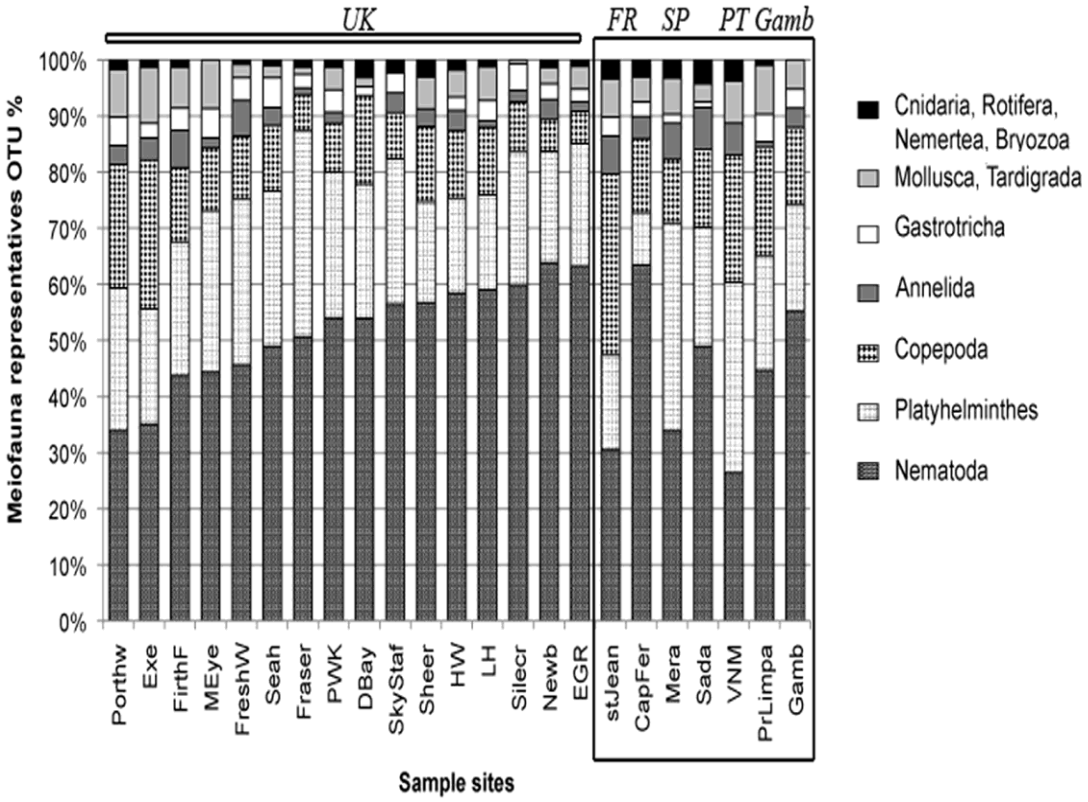


Figure 4.6. Taxonomic distribution of OTUs assigned to the main representatives of meiofauna phyla (with more than 5 OTUs) found in the 23 sampled sites. Clustering was performed at 99% similarity level on normalized data. Plotted are sixteen samples around the United Kingdom (UK), two samples from France (FR), Spain (SP) and Portugal (PT) and one sample from Gambia (Gamb). Skye Staffing (SkyStaf), Littlehampton (LH), Exe (Exe), Fraserburgh (Fraser), Egremont (EGR), Mogs Eye (MEye), Porthawan (Porthw), Sheerness (Sheer), Firth of Fourth (FirthF), Freshwater West (FreshW) and Silecroft (Silecr). France samples: St Jean Luz (stJean) and Cap Ferret (CapFer). Spain samples: Mera (Mera) and Sada (Sada). Portugal samples: Praia Limpa (PrLimpa) and Vila Nova de Milfontes (VNM).

The slope of OTU rarefaction curves at 99% cut-off for the main meiofauna phyla Nematoda, Platyhelminthes, Copepoda, Annelida and Gastrotricha did not reach an asymptote indicating an under sampling of these phyla (**Figure 4.7**). Thus, still 694,802 samples and ca. 2000 OTU-defined taxa fail to achieve saturation, even for low abundance phyla where rarefaction curves tend to converge (Tipper 1979). Richness estimates were higher for Nematoda and the lowest levels found within the main meiofauna representatives were the Annelida.

Almost 5% of the OTUs recovered no significant match to known ribosomal databases (OTUs <90% identity) and were named as not assigned (NAs)

(**Supplementary Table SIII**). These OTUs might represent previously unsampled diversity that is not available in public databases. Further manual analysis of these OTUs showed that the majority was placed within the Nematoda (43% OTUs, 5168 sequences) and almost 14% corresponded to taxa that are simply annotate as "Environmental Samples". Manual annotation of the 454 Roche reads annotated as environmental samples show that many are indeed nematodes.

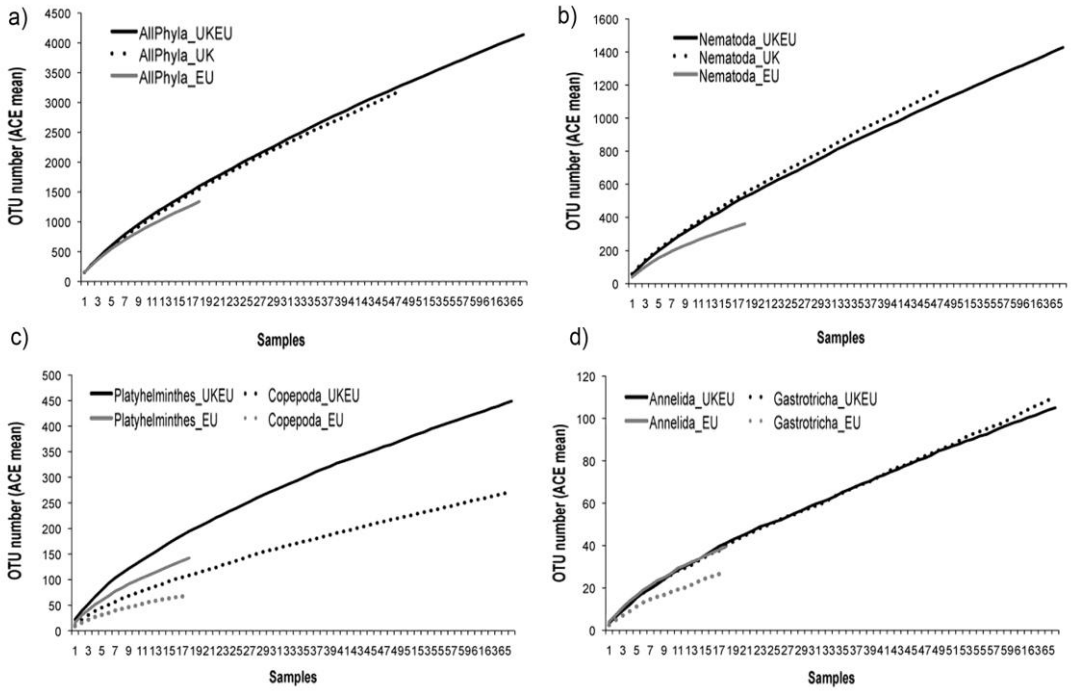


Figure 4.7. Rarefaction curves of the abundance-based coverage estimation (ACE) diversity estimator. Plots are shown for (a) All phyla, (b) Nematoda, (c) Platyhelminthes (full line) and Copepoda (dashed line), (d) Annelida (full line) and Gastrotricha (dashed line) at 99% identity OCTU cut-off all meiobenthic samples (UK, United Kingdom; EU, Europe), except Gambia. Curves were estimated from 100 randomizations, without replacement, using EstimateS, version 8.2.0.

Discussion

Deriving accurate identification of OTU richness

The introduction of pyrosequencing technologies has revolutionized our ability to explore the so-called "hidden diversity" (Foissner 1999; Slapeta *et al.* 2005). Nonetheless, major concerns focusing on the data analysis of the massive amount of data produced by these sequencing methods requires efficient and flexible

bioinformatics applications. Bioinformatic analysis has developed rapidly recently and new emerging tools are becoming available to test the veracity of unknown microbial communities. Despite using OCTUPUS on a previous environmental dataset using a very conservative approach, this pipeline when compared to a recently developed software AmpliconNoise (Quince *et al.* 2011) appears to overestimate OTU richness at higher levels of similarity cut-off. Although differences in OTUs richness at lower cut-offs were similar between pipelines, estimates diverge, whereby OCTUPUS generates five-times more OTUs than AmpliconNoise at stringent clustering similarity. Although sequencing of environmental DNA has allowed us to venture into a vast diversity of unknown microorganisms, much of this putative "new diversity" may derive from flawed bioinformatic approaches (Kunin *et al.* 2009; Reeder & Knight 2009; Huse *et al.* 2010; Quince *et al.* 2011).

Genomic approaches to taxon diagnosis exploit diversity among DNA sequences to identify organisms (Kurtzman 1994; Wilson 1995) and the choice of an accurate method to discriminate between different OTUs, defined as a *proxy* for species as long been an argument. Here all ecological comparisons analysis were performed on normalized datasets at both cut-offs 96% and 99% and the observed patterns were congruent, although richness estimates using AmpliconNoise differed slightly. OTU richness is dependent on the level of similarity used to define its clustering (Floyd *et al.* 2002; Huse *et al.* 2010; Kumar *et al.* 2011). In general a cut-off of 97% similarity is often considered as a *proxy* for species (Vandamme *et al.* 1996; Uilenberg *et al.* 2004) and in fact in eukaryotic organisms such as nematodes, a cut-off of 99.5% has been justified for chain termination sequencing datasets (Blaxter *et al.* 2005). Nonetheless it would be unadvisable to apply an identical biological limit to different taxa (Goldstein *et al.* 2000). Different copies of the rDNA gene exist and evidences (derived from personal data and Porazinska et al. 2010) indicate that in the case of the Nematoda most species have a dominating common sequence and variants represent a minority, that could either be real or sequencing artefacts. At 99% similarity the probabilities of recovering the maximum number of OTUs in a sample are higher (Porazinska *et al.* 2010). Further to this, Hebert *et al.* (2003) showed that levels of sequence divergence for the COI in eleven phyla of the metazoa, are regularly greater than 2% between species pairs with the exception of the Cnidaria with the majority of the species pairs having a sequence divergence less than 1%.

Avise (2000) also supported that intraspecific divergences are rarely greater than 2% and most are less than 1% for the COI gene. The COI gene has high evolutionary rates and even so it enables the discrimination of closely allied species with a sequence divergence of 1-2%. Nuclear genes like the 18S rDNA have a highly conserved sequence core but also possess rapidly evolving variable regions (Woese *et al.* 1990; Hillis & Dixon 1991; Gillespie *et al.* 2006) and also have an extreme variation in gene copy number in eukaryotes (Amaral-Zettler *et al.* 2009). Nonetheless, the nSSU is a slower evolving gene and intra-species variation should not be lower than 2%. Of course there are exceptions to the rule and intra- and inter-species variability should not be ruled out when defining OTU richness and diversity. Richness estimates should always encompass at least two different cut-offs whereas for ecological comparisons such as biogeography or community composition between different sites the use of a higher cut-off will better reflect the physical distribution of possible multiple genotypes derived from biological species.

Biogeography richness of marine benthic eukaryotes: local and regional richness

Microorganism diversity in the oceans has been hypothesized to be much greater than previous estimates based on conventional molecular techniques (Sogin *et al.* 2006) elevated estimates of diversity and the extended rare biosphere are likely to have been derived from misinterpretation of data (Pedros-Alio 2007; Elshahed *et al.* 2008; Hamp *et al.* 2009; Kunin *et al.* 2009; Huse *et al.* 2010; Youssef *et al.* 2010). In this study, rarefaction analysis revealed quite different levels of complexity and diversity across the several marine environments. Predicted estimates of richness were between 100 to 500 different OTUs (using a 99% identity cut-off). Nonetheless, for the majority of the samples eukaryotic diversity was not exhaustively characterized. Surprisingly, for the first time diversity levels were accomplished for some sites around Europe, including the UK coast. Sheerness represented one of the most rich sample site but with insufficient sampling effort, on the other hand Harwich that is just 80Km apart and also very OTU rich, covered true representation of local diversity, and both samples exhibited high community evenness patterns. At such a small spatial scale distance effect seems to be negligible showing clear evidences of habitat uniqueness and segregation.

In traditional molecular studies, dominant populations have masked the detection of low abundance OTUs, their overwhelming genetic diversity, and their individual distribution pattern in marine environments (Sogin *et al.* 2006). This could be the case of Gambia samples with low species richness but with the highest beta diversity or maybe simply because it is just very singular. Nevertheless, Gambia nutrient-rich coastal upwelling should provide available nutrients for higher meiobenthic richness. Contrarily, communities showing high alpha and beta diversity with an even species distribution, like Cap Ferret, Sheerness and Harwich suggest a high species turnover on this scale of sampling.

Information from estimates of diversity derived from rarefaction curves combined with OTUs richness suggests that there are disparate meiobenthic communities. This is indicative of specific local richness and diversity that probably depends upon local environmental factors [sediment grain size, salinity, temperature, organic matter and chlorophyll a (Chl-a)] and dominating species. In fact, local adaptation together with ubiquitous distribution will increase the probability of finding shared OTUs on the same habitat type, independently where that habitat exists (Finlay 2002). In this study 60-80% of OTUs were shared between the majority of the sampled sites and ca. 20%-40% were unique to a given site. The diversity of habitats found around the UK coast might also explain higher percentages of unique OTUs (endemic OTUs with narrow range size) in some samples that were independent of the sample size. Species-based measures of β diversity have been useful to study distributional patterns, because they can evaluate whether similar environments contain the same species despite distance and other geographic barriers (Noguez et al. 2005). Thus asserting that meiobenthic communities are ubiquitous but also confirm confined groups in marine coastal environments at the intertidal level.

Community similarity between sites

Muldimensional scaling of eukaryotic diversity between all sites suggests that there is a distinction in community composition into biogeographic areas between the UK coast and France/Spain/Portugal. The decrease in community similarity with geographic distance is a universal biogeographic pattern observed in communities from all domains of life (Nekola & White 1999; Green *et al.* 2004; Horner-Devine *et al.* 2004). Nonetheless, not always the closer regions had similar community

composition and vice-versa but there was a tendency for some main biogeographic patterns. Sites located in the North of the UK tended to be more similar whereas West and East coasts were more dissimilar, where samples located in the Southwest were more similar to the other European samples, probability due to geographic proximity. This suggests that long-term dispersion of environmental benthic communities would be expected to occur from Southwest UK to the European Atlantic coast (France, Spain and Portugal) due to the Gulf Stream currents. Contrarily, the North Atlantic drift currents could structure meiofauna communities due to homogenization effect of currents along the UK Northwest coast. There is also the existence of a surface current of North Atlantic water reaching as far as the Alboran Sea that might mark the main barrier between Atlantic and Mediterranean populations for some marine invertebrates (Maldonado & Young 1996). There are evidences that distinct populations are often difficult to detect in the marine environment (Palumbi 1992) in particular for marine invertebrates with high dispersal potential lacking a sharp geographical differentiation and genetic variation occurring almost exclusively within populations (Barber *et al.* 2000; Taylor & Hellberg 2003). In addition, Meiofauna have been observed to aggregate in clumps of 'marine snow' (Silver *et al.* 1978), and transport via these small organic parcels would presumably provide nutritional sustenance during pelagic journeys. Some authors have additionally suggested that meiofauna can be transported via sea ice (Giere 2009) and floating pieces of rubbish (Barnes & Milner 2005). In terms of anthropogenic transport, ballast water (and its associated sediment) and fouling on vessels are known to carry meiofaunal populations across oceans during shipping operations (Giere 2009). In this study, there was a correlation between distance and community similarity suggesting that dispersal is a limiting factor (Martiny *et al.* 2011) for the main representatives of eukaryotic marine meiofauna and local environmental and biotic factors will also affect community composition. Different combinations of environmental variables can influence meiobenthic communities structure (Coull 1999) such as temperature, water depth, salinity, Chl-a and silt-clay content (Vincx *et al.* 1990; Danovaro & Gambi 2002; Danovaro *et al.* 2008b). Despite no such parameters were measured in this thesis and samples are all from the same depth, it is possible that temperature, salinity and food availability (Chl-a) could greatly dictate the heterogeneity of the studied meiobenthic communities. Furthermore, high meiofauna richness has also been found in upwelling regions and

frontal areas (Pfannkuche 1985; Hua *et al.* 2009) due to the increase of available food resources in the sediment. Thus further suggesting that there are abiotic factors, such as tides and currents affecting meiobenthic community structure (Steyaert *et al.* 1999; Steyaert *et al.* 2003; Austen & Widdicombe 2006).

Assessing meiofauna geographic patterns and richness

Biodiversity assessments are fundamental for basic diversity science from an ecological, biogeographical and evolutionary perspective (Miloslavich *et al.* 2010) and high-throughput sequencing has further simplified this task. Relative abundances across the marine benthos offer strong insights of community patterns. The fine-scale community structuring indicated a general pattern where the Nematoda are always dominant and the most abundant in marine sediments, also reported in taxonomic (Blaxter 2004) or molecular studies. These usually are followed by the Platyhelminthes and Copepods, nonetheless derived from the disparity of the sampled sites there were some exceptions to the rule. In fact meiofauna community structure was more homogenous in the UK samples whereas France, Spain, Portugal and Gambia were clearly more heterogeneous with some samples being dominated by the Platyhelminthes within the meiofauna representatives. This shows clear evidence for a distinguishable community pattern between the UK and mainland Europe despite high percentages of widespread-shared OTUs between some samples. In fact, regional differences in the composition of the metazoan meiofauna along continental margins are hard to detect, because of pronounced local variations (Soltwedel 2000) and numerous biotic and abiotic factors (Soltwedel 2000; Soltwedel *et al.* 2005) that might hamper community assemblages.

The richness estimates derived from the rarefaction curves indicated that our high-throughput sequencing study was not sufficient to sample all meiobenthic diversity, thus asserting that meiobenthic diversity is extremely rich. Numbers of estimated richness from our study suggests the existence of more than 4500 OTUs of meiobenthic eukaryotes and more than 1600 OTUs of nematodes just for the UK coast and some samples around France, Spain and Portugal. Current estimates of marine benthos around the British Isles are of about 450 species of nematodes (Giere 2009) and Costello *et al.* (2006) estimates for the European marine nematodes is

about 1,837 species. Lambshead & Boucher (2003) reported that 30–40% of free-living Nematoda in European seas found in field surveys were new to science. Nonetheless, Costello (2011) estimate that the average number of species remained to be discovered in European seas is ca. 8500 and that only 3% of free-living Nematoda in Europe were described.

In microorganisms in general, the most abundant and dominant species have higher dispersal rates (Finlay 2002; Cermeno *et al.* 2010) and thus are more ubiquitous which may be the case of the Nematoda and Platyhelminthes. The observed significant association of Nematoda richness being inversely proportional to the Platyhelminthes richness suggests that these most abundant phyla might have an impact on how other meiobenthic organisms are structured in the community. Some members of the community might serve as keystone species whereas others could simply be the result of historical ecological impacts that would potentially become dominant if shifts in environmental conditions occur (Freckman *et al.* 1997; Hooper *et al.* 2005a). Some species of Nematoda might represent the "driver species" Walker (1992) refers to in environmental communities, whose roles in regulating community function is crucial. Thus, because nematodes are the most abundant metazoa in marine sediments (Lambshead & Boucher 2003; Lambshead 2004; Bhadury *et al.* 2006) it seems intuitive that they will directly model the dynamics and community assemblage patterns. Ecological grouping in nematodes is often related to substrate type (mud or sand) and feeding mode (microvores, predators) (Giere 2009). Despite there being no significant association with sediment type, apparent correlations between granulometry and diversity may well be caused by the intrusion of other factors (Boucher & Lambshead 1995). Nematodes and Platyhelminthes did not show a correlation with geographic distance supporting the cosmopolitanism of these phyla and further suggesting that specific differences might occur within each phylum. The results also indicate that community and/or taxa ecology and geographic distance could be determinative of a community diversity and richness. This could be the result of the existence of numerous microhabitats, microclimatic properties, soil chemical properties, and also organisms differentiated life cycle events in meiobenthic environments (Giere 2009). Further to this, some organisms may exist in quiescent or dormant stages (Coleman 2002), allowing for considerable niche space for the impressive meiobenthic diversity. The role of redundant species and their functional roles are crucial to

understand the relationship between biodiversity and ecosystem function. Without understanding the biology of species involved, it can be difficult to decide how many functional types are present in a system or determine the functional roles of individual species (Bolger 2001). High-throughput studies continue to give insights and predictions of richness but further refinement (seasonal variation, biotic parameters) is needed to understand the extent of community change and diversity across benthic ecosystems. It seems that even applying conservative approaches and eliminating chimeras and other artefacts from environmental datasets estimates of diversity and richness are still beyond our expectations.

CHAPTER 5 - Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses

Introduction

Utilizing second-generation pyrosequencing technologies in environmental DNA analyses have provided unique insights into prokaryotic (Sogin *et al.* 2006; Huber *et al.* 2007) and eukaryotic (Massana & Pedros-Alio 2008) molecular diversity and ecology. Massive parallel pyrosequencing has the potential to produce a large volume of data relatively cheaply and with an unprecedented read depth, generating millions of DNA sequences within a matter of hours (Margulies *et al.* 2005). One of the major data analysis challenges is determining if sequences produced from pyrosequencing-amplified regions of marker genes correspond to genuine biological diversity. Recently, studies have recognized that biodiversity levels have become inflated due to artefacts associated with sample processing including both the PCR amplification and the pyrosequencing itself (Kunin *et al.* 2009; Haas *et al.* 2011; Quince *et al.* 2011). PCR amplification with universal primers applied to genes conserved across phyla, such as the ribosomal nuclear small subunit (nSSU), is commonly used to identify microbial eukaryotes in natural environments. The extreme conservation of primer binding sites and the availability of extensive database resources (Pruesse *et al.* 2007) has resulted in the nSSU being the most widely used marker for studying the molecular taxonomy of a diverse range of eukaryotes. Target taxa range from all protist kingdoms (Pawlowski *et al.* 2011) to metazoan microorganisms, that are dominated by the Nematoda (Porazinska *et al.* 2009). In such analyses, one of the most commonly reported sources of sequence artefacts associated with highly homologous nSSU genes from environmental DNA samples is the formation of chimeric sequences during PCR amplification (von Wintzingerode *et al.* 1997; Huber *et al.* 2004; Quince *et al.* 2009; Reeder & Knight 2009; Quince *et al.* 2011).

Chimeric sequences, or chimeras, are generated when incomplete extension occurs during PCR amplification and the resulting amplicon re-anneals to a foreign DNA strand and is copied to completion in the following PCR cycles. Chimeras are composed of two or more phylogenetically distinct parental sequences and have been shown to occur in PCR-amplified nSSU datasets with frequencies of 30% to 70% (Wang & Wang 1997; Ashelford *et al.* 2006; Haas *et al.* 2011) thus leading to false diversity estimates and false novel taxa. The critical factors that seem to affect PCR-

generated recombination are the number of PCR cycles, PCR extension time, template concentration, *Taq* DNA polymerases and amplicon size (Wang & Wang 1997; Qiu *et al.* 2001; Lahr & Katz 2009; Engelbrektson *et al.* 2010). Chimera formation can be minimized experimentally by PCR optimization, nonetheless, though no method has yet proved to be entirely effective. The importance of detecting chimeras is such that a plethora of bioinformatic software has also been developed, such as Chimera_Check (Cole *et al.* 2003), Bellerophon (Huber *et al.* 2004), CCode (Gonzalez *et al.* 2005), Pintail (Ashelford *et al.* 2005), Mallard (Ashelford *et al.* 2006), Chimera Slayer (Haas *et al.* 2011) and Perseus (Quince *et al.* 2011). With the exception of Perseus, most of these approaches will only detect clear induced chimeras (Smyth *et al.* 2010) and their accuracy for chimera detection has not been rigorously tested (Haas *et al.* 2011) or is still at an early stage, especially given recent advances in environmental DNA sequencing approaches. Although metagenetic analyses are clearly based on complex and phylogenetically diverse assemblages, the roles of sample richness and phylogenetic diversity in driving chimera formation are largely unknown.

Wang and Wang (Wang & Wang 1996, 1997) tested how sequence similarity between cloned 16S rRNA genes or mixed bacteria genomic DNA can influence PCR-based chimera formation. Nonetheless, these investigations were performed on a very small scale, did not consider sample richness, and pre-dated the current second generation sequencing perspective of amplicon pool diversity. The overarching aim here is to (a.) analyze the effect that richness, evenness and genetic diversity play in chimera formation and link this to diversity estimates, and (b.) understand how chimeras are formed with respect to variable genetic diversity and secondary structure of the parent nSSU molecule. To this end a nSSU dataset was generated by 454 Roche pyrosequencing on control pools of closely and distantly related nematode mock communities of known identity and richness.

Material and Methods

Sample preparation

To test if chimera formation during PCR reactions was associated with taxon richness or with phylogenetic distance, 74 Sanger-sequenced single nematode species were blast aligned to a contemporary Nematoda phylogenetic framework (Meldal *et al.* 2007). Subsequently, the sequences were aligned using ClustalX and pairwise distances (p-distance) were calculated using MEGA 4.1 (Tamura *et al.* 2007). Based on the phylogenetic affinities of the nematode sequences, subsets of closely related (mean pairwise divergence [MPD] of 25%, refered to as "phylogenetically close") and distantly related (MPD of 40%, refered to as "phylogenetically distant") nSSU controls were generated by pooling the DNA extracts of 12, 24 or 48 individuals.

DNA extraction and preparation

DNA extraction from DESS-preserved (Yoder *et al.* 2006) single worms was performed using a DNeasy Blood & Tissue Kit (Qiagen Inc), following the manufacturer's instructions. After extraction all DNA was eluted in 40 µl of AE buffer and samples were stored at –20 °C until use. The DNA extracts from all single individuals were quantified using a Nanodrop spectrophotometer and diluted to 0.5 ng/ µl, and five replicates of the 12, 24 and 48 individuals selected for the closely and distantly related treatments.

PCR amplification and sequencing analysis

The primers SSUFO4 forward (5'-GCTTGTAAAGATTAAGCC-3') and SSUR26 reverse (5'-CATTCTTGGCAAATGCTTTCG-3') were used to amplify approximately 450bp of the nuclear small subunit rDNA (18S rDNA) region (Blaxter *et al.* 1998). Fusion primers were then developed according to Chapter 2. PCR amplification reactions and the thermocycle for the targeted nSSU region were optimised. Optimised reactions were performed using 0.25ng/ul of genomic DNA template in 3x40 µl reactions using *Pfu* DNA polymerase (Promega) for each of the closely and distantly related nematode pools (12, 24 and 48 individuals) and all individual DNA extracts. PCR thermocycle conditions consisted of a 2 min

denaturation step at 95 °C followed by 35 cycles (thus facilitating the generation of chimeras (Wang & Wang 1996, 1997; Haas *et al.* 2011)) of 1 min at 95 °C, 45 s at 55 °C, 3 min at 72 °C and a final extension of 10 min at 72 °C. Negative controls (ultrapure water only) were included for all amplification reactions. Electrophoresis of triplicate PCR products was undertaken on a 2% gel with Top Vision™ LM GQ Agarose (Fermentas), and the expected 450bp fragment was purified using the QIAquick Gel Extraction Kit (Qiagen), following the manufacturer's instruction. All purified PCR products were quantified with an Agilent Bioanalyser 2100 and diluted to the same concentration (10 ng/µl). PCR amplifications from single nematodes and pooled nematodes were sequenced in a single direction (A-Amplicon) on a quarter and three quarters of a plate, respectively, using a 454 Roche GSFLX (454 Life Sciences, Roche Applied Science) sequencing platform at Liverpool University's Centre for Genomic Research, UK.

Denoised reads and detection of chimeric PCR molecules

Pyrosequencing reads derived from 454 Roche data contain a substantial number of errors (referred to as *noise*), which includes sequencing errors, PCR single base substitutions and PCR chimeras (Quince *et al.* 2009). AmpliconNoise was used to remove noise from the pyrosequencing data, this comprises filtering, flowgram and sequence clustering steps. It has been shown to reduce noise by approximately 50% in environmental data sets (Quince *et al.* 2011). Subsequently, chimeras were identified using Perseus (Quince *et al.* 2011): this algorithm generates a Chimera Index (CI) for each read greater than or equal to zero with higher values corresponding to reads that are most likely to be chimeric. Perseus by pairwise alignments to all sequences of greater than equal abundance identifies the most likely parent sequences of the candidate read and the most likely break point. Logistic regression is then used to classify chimeras so that the pyrosequencing data output lists chimeric and non chimeric sequences. The lower the probability of the sequence evolving naturally, the higher the Chimera Index (Quince *et al.* 2011).

Perseus finds break point positions in the two parent sequences. To compare across the whole data set it is necessary to fix these positions relative to a reference sequence. To do this, a four-way alignment between each chimeric sequence, its two parents and the *C. elegans* reference sequence was formed. The most likely break

point was identified by minimising the number of differences between the sections of the parents contributing to the chimera and the chimera itself. The position of each break point on the reference sequence was then recorded and from this, the frequency breaks occurring at each position could be calculated. MFold, RNA-folding software was used to predict the potential role on chimera formation of the secondary structure of the 18S rDNA amplicon region (Markham & Zuker 2008).

Generation of Operational Taxonomic Units

Denoised mock nematode community data from which chimeras had been removed was used to identify Operational Taxonomic Units (OTUs). OTUs were calculated using a complete linkage-clustering algorithm, measuring the distance between the most distant members in each cluster, at a 99% identity cut-off. The number of OTUs generated was then used to determine the effect of taxa richness on chimera formation within a sample. Although numbers of reads within treatments varied this did not have a significant effect (ANOVA, $p>0.05$) on chimera frequencies, number of OTUs and/ or Shannon Index. However, all the analyses were also performed on a normalised dataset by subsampling equal read numbers from each treatment and observations were found to be congruent with the non-normalized data (data not shown).

Statistical analysis

Species richness or in this case OTU richness takes no account of the evenness of the distribution. An index with better properties is the Shannon index. This increases with more taxa but also as the distribution of abundances across taxa becomes more even. The Shannon Index of biodiversity was established for each sample using Vegan R (Dixon & Palmer 2003). To analyze the relationship between overall chimera percentage and the explanatory variables (e.g. phylogenetic relatedness, richness, diversity, number of reads) a linear model was fitted to the data, giving a multiplicative coefficient for each explanatory variable. An analysis of variance (ANOVA) was then performed to statistically determine which of the variables had an effect on the chimera percentage. Variables without a significant effect on

chimera percentage were removed and the model was refitted to give accurate ANOVA results. A probability (*P*-value) less than 0.05 was considered significant.

Results and Discussion

Here, it was possible for the first time to assess the effect of OTU richness, evenness and phylogenetic relatedness on chimera formation in a second generation sequencing environmental dataset. Moreover, a rigorous experimental design and bioinformatic analysis facilitated the identification of chimera breakpoint frequencies within the parent nSSU molecule. The 'mock community' contained equivalent concentrations of 18S rDNA genes of individual nematodes of known identity and richness. Nematodes are the most abundant phylum of meiofaunal environmental samples representing a major part of biodiversity and performing numerous essential roles in ecosystems processes (Snelgrove 1999; Blaxter *et al.* 2005). The nematodes that were chosen included both phylogenetically distant and closely related species to emulate a likely environmental assemblage. Nuclear small subunit rDNA amplicons (*aka*. 18S rDNA) were generated from 74 individual Sanger-sequenced nematodes that were pooled to form closely related and distantly related nematode assemblages. Amplicons were sequenced on a Roche 454 GSFLX platform and generated a total of 339,515 pyrosequence reads. AmpliconNoise (Quince *et al.* 2009) generated 236,406 reads after removing errors arising from PCR and pyrosequencing and truncating sequences to a uniform 200bp. Chimeras were detected in denoised 'mock community' data using Perseus algorithm (Quince *et al.* 2011) and ca. 42% of the sequences were classified as chimeric and were removed before taxon richness was assessed by clustering sequences into Operational Taxonomic Units (OTUs) at a 99% identity threshold for each dataset. A summary of the mean number of reads, denoised sequences, chimera percentages and OTUs for each dataset is given in **Table IV**.

Table IV – Mean numbers of OTUs, denoised sequences, chimera percentages and reads for the pools of close and distantly related nematodes with 48, 24 and 12 individuals, respectively.

Species phylogeny	Species number	OTUs at 99%	Denoised sequences	Chimera %	Read number
Close	48	87.6	138.4	35.60	13882
Close	24	40.4	63.2	34.55	3809
Close	12	35.8	42.8	14.57	6159.8
Distant	48	63.2	161	58.97	5657.8
Distant	24	53.6	119	53.57	10134
Distant	12	34.4	58.2	39.93	7638.2

Denoised sequences contained between ca. 15% to 60% of chimeras in some pools, confirming that 35 cycle PCRs do indeed generate numerous chimeras (Wang & Wang 1996, 1997; Qiu *et al.* 2001; Sipos *et al.* 2007) even within a small mock environmental dataset. The results stress the importance of a chimera removal step to allow an accurate estimation of OTU numbers and robust estimates of biodiversity levels in environmental samples.

Overall the mean OTU numbers was approximately double the number of unique nematode species in each pool. This could be associated not only with sequencing artefacts but also because organisms frequently contain multiple copies of heterogeneous nSSU genes (Clayton *et al.* 1995). To assess the impact on the dataset of multi-copy nSSUs, all single nematode PCR products were amplified with unique MID-tag sequences. Of the 74 MID-tagged single nematode amplifications, 61 were single copy 18S rDNA and 10 were double copy but all taxa were represented by a similar total number of sequences in PCR reactions (data not shown).

A striking observation was the difference in chimera formation between close and distantly related nematode assemblages. In the latter the mean percentage of sequences classified as chimeric was 55% for the 24 and 48 species pools and was significantly higher (ANOVA, $p<0.001$) than the equivalent pools of closely related nematode assemblages that had 35% chimeras. In line with the previous observation the mean percentage of chimeras was significantly lower (ANOVA, $p<0.01$) in the 12 species pools irrespective of the similarity or distance of the individuals in the nematode assemblages (**Table IV**). These results suggest that chimera formation is significantly higher in more phylogenetically diverse and richer data sets.

Although the studies were on a smaller scale, Qiu *et al.* (2001) and Wang and Wang (1996) analysed chimera formation with bacterial rRNA clones and also found

that PCR artefacts and chimera frequency increased as species diversity increased. To further confirm this, OTU diversity within the two nematode assemblages (close and distant pools) was expressed using the Shannon Index (Ricotta & Szeidl 2006). OTU diversity (Shannon Index) and OTU richness (OTUs numbers) showed a significant effect (ANOVA, $p<0.01$, $p<0.001$) on chimera frequency further supporting the hypothesis that more diverse and richer samples generate a higher frequency of chimeric molecules. Additionally, diversity (Shannon index) had a significant effect both on the closely ($p<0.05$, $P= 0.0324$) and distantly ($p<0.01$, $P= 0.0023$) related nematode assemblages, and had a positive relationship with chimera frequency (**Figure 5.1**).

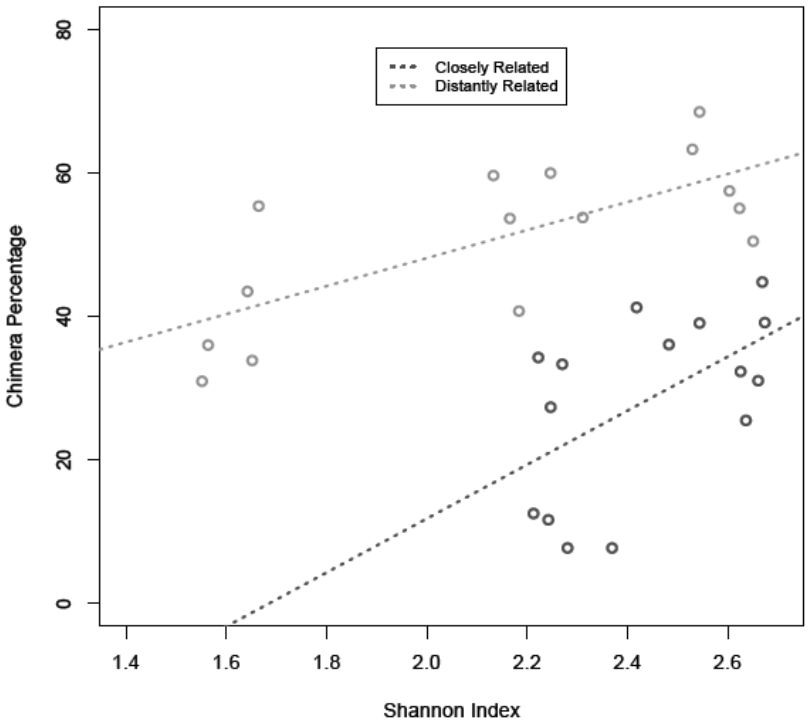


Figure 5.1- Chimera percentage and Shannon Index on close and distant related pools of nematodes.

Analysis of chimera breakpoint occurrence in nSSU amplicon sequences revealed that regions with higher nucleotide sequence similarity had significantly higher breakpoint frequencies ($p<0.001$, $P= 0.00039$) (**Figure 5.2a, 5.2b)**. Indeed, in studies with bacteria using 16S rRNA gene a large number of competing templates with fairly high sequence similarity generated more chimeras (Wang & Wang 1996; von

Wintzingerode *et al.* 1997; Haas *et al.* 2011). Presumably, one explanation for this phenomenon may be the priming of strand synthesis by prematurely terminated templates in the next PCR round.

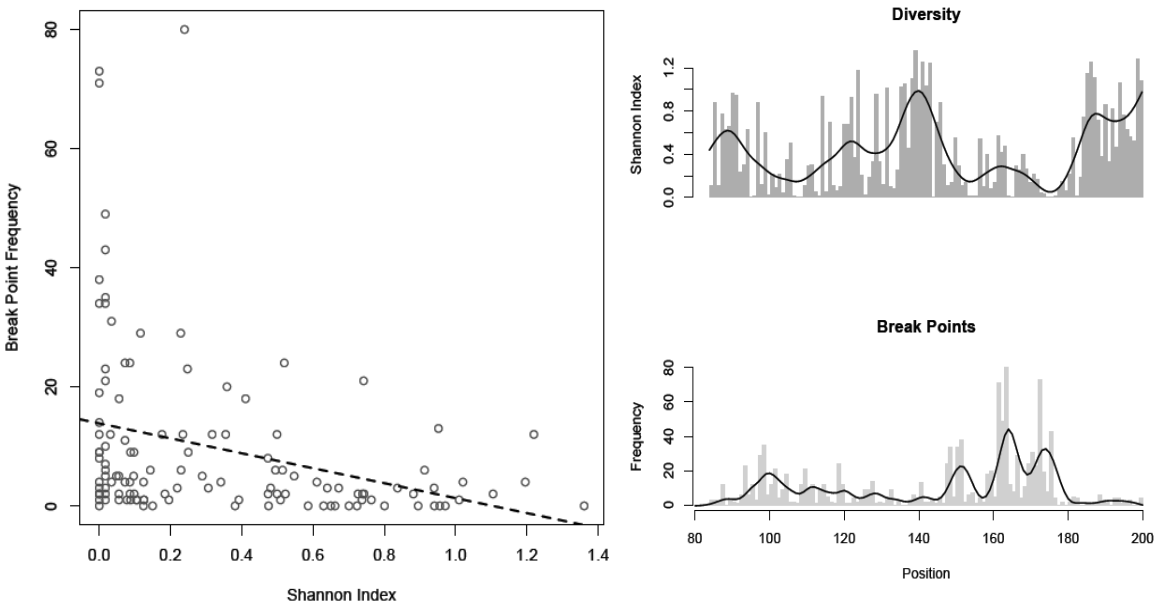


Figure 5.2- Nucleotide diversity (Shannon Index) and breakpoint frequencies occurrence in single nematodes and parental chimeric sequences, respectively.

Different copies of the nSSU genes from the same organism may differ by up to 6.5% (Clayton *et al.* 1995; Wang & Wang 1997) and in the present study alignment of close and distantly related nematodes indicated an overall sequence divergence of 10% and also enough to generate chimeras. To better reflect an environmental dataset, in the present study an alignment of ten representatives of each meiofauna phyla was performed and a 23% overall sequence distance (data not shown) was observed for the same nSSU region. In fact, Wang and Wang (1997) (Wang & Wang 1997) suggested that despite some degree of nucleotide mismatching, partly terminated heterologous 16S rDNA templates can often be completed in the subsequent polymerization step resulting in chimeras. Thus, the possibility that formation of chimeric sequences between different copies of the nSSU genes was also likely to occur (Wang & Wang 1996; von Wintzingerode *et al.* 1997; Haas *et al.* 2011) is now confirmed with this experiment. It is probably, the degree of sequence similarity within each individual that may determine chimera breakpoint

formation. This is an issue inherently associated with the fact that nSSU gene is a multicopy gene and intra-specific variability might have a determinant effect on chimera formation, especially when sample richness is quite high. Hass (2011) and Wang (1996) verified that more similar 16S rDNA genes more readily form chimeras but they did not discard the possibility of chimera formation among different species. In fact, the latter phenomenon is evident for the first time in the present study where chimeras are more often generated among richer, phylogeneticaly diverse samples, although the region where the chimera forms has to have sufficient conservation to favour hybridization and chimera formation.

Chimeras are generally composed of two true sequences, occasionally more (Quince *et al.* 2011), with a discrete break point where the transition from one sequence to another occurs. In the present dataset the distribution of chimera breakpoints showed a similar pattern across closely and distantly related nematode assemblages, with a mean peak of frequency at the first 140bp of the selected nSSU region (**see Supplementary Figure S5.1 of Appendice I**). Although GC content is thought to correlate with chimera formation due to inefficient strand separation and susceptibility to secondary structure formation, a detailed analysis of the parent chimeric sequences at the breakpoints did not reveal a significant correlation between GC rich regions and chimera frequencies. To further investigate the breakpoint region, the secondary structure of the amplified nSSU fragment was modeled in twelve single nematode sequences at 55°C and 65°C folding temperatures (**see Supplementary Table SI of Appendice II)**. Analysis of the nSSU secondary structure showed that the regions where the breakpoints occurred coincided with hairpin loop structures at both temperatures, although at 65°C regions of secondary structure were less abundant (**see Supplementary Table SI of Appendice II, Figure 5.3**).

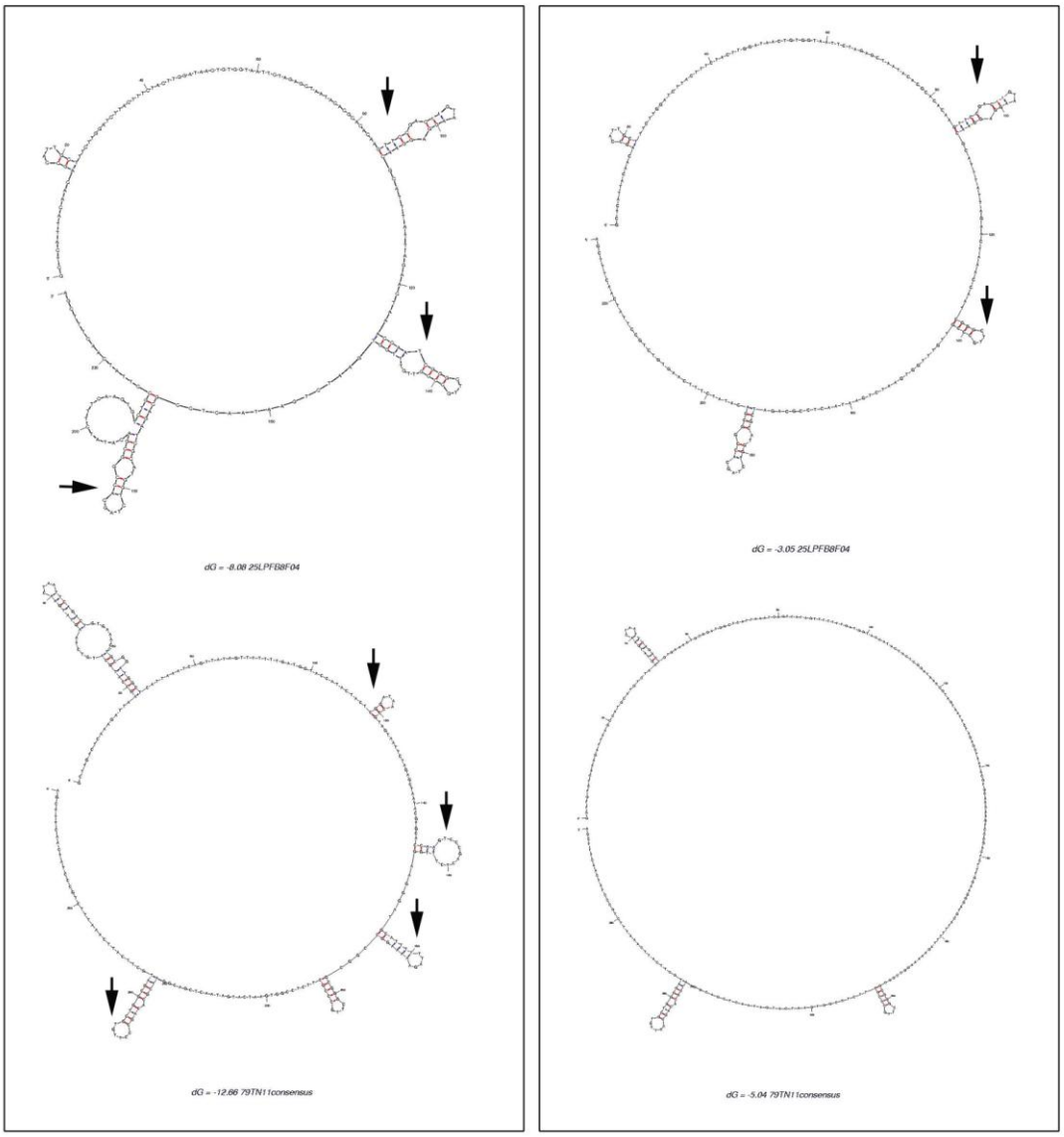


Figure 5.3 – Most frequent predicted secondary structures found on the 18S rDNA amplicon at (a) 55ºC and (b) 65ºC folding temperatures, using as an example two single nematodes. Arrows indicate where the most frequent breakpoints occur generally matching hairpin-loops. 79TN11consensus and 25LBFB8F04 are the names of each nematode. dG; free energy necessary for sequence stability at a given temperature.

Hairpin-loops are common motifs in nSSU gene secondary structure due to their importance in ribosome folding and function (Chen *et al.* 2004) and their presence requires greater energy for melting to occur during PCR and their maintenance will make it more difficult for DNA polymerases to read through. Modelling *in silico* revealed that the nSSU amplicon region in the present study retained some secondary

structure at the primer annealing temperature (55°C) and may have been one of the causes of premature termination of DNA synthesis. This is corroborated by results of multiple sequence alignments of PCR-induced chimeras (Wu *et al.* 2007) that reveal the recombinant regions were correlated with DNA template secondary structures. Fewer secondary structures of the nSSU amplicon were found at 65°C, suggesting that i) primers should be designed with a high annealing temperature and/or ii) genes chosen for environmental metagenetic analyses should be selected for a low tendency to secondary structure formation which should reduce the disposition of complex samples to form chimeras.

The investigative use of higher annealing temperatures in this study was not possible since the optimal thermocycling conditions used to amplify meiofaunal representatives (De Ley *et al.* 2005; Floyd *et al.* 2005) preclude more stringent annealing temperatures. In fact, an intuitive general rule for metagenetic studies would be to avoid high annealing temperatures to ensure the co-amplification of large ranges of taxa from disparate phyla. This implies not only having very high quality DNA samples with uncontaminated laboratory environment but also compulsory stringent analyses by using algorithms to remove artefacts and/ or putative chimeras after sequencing. In addition, the use of reference databases to detect chimeric molecules in environmental datasets is complicated by their unpredictable diversity, meaning that reference data may not be representative of the true diversity. On the other hand, the existence of chimeric sequences in public DNA databases is well known (Hugenholtzt & Huber 2003; Ashelford *et al.* 2005) and the risk of classifying chimeras as new organisms is becoming higher than the risk of neglecting non-chimeric ones.

Experience from recent studies where ca. 65% of the sequences generated from a 454 Roche environmental dataset were discarded leads us to suggest that metagenetic analyses are the ideal "breeding ground" for recombinant DNA molecules. DNA amplification by PCR has become the main crucial step used for next-generation sequencing technologies in the analysis of environmental samples and so PCR-derived artifacts are continuously increasing. Based on the analyses, the theory of chimera formation having a stochastic distribution (Jumpponen 2007) should probably be re-evaluated because their occurrence can be influenced by several factors, namely PCR conditions, amplicon nucleotide diversity, molecule folding structure and sequencing strategies. In fact, almost all steps of the molecular

approach can introduce biases or errors (von Wintzingerode *et al.* 1997). Overall, this study will significantly contribute to a better understanding of chimera formation and pyrosequencing strategies to be considered when conducting any studies focused on the PCR amplification of environmental DNA.

CHAPTER 6 – Final discussion and concluding remarks

Biodiversity studies on microorganisms using next-generation sequencing approaches have proved to be a challenging task. In this thesis it was possible to demonstrate levels of biodiversity and richness of meiobenthic communities that were previously unknown. Patterns of community assemblages within the European coast were assessed to reveal the distribution patterns of small-sized aquatic organisms. Further to this and with the accelerating pace of available bioinformatic pipelines it was also possible within four years to test several different *in silico* approaches to analyse optimally massive datasets. The field however, is still in its infancy and certainly many developments will emerge as the field of metagenetics matures. Acknowledging problems inherent in PCR-amplification and *in silico* data analysis derived from pyrosequenced environmental samples gave insights to further minimize and possibly to avoid future misinterpretation of diversity levels.

Distribution patterns of marine meiofauna

The factors determining the extent of a species geographic range have long interested biologists (Darwin 1859; Mayr 1963). Various environmental factors influence organism abundance and community composition (Soltwedel 1997; Wilson *et al.* 2004; Ingels *et al.* 2006) namely dispersal, birth and extinction (MacArthur & Wilson 1967; Holt *et al.* 2005), in addition to interspecific interactions such as competition and predation. The ability to disperse depends on many aspects that may be specific to particular groups of organisms (Holt *et al.* 2005). Suggesting that patterns of biodiversity are the result of past events that can provide important clues about the history of a species and help explain its current population structure (Feral 2002).

Although biological diversity processes may be similar in both marine and terrestrial environments the major differences appear largely as a result of the dispersive nature of marine larvae and the wide distribution of organisms and of habitats (Snelgrove *et al.* 1999; Feral 2002). About 70% of marine species have a planktonic larva phase, which is associated with very high fecundities, explosive reproductive potential and important dispersal and migratory capacities (Palumbi 1992). Accordingly, the wide, or even cosmopolitan distribution of marine

meiofauna has previously been considered a "paradox" because most meiofaunal organisms typically do not have a planktonic larval stage (Giere 2009) and so dispersal is primarily due to the suspension of adults and juveniles in the water column (Shanks & Edmondson 1990). The metapopulation concept in coastal marine benthic invertebrates relies on the fact that these taxa are likely to be restricted to a range of habitats with often pronounced ecological gradients (temperature, salinity, light, trophic abundance) (Reise 2003; Derycke *et al.* 2008a; Giere 2009) limiting distributions according to physical and biological constraints (Gray 1997; Giere 2009; Gray & Elliott 2009). Meiobenthic communities are known for having high numbers of cryptic species, morphologicaly very similar or even identical but genetically very divergent. This high species and genetic diversity might suggest that some meiobenthic faunal distribution has distinct ecological ranges because of intensive metapopulation dynamics (Derycke *et al.* 2008a). The data reported in the present thesis reveals a high level of common shared OTUs between sites and is congruent with a metapopulation concept, as previously proposed as characteristic of many coastal invertebrates (Harrison & Hastings 1996). In such a model each sample point is envisaged as a patch with exchange of individuals between them with local populations sustaining themselves over generations and exhibiting dynamics partially independent from other such population within a region (Harrison & Hastings 1996). However, in addition to the metapopulation concept there are local self-sustaining populations identified by the presence of unique OTUs that are not shared between sample sites, presumably representing specific locally adapted organisms.

Genetic pools of the majority of widely distributed species are rarely homogenous from throughout habitat ranges (Hilbish 1996; Avise & Johns 1999). This may be because of several nonrandom factors including neutral genetic drift and strong selection through local adaptation (Servedio 2004). In fact, several morphological features such as a panoply of highly specific feeding apparatus of some meiobenthic representative's supports the trophic niche-partitioning, as is exemplified by the narrow selection of bacterial strains consumed by stenophagous nematodes (Moens *et al.* 1999a) or harpaticoids that exhibit preference for habitat patches dominated by diatoms (de Troch *et al.* 2008). Furthermore specialized reproductive strategies (Michiels & Traunspurger 2005; Mittelbach *et al.* 2007) are also known in meiobenthic species that further allows them to colonize and persist successfully in

the benthos. Meiofauna are known to specialize and strategically adapt to their environment, representing animals with high potential for sympatric speciation, opening possibilities for new niches and evolutionary pathways without any geographic separation (Palumbi 2003; Giere 2009). The results provide evidence of a heterogenous meiobenthic local distribution in which 50% of the taxa identified were endemic that is similar to the 30% previously reported for protist distribution (Foissner 2006), reflecting high species turnover between sites (b-diversity) and thus high regional diversity. Nonetheless, large-scale distribution of some meiobenthic representatives is still unexplained as are the mechanisms behind the widespread distribution of meiobenthic species. An interesting analysis that could provide insights into this question may be the comparison with microorganism distribution.

The model of ubiquitous dispersal of microorganisms (Finlay & Clarke 1999; Finlay 2002; Fenchel & Finlay 2004; Finlay 2004; Finlay & Fenchel 2004) has contributed much to our understanding of microbial diversity and distribution. The idea that 'everything is everywhere' (Baas-Becking 1934) based on the assumption that the enormous dispersal capabilities of microorganisms allow them to expand into almost any habitat (Finlay 2002; Fenchel & Finlay 2004) still remains a controversial issue (Foissner 1999; Coleman 2002; Lachance 2004). Ubiquitous distributional patterns were demonstrated in different groups of protists (Finlay & Esteban 2001; Wilkinson 2001; Fenchel & Finlay 2004; Finlay *et al.* 2004; Finlay & Fenchel 2004), with microorganisms showing enormous population sizes, small body sizes and high dispersal rates with a cosmopolitan distribution and relatively low diversity. The theory further suggests that prokaryotes, unicellular eukaryotes and small multicellular organisms have a cosmopolitan distribution because of their minute sizes and their ability to form dormant stages (cysts, eggs, spores), which facilitate dispersal by air, dust, and migrating animals (Alongi 1990; Foissner 1999). However, a contrasting model to the generalization of the neutral model arises from the fact that the rate of dispersal of microorganisms is not sufficiently high to overcome historical dispersal limitations and human influence (Feral 2002; Foissner 2006; Foissner 2007). Even in the absence of any physical or genetic barriers for dispersal, the fact that some species or taxa have a ubiquitous or large-scale distribution does not imply that the majority will colonize and have time to adapt to the new environment (Kristiansen 2000). This facilitates the existence of endemic taxa, many of which remain to be discovered (Telford *et al.* 2006; Foissner 2007). In

fact, there are several microorganims with known restricted distributions, such as protists (Foissner 2006; Foissner 2007) and also protozoan (Bass *et al.* 2007; Smith & Wilkinson 2007). Thus suggesting that the ubiquity model might be bias because of under-sampling and/or misidentification of the samples (Mitchell & Meisterfeld 2005), representing a crucial problem for evaluation of distribution patterns (Finlay & Clarke 1999; Finlay *et al.* 2004). Nonetheless, the theory used to explain microorganism distribution do not fully explain marine meiofauna taxa distribution as many do not have a dormant stage but still have a wide dispersal distribution (Foissner 2007; Giere 2009).

Continental and even cosmopolitan distribution among meiobenthic taxa appear to result from a variety of dispersive mechanisms (Giere 2009) differing always between taxonomic groups (Brandt *et al.* 2007a). Meiofauna have been observed to aggregate in clumps of 'marine snow' (Silver *et al.* 1978), and transport via these small organic parcels would presumably provide nutritional sustenance during pelagic journeys. Some authors have additionally suggested that meiofauna can be transported via sea ice (Giere 2009) and floating pieces of rubbish (Barnes & Milner 2005). In terms of anthropogenic transport, ballast water (and its associated sediment) and fouling on vessels are known to carry meiofaunal populations across oceans during shipping operations (Giere 2009). Giere (2009) further suggests that these natural rafts, whether marine snow, sea ice or drifting islands, could have been responsible for repeated trans-oceanic long-distance transport resulting in little differentiation in meiofauna between distance regions. Although molecular data have revealed the vast scope of microbial diversity in virtually all habitats (Green *et al.* 2004; Slapeta *et al.* 2005), the fundamental conflict between ubiquitous and endemic distribution patterns, as basic to the 'everything is everywhere' dispute (Finlay 2002; Foissner 2007; Cermeno & Falkowski 2009; Nolte *et al.* 2010), is far from being solved. Environmental and biotic factors probably play a part in determining local patterns of meiobenthic abundance and diversity. In fact, environmental factors might explain more of the variance in community structure than geographic distance, especially when considering that habitats in close proximity are often similar with respect to their environmental characteristics (Fierer 2008). In this thesis, there was a correlation between distance and community similarity suggesting that dispersal could represent a limiting factor (Fuentes 2002; Martiny *et al.* 2011) thus for some representatives of eukaryotic marine meiofauna the neutral theory (Hubbell 2001)

should not apply. Bell (2000) emphasized that neutral patterns (i.e. random accumulations of differences between sites) may result from non-neutral processes and showed that metapopulation dynamics can provide neutral type patterns through sporadic populations at a landscape scale as local extinctions and re-colonization takes place. The direct assessment of life-history-traits such as individual reproductive success, that underpins the neutral model (Hubbell 2001) would be essential to decide which theory (neutral or niche) a given environmental community follows, but it is clear though that metapopulation dynamics apply to coastal meiobenthic communities. It is also likely that differences in meiobenthic communities might also result from environmental and abiotic factors correlated with distance, such as salinity, food availability and currents dynamic that is also closely linked to seasonality and this was not addressed in the studies in this thesis.

Diversity and composition of marine meiofauna

Species richness is the simplest way to describe community and regional diversity (Magurran 2005; Magurran & Henderson 2010) and thus it forms the basis of many ecological models of community structure (MacArthur & Wilson 1967). Quantifying species richness is essential, not only for basic comparisons among sites, but also to tackle the saturation of local communities colonized from regional source pools (Cornell 1999). Compilations of the global inventory of species are uncertain, mainly because many species may remain un-described and also because various taxonomic groups contain synonyms (species that have been given two or more names) or misidentified organisms (Alongi 1990; Foissner 1999; Costello & Wilson 2011), that contribute to inflate species number. In addition, different taxonomic experts differ in their estimates of the global number of species for particular taxa. For example estimates of overall taxon diversity of animal life on Earth are in the tens of millions, with ca. 1.8 million animal species formally described (May 1988; de Meeus *et al.* 2003). Coverage across different biological groups is very uneven with a known taxonomic deficit especially marked for microfauna and meiofauna due to inherent problems in identification and hyper abundant representatives.

Microorganism diversity in the oceans has been claimed to be much greater than previous estimates based on conventional molecular techniques (Sogin *et al.* 2006), but elevated estimates of diversity and the extended rare biosphere are likely to have

been derived from misinterpretation of data (Pedros-Alio 2007; Elshahed *et al.* 2008; Hamp *et al.* 2009; Kunin *et al.* 2009; Huse *et al.* 2010; Youssef *et al.* 2010). European marine biodiversity inventories, based on morphological analysis, reported a total 29713 species-level taxa that were catalogued from European seas, including the Arctic, deep-sea and Black-sea, where about 16056 were represented by marine meiofauna (Costello *et al.* 2006). This level of richness is proportionate with that of total estimates of 16000 protozoan species (Fenchel & Finlay 2006), albeit careful should be taken regarding these estimates since they are derived from morphological data only. In this thesis total estimates of taxon richness for meiobenthic taxa were of 1714 and 2700 at 96% and 99% cut-offs, respectively. Nonetheless, comparisons of richness should only be conducted between similar habitats because of known gradients of diversity (Lambshead *et al.* 2000; Gage *et al.* 2004; Hillebrand 2004; Allen & Gillooly 2006; Mittelbach *et al.* 2007; Danovaro *et al.* 2008b; Fuhrman *et al.* 2008; Gotelli *et al.* 2009) such as occur in deep-sea *vs.* coastal areas or tropics *vs.* Arctic (Gray 2000, 2001; Crame 2009; Clarke & Crame 2010). In addition, estimates derived from morphological analysis are not really comparable with molecular data due to constraints of taxon assessment that will be further addressed, especially given the swift development of sequencing approaches to delimit richness.

Species richness is a natural measure of biodiversity, but it is an elusive quantity that is difficult to measure (May 1988). A fundamental problem is that, for diverse taxa, as more individuals are sampled, more species will be recorded (Bunge & FitzPatrick 1993). The same, of course, is true for higher taxa, such as genera or families. The sampling curve rises relatively fast at first but slows down in later samples as increasingly singular taxa are added. In principle, for a survey of some well-defined spatial scope, an asymptote will eventually be reached and no further taxa will be added (Gotelli & Colwell 2001; Shaw *et al.* 2008). In this thesis the richness estimates derived from the rarefaction curves indicated that our high-throughput sequencing study was not sufficient to sample all biodiversity but also alludes to the fact that meiobenthic eukaryotic phylum diversity is extremely rich. Rarefaction estimators of richness produced in this study reveals the existence of more than 4500 OTUs of meiobenthic eukaryotes and more than 1600 OTUs just for nematodes, for the UK coast and also for some samples around France, Spain and Portugal, from a single habitat, i.e. a narrow marine littoral range. Thus the study

unmasked levels of intertidal meiofauna richness and diversity that clearly surpassed expectations.

Gotelli & Colwell (2001) also stress that species richness is not necessarily the ``correct'' way to measure diversity, but that patterns of diversity will be very sensitive to which measure is used. Currently it is widely established that patterns of species richness are greatly influenced by temperature and so there is a cline from the Arctic to the tropics (Crame 2009; Clarke & Crame 2010). The pattern of taxon richness along the sampled transects exhibited remarkable differences at a large and small-geographic scale, but no clear-cut north-south richness pattern was observed between the sampled sites. This could be associated with the absence of sharp and distinct latitudes and when present (for example Gambia) there was not enough sampling from those regions. Nonetheless, the fine-scale community structuring indicated a general meiobenthic community pattern ranking the Nematoda as the most abundant and dominant in marine sediments, as previously reported in taxonomic (Blaxter 2004) or small-scale molecular studies. These are usually followed in abundance by the Platyhelminthes and Copepods, nonetheless derived from the disparity of the sampled sites there were some exceptions to the rule. There was a clear community composition pattern between UK and Europe despite high percentages of widespread-shared OTUs between some samples. The UK samples showed a more homogenous community structure whereas France, Spain, Portugal and Gambia were clearly more heterogeneous and the Platyhelminthes were sometimes the most abundant phyla within the meiofauna representatives. In fact, regional differences in the composition of the metazoan meiofauna along continental margins are generally hard to detect, because of pronounced local variations (Soltwedel 2000) and numerous biotic and abiotic factors (Soltwedel 2000; Soltwedel *et al.* 2005) that probably hamper community assemblages. Although numerous environmental factors influence meiobenthic distribution and assemblage, grain size, is known to be the predominant driver of meiofaunal community structure and diversity (Giere 2009). Nonetheless, the fine scale community structuring also indicates that there are likely to be a host of additional biotic (e.g. prokaryote communities and organic matter) and abiotic (sediment grain shape, surface composition) micro-geographical factors responsible for community structuring within the benthos (Giere 2009). High-throughput studies continue to give insight and predictions of richness but further refinement (seasonal variation, biotic

parameters) is needed to understand the extent of community change and diversity across benthic ecosystems. The results reported in this thesis reveal that even using a conservative approach and eliminating chimeras and other artifacts from environmental datasets estimates of diversity and richness are still beyond our expectations.

Taxonomic assessment and OTU definition

For years morphological methods were the only method widely used to identify organisms. Moreover, the biosphere is so extraordinarily diverse that methodical cataloguing of biodiversity by traditional methods will probably never provide a complete 'species list' for the planet (May 1988). This recognition, coupled with real controversy over both how to define and how to diagnose 'species' (Adams 1998), has led to proposals for the development of molecular taxonomy based methods based on a defined part of the genome (Seberg *et al.* 2003; Tautz *et al.* 2003). For any taxon diagnosis system, heuristics are necessary for defining operational taxonomic units (OTU), based on sequence differences at short, orthologous marker gene sequences (Tautz *et al.* 2003; Blaxter 2004). This follows the general definition of OTUs as groups of organisms used in a taxonomic study without designation of taxonomic rank, which might be a *proxy* for species (Countway *et al.* 2005; Caron *et al.* 2009).

There is a mismatch between the total numbers of identified species when morphological taxonomic methods or molecular taxonomy are applied to identify species. This is perhaps unsurprising considering the time it takes for a taxonomist to identify a species using its morphology and the time it takes using gene sequences. Moreover, with the advent of next generation sequencing methods, which have the capacity to generate millions of sequences (so called pyro-sequences) in a few hours, this difference has become even more exaggerated. The advantages of DNA-based methods may be to improve both the accuracy and precision of species identification, helping to identify species hard to diagnose morphologically and also enabling non-taxonomist to identify species rapidly and reliably (Ebach & Holdrege 2005; Gregory 2005; Hebert & Gregory 2005; Schindel & Miller 2005; Smith 2005). Due to the large amount of sequencing data produced, it is anticipated that it will be possible to address the species richness of protists (Medinger *et al.* 2010) and also of other

microorganims. Such approaches have already shown that both eukaryotic and prokaryotic micro-organisms have large intraspecific genetic differences, even for conserved genes such as those coding for 18S rDNA (Fenchel & Finlay 2006).

Despite the great potential of molecular taxonomies there is a major impediment that is to establish the amount of genetic dissimilarity that should be accepted as intra-specific. This is complicated by the fact that rates of evolution for the same gene in different taxonomic groups may vary (Goldstein et al. 2000; Caron et al. 2009). In Bacteria for example, where the species concept is very conservative, species differ by values in the order of 1 to 2% in nSSU (16S or 18S rDNA) gene sequences (Hackstein 1997; Caron et al. 2009). In general for 18S rDNA genes, a 97% OTU similarity is often considered as a *proxy* for species (Vandamme et al. 1996; Uilenberg et al. 2004) and in fact in eukaryotic organisms such as nematodes, 99.5% similarity has been justified (Blaxter *et al.* 2005). Moreover, using 99% similarity the probabilities of recovering the maximum number of species' genotypes in a sample are higher (Porazinska *et al.* 2010). More stringent similarities are known to inflate estimates of taxon richness nonetheless this should be more appropriate for the study of environmental community patterns.

Hebert *et al.* (2003) showed that levels of sequence divergence for the COI (citochrome oxidase subunit 1) in eleven phyla of the metazoa, are regularly greater than 2% between species pairs with the exception of the Cnidaria with the majority of the species pairs having a sequence divergence less than 1%. The COI gene has high evolutionary rates and even so it has enabled the discrimination of closely allied species with a sequence divergence of 1-2%. Nuclear genes like the 18S rDNA have a highly conserved sequence core but also possess rapidly evolving variable regions (Woese *et al.* 1990; Hillis & Dixon 1991; Gillespie *et al.* 2006) and also have an extreme variation in gene copy number in eukaryotes (Amaral-Zettler *et al.* 2009). The choice of 18S rDNA as a marker for species has already been described in marine metazoa and is reported to have less intra-species variation than the COI gene (White 2011). However, there are exceptions to the rule and intra- and inter-species variability should not be ruled out when defining OTU richness and diversity. Richness estimates should always encompass at least two different cut-offs and for ecological comparisons such as biogeography or community composition between different sites the use of a higher cut-off will better reflect the distribution of genotypes within a biological species. Relating species to OTUs will always be a

contentious issue, but bioinformatic sequence comparisons and additional molecular assisted taxonomy will inevitably bridge the gap as further voucher specimens are linked to sequences (Blaxter 2003a; Blaxter *et al.* 2003; Markmann & Tautz 2005). OTU richness estimates provided by nuclear ribosomal genes (nSSU, 18S rDNA) from next generation sequences can vary according to the particular region surveyed, and absolute richness estimates based on different portions of the 18S rDNA gene could be used as complementary source of information.

The name of a species embodies a wealth of biological information, whereas 18S rDNA gene sequences do not, *per se*, provide any such information (Fenchel & Finlay 2006). Nonetheless, the use of a classical taxonomic approach, based on phenotypic characteristics combined with the increasing availability of large molecular datasets, could represent a means to determine local and global numbers of species. Even though it is considered desirable in community ecology that the individuals in a sample are identified to the species level (Gotelli 2004), it is not essential for the study of ecological patterns to assign species-level taxa (Somerfield & Gage 2000; Terlizzi *et al.* 2003; Quijón & Snelgrove 2006; Naser 2010). It would have thus not been mandatory to assign the inferred OTUs to species to perform ecological analyses (Pfenninger *et al.* 2007).

Molecular and in silico advances of pyro-tagged environmental samples

High-throughput techniques are not only being developed for single individuals but also for whole-populations, revolutionizing the study of ecology and evolution. These techniques include cheap, very fast and high capacity DNA sequencing, producing large amounts of data at a low cost (Ronaghi *et al.* 1998; Ronaghi 2001; Edwards *et al.* 2006; Rabouille *et al.* 2006; Turnbaugh *et al.* 2006; Meyer *et al.* 2007). The use of next generation sequencing to study environmental community biodiversity commences with an important phase of sample processing. Here, sampling strategies are known to influence diversity levels (Bett *et al.* 1994); one has to have sufficient representation of the study area to extrapolate biodiversity levels as close to reality as possible (Gotelli & Colwell 2001; Chao *et al.* 2009). In this study such explicit assumptions were clearly not achieved for all sampling areas. Further to this, sampling from the intertidal zone will likely result in different numbers of

richness and diversity levels compared to subtidal, neritic and oceanic zones. The same applies for differences between the depth sampling areas (Soltwedel *et al.* 2005) such as the benthos *versus* pelagic zones. Sample procedure and manipulation are also important for keeping sample integrity. Conceptual issues such as the choice of the gene(s) used in the study are also very important considerations. For example, some species possess multiple RNA gene copies with somewhat different base pair compositions (Scholin *et al.* 1993). These different sequences could conceivably produce multiple OTUs for a single specimen if the differences are large enough, although it appears to be relatively rare (Caron *et al.* 2009). Similarly, the use of rapidly evolving genes or intergenic spacer regions might result in the creation of multiple OTUs for individuals that would be grouped into a single species using other criteria (O'Mahony *et al.* 2007).

DNA manipulation (Medinger *et al.* 2010) and PCR conditions (Qiu *et al.* 2001; Wu *et al.* 2010) are probably one of the main drivers of diversity artefacts in pyro-sequenced samples. In fact, almost all steps of the molecular approach can introduce biases or errors (von Wintzingerode *et al.* 1997). Specifically, during PCR-amplification when incomplete extension occurs the resulting amplicon re-anneals to a foreign DNA strand and is copied to completion in the following PCR cycles, generating chimeric sequences (Wang & Wang 1996, 1997). These chimeras are then hugely amplified during the pyrosequencing step and when further analysed may be mistakenly identified as new species (Kunin *et al.* 2009; Huse *et al.* 2010; Quince *et al.* 2011), thus greatly inflating community richness. Although not entirely effective, chimera formation can be minimized experimentally by PCR optimization, by reducing the number of PCR cycles and increasing extension time, using lower template concentrations and a shorter amplicon size (Wang & Wang 1997; Qiu *et al.* 2001; Lahr & Katz 2009; Engelbrektson *et al.* 2010).

Massively parallel sequencing is used to analyse a population of PCR products arising from amplification of environmental samples and is followed by bioinformatic analysis in order to discriminate biodiversity levels. The first obstacle faced is the analysis of the huge amounts of data generated by these experiments. This starts with the fact that there are currently no unified data formats and the need to map millions to billions of sequences (Nowrousian 2010). The development of bioinformatics tools and databases to better cope with these types of data will be one of the main factors determining how effective next generation sequencing will be for

a wider research community. Caution is required when choosing a program to analyse the massive number of sequences produced. Several problems inherent to the analysis of 18S rDNA gene pyro-sequences are currently recognized and mostly focus on the generation of high levels of richness associated with chimeric sequences (Huber et al. 2004; Ashelford et al. 2005; Gonzalez et al. 2005; Ashelford et al. 2006; Haas et al. 2011; Quince et al. 2011) and the way sequences are clustered and also the algorithm used (Ashelford et al. 2006; Kunin et al. 2009; Quince et al. 2009; Huse et al. 2010). Further to this, the use of reference databases to detect chimeric molecules in environmental datasets also requires caution because environmental samples are very diverse and reference data is unlikely to be representative of the true diversity. Furthermore, the existence of chimeric sequences in public DNA databases is well known (Hugenholtzt & Huber 2003; Ashelford *et al.* 2005) and the risk of classifying chimeras as new organisms is becoming higher than the risk of neglecting non-chimeric ones.

The advent of next generation sequencing represents a major breakthrough in molecular biology, genetics, and beyond, as well as a great leap forward for genomics and systems biology analyses (Nowrousian 2010). Overall, this study will significantly contribute to a better understanding of chimera formation and pyrosequencing strategies that should be considered when conducting any study focused on the PCR amplification of environmental DNA.

CHAPTER 7 – Bibliography

AmpliconNoise Google Code Project. [http://code.google.com/p/ampliconnoise/].

Abebe E, Blaxter M (2003) Comparison of biological, molecular, and morphological methods of species identification in a set of cultured Panagrolaimus isolates. *Journal of Nematology* **35**, 119-128.

Abebe E, Mekete T, Thomas W (2011) A critique of current methods in nematode taxonomy. *African Journal of Biotechnology* **10**, 312-323.

Abebe E, Wilfrida D, Paul DL (2008) Global diversity of nematodes (Nematoda) in freshwater. *Hydrobiologia* **595**, 67-78.

Adams BJ (1998) Species concepts and the evolutionary paradigm in modem nematology. *J Nematol* **30**, 1-21.

Aguinaldo AM, Turbeville JM, Linford LS, *et al.* (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**, 489-493.

Allen AP, Brown JH, Gillooly JF (2002) Global biodiversity, biochemical kinetics, and the energetic-equivalence rule. *Science* **297**, 1545-1548.

Allen AP, Gillooly JF (2006) Assessing latitudinal gradients in speciation rates and biodiversity at the global scale. *Ecol Lett* **9**, 947-954.

Allen MJ, Martinez-Martinez J, Schroeder DC, Somerfield PJ, Wilson WH (2007) Use of microarrays to assess viral diversity: from genotype to phenotype. *Environ Microbiol* **9**, 971-982.

Aller R, Aller J (1992) Meiofauna and solute transport in marine muds. *Limnol Oceanogr* **37**, 1018-1033.

Alongi DM (1987) The Influence of Mangrove-Derived Tannins on Intertidal Meiobenthos in Tropical Estuaries. *Oecologia* **71**, 537-540.

Alongi DM (1990) The Ecology of Tropical Soft-Bottom Benthic Ecosystems. *Oceanography and Marine Biology* **28**, 381-496.

Altschul SF, Madden TL, Schäffer AA, *et al.* (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.". *Nucleic Acids Research* **25**, 3389-3402.

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE* **4**, e6372.

Amend AS, Seifert KA, Bruns TD (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology* **19**, 5555-5565.

Anderson M (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**, 32-46.

Andrassy I, Gibson JAE (2007) Nematodes from saline and freshwater lakes of the Vestfold Hills, East Antarctica, including the description of Hypodontolaimus antarcticus sp n. *Polar Biology* **30**, 669-678.

Angly FE, Felts B, Breitbart M, *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol* **4**, e368.

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology* **71**, 7724-7736.

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2006) New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl Environ Microbiol* **72**, 5734-5741.

Austen M (2004) Natural nematode communities are useful tools to address ecological and applied questions. *Nematol Monogr Perspect* **2**, 1-17.

Austen MC, Lambshead PJD, Hutchings PA, *et al.* (2002) Biodiversity links above and below the marine sediment-water interface that may influence community stability. *Biodiversity and Conservation* **11**, 113-136.

Austen MC, Widdicombe S (2006) Comparison of the response of meio- and macrobenthos to disturbance and organic enrichment. *Journal of Experimental Marine Biology and Ecology* **330**, 96-104.

Avise JC (1994) *Molecular markers, natural history and evolution* Chapman and Hall, New York.

Avise JC (2000) Stability, equilibrium and molecular aspects of conservation in marine species. *Hydrobiologia* **420**, Xi-Xii.

Avise JC, Johns GC (1999) Proposal for a standardized temporal scheme of biological classification for extant species. *Proc Natl Acad Sci U S A* **96**, 7358-7363.

Avise JC, Wollenberg K (1997) Phylogenetics and the origin of species. *Proc Natl Acad Sci U S A* **94**, 7748-7755.

Baas-Becking L (1934) *Geobiologie of inleiding tot de milieukunde* The Hague, the Netherlands.

Baermann G (1917) Eine einfache Methode zur Auffindung von Ancyclostomum (Nematode) Larven in Erdproben. *Nederlands Tijdscrift Voor Geneeskunde* **57**, 131-137.

Bahr M, Crump BC, Klepac-Ceraj V, *et al.* (2005) Molecular characterization of sulfate-reducing bacteria in a New England salt marsh. *Environ Microbiol* **7**, 1175-1185.

Balakrishnan R (2005) Species concepts, species boundaries and species identification: a view from the tropics. *Syst Biol* **54**, 689-693.

Baldauf SL (2003) The deep roots of eukaryotes. *Science* **300**, 1703-1706.

Barber PH, Palumbi SR, Erdmann MV, Moosa MK (2000) Biogeography. A marine Wallace's line? *Nature* **406**, 692-693.

Barnes D, Milner P (2005) Drifting plastic and its consequences for sessile organism dispersal in the Atlantic Ocean. *Marine Biology* **146**, 815-825.

Barnes DKA, Linse K, Enderlein P, *et al.* (2008) Marine richness and gradients at Deception Island, Antarctica. *Antarctic Science* **20**, 271-279.

Bass D, Richards TA, Matthai L, Marsh V, Cavalier-Smith T (2007) DNA evidence for global dispersal and probable endemicity of protozoa. *BMC Evol Biol* **7**, 162.

Bell G (2000) The distribution of abundance in neutral communities. *American Naturalist* **155**, 606–617.

Bell G (2001) Neutral macroecology. *Science* **293**, 2413-2418.

Bell SS, Watzin MC, Coull BC (1978) Biogenic structure and its effect on the spatial heterogeneity of meiofauna in a salt marsh. *J. Exp. Mar. Biol. Ecol.* **35**, 99-107.

Bellemain E, Carlsen T, Brochmann C, *et al.* (2010) ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Microbiol* **10**, 189.

Berg P, Risgaard-Petersen N, Rysgaard S (1998) Interpretation of measured concentration profiles in sediment pore water. *Lirnnol Oceanogr* **6**, 1500-1510.

Bett BJ, Vanreusel A, Vincx M, *et al.* (1994) Sampler Bias in the Quantitative Study of Deep-Sea Meiobenthos. *Marine Ecology-Progress Series* **104**, 197-203.

Bhadury P, Austen M, Bilton D, *et al.* (2005) Combined morphological and molecular analysis of individual nematodes through short-term preservation in formalin. *Molecular Ecology Notes* **5**, 965-968.

Bhadury P, Austen MC, Bilton DT, *et al.* (2006) Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. *Mar Ecol Prog Ser* **320**, 1-9.

Bhadury P, Austen MC, Bilton DT, *et al.* (2008) Evaluation of combined morphological and molecular techniques for marine nematode (Terschellingia spp.) identification. *Marine Biology* **154**, 509-518.

Bickford D, Lohman DJ, Sodhi NS, *et al.* (2007) Cryptic species as a window on diversity and conservation. *Trends Ecol Evol* **22**, 148-155.

Binladen J, Gilbert MT, Bollback JP, *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* **2**, e197.

Blanca J, Canizares J, Roig C, *et al.* (2011) Transcriptome characterization and high throughput SSRs and SNPs discovery in Cucurbita pepo (Cucurbitaceae). *BMC Genomics* **12**, 104.

Blaxter M (2003a) Counting angels with DNA. *Nature* **421**, 122-124.

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

Blaxter M, Elsworth B, Daub J (2004) DNA taxonomy of a neglected animal phylum: an unexpected diversity of tardigrades. *Philos Trans R Soc Lond B Biol Sci* **271**, S189-S192.

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

Blaxter M, Floyd R, Abebe E (2003) Molecular barcoding for nematode identification and diversity studies. *Journal of Nematology* **35**, 326.

Blaxter M, Mann J, Chapman T, *et al.* (2005) Defining operational taxonomic units using DNA barcode data. *Philos Trans R Soc Lond B Biol Sci* **360**, 1935-1943.

Blaxter ML (2004) The promise of a DNA taxonomy. *Philos Trans R Soc Lond B Biol Sci* **359**, 669-679.

Blaxter ML, De Ley P, Garey JR, *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**, 71-75.

Blouin MS (2000) Neutrality tests on mtDNA: unusual results from nematodes. *J Hered* **91**, 156-158.

Blouin MS (2002) Molecular prospecting for cryptic species of nematodes: mitochondrial DNA versus internal transcribed spacer. *Int J Parasitol* **32**, 527-531.

Blouin MS, Yowell CA, Courtney CH, Dame JB (1998) Substitution bias, rapid saturation, and the use of mtDNA for nematode systematics. *Mol Biol Evol* **15**, 1719-1727.

Blow N (2008) Exploring unseen communities. *Nature* **453**, 687-690.

Bohannon J (2007) Metagenomics - Ocean study yields a tidal wave of microbial DNA. *Science* **315**, 1486-1487.

Bolger T (2001) The functional value of species biodiversity – a review. *Proc. Irish Roy. Acad.* **101B**, 199–204.

Bongers T (1988) De Nematoden van Nederland. Koninklijke Nederlandse Natuurhistorische Vereniging, Utrecht

Borgonie G, Garcia-Moyano A, Litthauer D, *et al.* (2011) Nematoda from the terrestrial deep subsurface of South Africa. *Nature* **474**, 79-82.

Bostrom C, O'Brien K, Roos C, Ekebom J (2006) Environmental variables explaining structural and functional diversity of seagrass macrofauna in an archipelago landscape. *Journal of Experimental Marine Biology and Ecology* **335**, 52-73.

Boucher G, Clavier J (1990) Contribution of Benthic Biomass to Overall Metabolism in New-Caledonia Lagoon Sediments. *Marine Ecology-Progress Series* **64**, 271-280.

Boucher G, Lambshead PJD (1995) Ecological biodiversity of marine nematodes in samples from temperate, tropical, and deep sea regions. *Conserv Biol* **9**, 1594-1604.

Bouchet P (2006) The magnitude of marine biodiversity. In: *The exploration of marine biodiversity.Scientific and technological challenges* (ed. Duarte C), pp. 33-64. Fundacion BBVA.

Brandt A, De Broyer C, De Mesel I, *et al.* (2007a) The biodiversity of the deep Southern Ocean benthos. *Philos Trans R Soc Lond B Biol Sci* **362**, 39-66.

Brandt A, Gooday AJ, Brandao SN, *et al.* (2007b) First insights into the biodiversity and biogeography of the Southern Ocean deep sea. *Nature* **447**, 307-311.

Bremner J, Rogers SI, Frid CLJ (2003) Assessing functional diversity in marine benthic ecosystems: a comparison of approaches. *Marine Ecology-Progress Series* **254**, 11-25.

Bremner J, Rogers SI, Frid CLJ (2006) Matching biological traits to environmental conditions in marine benthic ecosystems. *Journal of Marine Systems* **60**, 302-316.

Brown J (1995) *Macroecology* University of Chicago Press, Chicago.

Brown MV, Philip GK, Bunge JA, *et al.* (2009) Microbial community structure in the North Pacific ocean. *ISME J* **3**, 1374-1386.

Brunel P (2006) Marine Biodiversity. In: *Oceanography* (eds. Chen C-TA, Nihoul JCJ). Eolss Publishers, Oxford, UK.

Bucklin A, Wiebe PH, Smolenack SB, *et al.* (2007) DNA barcodes for species identification of euphausiids (Euphausiacea, Crustacea). *Journal of Plankton Research* **29**, 483-493.

Bunge J, FitzPatrick M (1993) Estimating the number of species: a review. *Journal of the American Statistical Association* **88**, 364–373.

Burki F, Kudryavtsev A, Matz MV, *et al.* (2010) Evolution of Rhizaria: new insights from phylogenomic analysis of uncultivated protists. *BMC Evol Biol* **10**, 377.

Buyck B (1999) Taxonomists are an endangered species in Europe. *Nature* **401**, 321.

Carlson JK, Randall TA, Mroczka ME (1997) Feeding habits of winter flounder (Pleuronectes. americanus) in a habitat exposed to anthropogenic disturbance. *Science* **21**, 65-73.

Caron DA, Countway PD, Savai P, *et al.* (2009) Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl Environ Microbiol* **75**, 5797-5808.

Carvalho G, Creer S, Allen M, *et al.* (2009) *Genomics in the discovery and monitoring of marine biodiversity* Springer-Verlag. *In Press.*

Carvalho G, Creer S, Allen M, *et al.* (2010) *Genomics in the discovery and monitoring of marine biodiversity* Springer Netherlands.

Casiraghi M, Bazzocchi C, Mortarino M, Ottina E, Genchi C (2006) A simple molecular method for discriminating common filarial nematodes of dogs (Canis familiaris). *Veterinary Parasitology* **141**, 368-372.

Cavalier-Smith T (2004) Only six kingdoms of life. *Proc Biol Sci* **271**, 1251-1262.

Cavalier-Smith T (2006) Rooting the tree of life by transition analyses. *Biol Direct* **1**, 19.

Cavalier-Smith T (2010) Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biol Lett* **6**, 342-345.

Cermeno P, de Vargas C, Abrantes F, Falkowski PG (2010) Phytoplankton biogeography and community stability in the ocean. *PLoS ONE* **5**, e10037.

Cermeno P, Falkowski PG (2009) Controls on diatom biogeography in the ocean. *Science* **325**, 1539-1541.

Chao A, Chazdon RL, Colwell RK, Shen TJ (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* **62**, 361-371.

Chao A, Colwell RK, Lin CW, Gotelli NJ (2009) Sufficient sampling for asymptotic minimum species richness estimators. *Ecology* **90**, 1125-1133.

Chapin FS, Zavaleta ES, Eviner VT, *et al.* (2000) Consequences of changing biodiversity. *Nature* **405**, 234-242.

Chave J (2004) Neutral theory and community ecology. *Ecology Letters* **7**, 241-253.

Chazdon R, Colwell R, Denslow J, Guariguata M (1998) Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica. In: *Forest biodiversity research, monitoring and modeling: Conceptual background and Old World case studies* (eds. Dallmeier F, Comiskey J), pp. 285-309. Parthenon Publishing, Paris.

Chen G, Znosko BM, Jiao X, Turner DH (2004) Factors affecting thermodynamic stabilities of RNA 3 x 3 internal loops. *Biochemistry* **43**, 12865-12876.

Chou HH, Holmes MH (2001) DNA sequence quality trimming and vector removal. *Bioinformatics* **17**, 1093-1104.

Clarke A (1992) Is there a latitudinal diversity cline in the sea? *Trends Ecol Evol* **7**, 286-287.

Clarke A, Crame JA (2010) Evolutionary dynamics at high latitudes: speciation and extinction in polar marine faunas. *Philos Trans R Soc Lond B Biol Sci* **365**, 3655-3666.

Clarke K, Gorley R (2006) PRIMER v6: User Manual/Tutorial. *PRIMER-E, Plymouth*.

Clayton DA (1984) Transcription of the mammalian mitochondrial genome. *Annual Review of Biochemistry* **53**, 573-594.

Clayton RA, Sutton G, Hinkle PS, Jr., Bult C, Fields C (1995) Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. *Int J Syst Bacteriol* **45**, 595-599.

Coghlan A (2005) Nematode genome evolution. *WormBook*, 1-15.

Cole JR, Chai B, Marsh TL, *et al.* (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Research* **31**, 442-443.

Coleman AW (2002) Microbial eukaryote species. *Science* **297**, 337; author reply 337.

Colwell RK, Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. *Philos Trans R Soc Lond B Biol Sci* **345**, 101-118.

Cook AA, Bhadury P, Debenham NJ, *et al.* (2005) Denaturing gradient gel electrophoresis (DGGE) as a tool for identification of marine nematodes. *Mar Ecol Prog Ser* **291**, 103-113.

Cook AA, Lambshead PJD, Hawkins LE, Mitchell N, Levin LA (2000) Nematode abundance at the oxygen minimum zone in the Arabian Sea. *Deep-Sea Research Part Ii-Topical Studies in Oceanography* **47**, 75-85.

Coomans A (2000) Nematode systematics: past, present and future. *Nematology* **2**, 3-7.

Copley JTP, Flint HC, Ferrero TJ, Van Dover CL (2007) Diversity of meiofauna and free-living nematodes in hydrothermal vent mussel beds on the northern and southern East Pacific Rise. *Journal of the Marine Biological Association of the United Kingdom* **87**, 1141-1152.

Coppe A, Pujolar JM, Maes GE, *et al.* (2010) Sequencing, de novo annotation and analysis of the first Anguilla anguilla transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel. *BMC Genomics* **11**, 635.

Cornell H (1999) Unsaturation and regional infuences on species richness in ecological communities: a review of the evidence. *Ecoscience* **6**, 303±315.

Costello M, Emblow C, Bouchet P, Legakis A (2006) European marine biodiversity inventory and taxonomic resources: state of the art and gaps in knowledge. *Marine Ecology Progress Series* **316**, 257-268.

Costello MJ, Coll M, Danovaro R, *et al.* (2010) A census of marine biodiversity knowledge, resources, and future challenges. *PLoS ONE* **5**, e12110.

Costello MJ, Wilson SP (2011) Predicting the number of known and unknown species in European seas using rates of description. *Global Ecology and Biogeography* **20**, 319-330.

Coull B (1999) Role of meiofauna in estuarine soft-bottom habitats. *Australian Journal of Ecology* **24**, 327–343.

Countway PD, Gast RJ, Savai P, Caron DA (2005) Protistan diversity estimates based on 18S rDNA from seawater incubations in the Western North Atlantic. *J Eukaryot Microbiol* **52**, 95-106.

Covich A, Austen M, Barlocher F, *et al.* (2004) The Role of Biodiversity in the Functioning of Freshwater and Marine Benthic Ecosystems. *BioScience* **54**, 767-775.

Crame JA (2009) Evolution. Time's stamp on modern biogeography. *Science* **323**, 720-721.

Crist TO, Veech JA (2006) Additive partitioning of rarefaction curves and species-area relationships: unifying alpha-, beta- and gamma-diversity with sample size and habitat area. *Ecol Lett* **9**, 923-932.

Cronn R, Cedroni M, Haselkorn T, Grover C, Wendel JF (2002) PCR-mediated recombination in amplification products derived from polyploid cotton. *Theor Appl Genet* **104**, 482-489.

Danovaro R, Gambi C (2002) Biodiversity and trophic structure of nematode assemblages in seagrass systems: evidence for a coupling with changes in food availability. *Marine Biology* **141**, 667-677.

Danovaro R, Gambi C, Dell'Anno A, *et al.* (2008a) Exponential decline of deep-sea ecosystem functioning linked to benthic biodiversity loss. *Curr Biol* **18**, 1-8.

Danovaro R, Gambi C, Lampadariou N, Tselepides A (2008b) Deep-sea nematode biodiversity in the Mediterranean basin: testing for longitudinal, bathymetric and energetic gradients. *Ecography* **31**, 231-244.

Darwin C (1859) *On the Origin of Species by Means of Natural Selection* John Murray, London.

de Jonge V, Bouwman L (1977) A simple density separation technique for quantitative isolation of meiobenthos using the colloidal silica Ludox-TM. *Mar Biol* **42**, 143–148.

De Ley P, Blaxter M (2004) A new system for Nematoda: combining morphological characters with molecular trees, and translating clades into ranks and taxa. In: *Nematology Monographs and Perspectives 2* (eds. Cook R, Hunt DJ), pp. 633–653. Brill Publishing, Leiden.

De Ley P, De Ley IT, Morris K, *et al.* (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**, 1945-1958.

de Meeus T, Durand P, Renaud F (2003) Species concepts: what for? *Trends Parasitol* **19**, 425-427.

De Queiroz K (2007) Species concepts and species delimitation. *Syst Biol* **56**, 879-886.

de Troch M, Melgo-Ebarle JL, Angsinco-Jimenez L, Gheerardyn H, Vincx M (2008) Diversity and habitat selectivity of harpacticoid copepods from sea grass beds in Pujada Bay, the Philippines. *Journal of the Marine Biological Association of the United Kingdom* **88**, 515-526.

Derycke S, Backeljau T, Vlaeminck C, *et al.* (2007) Spatiotemporal analysis of population genetic structure in Geomonhystera disjuncta (Nematoda, Monhysteridae) reveals high levels of molecular diversity. *Marine Biology* **151**, 1799-1812.

Derycke S, Fonseca G, Vierstraete A, *et al.* (2008a) Disentangling taxonomy within the Rhabditis (Pellioditis) marina (Nematoda, Rhabditidae) species complex using molecular and morhological tools. *Zoological Journal of the Linnean Society* **152**, 1-15.

Derycke S, Remerie T, Backeljau T (2008b) Phylogeography of the *Rhabditis* (*Pellioditis*) marina species complex: evidence for long-distance dispersal, and for range expansions and restricted gene flow in the northeast Atlantic. *Molecular Ecology* **17**, 3306-3322.

Derycke S, Remerie T, Vierstraete A, *et al.* (2005) Mitochondrial DNA variation and cryptic speciation within the free-living marine nematode *Pellioditis marina*. *Marine Ecology-Progress Series* **300**, 91-103.

Desnues C, Rodriguez-Brito B, Rayhawk S, *et al.* (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452**, 340-345.

Dixon P, Palmer MW (2003) VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* **14**, 927-930.

Doi RH, Igarashi RT (1965) Conservation of Ribosomal and Messenger Ribonucleic Acid Cistrons in Bacillus Species. *J Bacteriol* **90**, 384-390.

Donn S, Griffiths BS, Neilson R, Daniell TJ (2008) DNA extraction from soil nematodes for multi-sample community studies. *Applied Soil Ecology* **38**, 20-26.

Dubnau D, Smith I, Morell P, Marmur J (1965) Gene conservation in Bacillus species. I. Conserved genetic and nucleic acid base sequence homologies. *Proceedings of the National Academy of Sciences USA* **54**, 491-498.

Dunn CW, Hejnol A, Matus DQ, *et al.* (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745-749.

Ebach MC, Holdrege C (2005) DNA barcoding is no substitute for taxonomy. *Nature* **434**, 697.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797.

Edwards RA, Rodriguez-Brito B, Wegley L, *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**, 57.

Elshahed MS, Youssef NH, Spain AM, *et al.* (2008) Novelty and uniqueness patterns of rare members of the soil biosphere. *Appl Environ Microbiol* **74**, 5422-5428.

Engelbrektson A, Kunin V, Wrighton KC, *et al.* (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J* **4**, 642-647.

Enquist BJ, Allen AP, Brown JH, *et al.* (2007) Biological scaling: does the exception prove the rule? *Nature* **445**, E9-10; discussion E10-11.

Esteves AM, Silva VMAPD (1998) The behavior of sugar flotation technique in meiofauna extraction from different samples. *Tropical Ecology* **39**, 283-284.

Fenchel T, Finlay BJ (2004) The Ubiquity of Small Species: Patterns of Local and Global Diversity. *BioScience* **54**, 777-784.

Fenchel T, Finlay BJ (2006) The diversity of microbes: resurgence of the phenotype. *Philos Trans R Soc Lond B Biol Sci* **361**, 1965-1973.

Feral JP (2002) How useful are the genetic markers in attempts to understand and manage marine biodiversity? *Journal of Experimental Marine Biology and Ecology* **268**, 121-145.

Ferguson A (1980) *Biochemical Systematics and Evolution* Blackie, Glasgow and London.

Fierer N (2008) Microbial biogeography: patterns in microbial diversity across space and time. In: *Accessing Uncultivated Microorganisms: from the Environment to Organisms and Genomes and Back* (ed. Zengler K), pp. 95-115. ASM Press, Washington DC.

Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. *Science* **296**, 1061-1063.

Finlay BJ (2004) Protist taxonomy: an ecological perspective. *Philos Trans R Soc Lond B Biol Sci* **359**, 599-610.

Finlay BJ, Clarke KJ (1999) Apparent global ubiquity of species in the protist genus Paraphysomonas. *Protist* **150**, 419-430.

Finlay BJ, Esteban GF (2001) Exploring Leeuwenhoek's legacy: the abundance and diversity of protozoa. *Int Microbiol* **4**, 125-133.

Finlay BJ, Esteban GF, Fenchel T (2004) Protist diversity is different? *Protist* **155**, 15-22.

Finlay BJ, Fenchel T (2004) Cosmopolitan metapopulations of free-living microbial eukaryotes. *Protist* **155**, 237-244.

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

Floyd RM, Rogers AD, Lambshead PJD, Smith CR (2005) Nematode-specific PCR primers for the 18S small subunit rRNA gene. *Molecular Ecology Notes* **5**, 611-612.

Foissner W (1999) Protist diversity: estimates of the near- imponderable. *Protist* **150**, 363—368.

Foissner W (2006) Biogeography and dispersal of micro-organisms: a review emphasizing protists. *Acta Protozool* **45**, 111-136.

Foissner W (2007) Protist diversity and distribution: some basic considerations. *Biodiversity and Conservation* **17**, 235–242.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoeck R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* **3**, 294-299.

Fontaneto D, Kaya M, Herniou EA, Barraclough TG (2009) Extreme levels of hidden diversity in microscopic animals (Rotifera) revealed by DNA taxonomy. *Molecular Phylogenetics and Evolution*.

Forest F, Grenyer R, Rouget M, *et al.* (2007) Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature* **445**, 757-760.

Freckman DW, Blackburn TH, Brussaard L, *et al.* (1997) Linking biodiversity and ecosystem functioning of soils and sediments. *Ambio* **26**, 556-562.

Frias-Lopez J, Shi Y, Tyson GW, *et al.* (2008) Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A* **105**, 3805-3810.

Fuentes M (2002) Seed dispersal and tree species diversity. *Trends Ecol Evol* **17**, 550.

Fuhrman JA (2009) Microbial community structure and its functional implications. *Nature* **459**, 193-199.

Fuhrman JA, Steele JA, Hewson I, *et al.* (2008) A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci U S A* **105**, 7774-7778.

Gage JD, Lambshead PJD, Bishop JDD, Stuart CT, Jones NS (2004) Large-scale biodiversity pattern of Cumacea (Peracarida : Crustacea) in the deep Atlantic. *Marine Ecology-Progress Series* **277**, 181-196.

Galamba A, Soetaert K, Buyssens P, *et al.* (2001) Molecular and biochemical characterisation of Mycobacterium smegmatis alcohol dehydrogenase C. *Fems Microbiology Letters* **196**, 51-56.

Garaizar J, Rementeria A, Porwollik S (2006) DNA microarray technology: a new tool for the epidemiological typing of bacterial pathogens? *FEMS Immunol Med Microbiol* **47**, 178-189.

Gaston KJ (2007) Latitudinal gradient in species richness. *Curr Biol* **17**, R574.

Gaston KJ (2009) Geographic range limits of species. *Proc Biol Sci* **276**, 1391-1393.

Gee JM, Warwick RM (1996) A study of global biodiversity patterns in the marine motile fauna of hard substrata. *Journal of the Marine Biological Association of the United Kingdom* **76**, 177-184.

Gheskiere T, Hoste E, Vanaverbeke J, Vincx M, Degraer S (2004) Horizontal zonation patterns and feeding structure of marine nematode assemblages on a macrotidal, ultra-dissipative sandy beach (De Panne, Belgium). *Journal of Sea Research* **52**, 211-226.

Gheskiere T, Magda V, Greet P, Steven D (2006) Are strandline meiofaunal assemblages affected by a once-only mechanical beach cleaning? Experimental findings. *Marine Environmental Research* **61**, 245-264.

Gheskiere T, Vincx M, Urban-Malinga B, *et al.* (2005a) Nematodes from wave-dominated sandy beaches: diversity, zonation patterns and testing of the isocommunities concept. *Estuarine Coastal and Shelf Science* **62**, 365-375.

Gheskiere T, Vincx M, Weslawski JM, Scapini F, Degraer S (2005b) Meiofauna as descriptor of tourism-induced changes at sandy beaches. *Marine Environmental Research* **60**, 245-265.

Giere O (2009) *Meiobenthology: The microscopic motile fauna of aquatic sediments* Springer-Verlag, Heidelberg.

Gillespie JJ, Johnston JS, Cannone JJ, Gutell RR (2006) Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of Apis mellifera (Insecta: Hymenoptera): structure, organization, and retrotransposable elements. *Insect Mol Biol* **15**, 657-686.

Goldstein PZ, DeSalle R, Amato G, Vogler AP (2000) Conservation genetics at the species boundary. *Conservation Biology* **14**, 120-131.

Gonzalez JM, Zimmermann J, Saiz-Jimenez C (2005) Evaluating putative chimeric sequences from PCR-amplified products. *Bioinformatics* **21**, 333-337.

Good IJ (1953) The population frequencies of species and the estimation of populaton parameters. *Biometrika* **40**, 237-264.

Gotelli NJ (2004) A taxonomic wish-list for community ecology. *Philos Trans R Soc Lond B Biol Sci* **359**, 585-597.

Gotelli NJ, Anderson MJ, Arita HT, *et al.* (2009) Patterns and causes of species richness: a general simulation model for macroecology. *Ecol Lett* **12**, 873-886.

Gotelli NJ, Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* **4**, 379-391.

Gotelli NJ, Dorazio RM, Ellison AM, Grossman GD (2010) Detecting temporal trends in species assemblages with bootstrapping procedures and hierarchical models. *Philos Trans R Soc Lond B Biol Sci* **365**, 3621-3631.

Gracey AY (2007) Interpreting physiological responses to environmental change through gene expression profiling. *J Exp Biol* **210**, 1584-1592.

Gray J (2000) The measurement of marine species diversity, with an application to the benthic fauna of the Norwegian continental shelf. *J Exp Mar Bio Ecol* **250**, 23-49.

Gray J (2001) Marine diversity: the paradigms in patterns of species richness examined. *Scientia Marina* **65**, 41-56.

Gray JS (1996) Is coastal diversity as high as that of the deep sea? In: *Biology and ecology of shallow coastal waters. 28th European Marine Biology Symposium.* (ed. Eleftheriou A, Ansell, A.D., Smith, C.J.), pp. 181–184, Hersonnissos, Greece.

Gray JS (1997) Marine biodiversity: patterns, threats and conservation needs. *Biodiversity and Conservation* **6**, 153-175.

Gray JS, Elliott M (2009) *Ecology of marine sediments. From science to management* Oxford University Press New York.

Green JL, Holmes AJ, Westoby M, *et al.* (2004) Spatial scaling of microbial eukaryote diversity. *Nature* **432**, 747-750.

Gregory TR (2005) DNA barcoding does not compete with taxonomy. *Nature* **434**, 1067.

Haas BJ, Gevers D, Earl AM, *et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21**, 494-504.

Hackstein J (1997) Eukaryotic molecular biodiversity: systematic approaches for the assessment of symbiotic associations. *Antonie van Leeuwenhoek* **72**, 63–76.

Hajibabaei M, deWaard JR, Ivanova NV, *et al.* (2005) Critical factors for assembling a high volume of DNA barcodes. *Philos Trans R Soc Lond B Biol Sci* **360**, 1959-1967.

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

Hajibabaei M, Singer GA, Clare EL, Hebert PD (2007a) Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. *BMC Biol* **5**, 24.

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007b) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics* **23**, 167-172.

Hall N (2007) Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* **210**, 1518-1525.

Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Research*, in press.

Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods* **5**, 235-237.

Hamp TJ, Jones WJ, Fodor AA (2009) Effects of experimental choices and analysis noise on surveys of the "rare biosphere". *Appl Environ Microbiol* **75**, 3263-3270.

Hao X, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* **27**, 611-618.

Harrison S, Hastings A (1996) Genetic and evolutionary consequences of metapopulation structure. *Trends Ecol Evol* **11**, 180-183.

Hebert PD, Gregory TR (2005) The promise of DNA barcoding for taxonomy. *Syst Biol* **54**, 852-859.

Hebert PD, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proc Natl Acad Sci U S A* **101**, 14812-14817.

Hebert PD, Ratnasingham S, deWaard JR (2003a) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Biol Sci* **270 Suppl 1**, S96-99.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003b) Biological identifications through DNA barcodes **270**, 313-321.

Heip C, Huys R, Vincx M, *et al.* (1990) Composition, Distribution, Biomass and Production of North-Sea Meiofauna. *Netherlands Journal of Sea Research* **26**, 333-342.

Heip C, Vincx M, Vranken G (1985) The Ecology of Marine Nematodes. *Oceanography and Marine Biology* **23**, 399-489.

Heip C, Warwick R, Carr M, *et al.* (1988) Analysis of community attributes of the benthic meiofauna of Frierfjord/Langesundfjord. *Mar Ecol Prog Ser* **46**, 171-180.

Hendry AP (2009) Evolutionary biology: Speciation. *Nature* **458**, 162-164.

Higgins RP, Thiel H (1988) *Introduction to the study of meiofauna* Smithsonian Institution Press.

Hilbish T (1996) Population genetics of marine species: The interaction of natural selection and historically differentiated populations. *J. Exp. Mar. Biol. Ecol.* **200**, 67-83.

Hillebrand H (2004) On the generality of the latitudinal diversity gradient. *Am Nat* **163**, 192-211.

Hillis D, Moritz C (1990) An overview of applications of molecular systematics. In: *Molecular systematics* (eds. Hillis D, Moritz C), pp. 502-515. Sinauer Associates, Sunderland, Massachusetts.

Hillis DM, Dixon MT (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. *Q Rev Biol* **66**, 411-453.

Holt R, Keitt T, Lewis M, Maurer B, Taper M (2005) Theoretical models of species' borders: single species approaches. *Oikos* **108**.

Hooper DU, Chapin FS, Ewel JJ, *et al.* (2005a) Effects of biodiversity on ecosystem functioning: A consensus of current knowledge. *Ecological Monographs* **75**, 3-35.

Hooper DU, F. S. Chapin I, Ewel JJ, *et al.* (2005b) Effects of biodiversity on ecosystem processes: Implications for ecosystem management. A position statement for the Ecological Society of America. *Ecological Monographs* **75**, 3-35.

Hooper DU, Solan M, Symstad A, *et al.* (2002) Species diversity, functional diversity and ecosystem functioning. In: *Biodiversity and ecosystem functioning: synthesis and perspectives* (eds. Inchausti P, Loreau M, Naeem S), pp. 195-208. Oxford University Press, Oxford.

Horner-Devine MC, Lage M, Hughes JB, Bohannan BJ (2004) A taxa-area relationship for bacteria. *Nature* **432**, 750-753.

Howeth JG, Leibold MA (2010) Species dispersal rates alter diversity and ecosystem stability in pond metacommunities. *Ecology* **91**, 2727-2741.

Hua E, Zhang ZN, Zhang Y (2009) Environmental factors affecting nematode community structure in the Changjiang Estuary and its adjacent waters. *Journal of the Marine Biological Association of the United Kingdom* **89**, 109-117.

Hubbell SP (2001) *The unified neutral theory of biodiversity and biogeography.* Princeton University Press, Princeton.

Huber JA, Butterfield DA, Baross JA (2002) Temporal changes in archeal diversity and chemistry in a mid-ocean ridge subseafloor habitat. *Applied and Environmental Microbiology* **68**, 1585-1594.

Huber JA, Mark Welch DB, Morrison HG, *et al.* (2007) Microbial population structures in the deep marine biosphere. *Science* **318**, 97-100.

Huber T, Faulkner G, Hugenholtz P (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**, 2317-2319.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources* **8**, 3-17.

Hugenholtz P, Tyson GW (2008) Microbiology: metagenomics. *Nature* **455**, 481-483.

Hugenholtzt P, Huber T (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int J Syst Evol Microbiol* **53**, 289-293.

Hughes J, Vogler AP (2006) Gene expression in the gut of keratin-feeding clothes moths (Tineola) and keratin beetles (Trox) revealed by subtracted cDNA libraries. *Insect Biochemistry and Molecular Biology* **36**, 584-592.

Hughes JB, Hellmann JJ (2005) The application of rarefaction techniques to molecular inventories of microbial diversity. *Methods Enzymol* **397**, 292-308.

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67**, 4399-4406.

Hurlbert SH (1971) The non-concept of species diversity: a critique and alternative parameters. *Ecology* **52**, 577-586.

Huse SM, Dethlefsen L, Huber JA, *et al.* (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* **4**, e1000255.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**, R143.

Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12**, 1889-1898.

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Research* **17**, 377-386.

Ingels J, Vanhove S, De Mesel I, Vanreusel A (2006) The biodiversity and biogeography of the free-living nematode genera Desmodora and Desmodorella (family Desmodoridae) at both sides of the Scotia Arc. *Polar Biology* **29**, 936-949.

Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN (2007) Universal primer cocktails for fish DNA barcoding. *Molecular Ecology Notes* **7**, 544-548.

Jenkins W (1964) A rapid centrifugal-flotation technique for separating nematodes from soil. *Plant Disease Research* **48**, 692.

Jensen PR (2010) Linking species concepts to natural product discovery in the post-genomic era. *J Ind Microbiol Biotechnol* **37**, 219-224.

Jonsson KI (2005) The evolution of life histories in holo-anhydrobiotic animals: a first approach. *Integr Comp Biol* **45**, 764-770.

Jost L (2007) Partitioning diversity into independent alpha and beta components. *Ecology* **88**, 2427-2439.

Jumpponen A (2007) Soil fungal communities underneath willow canopies on a primary successional glacier forefront: rDNA sequence results can be affected by primer selection and chimeric data. *Microb Ecol* **53**, 233-246.

Kemp PF, Aller JY (2004) Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. *FEMS Microbiology Ecology* **47**, 161-177.

Kennedy AD, Jacoby CA (1999) Biological indicators of marine environmental health: meiofauna-a neglected benthic component? *Environmental Monitoring Assessment* **54**, 47-68.

Knowlton N (1986) Cryptic and Sibling Species among the Decapod Crustacea. *Journal of Crustacean Biology* **6**, 356-363.

Knowlton N (1993) Sibling Species in the Sea. *Annual Review of Ecology and Systematics* **24**, 189-216.

Knowlton N (2000) Molecular genetic analyses of species boundaries in the sea. *Hydrobiologia* **420**, 73-90.

Koonin EV (2010) The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol* **11**, 209.

Kovach WL (1999) MVSP - a multivariate statistical package for windows, ver. 3.12d. Kovach Computing Services, Pentraeth, Wales, U.K.

Kristiansen J (2000) Cosmopolitan chrysophytes. *Syst. Geogr. Pl.* **70**, 291-300.

Kumar S, Carlsen T, Mevik BH, *et al.* (2011) CLOTU: An online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. *BMC Bioinformatics* **12**, 182.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2009) Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. *Environ Microbiol*.

Lachance M (2004) Here and there or everywhere? . *Bioscience* **54**, 884.

Lahr DJ, Katz LA (2009) Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* **47**, 857-866.

Lambert DM, Baker A, Huynen L, *et al.* (2005) Is a large-scale DNA-based inventory of ancient life possible? *J Hered* **96**, 279-284.

Lambshead PD, Boucher G (2003) Marine nematode deep-sea biodiversity - hyperdiverse or hype? *Journal of Biogeography* **30**, 475-485.

Lambshead PJD (1986) Sub-Catastrophic Sewage and Industrial-Waste Contamination as Revealed by Marine Nematode Faunal Analysis. *Mar Ecol Prog Ser* **29**, 247-260.

Lambshead PJD (1993) Recent developments in marine benthic biodiversity research. *Oceanis* **19**, 5-24.

Lambshead PJD (2004) Marine nematode biodiversity. In: *Nematode Morphology, Physiology and Ecology* (eds. Chen Z, Chen S, Dickson D), pp. 438-492. Tsinghua University Press, Tsinghua.

Lambshead PJD, Paterson GLJ (1986) Ecological Cladistics an Investigation of Numerical Cladistics as a Method for Analyzing Ecological Data. *Journal of Natural History* **20**, 895-909.

Lambshead PJD, Tietjen J, Ferrero T, Jensen P (2000) Latitudinal diversity gradients in the deep sea with special reference to North Atlantic nematodes. *Marine Ecology-Progress Series* **194**, 159-167.

Lawton JH, Bignell DE, Bolton B, *et al.* (1998) Biodiversity inventories, indicator taxa and the effects of habitat modification in tropical forest. *Nature* **391**, 72-76.

Lazarevic V, Whiteson K, Huse S, *et al.* (2009) Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods* **79**, 266-271.

Leibold MA, Holyoak M, Mouquet N, *et al.* (2004) The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters* **7**, 601-613.

Li J, Vincx M, Herman PMJ, Heip C (1997) Monitoring meiobenthos using cm-, m- and km-Scales in the Southern Bight of the North Sea. *Marine Environmental Research* **43**, 265-278.

Litvaitis M, Nunn G, Thomas W, Kocher T (1994) A molecular approach for the identification of meiofaunal turbellarians (Platyhelminthes, Turbellaria). *Marine Biology* **120**, 437-442.

Loreau M, Naeem S, Inchausti P, *et al.* (2001) Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* **294**, 804-808.

Loreau M, Oteng-Yeboah A, Arroyo MT, *et al.* (2006) Diversity without representation. *Nature* **442**, 245-246.

Loy A, Lehner A, Lee N, *et al.* (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Applied and Environmental Microbiology* **68**, 5064-5081.

Ludwig W, Strunk O, Westram R, *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Research* **32**, 1363-1371.

MacArthur R, Wilson E (1967) *The Theory of Island Biogeography* Princeton University Press, Princeton, N.J.

Macas J, Neumann P, Navratilova A (2007) Application of massively parallel pyrosequencing ("454-sequencing") to comprehensive analysis of repetitive DNA in higher plant genomes. *Chromosome Research* **15**, 6-6.

Magurran A (2004) *Measuring biological diversity* Blackwell Science, Oxford, UK.

Magurran AE (1998) Measuring richness and evenness. *Trends Ecol Evol* **13**, 165-166.

Magurran AE (2005) Biological diversity. *Curr Biol* **15**, R116-118.

Magurran AE, Henderson PA (2010) Temporal turnover and the maintenance of diversity in ecological assemblages. *Philos Trans R Soc Lond B Biol Sci* **365**, 3611-3620.

Maldonado M, Young CM (1996) Bathymetric patterns of sponge distribution on the Bahamian slope. *Deep Sea Research Part I: Oceanographic Research Papers* **43**, 897-915.

Mare MF (1942) A study of a marine benthic community with special reference to the microorganisms. *Journal of the Marine Biological Association of the United Kingdom* **25**, 517-554.

Margulies M, Egholm M, Altman WE, *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Markham NR, Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol* **453**, 3-31.

Markmann M, Tautz D (2005) Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. *Philos Trans R Soc Lond B Biol Sci* **360**, 1917-1924.

Martiny JB, Bohannan BJ, Brown JH, *et al.* (2006) Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**, 102-112.

Martiny JB, Eisen JA, Penn K, Allison SD, Horner-Devine MC (2011) Drivers of bacterial {beta}-diversity depend on spatial scale. *Proc Natl Acad Sci U S A* **108**, 7850-7854.

Massana R, Pedros-Alio C (2008) Unveiling new microbial eukaryotes in the surface ocean. *Curr Opin Microbiol* **11**, 213-218.

May R (1992) Bottoms up for the oceans. *Nature* **357**, 278-279.

May RM (1988) How Many Species Are There on Earth. *Science* **241**, 1441-1449.

May RM (1994) Biological Diversity: Differences between land and sea. *Philos Trans R Soc Lond B Biol Sci* **343**, 105-111.

Mayr E (1963) *Animal Species and Evolution* Belknap Press, Cambridge, MA.

McGill B, Collins C (2003) A unified theory for macroecology based on spatial patterns of abundance. *Evol. Ecol. Res.* **5**, 469– 492.

McIntosh RP (1967) An index of diversity and the relation of certain concepts to diversity. *Ecology* **48**, 392-404.

Medinger R, Nolte V, Pandey RV, *et al.* (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Mol Ecol* **19 Suppl 1**, 32-40.

Meldal BH, Debenham NJ, De Ley P, *et al.* (2007) An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. *Mol Phylogenet Evol* **42**, 622-636.

Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46.

Meyer M, Briggs AW, Maricic T, *et al.* (2008a) From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Research* **36**, e5.

Meyer M, Stenzel U, Hofreiter M (2008b) Parallel tagged sequencing on the 454 platform. *Nature Protocols* **3**, 267-278.

Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research* **35**, e96.

Meyerhans A, Vartanian JP, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids Res* **18**, 1687-1691.

Michiels IC, Traunspurger W (2005) Benthic community patterns and the composition of feeding types and reproductive modes in freshwater nematodes. *Nematology* **7**, 21-36.

Miloslavich P, Diaz JM, Klein E, *et al.* (2010) Marine biodiversity in the Caribbean: regional estimates and distribution patterns. *PLoS ONE* **5**, e11916.

Mirto S, Danovaro R (2004) Meiofaunal colonisation on artificial substrates: a tool for biomonitoring the environmental quality on coastal marine systems. *Marine Pollution Bulletin* **48**, 919-926.

Mitchell E, Meisterfeld R (2005) Taxonomic confusion blurs the debate on cosmopolitanism versus local endemism of free-living protists. *Protist* **156**, 263–267.

Mitreva M, Blaxter ML, Bird DM, McCarter JP (2005) Comparative genomics of nematodes. *Trends Genet* **21**, 573-581.

Mittelbach GG, Schemske DW, Cornell HV, *et al.* (2007) Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecol Lett* **10**, 315-331.

Moens T, Verbeeck L, de Maeyer A, Swings J, Vincx M (1999a) Selective attraction of marine bacterivorous nematodes to their bacterial food. *Marine Ecology-Progress Series* **176**, 165-178.

Moens T, Verbeeck L, Vincx M (1999b) Feeding biology of a predatory and a facultatively predatory nematode (Enoploides longispiculosus and Adoncholaimus fuscus). *Marine Biology* **134**, 585-593.

Moon-van der Staay SY, De Wachter R, Vaulot D (2001) Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607-610.

Moreira D, Lopez-Garcia P (2002) The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends in Microbiology* **10**, 31-38.

Mou X, Sun S, Edwards RA, Hodson RE, Moran MA (2008) Bacterial carbon processing by generalist species in the coastal ocean. *Nature* **451**, 708-711.

Müller WEG, Brümmer F, Batel R, Müller IM, Schröder HC (2003) Molecular biodiversity. Case study: Porifera (sponges). *Naturwissenschaften* **90**, 103-120.

Murrell M, Fleeger J (1989) Meiofauna abundance on the Gulf of Mexico continental shelf affected by hypoxia. *Continental Shelf Research* **9**, 1049-1062.

Naeem S, Wright JP (2003) Disentangling biodiversity effects on ecosystem functioning: deriving solutions to a seemingly insurmountable problem. *Ecology Letters* **6**, 567-579.

Naser HA (2010) Testing taxonomic resolution levels for detecting environmental impacts using macrobenthic assemblages in tropical waters. *Environ Monit Assess* **170**, 435-444.

Nassonova E, Smirnov A, Fahrni J, Pawlowski J (2010) Barcoding amoebae: comparison of SSU, ITS and COI genes as tools for molecular identification of naked lobose amoebae. *Protist* **161**, 102-115.

Nekola J, White P (1999) The distance decay of similarity in biogeography and ecology. *J Biogeogr* **26**, 867–878.

Noguez A, Arita H, Escalante A, *et al.* (2005) Microbial macroecology: highly structured prokaryotic soil assemblages in a tropical deciduous forest. *Global. Ecol. Biogeogr.* **14**, 241–248.

Nolte V, Pandey RV, Jost S, *et al.* (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Molecular Ecology* **19**, 2908-2915.

Noonan JP, Coop G, Kudaravalli S, *et al.* (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science* **314**, 1113-1118.

Norse E (1993) *Global Marine Biological Diversity: A Strategy for Building Conservation into Decision Making* Island Press, Washington, D.C.

Nowrousian M (2010) Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot Cell* **9**, 1300-1310.

Nybakken J (1997) *Marine biology: an ecological approach*, 4th edn. edn. Harper Collins College Publishers, New York.

O'Mahony EM, Tay WT, Paxton RJ (2007) Multiple rRNA variants in a single spore of the microsporidian Nosema bombi. *J Eukaryot Microbiol* **54**, 103-109.

Orsi W, Edgcomb V, Jeon S, *et al.* (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. II. Habitat specialization. *ISME J*.

Pace B, Campbell LL (1971) Homology of ribosomal ribonucleic acid diverse bacterial species with Escherichia coli and Bacillus stearothermophilus. *J Bacteriol* **107**, 543-547.

Palumbi SR (1992) Marine speciation on a small planet. *Trends Ecol Evol* **7**, 114-118.

Palumbi SR (2003) Ecological subsidies alter the structure of marine communities. *Proc Natl Acad Sci U S A* **100**, 11927-11928.

Palumbi SR, McLeod KL, Grunbaum D (2008) Ecosystems in Action: Lessons from Marine Ecology about Recovery, Resistance, and Reversibility. *BioScience* **58**, 33-42.

Parameswaran P, Jalili R, Tao L, *et al.* (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research* **35**.

Parfrey LW, Barbero E, Lasser E, *et al.* (2006) Evaluating support for the current classification of eukaryotic diversity. *PLoS Genet* **2**, e220.

Pawlowski J, Christen R, Lecroq B, *et al.* (2011) Eukaryotic richness in the abyss: insights from pyrotag sequencing. *PLoS ONE* **6**, e18169.

Pedros-Alio C (2007) Ecology. Dipping into the rare biosphere. *Science* **315**, 192-193.

Peplies J, Lachmund C, Glockner FO, Manz W (2006) A DNA microarray platform based on direct detection of rRNA for characterization of freshwater sediment-related prokaryotic communities. *Appl Environ Microbiol* **72**, 4829-4838.

Petchey OL, Gaston KJ (2006) Functional diversity: back to basics and looking forward. *Ecology Letters* **9**, 741-758.

Pfannkuche O (1985) The deep-sea meiofauna of the Porcupine seabight and abyssal plain (NE Atlantic): population structure, distribution, standing stocks. *Oceanologica Acta* **8**, 343–353.

Pfenninger M, Nowak C, Kley C, Steinke D, Streit B (2007) Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic Chironomus (Diptera) species. *Mol Ecol* **16**, 1957-1968.

Pfunder M, Holzgang O, Frey JE (2004) Development of microarray-based diagnostics of voles and shrews for use in biodiversity monitoring studies, and evaluation of mitochondrial cytochrome oxidase I vs. cytochrome b as genetic markers. *Mol Ecol* **13**, 1277-1286.

Philippe H, Derelle R, Lopez P, *et al.* (2009) Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* **19**, 706-712.

Picard C, Ponsonnet C, Paget E, Nesme X, Simonet P (1992) Detection and enumeration of bacteria in soil by direct DNA extraction and polymerase chain reaction. *Applied and Environmental Microbiology* **58**, 2717-2722.

Piganeau G, Gardner M, Eyre-Walker A (2004) A broad survey of recombination in animal mitochondria. *Molecular Biology and Evolution* **21**, 2319-2325.

Platt HM (1994) *The Phylogenetic Systematics of Free-living Nematodes* The Royal Society, London.

Platt HM, Shaw KM, Lambshead PJD (1984) Nematode Species Abundance Patterns and Their Use in the Detection of Environmental Perturbations. *Hydrobiologia* **118**, 59-66.

Platt HM, Warwick RM (1983) *Free-living marine nematodes. Part I. British Enoplids* Cambridge University Press, Cambridge.

Platt HM, Warwick RM (1988) *Free-living marine nematodes* Cambridge University Press, Leiden, NewYork.

Poore CB, Wilson GDF (1993) Marine species richness. *Nature* **361**, 597-598.

Porazinska D, Giblin-Davis R, Faller L, *et al.* (2009) Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Mol Ecol Resour* **9**, 1439 - 1450.

Porazinska DL, Giblin-Davis RM, Esquivel A, *et al.* (2010) Ecometagenetics confirm high tropical rainforest nematode diversity. *Molecular Ecology* **19**, 5521-5530.

Porebski S, Bailey L, Baum B (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Molecular Biology Reporter* **15**, 8-15.

Porteous LA, Seidler RJ, Watrud LS (1997) An improved method for purifying DNA from soil for polymerase chain reaction amplification and molecular ecology applications. *Mol. Ecol.* **6**, 787-791.

Porteus LA, Seidler RJ, Watrud LS (1997) An improved method for purifying DNA from soil for polymerase chain reaction amplification and molecular ecology applications. *Molecular Ecology* **6**, 787-791.

Powers T (2004) Nematode molecular diagnostics: From bands to barcodes. *Annual Review of Phytopathology* **42**, 367-383.

Powers TO, Neher DA, Mullin P, *et al.* (2009) Tropical nematode diversity: vertical stratification of nematode communities in a Costa Rican humid lowland rainforest. *Molecular Ecology*.

Pruesse E, Quast C, Knittel K, *et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**, 7188-7196.

Purvis A, Hector A (2000) Getting the measure of biodiversity. *Nature* **405**, 212-219.

Qiu X, Wu L, Huang H, *et al.* (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl Environ Microbiol* **67**, 880-887.

Quaiser A, Zivanovic Y, Moreira D, Lopez-Garcia P (2011) Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. *ISME J* **5**, 285-304.

Quijón PA, Snelgrove PVR (2006) The use of coarser taxonomic resolution in studies of predation on marine sedimentary fauna. *Journal of Experimental Marine Biology and Ecology* **330**, 159-168.

Quince C, Lanzen A, Curtis TP, *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6**, 639-641.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12**, 38.

Rabouille S, Staal M, Stal LJ, Soetaert K (2006) Modeling the dynamic regulation of nitrogen fixation in the cyanobacterium Trichodesmium sp. *Applied and Environmental Microbiology* **72**, 3217-3227.

Ramette A, Tiedje JM (2007) Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microb Ecol* **53**, 197-207.

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes* **7**, 355-364

Reeder J, Knight R (2009) The 'rare biosphere': a reality check. *Nat Methods* **6**, 636-637.

Reise K (1988) Plathelminth diversity in littoral sediments around the island of Sylt in the North Sea. *Prog Zoo1* **36**, 469-480.

Reise K (2003) Metapopulation structure in the lagoon cockle Cerastoderma lamarcki in the northern Wadden Sea. *Helgol Mar Res* **56**, 252–258.

Rice W (1988) Analyzing tables of statistical test. *Evolution* **43**, 223-225.

Ricotta C (2010) On beta diversity decomposition: trouble shared is not trouble halved. *Ecology* **91**, 1981-1983.

Ricotta C, Szeidl L (2006) Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. *Theor Popul Biol* **70**, 237-243.

Riga E, Karanastasi E, Oliveira CMG, Neilson R (2007) Molecular identification of Two stubby root nematode species. *American Journal of Potato Research* **84**, 161-167.

Rodriguez-Brito B, Rohwer F, Edwards RA (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**, 162.

Roger AJ, Simpson AG (2009) Evolution: revisiting the root of the eukaryote tree. *Curr Biol* **19**, R165-167.

Ronaghi M (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Research* **11**, 3-11.

Ronaghi M, Uhlen M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. *Science* **281**, 363-365.

Roossinck MJ, Saha P, Wiley GB, *et al.* (2010) Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. *Mol Ecol* **19 Suppl 1**, 81-88.

Rundell RJ, Leander BS (2010) Masters of miniaturization: convergent evolution among interstitial eukaryotes. *Bioessays* **32**, 430-437.

Rysgaard S, Christensen PB, Sorensen MV, Funch P, Berg P (2000) Marine meiofauna, carbon and nitrogen mineralization in sandy and soft sediments of Disko Bay, West Greenland. *Aquatic Microbial Ecology* **21**, 59-71.

s'Jacob JJ, van Bezooijen J (1984) A manual for practical work in nematology, p. 77 pp, Wageningen Agricultural University, Wageningen, The Netherlands.

Sala O, Knowlton N (2006) Global marine biodiversity trends. *Annual Review of Environment and Resources* **31**, 93-122.

Sala OE, Chapin FS, 3rd, Armesto JJ, *et al.* (2000) Global biodiversity scenarios for the year 2100. *Science* **287**, 1770-1774.

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning. A laboratory manual. 2nd ed.* Cold Spring Harbor Laboratory Press, New York.

Sanders H (1968) Marine benthic diversity: a comparative study. *American Naturalist* **102**, 242–282.

Sanger F, Nicklen S, Coulsen AR (1977) DNA sequencing with chain terminating inhibitors. *Proceedings of the National Academy of Science, USA* **74**, 5463-5467.

Schauer R, Bienhold C, Ramette A, Harder J (2010) Bacterial diversity and biogeography in deep-sea surface sediments of the South Atlantic Ocean. *ISME J* **4**, 159-170.

Schindel DE, Miller SE (2005) DNA barcoding a useful tool for taxonomists. *Nature* **435**, 17.

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

Scholin C, Anderson D, Sogin M (1993) Two distinct smallsubunit ribosomal RNA genes in the North American toxic dinoflagellate Alexandrium fundyense (Dinophyceae). *J. Phycol* **29**, 209–216.

Schratzberger M, Rees H, Boyd S (2000) Effects of simulated deposition of dredged material on structure of nematode assemblages-the role of contamination. *Mar Biol Bull* **137**, 613-622.

Seberg O, Humphries C, Knapp S, *et al.* (2003) Shortcuts in systematics? A commentary on DNA-based taxonomy. *Trends in Ecology & Evolution* **18**, 63-65.

Servedio MR (2004) The evolution of premating isolation: local adaptation and natural and sexual selection against hybrids. *Evolution* **58**, 913-924.

Seutin G, White B, Boag P (1991) Preservation of avian blood and tissue samples for DNA analyses. *Canadian Journal of Zoology* **69**, 82-90.

Shafikhani S (2002) Factors affecting PCR-mediated recombination. *Environ Microbiol* **4**, 482-486.

Shanks A, Edmondson E (1990) The vertical flux of metazoans (holoplankton, meiofauna, and larval invertebrates) due to their association with marine snow. *Limnol Oceanogr* **35**, 455-463.

Shaw AK, Halpern AL, Beeson K, *et al.* (2008) It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10**, 2200-2210.

Shaw KL (2002) Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proc Natl Acad Sci U S A* **99**, 16122-16127.

Shinn AP, Banks BA, Tange N, *et al.* (2000) Utility of 18S rDNA and ITS sequences as population markers for Lepeophtheirus salmonis (Copepoda : Caligidae) parasitising Atlantic salmon (Salmo salar) in Scotland. *Contributions to Zoology* **69**, 89-98.

Silver MW, Shanks AL, Trent JD (1978) Marine snow: microplankton habitat and source of small-scale patchiness in pelagic populations. *Science* **201**, 371-373.

Simpson AG, Roger AJ (2004) The real 'kingdoms' of eukaryotes. *Curr Biol* **14**, R693-696.

Sipos R, Szekely AJ, Palatinszky M, *et al.* (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol* **60**, 341-350.

Slapeta J, Moreira D, Lopez-Garcia P (2005) The extent of protist diversity: insights from molecular ecology of freshwater eukaryotes. *Proc Biol Sci* **272**, 2073-2081.

SMEBD (2009) World Register of Marine Species. *Accessed at http://www.marinespecies.org on [2009-07-27]*.

Smith B, Wilson J (1996) A consumer's guide to evenness indices. *Oikos* **76**, 70–82.

Smith H, Wilkinson D (2007) Not all free-living microorganisms have cosmopolitan distributions: the case of Nebela (Apodera) vas Certes (Protozoa: Amoebozoa: Arcellinida). *J Biogeogr* **34**, 1822–1831.

Smith MA, Woodley NE, Janzen DH, Hallwachs W, Hebert PD (2006) DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). *Proc Natl Acad Sci U S A* **103**, 3657-3662.

Smith VS (2005) DNA barcoding: perspectives from a "Partnerships for Enhancing Expertise in Taxonomy" (PEET) debate. *Syst Biol* **54**, 841-844.

Smyth RP, Schlub TE, Grimm A, *et al.* (2010) Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene* **469**, 45-51.

Snelgrove P, Blackburn TH, Hutchings PA, *et al.* (1997) The importance of marine sediment biodiversity in ecosystem precesses. *Ambio* **26**, 578-583.

Snelgrove PVR (1999) Getting to the bottom of marine biodiversity: Sedimentary habitats - Ocean bottoms are the most widespread habitat on Earth and support high biodiversity and key ecosystem services. *Bioscience* **49**, 129-138.

Snelgrove PVR, Grassle JP, Grassle JF, Petrecca RF, Ma HG (1999) In situ habitat selection by settling larvae of marine soft-sediment invertebrates. *Limnology and Oceanography* **44**, 1341-1347.

Snelgrove PVR, Smith CR (2002) A riot of species in an environmental calm: The paradox of the species-rich deep-sea floor. *Oceanography and Marine Biology, Vol 40* **40**, 311-342.

Soetaert K, Vincx M, Wittoeck J, Tulkens M (1995) Meiobenthic Distribution and Nematode Community Structure in 5 European Estuaries. *Hydrobiologia* **311**, 185-206.

Soetaert K, Vincx M, Wittoeck J, Tulkens M, Vangansbeke D (1994) Spatial Patterns of Westerschelde Meiobenthos. *Estuarine Coastal and Shelf Science* **39**, 367-388.

Sogin ML, Morrison HG, Huber JA, *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* **103**, 12115-12120.

Solan M, Cardinale BJ, Downing AL, *et al.* (2004) Extinction and ecosystem function in the marine benthos. *Science* **306**, 1177-1180.

Soltwedel T (1997) Meiobenthos distribution pattern in the tropical East Atlantic: indication for fractionated sedimentation of organic matter to the sea floor? *Marine Biology* **129**, 747-756.

Soltwedel T (2000) Metazoan meiobenthos along continental margins: a review. *Prog Oceanogr* **46**, 59-84.

Soltwedel T, Hasemann C, Queric NV, von Juterzenka K (2005) Gradients in activity and biomass of the small benthic biota along a channel system in the deep Western Greenland Sea. *Deep-Sea Research Part I-Oceanographic Research Papers* **52**, 815-835.

Somerfield PJ, Fonseca-Genevois VG, Rodrigues ACL, Castro FJV, Santos GAP (2003) Factors affecting meiofaunal community structure in the Pina Basin, an urbanized embayment on the coast of Pernambuco, Brazil. *Journal of the Marine Biological Association of the United Kingdom* **83**, 1209-1213.

Somerfield PJ, Gage JD (2000) Community structure of the benthos in Scottish Sea-lochs. IV. Multivariate spatial pattern. *Marine Biology* **136**, 1133-1145.

Somerfield PJ, Warwick RM, Moens T (2005) *Meiofauna techniques*, 3 edn. Blackwell Science, Oxford.

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

Steyaert M, Garner N, van Gansbeke D, Vincx M (1999) Nematode communities from the North Sea: environmental controls on species diversity and vertical distribution within the sediment. *Journal of the Marine Biological Association of the United Kingdom* **79**, 253-264.

Steyaert M, Moodley L, Nadong T, *et al.* (2007) Responses of intertidal nematodes to short-term anoxic events. *Journal of Experimental Marine Biology and Ecology* **345**, 175-184.

Steyaert M, Vanaverbeke J, Vanreusel A, *et al.* (2003) The importance of fine-scale, vertical profiles in characterising nematode community structure. *Estuarine, Coastal and Shelf Science* **58**, 353-366.

Stoeck T, Behnke A, Christen R, *et al.* (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol* **7**, 72.

Stoeck T, Kasper J, Bunge J, *et al.* (2007) Protistan diversity in the Arctic: a case of paleoclimate shaping modern biodiversity? *PLoS ONE* **2**, e728.

Subbotin SA, Ragsdale EJ, Mullens T, *et al.* (2008) A phylogenetic framework for root lesion nematodes of the genus Pratylenchus (Nematoda): Evidence from 18S and D2-D3 expansion segments of 28S ribosomal RNA genes and morphological characters. *Molecular Phylogenetics and Evolution* **48**, 491-505.

Sun Y, Cai Y, Liu L, *et al.* (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res* **37**, e76.

Sutherland W, Armstrong-Brown S, Armsworth P, *et al.* (2006) The identification of 100 ecological questions of high policy relevance in the UK. *Journal of Applied Ecology* **43**, 617-627.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596-1599.

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2002) DNA points the way ahead of taxonomy - In assessing new approaches, it's time for DNA's unique contribution to take a central role. *Nature* **418**, 479-479.

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003) A plea for DNA taxonomy. *Trends in Ecology & Evolution* **18**, 70-74.

Taylor MS, Hellberg ME (2003) Genetic evidence for local retention of pelagic larvae in a Caribbean reef fish. *Science* **299**, 107-109.

Telford RJ, Vandvik V, Birks HJ (2006) Dispersal limitations matter for microbial morphospecies. *Science* **312**, 1015.

Terlizzi A, Bevilacqua S, Fraschetti S, Boero F (2003) Taxonomic sufficiency and the increasing insufficiency of taxonomic expertise. *Marine Pollution Bulletin* **46**, 556-561.

Thomas RK, Nickerson E, Simons JF, *et al.* (2006) Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nature Medicine* **12**, 852-855.

Thomas W, Vida J, Frisse L, Mundo M, Baldwin J (1997) DNA sequences from formalin-fixed nematodes: integrating molecular and morphological approaches to taxonomy. *Journal of Nematology* **29**, 250–254.

Thompson R, Townsend C (2006) A truce with neutral theory: local deterministic factors, species traits and dispersal limitation together determine patterns of diversity in stream invertebrates. *J Anim Ecol* **75**, 476-484.

Thrush SF, Hewitt JE, Cummings VJ, Norkko A, Chiantore M (2010) beta-diversity and species accumulation in antarctic coastal benthos: influence of habitat, distance and productivity on ecological connectivity. *PLoS ONE* **5**, e11899.

Tipper J (1979) Rarefaction and rarefiction - the use and abuse of a method in paleoecology. *Paleobiology* **5**, 423-434.

Tita G, Desrosiers G, Vincx M, Clement M (2002) Intertidal meiofauna of the St Lawrence estuary (Quebec, Canada): diversity, biomass and feeding structure

of nematode assemblages. *Journal of the Marine Biological Association of the United Kingdom* **82**, 779-791.

Tobler NE, Pfunder M, Herzog K, Frey JE, Altwegg M (2006) Rapid detection and species identification of Mycobacterium spp. using real-time PCR and DNA-microarray. *J Microbiol Methods* **66**, 116-124.

Tsaousis A, Martin D, Ladoukakis E, Posada D, Zouros E (2005) Widespread recombination in published animal mtDNA sequences. *Molecular Biology and Evolution* **22**, 925-933.

Turnbaugh PJ, Ley RE, Mahowald MA, *et al.* (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027-1031.

Ugland KI, Gray JS, Lambshead PJD (2005) Species accumulation curves analysed by a class of null models discovered by Arrhenius. *Oikos* **108**, 263-274.

Uilenberg G, Thiaucourt F, Jongejan F (2004) On molecular taxonomy: what is in a name? *Exp Appl Acarol* **32**, 301-312.

Unterseher M, Jumpponen ARI, ÖPik M, *et al.* (2011) Species abundance distributions and richness estimations in fungal metagenomics – lessons learned from community ecology. *Molecular Ecology* **20**, 275-285.

Urban-Malinga B, Hedtkamp SIC, van Beusekom JEE, Wiktor J, Weslawski JM (2006) Comparison of nematode communities in Baltic and North Sea sublittoral, permeable sands - Diversity and environmental control. *Estuarine Coastal and Shelf Science* **70**, 224-238.

Vanaverbeke J, Gheskiere T, Vincx M (2000) The meiobenthos of subtidal sandbanks on the Belgian Continental Shelf (Southern Bight of the North Sea). *Estuarine Coastal and Shelf Science* **51**, 637-649.

Vandamme P, Pot B, Gillis M, *et al.* (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev* **60**, 407-438.

Vanreusel A, Clough L, Jacobsen K, *et al.* (2000) Meiobenthos of the central Arctic Ocean with special emphasis on the nematode community structure. *Deep-Sea Research Part I-Oceanographic Research Papers* **47**, 1855-1879.

Velasco JD (2008) Species concepts should not conflict with evolutionary history, but often do. *Stud Hist Philos Biol Biomed Sci* **39**, 407-414.

Venter JC, Remington K, Heidelberg JF, *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74.

Vincx M, Bett BJ, Dinet A, *et al.* (1994) Meiobenthos of the Deep Northeast Atlantic. *Advances in Marine Biology, Vol 30* **30**, 1-88.

Vincx M, Meire P, Heip C (1990) The distribution of nematodes communities in the Southern Bight of the North Sea. *Cahiers de Biologie Marine* **31**, 107–129.

Vogler AP, Monaghan MT (2007) Recent advances in DNA taxonomy. *Journal of Zoological Systematics and Evolutionary Research* **45**, 1-10.

von Wintzingerode F, Gobel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* **21**, 213-229.

Walker B (1992) Biological diversity and ecological redundancy. *Conservation Biology* **6**, 18-23.

Wang GC, Wang Y (1996) The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology* **142**, 1107-1114.

Wang GC, Wang Y (1997) Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl Environ Microbiol* **63**, 4645-4650.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PD (2005) DNA barcoding Australia's fish species. *Philos Trans R Soc Lond B Biol Sci* **360**, 1847-1857.

Warwick RM, Clarke KR (1998) Taxonomic distinctness and environmental assessment. *Journal of Applied Ecology* **35**, 532-543.

Warwick RM, Dashfield SL, Somerfield PJ (2006) The integral structure of a benthic infaunal assemblage. *J Exp Mar Biol Ecol* **330**, 12-18.

Warwick RM, Platt HM, Somerfield PJ (1998) *Free-living marine nematodes. Part III. British Monhysterids*, 53 edn. Field Studies Council, Shrewsbury.

Warwick RM, Somerfield PJ (2008) All animals are equal, but some animals are more equal than others. *Journal of Experimental Marine Biology and Ecology* **366**, 184-186.

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the Arabidopsis Transcriptome with Massively Parallel Pyrosequencing. *Plant Physiol* **144**, 32-42.

White K (2011) Nuclear 18S rDNA as a species-level molecular marker for Leucothoidae (Amphipoda) *Journal of Crustacean Biology* **31**, *in press*.

Whitehead A, Hemming J (1965) A comparison of some quantitative methods of extracting small vermiform nematodes from soil. *Annals of Applied Biology* **55**, 25-38.

Whittaker RH (1972) Evolution and Measurement of Species Diversity. *Taxon* **21**, 213-251.

Wilkinson D (2001) What is the upper size limit for cosmopolitan distribution in free-living microorganisms? *J Biogeogr* **28**, 285–291.

Willig M, Kaufman D, Stevens R (2003) Latitudinal gradients of biodiversity: pattern, process, scale, and synthesis. *Annual Review of Ecology, Evolution, and Systematics* **34**, 273–309.

Wilsey BJ, Chalcraft DR, Bowles CM, Willig MR (2005) Relatioships among indices suggest that richness is an incomplete surrogate for grassland biodiversity. *Ecology* **86**, 1178-1184.

Wilson AC, Cann RL, Carr SM, *et al.* (1985) Mitochondrial DNA and two perspectives on evolutionary genetics. *Biological Journal of the Linnean Society* **26**, 375-400.

Wilson K, Thorndyke M, Nilsen F, Rogers A, Martinez P (2005) Marine systems: moving into the genomics era. *Marine Ecology-an Evolutionary Perspective* **26**, 3-16.

Wilson RJ, Thomas CD, Fox R, Roy DB, Kunin WE (2004) Spatial patterns in species distributions reveal biodiversity change. *Nature* **432**, 393-396.

Woese CR (1987) Bacterial evolution. *Microbiological Reviews* **51**, 221-271.

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A* **87**, 4576-4579.

Worm B, Barbier EB, Beaumont N, *et al.* (2007) Biodiversity loss in the ocean: How bad is it? Response. *Science* **316**, 1282-1284.

Wu JY, Jiang XT, Jiang YX, *et al.* (2010) Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. *BMC Microbiol* **10**, 255.

Wu L, Tang T, Zhou R, Shi S (2007) PCR-mediated recombination of the amplification products of the Hibiscus tiliaceus cytosolic glyceraldehyde-3-phosphate dehydrogenase gene. *J Biochem Mol Biol* **40**, 172-179.

Ye W, Giblin-Davis R, Braasch H, Morris K, Thomas W (2007) Phylogenetic relationships among Bursaphelenchus species (Nematoda: Parasitaphelenchidae) inferred from nuclear ribosomal and mitochondrial DNA sequence data. *Molecular Phylogenetics and Evolution* **43**, 1185-1197.

Yeates GW, Bongers T (1999) Nematode diversity in agroecosystems. *Agriculture Ecosystems & Environment* **74**, 113-135.

Yoder M, De Ley IT, King IW, *et al.* (2006) DESS: a versatile solution for preserving morphology and extractable DNA of nematodes. *Nematology* **8**, 367-376.

Youssef NH, Couger MB, Elshahed MS (2010) Fine-scale bacterial beta diversity within a complex ecosystem (Zodletone Spring, OK, USA): the role of the rare biosphere. *PLoS ONE* **5**, e12414.

Yu Y, Breitbart M, McNairnie P, Rohwer F (2006) FastGroupII: a web-based bioinformatics platform for analyses of large 16S rDNA libraries. *BMC Bioinformatics* **7**, 57.

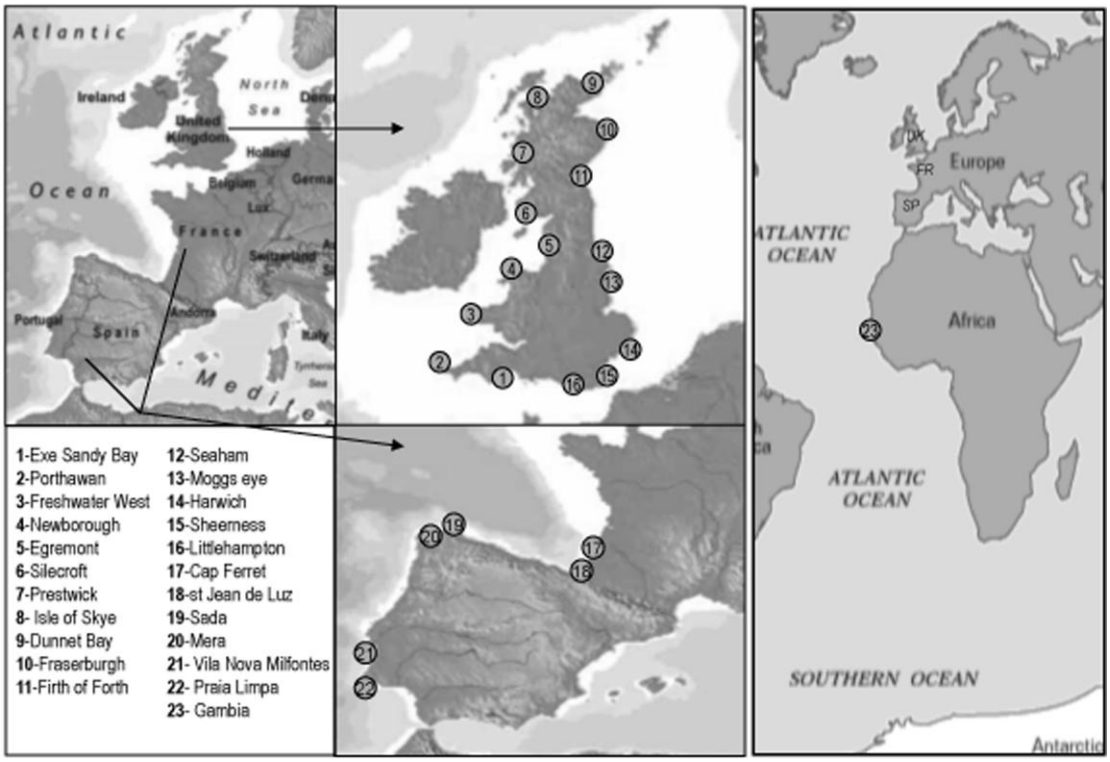
Zhang D-X, Hewitt GM (1997) The long and short of nuclear mitochondrial DNA (Numt) lineages - Reply. *Trends in Ecology and Evolution* **12**, 114.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**, 203-214.

Zhou J, Bruns MA, Tiedje JM (1996) DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology* **62**, 316-322.

APPENDICES

Appendix I



Supplementary Figure S1- Geographic localization of the 23 sampled sites. Each sample comprised three replicates and sampled sites comprised 16 samples from the United Kingdom (1- 16), two samples from France (17, 16), Spain (19, 20), Portugal (21, 22) and one sample from Gambia (23).

Appendix II

Supplementary Table SI- Geographic coordinates of the 23 sampled sites in the United Kingdom (UK), France (FR), Spain (SP), Portugal (PT) and Africa. Abbreviations for each site are given (brackets).

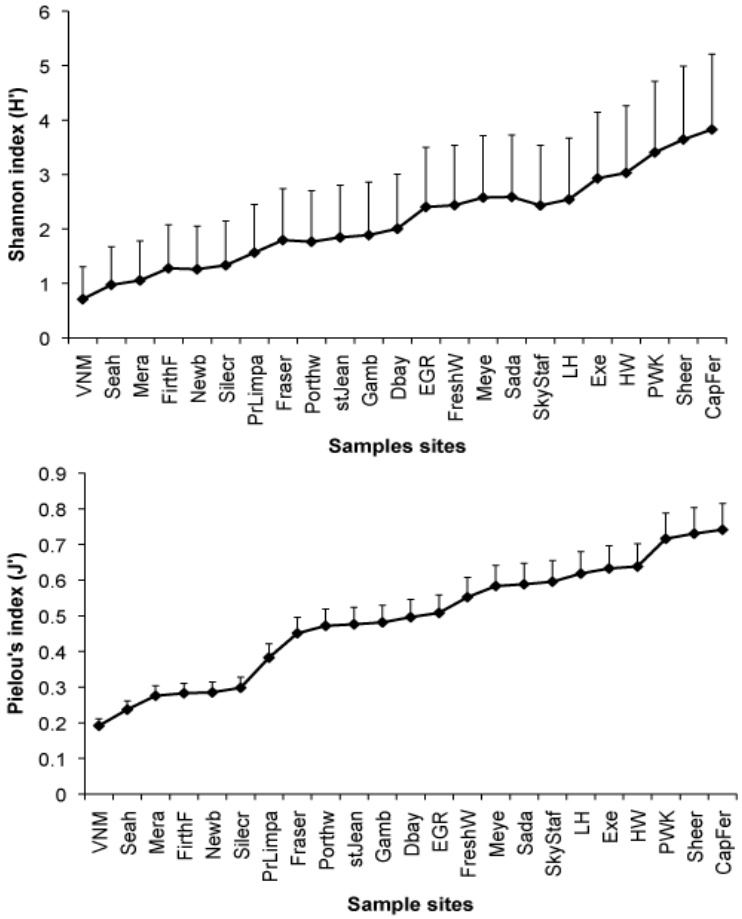
Site	Location	
	N	W
UK, Prestwick (PWK)	55^0 30.481	004^0 37.489
UK, Littlehampton (LH)	50° 48.126	000º 32.385
UK, Mersey Egremont (EGR)	54° 29.188	003° 36.293
UK, Moggs Head (Meye)	54° 54.309	001° 21.237
UK, Statfin (SkyStaf)	57° 38.154	006° 13.742
UK, Dunnet Bay (Dbay)	58° 36.868	003° 21.039
UK, Seaham (Seah)	54^0 51.281	001^0 20.667
UK, Exe (Exe)	50° 36.465	003° 22.488
UK, Harwich (HW)	51°56.225	001° 17.428
UK, Sheerness (Sheer)	51°26.411	000° 45.844
UK, Porthtowan (Porthw)	50° 28.024	005° 02.148
UK, Newborough (Newb)	53° 08.613	004° 24.383
UK, Firth of Forth (FirthF)	55° 52.372	002° 4.892
UK, Fraserburgh (Fraser)	57^0 40.594	001° 59.873
UK, Fresh water west (FreshW)	51^0 39.452	005^0 03.841
UK, Silecroft (Silecr)	54^0 12.961	003^0 21.285
PT, Praia Limpa (PrLimpa)	37° 05.458	008° 27.320
PT, Vila Nova de Milfontes (VNM)	37^0 43.445	008^0 47.556
SP, Mera (Mera)	43^0 22.698	008^0 20.275
SP, Sada (Sada)	43^0 20.567	008^0 14.371
FR, Cap Ferret (CapFer)	44° 20.672	001° 16.565
FR, St. Jean (stJean)	43° 23.668	001^0 39.617
Africa, Gambia (Gamb)	13° 28.142	016° 39.862

Appendix III

PWK	PWK																					
LH	1971	LH																				
EGR	340	1631	EGR																			
MEye	1707	874	2047	MEye																		
SkyStaf	357	2328	697	1350	SkyStaf																	
HW	2297	326	1957	632	1940	HW																
DBay	903	2874	1243	804	546	1352	DBay															
Seah	1716	865	2056	9	1359	539	813	Seah														
Exe	1715	256	1679	1130	2072	582	1934	1121	Exe													
Porthw	1431	540	1395	1414	1788	866	2334	1405	284	Porthw												
Sheer	2213	242	1873	632	2570	84	1436	640	498	782	Sheer											
Newb	681	1290	341	2388	1038	1616	1584	2396	1034	750	1532	Newb										
FirthF	1574	1007	1914	133	1217	681	671	141	1263	1547	765	2255	FirthF									
Fraser	1229	1352	1569	478	872	1026	326	486	1608	1892	1110	1910	345	Fraser								
FreshW	995	976	655	1850	1350	1302	1898	1858	720	436	1218	314	1983	2224	FreshW							
Silecr	376	1595	36	2200	733	1921	1279	2208	1339	1055	1837	305	1950	1605	619	Silecr						
PrLimpa	4632	3173	4292	3800	4989	3499	5535	3808	2917	3201	3415	3338	3933	4278	3637	4256	PrLimpa					
VNM	4494	3035	4154	2663	4851	3361	5397	2671	2779	3063	3277	3200	2796	3141	3499	4118	138	VNM				
Mera	3663	2204	3323	2832	4020	2530	4566	2240	1948	2232	2446	2369	2965	3310	2668	3287	969	831	Mera			
Sada	3648	2189	3308	2817	4005	2515	4551	2825	1933	2217	2431	2354	2950	3295	2653	3272	983	846	15	Sada		
CapFer	2892	1433	2552	2061	3249	1759	3795	2069	1177	1461	1675	1598	2194	2539	1897	2516	1740	1602	771	756	CapFer	
stJean	3000	1541	2660	2169	3357	1867	3903	2177	1285	1569	1783	1706	2302	2647	2005	2624	1632	1494	663	648	108	stJean
Gambia	7407	6315	7067	7189	7764	6641	7993	7180	6305	5976	6557	6726	7322	7667	6412	7031	3388	3526	4357	4372	5128	5020	Gambia

Supplementary Figure S2- Distance matrix of the 23 sampled sites from United Kingdom, France, Spain, Portugal and Gambia. Distances are given in Km.

APPENDIX IV



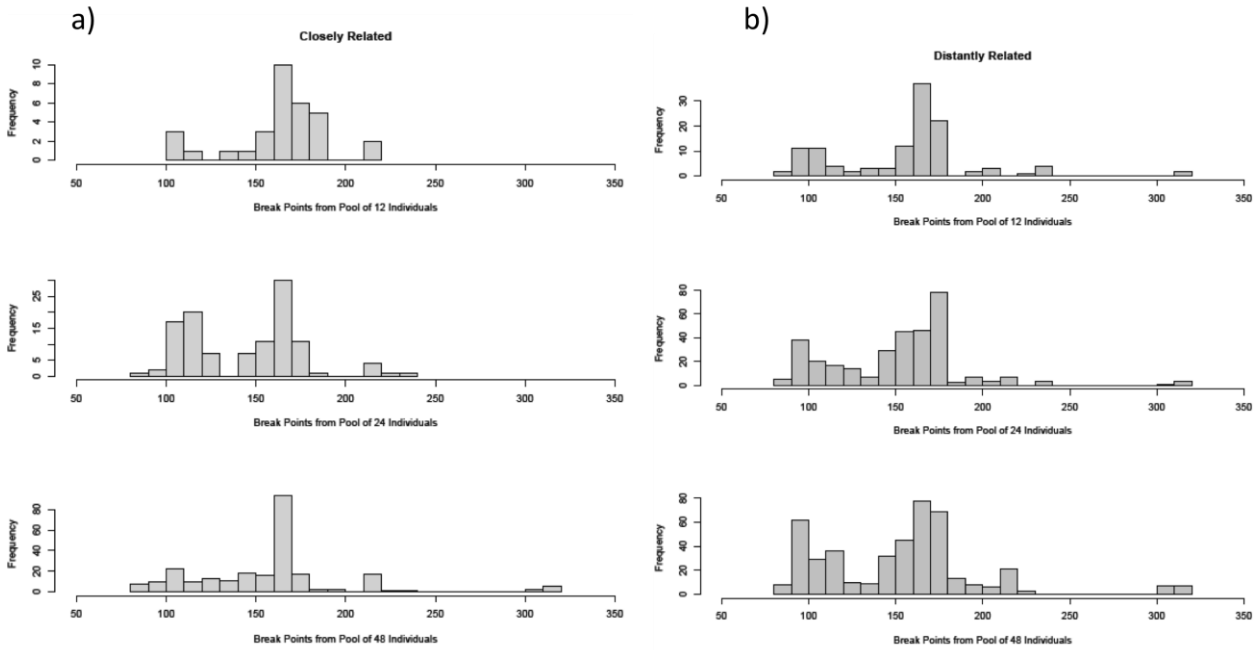
Supplementary Figure S3- The Shannon (H') and Pielou's (J') diversity measures for all sampled sites on normalized dataset at 99% sequence similarity level. UK samples: Skye staffin, Littlehampton, Exe, Fraserburgh, Egremont, Mogs eye, Porthawan, Sheerness, Firth of Fourth, Freshwater west and Silecroft. France samples: st Jean Luz and Cap Ferret. Spain samples: Mera and Sada. Portugal samples: Praia limpa and Vila Nova de Milfontes.

APPENDIX V

Supplementary Table SII- Number of total OTUs for all metazoan phyla found in the 23 sampled sites. OTUs clustering were performed on non-normalized data at 96% and 99% sequence similarity cut-off.

Phyla	Total OTUS 96%	Total OTUS 99%
Nematoda	525	888
Platyhelminthes	220	320
Stramenopiles	150	258
Arthropoda	123	198
Alveolata	72	112
Annelida	47	65
Gastrotricha	47	67
Rhizaria	42	54
Plantae	33	76
Mollusca	34	49
Tardigrada	24	36
Cnidaria	6	11
Rotifera	6	11
Bryozoa	4	6
Brachiopoda	4	5
Nemertea	3	6
Apusozoa	3	6
Kinorhyncha	3	4
Orthonectida	2	3
Tunicata	2	3
Gnathostomulida	1	2
Porifera	1	1
Excavata	1	2
Env samples	264	394
NA'S	97	123
TOTAL	1714	2700

APPENDIX VI



Supplementary Figure S4 - Chimera breakpoint frequencies on closely (a) and distantly (b) related pools of 48, 24 and 12 individual nematodes, respectively.

APPENDIX VII

Supplementary Table SIII- Predicted secondary structure motifs found on the selected 18S rDNA amplicon of single nematodes (hairpin-loops location and dG) and the correspondent breakpoint position on the chimeric sequences.

sp ID	hairpin loops 55C	dG (Kcal/mol)	break point position
79TN11consensus	G201-C206,T180-A186, A151-T163, G114-C119, A40-T45	-12.66	
11AH11R09	G237-C243,C190-G195, A148-T153, G125-C130, G102-C109, A73-T83, G39-T44	-14.93	
14LB1R09	G236-C242, C189-G194, A147-T152, C95-G101, T59-A69, G35-T40	-12.54	
15SLH4R09	C190-G195, C150-G155, T59-A69, C17-G26	-13.48	
23FF5R09	C191-G196, C151-G157, T60-A70, C18-G27	-12.65	
25LPFB8F04	T181-A187, G134-C139, t95-A99, C15-G20	-8.08	90, 100, 110, 130, 140,
45HW4VF	T175-A182, A153-T159, C128-G132, T64-A78, C35-G40	-4.71	150, 160, 170, 180, 190
58LPCFF4F04	T184-A191, C136-G141, C95-G100, C15-G20	-9.37	
65LPFB10F04	T183-A190, C135-G140, C94-G100, C15-G20	-8.21	
91LPCFF11F04	C148-G152, T109-A115, T84-A99, C29-G34, G11-T16	-8.23	
120PRESS8VF	G165-C173,G131-C136, C92-G96, C13-G18	-8.3	
7PRESS10VF	T202-A209, C146-G151, C113-G118, C34-G39, G16-T21	-11.38	
sp ID	**hairpin loops 65C**	**dG (Kcal/mol)**	**break point position**
79TN11consensus	C249-G256, G201-C206, A40-T45	-5.04	
11AH11R09	C190-G195, A148-T153, G125-C130, G102-C109, G39-T44	-6.37	
14LB1R09	G236-C242, C189-G194, A147-T152, T59-A69, G35-T40	-5.03	
15SLH4R09	T237-A243, C190-G195, C150-G155, G50-T55	-6.27	
23FF5R09	T237-A243, C191-G196, G51-T56	-5.43	
25LPFB8F04	T181-A187, G134-C139, t95-A99, C15-G20	-3.05	90, 100, 110, 130, 140,
45HW4VF	C128-G132	-0.53	150, 160, 170, 180, 190
58LPCFF4F04	C136-G141, C95-G100, C15-G20	-2.69	
65LPFB10F04	C135-G140, C98-G104, C15-G20	-2.78	
91LPCFF11F04	T109-A115, T84-A99, C29-G34, G11-T16	-2.72	
120PRESS8VF	G165-C173, G131-C136, C13-G18	2.87	
7PRESS10VF	C146-G151, C117-G123, C34-G39, G16-T21	-5.42	