A Thesis Submitted for the Degree of Doctor of Philosophy at

Harper Adams University

Harper Adams University

Student

Mohammed Ahmed, BSc. (Hons), MSc.

Title

Free-living nematodes detection using next-generation sequencing technology

A dissertation submitted in partial fulfilment of the requirements for the degree of:

Doctor of Philosophy

Date

21st November 2018

Abstract

The role of nematodes as biological indicators and as key players in nutrient cycling is well recognized. Others have been shown to cause immense losses in food production. Despite their importance, identification of nematodes has been hindered by difficulties in species identification using classical morphology. Alternative molecular-based methods including AFLP, PCR-RFLP and DNA barcoding used alone or in combination with the traditional methods can be time consuming when analysing multiple specimens in a sample. Metabarcoding provides the possibility to identify an array of individuals from many samples simultaneously. The challenge of this approach has been how to identify the most suitable DNA marker(s) as well as the lack of robust analysis pipeline for the sequence data.

An evaluation of the performance of four candidate DNA markers (NF1-18Sr2b, SSUFO4-SSUR22, D3Af-D3Br and JB3-JB5) on a mock community of nematodes showed NF1-18Sr2b is most suitable in terms of coverage and availability of reference sequences. Assessment of the most common bioinformatic tools (QIIME, MOTHUR and USEARCH) showed USEARCH had the best clustering algorithm, was the fastest, had best operational taxonomic units (otus) to actual diversity ratio and ranked the best in user-friendliness. In another mock community experiment, read numbers of taxa showed no correlation with their actual abundance in the community largely due to bias in amplification and copy numbers of the marker region.

Analysis of samples collected from a tillage and traffic experiment using morphological approach showed strong inhibition of herbivores by deep tillage and zero traffic. Bacterivores in general were inhibited by traffic and not affected by tillage. Appraisal of the metabarcoding approach using samples from the same experiment showed at broader classification levels (trophic groups and functional guild), abundance biases associated with the mock community experiment were minimal, the broad implication being metabarcoding data may be useful for assessing quality soil based on the structure of its nematode community.

Declaration

I declare that this thesis has been written by myself, that it describes the work carried out by myself, unless otherwise stated, and that it has not been submitted, in whole or in part, in any previous application for a degree. Information from other sources has been fully acknowledged and referenced in the text.

Mohammed Ahmed

November, 2018

Acknowledgements

My profound gratitude goes to Dr Matthew Back, my director of studies for his invaluable assistance and guidance throughout this PhD. I also wish to thank Dr Melanie Sapp for taking out of her own time to provide support, advice and feedback on this thesis. My appreciation also goes to Thomas Prior and Prof. Gerrit Karssen for helping advance my expertise in nematode taxonomy and their feedback on this work. I am also grateful to Dr Ian Adams for providing a good environment for setting up and carrying out some of my experiments and providing helpful feedback.

I am very grateful to the European Phytosanitary Research Coordination (EUPHRESCO) for funding this studentship in collaboration with Harper Adams University, Fera Science Ltd and the Dutch National Plant Protection Organization. I would like to express my gratitude to Dr Paula Misciewicz, Anthony Millington, Dr David White and Prof Dick Godwin for making it possible for me to use the large marsh site for my experiment.

My appreciation also goes to all the members of the nematology group at Harper Adams University, Dr Ivan Grove, Dr Simon Edwards, William Watts, Alex McCormack, Katarzyna Dybal, Victoria Taylor, Valeria Orlando, Nasamu Musa and Ana Morais-Natalio for their support.

I also wish to thank the following individuals at Fera Science Ltd. who assisted me during my time there, Rebecca Lawson, Erin Lewis, Ummey Hany and Paul Taylor.

Very special thanks to my mother, Fatimatu Hamidu and my brother Ibrahim Ahmed both for their support, I owe where I am now to them. I also wish to thank everyone else in the family for their help and prayers.

Table of Contents

List of Tables

List of Figures

1 Literature Review

1.1 Introduction

The phylum Nematoda is a species-rich taxonomic group that has been reported in abundant numbers across a wide range of habitats (Cobb, 1914; Holterman *et al.*, 2006), from aquatic marine and freshwater to terrestrial environments (van Megen *et al.*, 2009). They represent one of the most dominant metazoans on the surface of the earth in terms of abundance and diversity (Groombridge, 1992; Wilson & Khakouli-Duarte, 2009), with densities of up to 10^8 individuals per square meter and species richness of up to 60 morpho-species (species delineated based on morphology) per 75 cm^3 of sediment (Lambshead, 2004) reported in marine environments. According to Bongers and Bongers (1998), approximately four out of every five metazoans are estimated to be nematodes. They have also been shown to exhibit a diverse range of feeding behaviour (Yeates *et al.*, 1993) and life history strategies (Bongers, 1990). In terms of feeding groups, there are bacterivores, fungivores, herbivores, omnivores and predators. Life strategies span from the small-bodied highly fecund r-strategists, such as the bacterivorous rhabditids to the large-bodied less fecund K-strategists, such as the omnivorous dorylaimids (Johnson, Ferris, & Ferris, 1974).

Previous studies have shown that prevailing physical characteristics such as soil texture, climate, biogeography, as well as enrichment and disturbance events can be reflected through species composition of the local nematode community (Cobb, 1914; Yeates, 1984; Tietjen, 1989; Neher, 2001). In other words, depending on the state of the environment- for example whether soil is stable or has undergone some recent perturbation, the soil nematode community is likely to differ from one point in time to another. The contribution of nematodes to nutrient cycling (Bardgett *et al.*, 1999; Bongers & Ferris, 1999; Wardle *et al.*, 2006) and maintaining a balance in the functioning of ecosystem (Baxter *et al.*, 2013) are also well documented. As permanent community members (being unable to escape habitat disturbance), they serve as important biological indicators of sediment quality (Bongers & Ferris, 1999; Sochová, Hofman, & Holoubek, 2006; Wilson & Khakouli-Duarte, 2009; Höss *et al.*, 2011; Baxter *et al.*, 2013).

Nematode indices used to assess soil quality are based mostly on grouping, into nematodes feeding groups, reproductive strategies and general responses to physical and organic disturbances (Bongers, 1990; Bongers & Ferris, 1999). Classifications into such functional groups are often means of simply lumping together individuals considered to have similar influence on ecosystem functioning, and the validity of such grouping depends mainly on the underlying research objectives (Bongers & Bongers, 1998). Therefore, individuals within a group may not necessarily have close phylogenetic relationship. Although the family or genus level identification is often sufficiently

informative enough for understanding nematodes' role in soil functioning and the effect they have on the ecosystem, species level identification can potentially unravel more information pertaining to such key ecological concepts as biodiversity and functional redundancy (Bongers & Bongers, 1998; Yeates, 2003). The drawback, however, is that their high abundance, minute size and conserved morphology (Decraemer & Hunt, 2006) often obstruct the rapid and accurate identification of their species. Consequently, this has severely affected the fraction of environmental samples analysed in nematode community studies by limiting the scale and resolution of many important ecological studies (Porazinska *et al.* 2010).

In terms of the need for accurate identification of nematodes to species level, research has largely focused on plant-parasitic taxa, due mainly to the magnitude of direct economic losses they inflict on agriculture. They have been responsible for an estimated annual loss of £80 billion (McCarter, 2009). Their management in field crops has for a long time been dependent on the use of nematicides (Hague & Gowen, 1987) which are being gradually phased out following the realisation of the dangers that they pose to the environment (Akhtar & Malik, 2000). The EU some years ago made some very important modifications to its policy on the use of pesticides to make it more sustainable and to reduce the risk that chemical usage on agricultural fields pose to human health and the environment. This involved the re-evaluation (Regulation EC No 1107/2009 and Directive 2009/128/EC) of various synthetic pesticides leaving only a few nematicides available for use by growers (Ntalli & Menkissoglu-Spiroudi, 2011). This led to active search for alternative non-chemical options to replace the synthetic products (Kerry, 2000). Examples of such methods include crop rotation which is a cultural approach of ensuring a host plant is followed by a non-host and planting of resistant cultivars. Effective implementation of such strategies and cropping systems requires a good understanding of the taxonomy and biology of the specific plant-parasitic nematode species or group being targeted (Powers, 2004). Host plant resistance genes are often effective against a specific race or species of plant-parasitic nematodes (Gururani *et al.*, 2012). Therefore, knowing the targeted parasitic species or population can aid the selection of plant genotype to introduce into the field (Blok, 2005).

The existence of character variation and physiological races within species are among the problems associated with, but not limited to the taxonomy of plant-parasitic nematodes (Allen & Sher, 1967). Such complications, among other factors, became the main catalysts for the search for alternative non-morphological approaches to identifying nematode species. Within the genus *Meloidogyne*, a taxon that has received by far more attention than any other group of plant-parasitic nematodes (Sasser, Carter, & Hartman, 1984), techniques such as the differential host test (Sasser *et al.*, 1984), scanning

electron microscopy (Eisenback & Hirschmann, 1981; Charchar & Eisenback, 2000; Eisenback & Hunt, 2009), biochemical approaches such as isozyme electrophoresis (Bergé & Dalmasso, 1975; Esbenshade & Triantaphyllou, 1985, 1990; Carneiro, Almeida, & Quénéhervé, 2000; Tastet *et al.*, 2001) as well as molecular techniques (Harris, Sandall, & Powers, 1990; Hyman, 1990; Petersen & Vrain, 1996; Powers *et al.*, 2005) have been used to complement light microscopic approach for identification. Each of the above-mentioned techniques have certain constraints that limit their adoption as quick, accurate and simple tool for nematode identification across the phylum. Molecular-based methods, however, have maintained their reputation for being fast, reliable and an easy diagnostic approach across many taxa within the phylum Nematoda (Zijlstra, 1997; Blok, 2002, 2005; Powers, 2004).

It is important to mention that most of the pioneering works on molecular-based nematode detection were developed on plant-parasitic nematodes. And as evidence of how important molecular data has become in recent times for nematode taxonomy, it is becoming increasingly customary that most taxonomic descriptions comprise both morphology and morphometric studies as well as molecular analysis of the taxon's relatedness to other species (Handoo *et al.*, 2004; Vovlas *et al.*, 2011; Cantalapiedra-Navarrete *et al.*, 2013). Over the past two decades, there have been a number of published reviews on molecular methods of plant-parasitic nematode identification discussing in depth the different markers and DNA target regions used for discriminating species, their future prospects and limitations (Blok, 2002, 2005; Powers, 2004). High-throughput species identification using next generation sequencing (NGS) technology has also been applied for large scale nematode community studies to enhance better understanding of their diversity (Fonseca *et al.*, 2010; Porazinska *et al.*, 2010b, 2012a; Treonis *et al.*, 2018). This technique, known as metabarcoding has also been applied in the area of plant nematology as a means of analysing very large samples of important plant-parasitic nematode groups for improved understanding of their distribution and diversities (Eves-Van Den Akker *et al.*, 2015). The current and subsequent chapters of this thesis will discuss some of the past and most current approaches to nematode identification and classification with special emphasis on the use of high throughput species identification for unlocking the potential of using nematode communities for evaluation of management strategies and assessments of ecosystem health.

1.2 Classical taxonomy

The need for diagnosticians with the skills for routine identification of taxa based on morphological differences is a problem well acknowledged across many areas of plant pathology, of which nematology is no exception (Blok, 2005). According to Coomans

(2002), morphology can still provide useful diagnostic characters, especially if we are able to overcome the limited resolution light microscopy provides. And despite all its limitations, morphology-based study when carried out diligently can be as good as any biochemical or molecular method used in identifying taxa (Mayr & Ashlock, 1991; De Ley, 2006; Agatha & Strüder-Kypke, 2007). However, what is lacking according to Abebe *et al.* (2011) is the technical and taxonomic expertise required to correctly utilize phenotypic characters and use this to effectively make a decision about the identity of an organism. To worsen the current situation, there has been a continuous decline in the number of taxonomists which according to Coomans (2002), may also be detrimental even to the quality of published taxonomic research, since fewer qualified referees review manuscripts.

Prior to the introduction of molecular data, studies on phylogenetic relationships within nematology have been based on morphological characters. A notable challenge then to the use of morphological characters for achieving more natural classification was how to recognise homologs. That is not to say that such problems are not encountered with the use of molecular data (Abebe *et al.*, 2011). And although it is evidently much easier to identify and quantify sequence evolution than morphological evolution (De Ley, 2000), DNA data when used alone may be subject to some amount of noise and artefact (Dorris, De Ley, & Blaxter, 1999). In view of this, there has been a proposal of a more holistic approach to describing biodiversity which involves the integration of as much data about the organism as possible. According to Dayrat (2005), it is better that morphological and molecular approaches are not seen as competing with each other but rather, used to complement one another. For example, (Sites Jr & Marshall, 2004), in their review of twelve delimitation methods, cautioned against adherence to the use of one method to singly delimit species, since all of the approaches can possibly fail at some point when used in isolation. This integrative approach has been successfully applied in some studies for examining species diversity (Boisselier-Dubayle & Gofas, 1999; Shaw & Allen, 2000; Williams, 2000; Drotz, Saura, & Nilsson, 2001; Marcussen, 2003; De Ley *et al.*, 2005; Ferri *et al.*, 2009)

Integrative taxonomy is without a doubt an excellent approach to species delimitation, especially with the existence of several species concepts, and the fact that each of the species delineation approaches when used singly only constitutes one of the multiple aspects of life's diversity (Dayrat, 2005). However, a key constraint to the widespread adoption of this method is the time and expertise involved. One of the major goals of modern taxonomy is to find identification methods which are fast, accurate, reliable, affordable and perhaps even capable of characterizing undescribed specimens (Powers, 2004). In the identification of regulated pest species, for example, speed and accuracy are very important (Holterman *et al.*, 2012; Kiewnick *et al.*, 2014). Therefore, although reliable and probably more accurate than any of the individual approaches, integrative taxonomy

may lack the speed and simplicity which are equally important in certain situations. The best option therefore, remains to improve and optimize the process of collecting and analysing molecular data to make them singly powerful for species delineation.

1.3 Biochemical methods for nematode identification

Several biochemical and molecular approaches have been used for identification of nematodes. Genomic information at all levels have been utilized for identifying nematodes from DNA sequence, the structure of molecules, genetic mutations to the presence versus absence of genes (Subbotin & Moens, 2006). At the protein level, isozyme analysis (Esbenshade & Triantaphyllou, 1990; Payan & Dickson, 1990), two-dimensional sodium dodecyl sulphate polyacrylamide gel electrophoresis (2-D SDS-PAGE) (Ferris *et al.*, 1994), monoclonal or polyclonal antibodies-based serological techniques (Jones, Ambler, & Robinson, 1988; Schots *et al.*, 1990) (Table 1.1) and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOFMS) (Perera *et al.*, 2009) are the methods that have been utilized for distinguishing nematodes at species or sub-specific levels.

The use of molecular data in the form of DNA has also been widely accepted for identification of taxa, largely because of its inherent ability to overcome most limitations associated with traditional morphology-based nematode identification. Most molecular diagnostic methods are PCR-based and rely on DNA sequence variations. The DNA regions often specifically targeted include the nuclear ribosomal DNA, satellite DNAs and various protein coding genes within the mitochondrial genome (Blok, 2005).

Other approaches are based on random amplification of DNA sequences. Examples include the randomly amplified polymorphic DNA (RAPD) (Cenis, 1993; Castagnone-Sereno, Vanlerberghe-Masutti, & Leroy, 1994), amplified fragment length polymorphism (AFLP) (Semblat *et al.*, 1998; Marché *et al.*, 2001), restriction fragment length polymorphism (RFLP) (Curran, McClure, & Webster, 1986; Carpenter *et al.*, 1992), terminal restriction fragment length polymorphism (T-RFLP) (Donn *et al.*, 2012) and sequence characterized amplified DNA regions (SCAR) (Zijlstra, 2000; Zijlstra, Donkers-Venne, & Fargette, 2000; Carrasco-Ballesteros *et al.*, 2007) (Table 1.2). These random DNA target-based markers have the advantage of having a higher multiplex ratio, a feature which is particularly useful when there is insufficient sequence divergence in the targeted DNA regions (Blok, 2005). However, being random means that they may lack reproducibility. Moreover, approaches such as AFLP can be complex and expensive.

Table 1.1. Commonly used protein-based techniques for distinguishing between species/population of nematodes, their advantages, disadvantages and applications

Approach	Principle	Advantages	Disadvantages	Applications
Isozyme analysis	Patterns of gel-separated isoenzyme bands used to identify species	1. Robust and easy to carry out. 2. To date, offers an excellent means of identifying tropical root-knot nematode species. 3. Extracts from a single sedentary female sufficient for reliable identification	1. Dependent on a particular life-stage of the nematode (young female). 2. Being protein-based subjects this method to influence of environmental conditions (e.g. type of host)	Widely used to separate species of cyst and root-knot nematodes (Esbenshade & Triantaphyllou, 1990; Karssen *et al.*, 1995)
Two-dimensional polyacrylamide gel electrophoresis	Soluble proteins separated on the basis of their charges and masses on a gel	This method allows the separation of proteins with an even better resolution.	Subject to environmental variations.	Used to compare *Heterodera avenae* isolates (Ferris *et al.*, 1994)
Antibody-based serological techniques	Antibodies are raised against species of nematodes and used to detect them	1. Can provide good specificity and sensitivity. 2. Can reliably distinguish between the two species of potato cyst nematodes.	Occasional cross-reactivity can affect specificity.	Monoclonal antibody used to test major *Meloidogyne* species (Ibrahim, Davies, & Perry, 1996).

Table 1.2. Commonly used DNA-based techniques for distinguishing between species/population of nematodes, their advantages, disadvantages and applications

Markers	Principle	Advantages	Disadvantages	Applications
Restriction fragment length polymorphism (RFLP/PCR-RFLP)	Sequence polymorphism between species results in distinct cleaving sites for restriction enzymes, thus resulting in variable number of fragments with diverse sizes	1. The technique is fairly reproducible 2. Simple and inexpensive	Requires prior knowledge of the sequence of DNA region for design of primers or probes.	Using this technique, Carpenter *et al.* (1992) distinguished between three populations of a *Meloidogyne arenaria* race called race 2. Used by Oliveira *et al.* (2006) to separate species of *Xiphinema*
Random amplification of polymorphic DNA (RAPD)	A short primer set is used which anneal to several sites on the DNA. If two of the annealed short primer happen to be close and opposite to each other, they will produce an amplicon. Difference in the gel fingerprints of amplicons separates species or populations.	1. Sequence information of DNA region not a prerequisite. 2. Simple and inexpensive	Technique may lack reproducibility.	Used to distinguish between species and populations of *Meloidogyne* from different origins. (Castagnone-Sereno *et al.*, 1994). Used to separate three *Bursaphelenchus* spp. (Braasch, Burgermeister, & Pastrik, 1995)
Amplified fragment length polymorphism (AFLP)	This involves a series of PCR steps in which separate sets of primers are used to selectively amplify some subsets of products of each preceding PCR step. All selected fragments are run on a gel to product unique fingerprints.	1. Requires no prior knowledge of the sequence of the DNA region. 2. Highly reproducible.	1. Complex technique to carry out. 2. Expensive	Used to typify the genetic variability within the tobacco cyst nematode (TCN) complex (Marché *et al.*, 2001)
Sequence Characterised Amplified Region (SCAR)	A specific distinguishing marker from the fingerprint of a specific taxon or life stage of a species is isolated and amplified. This becomes a SCAR by which that taxon or life stage is identified.	1. Provides a rapid means of screening individuals. 2. Can be highly specific	May be labour-intensive.	Successfully used for identifying species of root-knot nematodes (Zijlstra *et al.*, 2000; Fourie, Zijlstra, & McDonald, 2001)

Table 1.2. Commonly used DNA-based techniques for distinguishing between species/population of nematodes, their advantages, disadvantages and applications. (Continued)

Markers	Principle	Advantages	Disadvantages	Applications
PCR and real-time qPCR	qPCR utilizes PCR to detect target species by using species-specific primers.	1. Highly robust, sensitive, specific and quantifiable 2. Can be multiplexed to detect multiple targets	Requires prior knowledge of the sequence of DNA region for design of primers or probes.	Successfully used for identifying *Xiphinema* spp. (Hübschen *et al.*, 2004b,a)
Terminal restriction fragment length polymorphism (T-RFLP)	This involves the use of fluorescently labelled primers to generate amplicons usually from bulk sample DNA. Amplicons are fragmented and fingerprints from only the labelled terminal fragments are visualized.	1. Because only terminal fragments are used, method is highly reproducible. 2. Method has very high throughput. 3. Semi-quantitative.	In complex communities, changes at the generic or species level may be masked by the method's limited resolution.	Successfully applied in field trial of the effect of tillage on nematode communities (Griffiths *et al.*, 2012)

1.3.1 DNA barcoding

Molecular diagnostics of nematodes has over the years seen enormous progress. Advancements in technology, particularly in the areas of DNA amplification and sequencing, have been the main driving forces towards achieving this. They have made it possible to accumulate substantial amounts of genetic data with sufficient information on sequence divergence that can aid in reliable and easy identification of nematodes (Blok, 2005). Data provided by molecular diagnostics have also enhanced our understanding of nematode systematics and biology in general, by demonstrating whether or not a targeted DNA region will be suitable for species identification (Holterman *et al.*, 2009). According to Powers and Fleming (1998), molecular approaches have enabled the validation of most of the classically delineated nematode taxa while providing clarification in areas where the classical approach has failed. For example, molecular approaches may provide the only practical means of discriminating between cryptic species (valid species that are morphologically indistinguishable) (Powers, 2004). They are also fast, relatively simple, applicable to all nematode life stages, provide highly specific means of identifying taxa, (Powers, 2004) and most of all provide substantial number of differential characteristics in the form of sequence divergence (Blok, 2005).

Like all DNA based identification methods, DNA barcoding was designed for situations where the morphology-based approach proved problematic. It is defined as the use of standardized DNA regions as markers for rapid and accurate species identification (Blaxter *et al.*, 2005; Hebert & Gregory, 2005). The key distinguishing feature between DNA barcoding and other molecular diagnostic methods is the use of standardized markers in the former. Therefore, one of the aims of the barcoding consortium when it was established was to build taxonomic reference libraries with sequences of standardized markers from different organisms (Taberlet *et al.*, 2012). With this amount of information available, a simple search of the sequences of such markers from unidentified organisms against these reference sequences can reveal the identities of these unidentified organisms.

DNA barcoding has proven useful in understanding the degree of variation there is within certain species and how these variations can obscure identification. For example, the concept of cryptic species shows how morphology alone cannot be relied on for discriminating phenotypically identical but valid species. Studies have shown that there are several examples of cryptic species (e.g. *Tobrilus gracilis* (Ristau, Steinfartz, & Traunspurger, 2013) and *Xiphinema krugi* (Oliveira *et al.*, 2006)) within the phylum Nematoda that were previously considered to be the same species (Chilton, Gasser, & Beveridge, 1995; Derycke *et al.*, 2005; Oliveira *et al.*, 2006; Fonseca, Derycke, & Moens, 2008). Barcoding also provides a means of identifying rare species or specimens with

limited availability (Powers, 2004).

DNA barcoding may also be the only option available for identifying an organism when the required life stage or specific sex for morphological identification is lacking or the morphology of the specimen being studied is badly distorted. And finally, on the control of pest movement within trade where speed and accuracy of species identification is critical, barcoding offers a quick and reliable means of detecting quarantine nematode species (Powers, 2004).

Hebert *et al.* (2003) proposed the use of COI of the mitochondrial DNA as a molecular marker for DNA barcoding. As a result, COI has been widely used as standard barcode marker for metazoans (Ferri *et al.*, 2009). Different markers have been proposed for other groups of cellular organisms. Markmann and Tautz (2005) used the nuclear rDNA to study the diversity of meiobenthos (small meiofauna that live in marine and freshwater sediments).

With current advancements in sequencing technology resulting in increasingly wide usage of next generation sequencing, a form of barcoding which has recently gained much popularity is DNA metabarcoding. Taberlet *et al.* (2012) defined metabarcoding as the automated identification of several species from a single bulk sample containing multiples of different taxa. Using this approach, it is possible to carry out high throughput identification of several species in a parallel fashion. DNA metabarcoding typically involves the analysis bulk DNA derived from environmental samples (Taberlet *et al.*, 2012).

Applying the environmental metabarcoding approach, Fonseca *et al.* (2010) used the nuclear rDNA to study marine metazoan biodiversity. In plants, on the other hand, the preferred barcode markers are ones found within the chloroplast genome, and identification often entails the use of combination of two or more regions of this genome (Lahaye *et al.*, 2008; Hollingsworth *et al.*, 2009) or with the ITS region of the nuclear ribosomal RNA gene (Li *et al.*, 2011; Tripathi *et al.*, 2013). The nuclear small subunit ribosomal DNA has also been successfully used as marker for studies involving nematodes (Floyd *et al*., 2002; Porazinska *et al*. 2010a).

The rDNA (18S and 28S) are preferred over the mitochondrial COI gene for nematological studies due to the availability of more conserved regions for universal primer design (Floyd *et al.*, 2002; Carvalho *et al.*, 2010). Moreover, the abundance of sequences of these two genes from described taxa in public databases make matching sequences for identification an easier job than when using COI. In terms of resolution, however, COI is capable of discriminating between species more than either of the rDNA genes (Derycke *et al.*, 2010). A combination of the 18S and 28S genes has been shown to be able to

significantly improve the resolution, thereby achieving better detection levels (Porazinska *et al.*, 2009).

A typical metabarcoding approach proceeds as follows (i) extracting bulk DNA from the organisms or directly from a sediment or soil (ii) amplifying a selected DNA marker region using a set of universal primers (iii) sequencing the amplicons all in parallel via a next generation sequencing platform (iv) clustering of sequences into operational taxonomic units (otus) using a certain similarity cut-off percentage and (v) matching each otu against sequences of identified organisms in a reference database (Valentini, Pompanon, & Taberlet, 2009). Metabarcoding like standard barcoding is based on the assumption that with appropriate barcode marker(s), each operational taxonomic unit can be assigned to a described species through its DNA sequence (Orgiazzi *et al.*, 2015) or identified as unknown if not yet described to assist with the discovery of unknown biodiversity.

Almost all DNA metabarcoding applications in nematology have mainly been based on the analysis of bulk samples of entire organisms already isolated from the containing substrates such as soil, water, plant material etc. (Creer *et al.*, 2010; Porazinska *et al.*, 2010c; Bik *et al.*, 2012b). Beyond multispecies identification from bulk samples of entire extracted organisms, metabarcoding also may comprise the use of total and typically degraded DNA extracted directly from environmental samples without prior isolation of organisms (Taberlet *et al.*, 2012). This approach, if successfully applied in nematology, could help overcome the inconsistencies and poor recovery rates associated with various nematode extraction methods (see, den Nijs and van den Berg 2013). Sapkota and Nicolaisen (2015) also tested and developed new amplification approach to enable high throughput analysing of soil samples by directly extracting the DNA without a nematode extraction step. The authors reported very good coverage of the nematode diversity within the tested soils. However, detailed assessment of the efficiency of DNA recovery from the soil is generally lacking. Also, such a method will usually only allow for analysis of soil samples much smaller in volume than recommended for optimum representation of the nematodes community structure (Wiesel *et al.*, 2015). Moreover, since most meiofaunal organisms are often found in substrates with volumes profoundly larger than the total biomass of the organisms themselves, it becomes eminent that they are separated first before DNA can successfully be extracted (Creer *et al.*, 2010). Nonetheless, with sufficient testing and validation, this approach can be immensely beneficial in the long run.

1.3.1.1 Limitations of high throughput DNA barcoding

There are a number of challenges associated with DNA metabarcoding analysis of environmental DNA. The most notable of these is the identification of a suitable marker to provide the required taxonomic coverage and species resolution. This problem is not

unique to only metabarcoding but associated with standard barcoding as well. As mentioned in earlier paragraphs, the 18S rDNA has been the most commonly used marker in nematode barcoding due to the availability of extensive database resources and the possibility of using conserved regions for designing versatile primers. The latter are continuously being improved to allow coverage of newly discovered taxa (Sapkota & Nicolaisen, 2015). In contrast, it has been shown to have limited taxonomic resolution among certain taxa within the phylum Nematoda (Creer *et al.*, 2010). Nonetheless SSU rDNA region is still the marker of choice for DNA metabarcoding of environmental samples where wider coverage remains essential and species level identification not strictly important.

Another issue with DNA metabarcoding is its reliance on PCR (Taberlet *et al.*, 2012). Significant amount of errors has been shown to accrue during amplification (Haas *et al.*, 2011; Porazinska *et al.*, 2012b). These errors often lead to misinterpretation of diversity within samples, mainly due to the formation of chimeras (Huber, Faulkner, & Hugenholtz, 2004; Edgar *et al.*, 2011). While most of these errors have been attributed to technical factors such as PCR and sequencing errors, inappropriate protocols such as incorrect annealing temperatures and cycle numbers as well as human errors can contribute to the formation sequence artefacts.

According to Porazinska *et al*. (2012b), up to 14% of raw sequence data can be made up of chimeras and in clustered otu datasets, they can constitute up to 40% of dataset. Considering how rampant they may be in sequence dataset, there is always the risk of such hybrid sequences being classified as new taxa or unknown to science. Stringent approaches to removing them from sequence data may sometimes be necessary during data analysis. Several bioinformatic tools designed to identify and discard such hybrid sequences from the reads generated from high throughput sequencing platforms are available (Beccuti *et al.*, 2013). For biodiversity studies, the most commonly used ones are Bellerophon (Huber *et al.*, 2004), ChimeraSlayer (Haas *et al.*, 2011), Perseus (Quince *et al.*, 2011) and uchime (Edgar *et al.*, 2011). Persues and uchime both operate on the assumption that chimeric sequences should be less frequent than the parental sequences (Bik *et al.*, 2012a). In other words, the assumption is that chimeras are less abundant than their parents because they have undergone fewer cycles of amplification compared to their parents. Another method of chimera picking which is incorporated within the QIIME analysis pipeline, is the blast fragment method which is based on the BLAST taxonomic-assignment (Altschul *et al.*, 1990).

One other constraint to DNA barcode-based identification is the need for a huge repository of sequences of characterized species. This data generation process is arguably the most important step, as the success of any future identification will depend

on the accuracy of sequence information in the database. Without any sequence from described taxa to match the obtained sequences with, they may convey limited biological or taxonomic meaning to the investigator. This need for existing sequence information for specific applications has been the main hindrance to many efforts to widen the choices of potential barcode markers, since that would mean channelling a substantial amount of effort into building databases with sequence information from as many characterized species as possible. It also explains why almost all metabarcoding studies involving nematodes tend to use only the 18S rDNA as barcode (Besansky, Severson, & Ferdig, 2003; Porazinska *et al.*, 2009; Creer *et al.*, 2010; Bik *et al.*, 2012b).

1.4 Next generation sequencing technology

In spite of the immense improvements made to the capillary electrophoresis sequencing method, cost of sequencing, time and labour needed were still too high for the growing demands for DNA sequence information (Metzker, 2005). The introduction of the various next generation sequencing (NGS) platforms offered alternative approaches that were cheaper and faster than the traditional capillary sequencing method. (Zhou *et al.*, 2013). The run time for these sequencers can range from just minutes to weeks (Glenn, 2011). There are currently a number of platforms available, all based on some common basic principles, such as their streamlined library preparation steps, and the simultaneity of sequencing and detection processes. They each employ complex interactions of enzymology, chemistry, high-resolution optics, hardware, and software engineering (Mardis, 2008).

The following are some of the next generation sequencing platforms that surfaced into the market some years ago: The Roche 454 genome sequencer, the Illumina Solexa technology, the SMRT sequencing technology by Pacific Biosciences, the Ion Torrent and the ABI SOLiD platform. Other platforms included the Polonator and the HeliScope single molecule sequencer technology. Both the Polonator and the HeliScope are single molecule (shotgun) sequencing platforms; hence, no amplification step is needed. These have the advantage of eliminating biodiversity inflation or artefacts often associated with PCR-based sequencing methods. The absence of PCR in their sequencing pipelines also means abundant information of taxa in samples, which are often obscured by amplification, can be revealed (Zhou *et al.*, 2013). There have been several review articles that have covered in detail how each of these platforms operate including the chemistry and the instrumentations involved (Metzker, 2005; Mardis, 2008, 2013). And more recently, to mark the ten years of NGS and 40 years of DNA sequencing, respectively, Goodwin *et al.* (2016) and Shendure *et al*. (2017) reviewed the different approaches used in NGS and how these technologies are impacting on genetic research. The current

review will, therefore, only touch on a few basics and key features of these platforms.

1.4.1.1 NGS Platforms

The Roche 454 pyrosequencer was the first next generation sequencing platform to become commercially available. It was introduced into the market in 2004 (Mardis, 2008). The platform is based on the pyrosequencing approach which was first described by (Hyman, 1988) (Figure 1.1). The main advantage to using this platform is the relatively long read lengths of the sequences, which means assembling of contigs is easier even in the absence of reference genomes. It does, however, have shallow sequencing coverage due to the few reads it generates per run (1 million sequences). It also has higher error rates, especially when it encounters homopolymer repeats within the sequence (Ekblom & Galindo, 2011). These characteristics are some of the reasons why the technology has since been superseded by other approaches described below. Roche shut down production for the platform in 2013 and finally withdrew support for the platform in 2016.

The Solexa/Illumina technology soon followed 454 technology in becoming the second NGS platform to be available commercially (Figure 1.2). Solexa sequencing has a far more superior sequencing output and depth of coverage than the 454 pyrosequencer. It records fewer incidences of errors in homopolymer regions compared to its 454 predecessor. One of its platforms, the Miseq series currently can produce read lengths of up to 2x300 bp (www.illumina.com/systems/miseq.html) which is an improvement over the 35 bp read lengths of the early Solexa platforms. That said, Illumina has its own unique base calling errors. For instance, it has been observed that accumulation of errors tends to be higher towards the 3' end than at the 5' end (Schröder *et al.*, 2010). There has also been an observed association between increased single-base errors and GGC sequence motifs (Nakamura *et al.*, 2011).

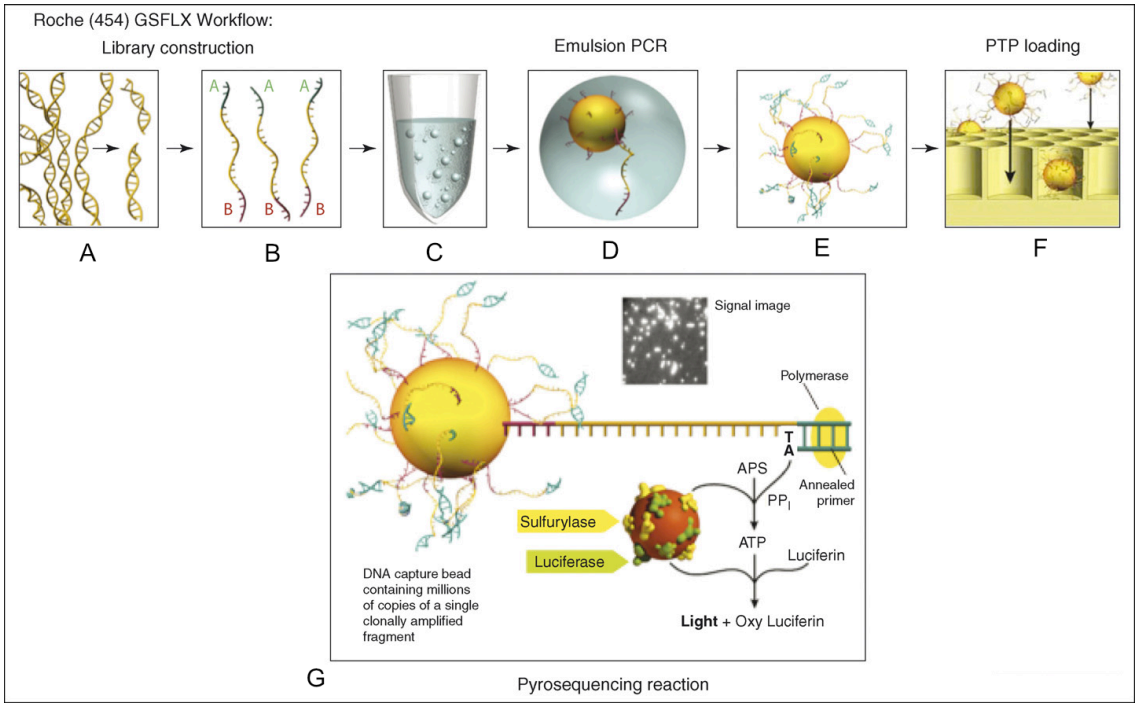

Figure 1.1. The 454 Pyrosequencing method. (A-F) Library preparation: (B) 454 pyrosequencing adapters are ligated to 5' and 3' of DNA fragments, (C) Each adapter-ligated DNA fragment is paired with an agarose bead in an aqueous micelle. (D) On the surface of each bead is a lawn of oligonucleotides designed to be complementary to the DNA fragment adapters to ensure that each fragment attaches to a bead. (E) Copies of each fragment are made through emulsion PCR. (F) Each bead with several copies of DNA fragment on its surface is transferred onto a picotitre plate where sequencing reaction takes place. The sequencing reaction (G) proceeds with the introduction of one of the four bases, at a time, until one is incorporated. Base incorporation is detected through the chain of reactions that follow, which ultimately leads to the production of light. Modified from (Mardis 2008)

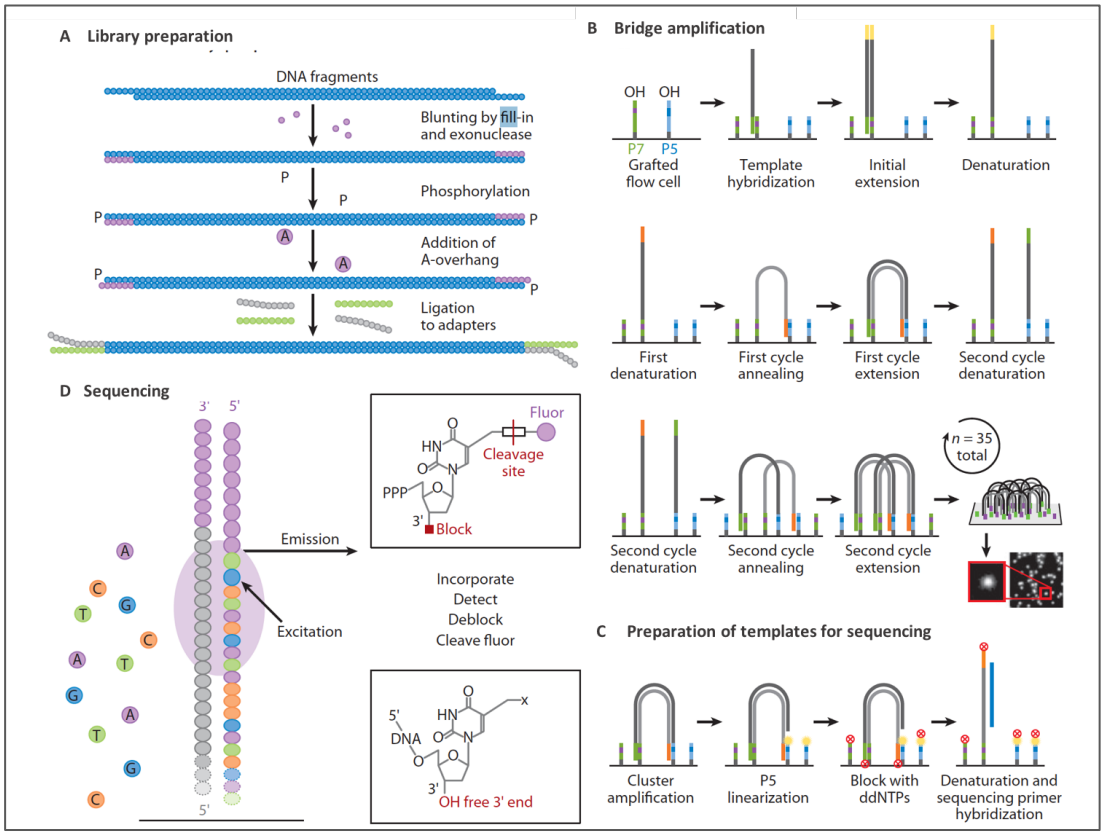

Figure 1.2. Illumina sequencing method. (A) Adapters are ligated to the ends of DNA fragments to be sequenced during library preparation. (B) Complementarity between adapters and oligonucleotides on the surface of flow cells ensures each denatured fragment is immobilised on the surface of the flow cell before bridge amplification which makes sufficient copies so that the fluorescence can be detected during sequencing. (C) The template copies are linearised and the reverse strands cleaved off leaving just the forward strand. The 3' ends of the forward strands are blocked and a sequencing primer hybridized to the template (D) Sequencing proceeds with the extension of the sequencing primers. A base is sequenced during each cycle in a way that the number of completed cycles equal the length of DNA sequenced. All four bases are terminally blocked and are introduced simultaneously during each cycle. The base that is incorporated produces a base-specific fluorescence detectable by a camera. Until the block is removed after the detection, no additional base can be incorporated. Hence, following detection, the incorporated base has to be deblocked and defloured before the next cycle. (Modified from (Mardis 2013))

The SOLiD platform from Applied Biosystems employs a similar library preparation as the previously mentioned NGS platforms. However, unlike the other platforms, it uses ligation to determine sequences. Because each base pair is essentially sequenced twice, the error rates encountered tend to be less in this platform (Mardis, 2008; Ekblom & Galindo, 2011).

The HeliScope was the first NGS platform to introduce the single-molecule sequencing approach. Although this platform has the advantage of being less prone to errors especially those related to amplification artefacts, it produced read lengths that are short compared to any of the previous technologies. For this reason and the high cost of the instrument, the HeliScope is no longer being sold (Glenn, 2011).

The Ion Torrent platform operates in a similar fashion as the 454 technology in that they both involve similar library preparation steps and sequential introduction of each of the four bases. However, instead of registering base incorporation by fluorescent emission, H^+ are released and a signal in proportion to the number of incorporated bases is detected (Rothberg *et al.*, 2011). The PGM (Personal Genome Machine) of Ion Torrent was evaluated together with other platforms such as Illumina and Pacific Biosystem by (Quail *et al.*, 2012). The results indicated that the PGM gave an excellent coverage for those sequences with high GC content to moderate AT richness. However, sequencing of AT-rich genomes resulted in substantial amount of bias with coverage for only about 70% of the genome. On its ability to detect variants, it slightly outperformed the MiSeq, but in doing so recorded significant number of false positives as well.

The SMRT sequencing technology by Pacific Biosciences is based on the natural process of DNA replication by DNA polymerase for real time sequencing of individual DNA molecules (Eid *et al.*, 2009). Each dNTP has a specific fluorescence label attached to its terminal phosphate, which upon incorporation of a nucleotide is detected immediately before it is cleaved off (www.pacificbiosciences.com/products/smrt-technology/). Features such as high speed, long read lengths, high fidelity and low cost per experiment have made this technology a desirable investment (Glenn 2011; https://genohub.com/ngs-instrument-guide/). However, in comparison with the Ion Torrent and MiSeq sequencers, higher depth of coverage is required for calling of variants (Quail *et al.*, 2012).

Another promising technology associated with single molecule sequencing is the nanopore sequencing. In addition to being used for high-throughput sequencing, nanopores technology has been used for sensing and detecting a wide range of small and large molecules (Haque *et al.*, 2013). Like the SMRT sequencing platform, nanopore technology has the capacity to sequence very long DNA or RNA molecules. Additionally, the technology requires no expensive fluorophore labelling (Schneider & Dekker, 2012). Oxford Nanopore technologies is currently the only manufacturer of commercially available platforms that implement nanopore sequencing. One of its platforms, the pocket-sized MinION is currently the most popular one. With accuracy currently around 92%, this piece of technology has already been applied in such areas of research as on-site analysis of disease pathogens, aneuploidy detection, and due to its portability, serves as a good teaching tool (Jain *et al.*, 2016).

1.4.2 Applications of metabarcoding

Most NGS-based nematode community studies have been based on the pyrosequencing method of the Roche 454 platform. The relatively longer read lengths generated with this platform at that time made it even more suitable for metabarcoding analysis. Porazinska

et al. (2009) reported one of the early studies to evaluate the suitability of NGS for nematode metagenetic analysis when they compared two potential barcode regions from the 18S and the 28S genomic regions. Using a combination of the two genomic regions, up to 97% of the species in the tested community were detected in this study. Using either of these markers alone could only recover approximately 90% of the diversity in the sample. The study demonstrated the tremendous potential of this approach in studying nematode communities. Unfortunately, the authors were not able to find any correlation between the number of reads generated for each of the sampled taxa and their abundances. In fact, some of the low abundant taxa produced the highest number of reads. Later, (Creer *et al.*, 2010) reported a case study of meiofaunal diversity in marine littoral benthos and tropical rainforest habitats. Out of eleven classified taxonomic groups recovered from each of the case studies, nematodes emerged as the most dominant taxonomic group in both environments through the proportion of the total number of operational taxonomic units (otus) that matched sequences of nematodes.

Using metabarcoding, Lallias *et al*. (2015) examined the variation in diversities of protists and microbial metazoans including nematodes across two distinct estuaries in UK. They utilized the same 18S rDNA region as the one used by Fonseca *et al*. (2010) in a similar study on marine microbial eukaryotes. One of the key aspects of their finding was the association between the patterns of marine meiofauna diversity and specific factors such as hydrodynamics, salinity range and granulometry depending on their life-history characteristics. In phytonematology, metabarcoding approach targeting a region within the mitochondrial genome was also used in a study to characterise populations of potato cyst nematodes from several Scottish soils (Eves-Van Den Akker *et al.*, 2015). In addition to describing the distribution of *Globodera pallida* mitotypes across Scotland, that study also outlined how to carry out an accurate, high throughput and quantitative means of characterizing up to a thousand fields at the same time.

Next Generation Sequencing methods have also been applied in sequencing of complete mitochondrial genomes (Jex *et al.*, 2008a; Jex, Littlewood, & Gasser, 2010). The process involved an initial amplification step referred to as Long PCR, which is important to provide enough copies of the mitochondrial genome for sequencing. This step basically amplifies the entire mitochondrial genome as two overlapping fragments of approximately 5 and 10 kb sizes (Hu *et al.*, 2007) which then were subsequently bulked and sequenced using the Roche 454 platform. Prior to the use of NGS for whole mitochondrial genome sequencing, the sequencing step was carried out by “primer walking” on capillary sequencers (Jex *et al.*, 2008b). Complete mitochondrial genome sequencing has the potential of enhancing our understanding of nematode relationships through the use of complete mitochondrial genome for inferring phylogeny between related taxa. At the moment, this area remains to be properly exploited. Although most widely adopted

phylogenetic relationships derived from molecular data are based on the small subunit ribosomal RNA gene (Blaxter *et al.*, 1998; Holterman *et al.*, 2006; van Megen *et al.*, 2009), information relating to phylogeny from the mitochondrial genome may increase greatly our understanding of relationships between nematodes.

1.5 Concluding remarks

DNA barcoding is a tool with great potential in the field of taxonomy. It can serve as a rapid identifying feature of organisms written simply as sequence of four distinct bases, thus providing an unambiguous reference for rapid identification (Bucklin, Steinke, & Blanco-Bercial, 2011). The application of this tool will allow non-experts to carry out some of the routine tasks of identifying species, thus equipping scientists with tools for identifying known organisms and recognition of new species. It can facilitate the recognition and discrimination of cryptic species. Moreover, unlike classical taxonomy, DNA barcoding makes it possible to determine the identity of a species from any life stage available. And this becomes particularly useful when analysing samples intercepted in trade, where diagnosticians are often confronted with the problem of having very limited material to work with.

Although the ultimate goal in DNA barcoding is the development of molecular tool(s) capable of profiling as wide diversity of the phylum as possible, for now, at least in nematology, both the classical and molecular fields are needed for a better understanding of the biology and diversity of nematodes. With the speed and higher output that molecular approaches offer, they make nematode community analysis less laborious and thus, facilitate the use of nematodes as bioindicators.

Despite the many published studies on nematode metabarcoding, only one study thus far has been published that outlines a robust workflow for carrying this out (Treonis *et al.*, 2018). In addition, as mentioned earlier, nematode communities are routinely used to assess the quality of soil. However, no real community study to date has been published on the performance of metabarcoding as a substitute for the standard morphology-based approach. The main objectives of the current research, therefore, was to determine the best analysis workflow–from selection of DNA marker to sequence analysis–for metabarcoding of soil nematodes and to identify the most appropriate means of applying this in assessing the effect of management practices on the soil community. Specific objectives are detailed under each experimental chapter.

2 General materials and methods

2.1 Mock community experiment

To determine the most suitable barcode marker(s) to use for metabarcoding soil nematodes, a mock community of different genera/species of nematodes was assembled and used to test these markers. The idea was to sample across a wide diversity of soil nematodes so that the two classes, Chromadorea and Enoplea (*sensu* De Ley and Blaxter 2002) would be well represented. Twenty-three genera were targeted. While a greater number of genera could have been included, it would have been challenging to replicate three times. Most of the genera were included based on how obtainable they were across different types of soil. Some, particularly the plant-parasitic species, were included because their cultures were already available at Fera Science Ltd.

2.1.1 Soil sampling for mock community experiment

Most of the assembled taxa were obtained from samples collected near and around the lake inside Fera Science Ltd. at Sand Hutton (York, UK). Some samples were also taken in an area of mixed woodland just outside of the Fera campus (Figure 2.1). Five bagged samples weighing about 1.5 kg were obtained. Four samples were collected around the lake and three from the area of woodland that were combined into one. At the end of the sampling, the samples were taken to the laboratory for storage at 6 °C.

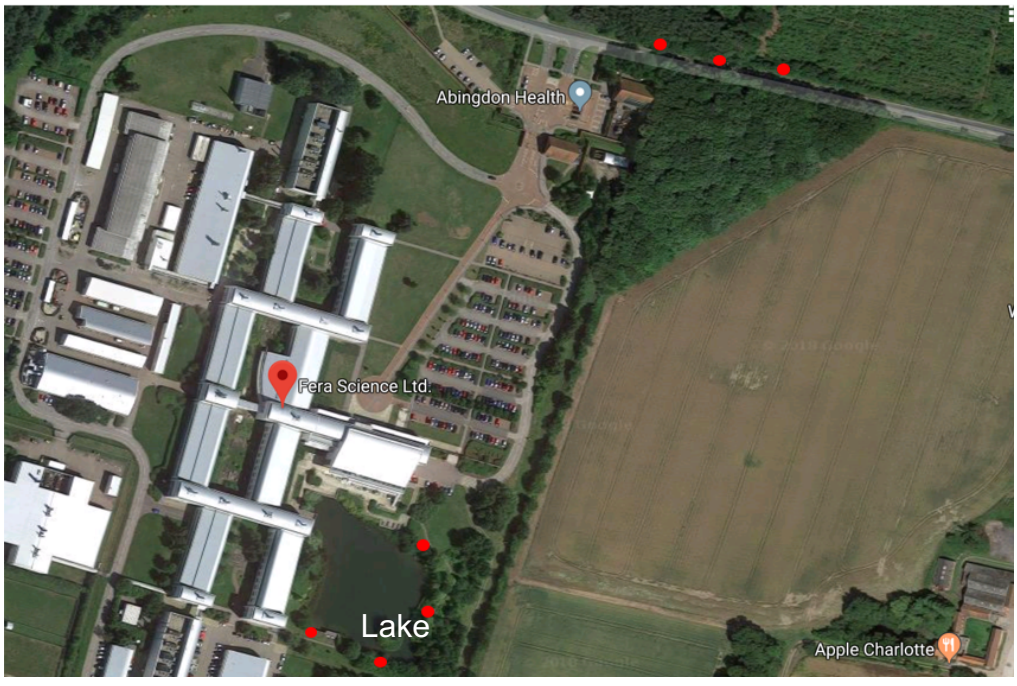

Figure 2.1. Locations within and close to Fera Science Ltd. where samples were collected. Red spots indicate locations where samples were taken. At each location, several soil cores were taken to make up to ~1 kg. (Image from Google Maps on 12/03/18)

2.1.2 Nematode extraction from soil

Nematodes were extracted using two different methods. For most of the samples, the method used was the Whitehead method (Whitehead & Hemming, 1965). This method is based on the motility and the tendency of nematodes to move towards moisture. The set up used in this study included a rectangular plastic tray, plastic-coated wire letter trays, wire mesh and a standard facial tissue (Figure 2.2). The other method used for extraction was the Seinhorst Two-Flask method (Seinhorst, 1955) (Figure 2.3) and it was used only to extract Trichodorids. While the Whitehead method extracts nematodes based on their active movement from the soil into the suspension, the two-flask approach extracts both active and inactive (including dead) ones.

At the lab, each of the bagged samples was homogenised by first emptying the soil into a tray and then gently mixing them. Lumps of soils formed by the soil auger during sampling were also loosened in the process. A subsample of 300 g was used for extraction in both methods. The sample collected from the woodland was the only one extracted using the Seinhorst Two-Flask method in order to maximise the chances of extracting taxa with low motility such as the Trichodoridae. The other four were extracted using the Whitehead Tray method.

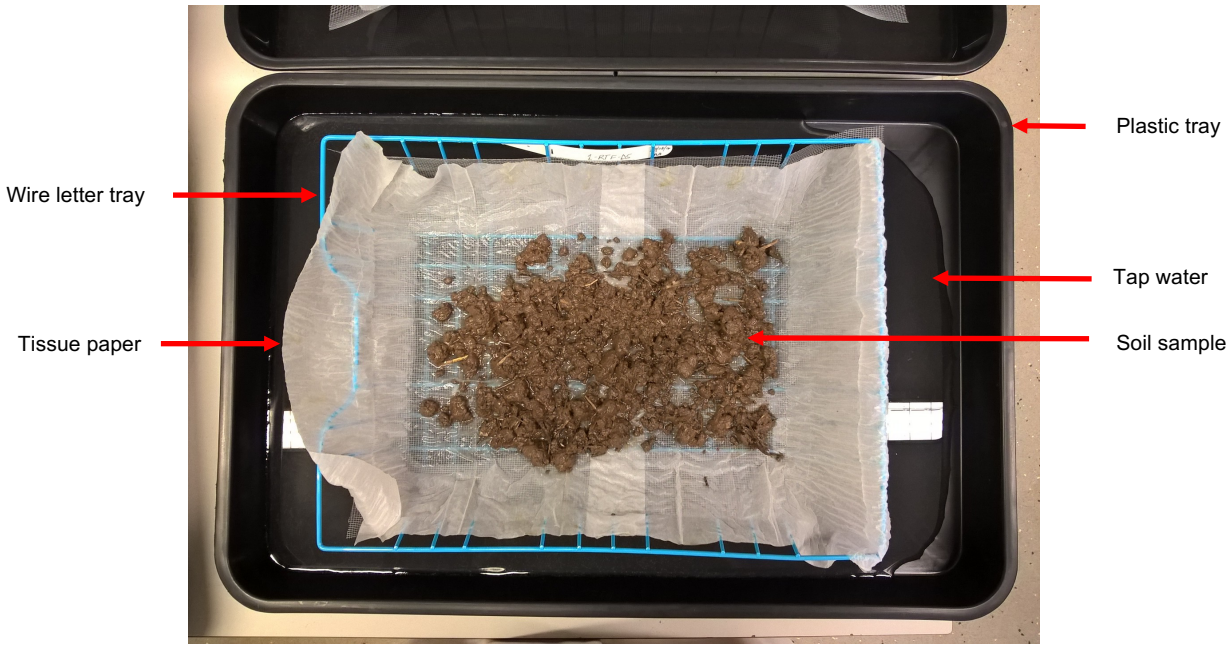

Figure 2.2. Setup for the Whitehead nematode extraction method. The technique depends on the nematodes' active migration towards a higher moisture gradient.

Figure 2.3. Seinhorst two-flask method showing the sedimentation process. (Courtesy, Nasamu Musa, Harper Adams University)

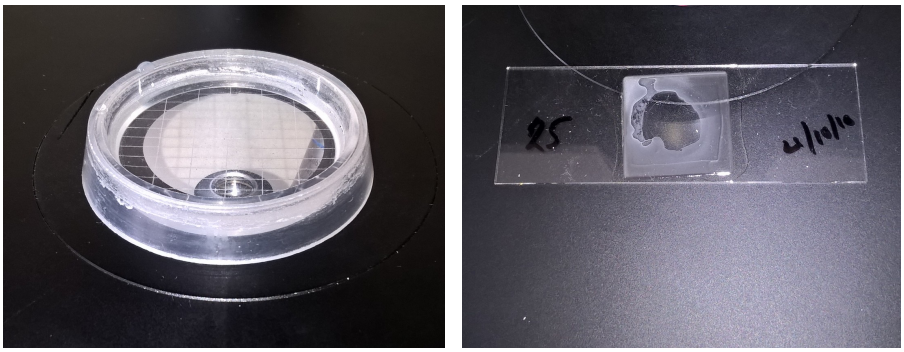

Figure 2.4. (Left) Counting dish with rectangular grid on the base. (Right) Temporary slide showing nail varnish sealant and the melted wax ring

2.1.3 Identification of specimens

The nematode suspensions obtained for each sample after extraction were concentrated to 10 ml. This was done by first allowing the suspension to settle for at least two hours and then pipetting out the supernatant until the volume was down to 10 ml. For each sample the suspension was emptied into a gridded round counting dish (Figure 2.4). The suspension was then examined for the presence of any of the taxa in the list for the mock communities using Leica M50 stereomicroscope (Leica microsystems Wetzlar, Germany) at 40 X magnification. At this magnification, any specimen that resembled any of the targeted taxa was picked and placed in a drop of water surrounded by a wax ring on a glass slide (Figure 2.4). About 3-5 individuals were placed on a slide at a time. The specimens were then covered with a glass slip and then sealed by placing the glass slide on a heat block to melt the wax ring. The seals were reinforced by lining the edges of the of glass cover slips with nail varnish.

The slides were then examined under high power microscope (Zeiss Axio Imager 2, ZEISS, Germany) and identified to the genus or species level using various taxonomic keys and monographs (Andrássy, 2005, 2007; Holovachov *et al.*, 2009; Ahmad & Jairajpuri, 2010; Holovachov & Boström, 2010). After the identities were confirmed, the slides were unmounted to rescue specimens. Those that matched the target taxa were included in the mock community samples. This procedure was repeated until the required numbers each of the sampled taxa were obtained for each of the mock community samples. The identifications were confirmed by taking a specimen each of the different taxa included in the mock community and sequencing it using the Sanger method. There is a more detailed description of how this was achieved in 3.2.3.

3 Evaluating short genomic regions for use in nematode metabarcoding

3.1 Introduction

Fundamental to any DNA sequence-based identification method is the choice of barcode marker(s) (Wilkinson *et al.*, 2017). The chosen marker has to meet most of the criteria outlined by Floyd *et al*. (2002), according to whom the targeted region must be a mosaic of conserved and variable motifs. Conservation of the region, particularly the flanks is necessary for designing universal primers. Within the region, conservation may also be necessary to ensure similarity between individuals of the same species. Likewise, a certain degree of variability within the sequence is required for distinguishing between species. A prerequisite for assigning taxonomy to a sequence is the availability of sequence reference libraries of that marker from a large number of taxa. Depending on the organism being studied, this marker may occur within the nuclear (nDNA), mitochondrial (mtDNA) or plastid DNA (cpDNA short for chloroplast DNA).

Within the mitochondrial DNA, the cytochrome *c* oxidase I (COI) protein coding gene has been the most widely used region, especially for metabarcoding of animals. Most studies involving insects and birds have utilized a region of this gene (Hebert *et al.*, 2003, 2004, 2016; Ramage *et al.*, 2017). The region has been shown to be unique among at least 95% of species of a diverse group of organisms including birds, insects, fishes and plants (Hajibabaei *et al.*, 2007), thus capable of delineating most species of these groups. The use of this region has also been facilitated largely by the enormity of the number of its sequences within the Barcode of Life Datasystems database (BOLD) (Ratnasingham & Hebert, 2007) with the collection numbering over 4.2 million validated sequences as at early 2016 (Coissac *et al.*, 2016).

Despite its success as a barcode marker for most animals, attempts to utilise the COI as a phylum-wide marker for nematodes has not been equally successful for a number of reasons (Creer *et al.*, 2010). First is the extremely high mutation rate of the mtDNA (Blouin *et al.*, 1995; Anderson, Blouin, & Beech, 1998) making the design of universal primers difficult. The hybridization sites for the most widely used primer set LCO1490 – HCO2198 (Folmer *et al.*, 1994) are poorly conserved across the nematode phylum (Blouin *et al.*, 1998) which is likely to result in poor recovery of taxa in bulk community analysis. The primers amplify approximately 710 bp of the COI gene which typically gives 651 bp of readable sequences (Folmer *et al.*, 1994).

Secondly, for metabarcoding on a platform like the Illumina MiSeq, the sheer size of this region precludes it from being utilizable, as this platform, which is currently the most widely used (Shendure *et al.*, 2017), can only sequence up to 300 bp in the forward and 300 bp in the reverse direction while allowing for a certain degree of overlap. This overlap

is to allow for merging of the reads during data analysis, which means that a target actually needs to be shorter than 600 bp. An alternative approach could be to use a shorter COI-based region such as the one amplified by the primer pairs JB3-JB4.5 (Bowles, Blair, & McManus, 1992). The size of the amplified fragment is approximately 366 bp, making it an excellent candidate for an NGS marker. In general, besides being used for discriminating species of certain genera of nematodes (Janssen *et al.*, 2016), most mtDNA-based markers have been largely overlooked in terms of metabarcoding due to their limited taxonomic coverage of the nematode phylum. A similar idea was argued by Deagle *et al*. (2014) who demonstrated how narrow the taxonomic coverage of some of the markers situated within the mitochondrial genome are. Finally, a significant proportion of the many COI nematode sequences in the BOL database belong to parasites of vertebrates, insects and plants, making this region not particularly appealing for nematode community analysis. This is why the most widely used markers for metabarcoding to date are ones associated with the nuclear ribosomal RNA gene repeats (rDNA).

In eukaryotes, rDNA units are known to occur in copies of up to several hundred tandem repeats per genome (Hillis & Dixon, 1991). Each unit consists of two internal transcribed spacers (ITS-1 and ITS-2) separating the 18S, 5.8S and 28S rDNA and an external transcribed spacer (ETS) is located upstream of the 18S gene. The gene is involved in the production of the RNA component of the ribosome, which is responsible for protein synthesis in all living cells. Because of this, the rDNA is well conserved across all life. In *Caenorhabditis elegans*, the array contains approximately 55 copies of the complete unit (Ellis, Sulston, & Coulson, 1986). And within the nematode phylum, copy numbers range from 50 to 100 (Floyd *et al.*, 2002). According to Creer *et al*. (2010), 18S and 28S subunits of this gene offer an excellent choice of barcode regions due to their multiple copies and the concerted evolution events they have been reported to undergo (Markmann & Tautz, 2005). In other words, the selection pressure on these genes helps maintain nearly identical copies within each species. As mentioned earlier, a significant proportion of published studies on nematode barcoding have used markers within this region (Porazinska *et al.*, 2009, 2010b; Creer *et al.*, 2010; Fonseca *et al.*, 2010; Holovachov *et al.*, 2017).

Despite its wide taxonomic coverage, the 18S rDNA-base markers like most markers mentioned here have certain limitations. Aside from the fact that some 18S rDNA markers lack the resolution to distinguish certain species of nematodes, the primers used to amplify them are often not specific. Using the primer pair described by Porazinska *et al*. (2009) on DNA extracted directly from the soil, Peham *et al*. (2017) showed that only 2.5% of sequenced reads were from nematodes. To make it useable for environmental DNA Sapkota and Nicolaisen (2015) had earlier proposed the use of a semi-nested amplification approach which they showed could increase the percentage of nematode

reads from just 3% to 64%. It is worth mentioning that the lack of specificity of these primers may not be such a pressing issue as long as nematodes are isolated from the soil prior to DNA extraction. Given the current capacities of the direct soil extraction kits available at the moment, which can handle only a very small fraction of recommended soil volume of 250 ml for nematode extraction (Peham *et al.*, 2017), it is safe to say that the traditional extraction methods will continue to be widely used until such a time when the capacities of these direct DNA extraction kits are improved. That said, the evaluation of markers, for now, will mostly rest on their taxonomic coverage. Peham *et al*. (2017) in their paper also made a case for the ITS region to be considered as a barcode candidate given the large collection of sequence reference libraries of this available in public databases. However, as stated by Floyd *et al*. (2002), attributes of the ITS region such as it being difficult to align, variable even within species and showing extreme length variation between diverse nematode taxa make it a difficult marker to use.

There are two regions within 18S rDNA that are commonly used in metabarcoding studies involving nematodes; (I) a region amplified using the primer sets NF1-18Sr2b as used by Porazinska *et al*. (2009) which will from hereon be called NF1-18Sr2b marker, and (II) a region amplified with primers, SSU04F-SSUR22 (Blaxter *et al.*, 1998) and has been used in a number of previous studies on metabarcoding of marine nematodes (Porazinska *et al.*, 2009, 2010b; Creer *et al.*, 2010; Fonseca *et al.*, 2010, 2014; Bik *et al.*, 2012b; Holovachov *et al.*, 2017). This second marker, which will hereon be referred to as SSUF04-SSUR22, is located very close to the 5' end of the 18S rDNA. Although few reasons have been provided for this, it appears throughout literature that the NF1-18Sr2b has been the preferred choice for soil nematodes and the SSUF04-SSUR22 for aquatic nematodes. Another rDNA-based marker located within the D3 expansion segment of the 28S rDNA is one that has also been previously tested on mock communities of nematodes (Porazinska *et al*. 2009). This region is amplified using the primer set D3Af-D3Br (Nunn, 1992). Porazinska *et al*. (2009) showed that it was able to detect ~90% of the sampled species, and when combined with NF1-18Sr2b, enhanced the recovery of the sampled taxa to 95%. The COI-region amplified using the JB3-JB5GED (Derycke *et al.*, 2010) primer pairs is also included in this comparison. Even though the COI region has been known to lack truly universal PCR primers (Deagle *et al.*, 2014), its inclusion in this comparison was aimed to give a demonstration of its performance on a bulk soil nematode community. In summary, the objective of this study was to compare the overall suitability of these markers for metabarcoding of soil nematodes and to clarify whether, given their limitations it will be better to look for new candidate markers for metabarcoding of nematodes. Specifically, they were compared on basis of the quality of obtained sequences, availability of reference libraries for assigning taxonomy, ease of assigning taxonomy, resolution and recovery of taxa.

3.2 Materials and methods

3.2.1 Sample extraction and microscopy

The taxa represented in the mock community were extracted from commercial samples strored at 6 °C, intercepted soil and plant samples as well as soil samples taken from within the grounds of Fera Science Ltd in Sand Hutton. Details of sample extraction and nematode identification are described in sections 2.1.2 and 2.1.3. Some of the specimens had already been previously identified and cultured on their plant hosts.

3.2.2 Mock community

Three replicates of artificial assemblages of nematodes were used as mock communities. For each replicate 23 different genera of known abundances were placed in Eppendorf tubes containing 20µl of molecular grade water. The mock communities were assembled to consist of taxa spanning as much diversity across the phylum as possible. In total, 19 different families belonging to six orders within Nematoda were represented (Table 3.1). The communities were dominated by members of the order Rhabditida due their numerical dominance in the natural soil environment.

3.2.3 Molecular identification of specimens using Sanger sequencing

Sequences of single specimens for twenty-one of the taxa represented in the mock community were analysed separately using the Sanger sequencing method for confirmation of their identities based on three distinct genomic regions. The only taxa left out were *Meloidogyne hapla* and *Laimaphelenchus* sp. because the former had been kept in culture for a long time at Fera and the latter previously studied and identified also at Fera. A specimen of each of the taxa included in the community was sequenced using the Sanger sequencing method. Each was picked into a separate Eppendorf tube and sequences of the three different regions were analysed. These regions were a nearly complete 18S rDNA region, the D2-D3 segment of the 28S rDNA region and the COI region.

3.2.4 DNA extraction

Extraction of DNA from the mock community replicates and the single specimens were performed using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Manchester, UK). Each samples (single-specimen samples and the three mock community replicates) were placed in 1.5 ml microcentrifuge tubes containing 20 µl of molecular biology grade water (MGW). A 160 µl volume of Qiagen ATL buffer was added to each sample, followed by

20 µl proteinase K before being incubated overnight at 56ºC. The lysed samples were further processed to obtain pure DNA according to the manufacturer's instructions for genomic DNA extraction.

3.2.5 Amplification of single specimen samples

For the 18S rDNA, an approximately 1800 bp long region was amplified as two overlapping fragments using two primer sets 988-1912R and 1813-2646R for the first and second fragments respectively (Holterman *et al.*, 2006).

The polymerase chain reaction (PCR) of both fragments of the 18S rDNA region was carried out in 25 µl reaction mixture containing 12.5 µl of BIO-X-ACT short mix 2x (Bioline reagents Limited, London), 0.6 µl of each primer namely 988F (5'-CTCAAAGATTAAGCCATGC-3') and 1912R (5'-TTTACGGTCAGAACTAGGG-3') for the first fragment; 1813F (5'-CTGCGTGAGAGGTGAAAT-3') and 2646R (5'-GCTACCT GTTACGACTTTT-3') for the second fragment, and 6.3 µl molecular grade water (MGW). The PCR conditions were 15 min at 95 °C; 5 cycles of (94 °C for 30 sec, 45 °C for 30 s and at 72 °C for 30 sec); 35 cycles of (94 °C for 30 sec, 54 °C for 30 s and 72 °C for 30 s); final extension for 5 min at 72 °C.

The approximately 650 bp long 28S rDNA region was amplified using the primers D2Af and D3Br (Nunn, 1992). The 25 µl reaction mix was made up of 5 µl template DNA, 12.5 µl of BIO-X-ACT short mix 2x, 0.6 µl of each of primers D2Af (5'-ACAAGTACCGTGAGGGAAAGTTG-3') and D3Br (5'-TCGGAAGGAACCAGCTACTA-3') and 6.3 µl molecular grade water (MGW) using the following PCR conditions: 4 min at 94 °C; 35 cycles of (94 °C for 60 s, 54 °C for 90 s and 72 °C for 2 min); final extension for 10 min at 72 °C.

The 400 bp region of the COI gene was amplified using the JB3-JB5 primers (Bowles *et al.*, 1992; Derycke *et al.*, 2005). Amplification was carried out in 25 µl reaction mixture containing the same components as with the other markers. The cycle programmeconsisted of an initial denaturation at 95 °C for 15 min, followed by 40 cycles of denaturation at 95 °C for 1 min, primer annealing at 41 °C for 30 s and extension at 72 °C for 2 min; then a final extension at 72 °C for 10 min.

The PCR amplicons were purified using the QIAquick PCR Purification Kit (Qiagen) before being sent to Eurofins genomics (https://eurofinsgenomics.eu) for sequencing using the same primers used for the PCR.

3.2.6 Amplification and library preparation of mock community samples

For each target barcode marker, four separate polymerase chain reactions (PCR) were set up, representing the three replicates and a blank sample spiked with molecular grade water (MGW). The 5' ends of each of the primers were ligated with MiSeq adapter sequences (Table 3.2). The reaction conditions were different for each marker as detailed in Table 3.3. For all the samples, PCR was performed in 25 µl reaction mix containing 1X Phusion HF buffer (New England Biolabs, Ipswich, MA, USA), 0.2 mM dNTPs, 0.5 µM each of adapter-ligated forward and reverse primers, 1 U of Phusion DNA polymerase (New England Biolabs) and 5 µl of template DNA was used.

Following the initial PCR reaction, the amplicons were all purified using the Ampure XP Beads (Beckman Coulter, Inc. USA). The purified products were quantified using a Qubit® Fluorometer (Thermo Fisher Scientific, Wilmington, DE, USA). This was then followed by the index PCR step where unique dual indices and the Illumina sequencing adapters were attached to each amplicon using Nextera XT index primers (Illumina, San Diego, CA, USA) for amplification (Illumina's 16S Metagenomic Sequencing Library Preparation protocol). The PCR was performed in 50 µl reaction volume containing 5 µl each of Nextera XT Index primers 1 and 2, 5 µl of template DNA, 1X HF buffer, 0.2 mM dNTPs, 1 µl $MgCl_2$, 0.5 U Phusion polymerase and 22 µl MGW. The PCR programme was set at 98ºC for 3 min, 8 cycles of 98ºC for 30 s, 55ºC for 30 s, 72ºC for 30 s and a final extension step at 72ºC for 5 mins. A list of samples and the combination of indexes used are provided in the appendix (Appendix 10.3).

The indexed products were then purified using Ampure XP Beads, quantified and pooled according to their molarity. After that, the pooled sample was run on an Agilent 2200 TapeStation system (Agilent Technologies, Santa Clara, CA, USA) to verify the size of the pooled amplicons. The pool was quantified and diluted to 4 nM concentration. Using the Illumina protocol, the sample was denatured by mixing with 0.2N NaOH. Ten percent denatured PhiX control library was added to the denatured sample to introduce diversity. The mixture was incubated for 2 min at 96 ºC and immediately put on ice, before being loaded in a MiSeq machine (Illumina) for sequencing. The sample was sequenced at Fera (York, UK) in a paired-end approach using 2 x 300 cycles V3 run kit.

3.2.7 Analysis of sanger sequence data from single-specimen samples

Sequences were received as both ABI and SEQ files. Both sequence file formats were visualized using BioEdit Sequence Alignment Editor (Hall, 1999). The ABI files provided the chromatographs for the base calls. Based on this, each sequence was visually edited to high quality by removing areas of ambiguous base calling inside BioEdit. Some of the edited forward and reverse reads could not be merged because for some of the data there was no overlap between the two pairs after editing. NCBI reference database accessed on 1st February 2018 was used for the blast search (NCBI Resource Coordinators, 2016). The blast hits from the three sequenced regions were used to complement one another.

3.2.8 Analysis of NGS data from mock community samples

Sequence analyses were performed using the USEARCH environment (Edgar, 2010) version 8.1.1861. For each of the barcode markers, the paired reads were merged using the *fastq_mergepairs* command, allowing 15 base mismatches in the aligned region. The merged reads were quality filtered and all reads with more than one base expected errors were removed (Edgar & Flyvbjerg, 2015). Reads shorter than 250 bp were also discarded using the USEARCH command *fastq_filter*. The filtered reads were dereplicated via *fastq_uniques* and then clustered into operational taxonomic units (otus) at 97% similarity cut-off using UPARSE (Edgar, 2013) applying the command *cluster_otus* which removes chimeric reads in the process. Otu IDs are assigned based on the number of otu clusters generated for each marker. It is therefore possible, in this chapter as well as subsequent ones, for the same ID to be associated with different taxa as long as the otus/taxa are from different markers.

3.2.8.1 Taxonomy assignment

Operational taxonomic units (Otus) were assigned taxonomy based on the *utax* method within USEARCH. Sequences used as reference database and how they were compiled are described in 3.2.8.2. As an alternative to the *utax* approach, otus of each marker were assigned taxonomy using the BLAST method (Zhang *et al.*, 2000) against NCBI nucleotide database. The blast search was performed using the blast command line tools with all parameters left at default settings. A phylogeny-based assignment method was also performed. For this approach, the reference sequences were first truncated to remove leading and trailing regions outside the primer annealing site of the markers using the USEARCH command *search_pcr* in order to facilitate alignment. The reference sequences were then combined with the otus and aligned using MUSCLE (Edgar, 2004), leaving all parameters at default settings. The aligned sequences were trimmed to the

length of markers inside MEGA 7 (Kumar, Stecher, & Tamura, 2016). The alignments were used to construct maximum likelihood trees using RAxML version 8.2.10 on CIPRES science gateway web portal (Miller, Pfeiffer, & Schwartz, 2010) with GTR as the substitution model at gamma rates distribution. Bootstrap was set to 1000 replicates. Trees were visualized, labelled and coloured within the interactive tree of life (iTOL) web-based tool (Letunic & Bork, 2016).

3.2.8.2 Reference databases for taxonomy

Reference library for assigning taxonomies to the otus generated from the two 18S rDNA markers was obtained from the Protists Ribosomal Reference database, PR^2 v 4.72 (Guillou *et al.*, 2012). The database consists of 18S ribosomal RNA and DNA sequences, with curated taxonomy of protists and other metazoans including nematodes. The version used contained 4910 nematode sequences and was last curated on 7th October 2017. Some of the sequences span the locations of both 18S rDNA markers used.

For the 28S rDNA, reference sequences were obtained from the SILVA ribosomal RNA gene database (Quast *et al.*, 2013) downloaded on 25th January 2018. Thirteen of the sampled taxa could not be found in this database, and so for those with available sequences in NCBI, they were downloaded and added to make a more complete reference database (Table 3.4). A custom python script was used to convert the taxonomies to *utax*-compatible format as instructed in (Edgar, 2015).

A search through the BOLD database for nematode COI sequences revealed that only nine of the taxa included in the mock community have sequences available for comparison. Therefore, as was done with the 28S rDNA, reference sequences of nematode COI from the Barcode of life project within NCBI were obtained (on 25th January 2018) using a command within the statistical assignment package (SAP 1.9.8) (Munch *et al.*, 2008b,a) and formatted for *utax* taxonomy assignments. After formatting the sequences, a few of the sequence entries were filtered out. The ones remaining accounted for only fourteen of the sampled taxa. For the blast taxonomy assignments sequences from NCBI database were used as references.

Table 3.1. Nematode taxa included in the mock community, their families and abundances. Classifications here are based on (De Ley and Blaxter 2002).

Family	Species	NCBI Accessions	Abundance
Alaimidae	*Alaimus* sp.	MG994936	2
Trichodoridae	*Trichodorus primitivus*	MG994943	1
Tripylidae	*Tripyla glomerans*	MG994928	2
Longidoridae	*Longidorus caespiticola*	MG994935	1
Longidoridae	*Xiphinema diversicaudatum*	MG994934	1
Aporcelaimidae	*Aporcelaimellus* sp.	MG994940	1
Mononchidae	*Prionchulus punctatus*	MG994945	2
Anatonchidae	*Anatonchus tridentatus*	MG994941	1
Plectidae	*Anaplectus* sp.	MG994930	1
Plectidae	*Plectus* sp.	MG993558	2
Neodiplogasteridae	*Pristionchus* sp.	MG994929	3
Anguinidae	*Ditylenchus dipsaci*	MG994937	3
Rhabditidae	*Rhabditis* sp.	MG994944	3
Steinernematidae	*Steinernema carpocapsae*	MG994932	12
Cephalobidae	*Acrobeles* sp.	MG994931	1
Cephalobidae	*Acrobeloides* sp.	Failed	2
Tylenchidae	*Tylenchus* sp.	Too short	3
Aphelenchoididae	*Laimaphelenchus penardi*	Not sequenced	8
Aphelenchoididae	*Aphelenchoides* sp.	MG994938	2
Hemicycliophoridae	*Hemicycliophora* sp.	MG994927	3
Criconematidae	*Criconema* sp.	MG994946	1
Heteroderidae	*Globodera rostochiensis*	MG994942	10
Meloidogynidae	*Meloidogyne hapla*	Not sequenced	7

Table 3.2. Primers used for amplification of the target barcode markers. Underlined sections of the sequences represent the Illumina overhang adapters.

Primer	Sequence (from 5' end)	Source
Nex_NF1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGTGGTGCATGGCCGTTCTTAGTT	Porazinska *et al*. 2009
Nex_18Sr2b	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTACAAAGGGCAGGGACGTAAT	
Nex_SSUF04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCTTGTCTCAAAGATTAAGCC	Blaxter *et al*. 1998
Nex_SSUR22	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCCTGCTGCCTTCCTTGGA	
Nex_D3FA	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GACCCGTCTTGAAACACGGA	Nunn 1992
Nex_D3BR	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGGAAGGAACCAGCTACTA	
Nex_JB3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTTTTGGGCATCCTGAGGTTTAT	Bowles *et al*. 1992
Nex_JB5GED	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG AGCACCTAAACTTAAAACATARTGRAARTG	Derycke *et al*. 2010

Table 3.3. PCR conditions for the primers used. Denaturation temperatures are based on manufacturers' recommendations for Phusion polymerase. Annealing temperatures vary between primers and were selected based on estimates from NEB Tm Calculator for each primer pair.

Primers	**Step**	**Temperature**	**Time**
Nex_NF1-Nex_18Sr2b	Denaturation	98ºC	2 min
	30 cycles	98ºC	20 s
		66ºC	45 s
		72ºC	45 s
	Final extension	72ºC	5 min
	Hold	12ºC	∞
Nex_SSUF04-Nex_SSUR22	Denaturation	98ºC	2 mins
	30 cycles	98ºC	20 s
		58ºC	45 s
		72ºC	45 s
	Final extension	72ºC	5 min
	Hold	12ºC	∞
Nex_D3Af-Nex_D3Br	Denaturation	98ºC	2 mins
	30 cycles	98ºC	20 s
		55ºC	45 s
		72ºC	45 s
	Final extension	72ºC	5 min
	Hold	4ºC	∞
Nex_JB3-Nex_JB5GED	Denaturation	98ºC	2 min
	5 cycles	98ºC	20 s
		55ºC	45 s
		72ºC	60 s
	35 cycles	98ºC	20 s
		50ºC	45 s
		72ºC	60 s
	Final extension	72ºC	10 min
	Hold	12ºC	∞

Table 3.4. List of taxa and the accessions of the 28S rDNA records imported from NCBI into the reference library.

Taxa	Accessions
Alaimus	DQ077791.1, JN1233432.1
Trichodorus	MF979186.1, MF979185.1, MF979182.1, KJ002504.1, JQ716462.1, JN123426.1, KX522761.1, JN123408.1, AM180729.1
Tripyla	MF325357.1, MF325355.1, MF125677.1, KU921604.1, KU921601.1
Prionchulus	MF325333.1, MF325332.1, MF325330.1, MF325320.1
Anatonchus	AY593065.1
Anaplectus	MF325171.1, MF325169.1, MF325170.1
Plectus	MF325286.1, MF325285.1, MF325281.1, MF325278.1, MF325273.1, MF325265.1
Ditylenchus	KF534514.1, KF534513.1, FJ707364.1, FJ707361.1, HQ219224.1
Tylenchus	KM058573.1, KM047508.1
Laimaphelenchus	KX580741.1, KX580740.1, KF881746.1
Aphelenchoides	KX357652.1, DQ328683.1, KT003987.1, MF325174.1, KY684030.1
Hemicycliophora	KC329574.1, KF430522.1, KF430521.1, KF430520.1, KF430519.1
Criconema	AY780954.1, AY780953.1, KU722385.1, FN433874.1

3.3 Results

3.3.1 Sanger sequences

Blast search of the sequences of all of the three genomic regions that were sequenced using the Sanger method confirmed the morphological identifications of almost all the specimens. The only taxa that could not be confirmed were the specimens identified by morphology to be *Criconema* and *Acrobeloides* sp. Sequences were not obtained from all the markers for these taxa, which was not surprising because none of the markers produced PCR bands for these taxa on the gel. Due to insufficient sequences of the COI region in NCBI to search the sequences against, only *Plectus* could be confirmed by this region. The D2-D3 region produced amplicons for all except the two missing taxa and was able to confirm the identities of all the taxa. The two fragments of the 18S rDNA together also identified fifteen of the twenty-one specimens. All sequences have been deposited in NCBI (Table 3.5).

Table 3.5. Confirmed identities of individuals included in the mock community by the three DNA regions sequenced using the Sanger method. ☑ denotes positive identification.

Samples	Morphology	18S region	28S region	COI region
Specimen_1	*Hemicycliophora* sp.		☑	
Specimen_2	*Ditylenchus dipsaci*	☑	☑	
Specimen_3	*Aporcelaimellus* sp.	☑	☑	
Specimen_4	*Anatonchus tridentatus*	☑	☑	
Specimen_5	*Globodera rostochiensis*	☑	☑	
Specimen_6	*Trichodorus primitivus*	☑	☑	
Specimen_7	*Rhabditis* sp.	☑	☑	
Specimen_8	*Prionchulus punctatus*	☑	☑	
Specimen_9	*Criconema* sp.			
Specimen_10	*Tripyla* sp.	☑	☑	
Specimen_11	*Pristionchus* sp.	☑	☑	
Specimen_12	*Anaplectus* sp.		☑	
Specimen_13	*Acrobeles* sp.		☑	
Specimen_14	*Acrobeloides* sp.			
Specimen_15	*Steinernema carpocapsae*	☑	☑	
Specimen_16	*Plectus* sp.	☑	☑	☑
Specimen_17	*Xiphinema diversicaudatum*	☑	☑	
Specimen_18	*Longidorus caespiticola*	☑	☑	
Specimen_19	*Alaimus* sp.	☑	☑	
Specimen_20	*Tylenchus* sp.		☑	
Specimen_21	*Aphelenchoides* sp.	☑	☑	

3.3.2 Sequence reads from mock community

The sequence reads were demultiplexed by the MiSeq Reporter software using default settings (allowing one mismatch in the indexes). The Illumina nextera indexes are designed such that each pair differed from the other by at least three bases, therefore

allowing a single mismatch should not result in assignment of reads to the wrong samples (Illumina Document # 15042322 v01, 2015). A summary of the number of reads generated for each marker from each of the three replicates is presented in Table 3.6. The read numbers between the replicates of the NF1-18Sr2b samples were similar. The highest variability between the replicates was found for the markers SSUF04-SSUR22 and JB3-JB5GED.

Table 3.6. Number of sequence reads generated for each of the markers across the three mock community replicates with standard error of means of the replicate samples.

Samples	Number of reads			
	NF1-18Sr2b	SSUF04-SSUR22	D3Af-D3Br	JB3-JB5GED
Replicate 1 (MC1)	2,483,453	3,162,379	3,897,994	1,236,201
Replicate 2 (MC2)	2,349,364	2,790,363	4,228,233	2,160,885
Replicate 3 (MC3)	2,435,278	1,953,138	4,309,817	1,204,900
Standard error of mean	39,216	357,585	125,899	377,501

3.3.3 Quality information

The average length for both forward and reverse reads before merging was 300 bp for all samples. Base quality distributions were similar across all four markers. The forward reads were generally of higher quality than the reverse reads especially toward the 3' end. Quality for both directions dropped significantly towards the 3' end and even more so for the reverse reads. After assembling the paired reads, the base qualities of the merged reads for all the markers were mostly higher than Q30 up to the 400th base position.

3.3.3.1 NF1-18Sr2b

Out of the ~7.3 million paired reads generated using this primer set for the three replicates combined, 47% (~3.5 million) were successfully merged. The average length of the overlap was 236 bp and the length of the merged reads was 362 bp. After the filtering step, 75,234 merged reads were eliminated for having expected error of more than one base. Of the reads that passed the quality check, 591,418 were identified as unique sequences and out of this, 456,425 were singletons. At 97% similarity cut-off, the unique reads were assigned to 138 otus. Chimeric sequences (5,677 reads) and the singletons were all discarded at this stage.

3.3.3.2 SSUF04-SSUR22

For this region, 38% of the 7.9 million paired reads generated from the three mock community replicates were successfully merged. Exactly 120,865 of the merged reads

were discarded after quality filtering for having an expected error of more than one base. The remaining reads were grouped into 721,450 unique sequences, of which 581,220 were singletons and of those that were not singletons, 6,813 were chimeras. The singletons and chimeras were all discarded and the rest clustered into 161 otus.

3.3.3.3 D3Af-D3Br

There were 5.3 million successfully merged reads, representing 42% of the total paired reads produced for this marker. The average length of the overlap was 261 bp and length of the merged reads on average was 336 bp. For this marker, only 9,763 of the merged reads were filtered out for having more than one base expected error. The quality checked reads grouped in 566,284 unique sequences. There were 466,283 singletons and 3,295 chimeras, all of which were removed before clustering which resulted in 144 otus.

3.3.3.4 JB3-JB5GED

Of the 4.6 million paired reads, 57% (2.6 million reads) were merged successfully. The mean overlap length was 124 bp and the merged reads were on average 336 bp long. 64.3% passed the filtering step, resulting in the removal of 227,013 reads. Exactly 355,543 sequences were identified as unique, and 275,258 of these were singletons. Chimeras (1,830 sequences) and the singletons were discarded. Clustering of the non-singletons produced 69 otus at 97% similarity cut-off.

3.3.4 Taxonomy assignment via utax

With the utax method only those genera assigned with support values of 0.5 (arbitrarily chosen) or higher were considered valid in this study (Table 3.7). For the NF1-18Sr2b marker, twenty-three otus produced such valid assignments and they accounted for fourteen of the sampled genera. The results also revealed a phenomenon encountered in some of the curated public database. This is to do with incomplete taxonomies or ambiguous description lines of some of the entries, as pointed out by (Holovachov *et al.*, 2017). Several of the otus could not be assigned names because their best hits were either ‘uncultured eukaryote’, ‘environmental nematode’, ‘Chromadorea_X’ or ‘Enoplea_X’.

Only eight otus of the SSUF04-SSUR22 marker were identified as nematodes which accounted for only five of the sampled taxa. The majority of otus were not given assignments, at least not with sufficient support for them to be considered valid.

For the D3Af-D3Br, only twenty-two of the total 144 otus were successfully assigned nematode identities which corresponded with eight of the sampled taxa.

The JB3-JB5GED marker was the only marker for which no successful assignments were achieved at the genus level. Only three otus were identified as nematodes and could only be confidently identified to the order rank. Two of the otus matched Rhabditida and the other one Tylenchida (according to classification by Siddiqi (2000)).

With exception of a few, most of the recovered taxa occurred in all three replicates for all the markers (Appendix 10.1). For NF1-18Sr2b, only two of the fourteen recovered taxa failed to occur in each of the replicates; only one for SSUF04-SSUR22, two for D3Af-D3Br and for JB3-JB5GED, only one.

Table 3.7. List of taxa recovered based on utax taxonomy assignment. For NF1-18Sr2b and SSUF04-SSUR22, sequences from PR^2 database were used as reference database and for D3Af-D3Br, combined nematode sequences from NCBI and SILVA were used. For JB3-JB5GED, sequences from NCBI database and Barcode of Life project were used.

NF1-18Sr2b	**SSUF04-SSUR22**	**D3Af-D3Br**	**JB3-JB5GED**
Alaimus sp.	*Globodera*	*Aphelenchoides gorganensis*	Rhabditida
Anaplectus sp.	*Longidorus*	*Ditylenchus dipsaci*	Tylenchida
Aphelenchoides ritzemabosi	*Prionchulus*	*Globodera ellingtonae*	
Globodera	*Rhabditis*	*Hemicycliophora wyei*	
Hemicycliophora conida	*Steinernema*	*Longidorus macrosoma*	
Laimaphelenchus penardi		*Meloidogyne hapla*	
Longidorus		*Rhabditis* sp.	
Meloidogyne hapla		*Trichodorus primitivus*	
Pristionchus			
Rhabditis			
Steinernema			
Tripyla sp.			
Tylenchus arcuatus			
Xiphinema			

3.3.5 Taxonomy assignment via blast search against NCBI database

The otus that were generated for each of the markers were used to perform a blast search against reference sequences from NCBI on 16th July 2017. Only alignments with expect (*E*) values less than 0.01 were considered. The top hits were examined for matches that had complete taxonomies. Also, only matches with an identity ≥ 95% were considered. Based on these criteria, otus of NF1-18Sr2b marker matched named sequences in NCBI. All sampled taxa were recovered in the blast output for this marker using this blast approach (Table 3.8). For most of the otus the *E* values of the alignments were 0.00. All non-nematode matches were ignored. For the SSUF04_SSUR22 marker, there were nine of the sampled taxa that were not recovered, due to no match or identities below 95%. These taxa were *Acrobeles, Acrobeloides, Plectus, Laimaphelenchus, Aphelenchoides, Hemicycliophora, Criconema, Aporcelaimellus* and *Tylenchus*. This marker also produced a couple of non-nematodes hits. Operational taxonomic units of the D3Af-D3Br marker had matches for all of the sampled taxa except *Anaplectus, Tylenchus* and *Criconema*. As with all the markers, there were some non-nematode hits. The JB3-JB5GED otus had a slight improvement with this method over the *utax* assignment. Unlike the *utax* taxonomy assignment which gave no assignments below the order level for this marker, the blast method was able to recover five of the sampled taxa. However, three of recovered taxa turned out to be products of sample cross-talk, a phenomenon whereby reads from one sample end up in a different sample within a multiplexed sample (Edgar, 2016). In general, there was a marked improvement in the assignment with the blast search against the NCBI nucleotide database in comparison with *utax* against either PR2, SILVA, or the COI databases. Almost all taxa recovered by the markers were detected across all three replicates (Appendix 10.2). In the NF1-18Sr2b samples, only four out of the twenty-three taxa failed to appear in all three replicates; *Criconema* and *Anaplectus* occurred in two replicates while *Alaimus* occurred only in one. For SSUF04-SSUR22 and JB3-JB5GED all recovered taxa were found in each of the replicates. Of the twenty recovered taxa in the D3Af-D3Br samples, there were only two taxa that failed to occur in all replicates, one of these taxa *Hemicycliophora wyei* was found only in one while *Acrobeles complexus* occurred in two.

Table 3.8. List of taxa recovered from blast search of all otus against the NCBI reference database. Only names appearing in top five hits and had similarities ≥ 95%, e value < 0.001 were considered.

NF1-18Sr2b	SSUF04-SSUR22	D3Af-D3Br	JB3-JB5GED
Alaimus sp.	*Alaimus* sp.	*Alaimus* sp.	*Anatonchus tridentatus*
Anaplectus sp.	*Anaplectus* sp.	*Aphelenchoides ritzemabosi*	*Meloidogyne hapla*
Anatonchus tridentatus	*Anatonchus tridentatus*	*Anatonchus tridentatus*	*Longidorus caespiticola*
Aphelenchoides ritzemabosi	*Ditylenchus dipsaci*	*Ditylenchus dipsaci*	*Steinernema carpocapsae*
Aporcelaimellus obtusicaudatus	*Globodera rostochiensis*	*Aporcelaimellus obtusicaudatus*	*Trichodorus primitivus*
Acrobeles sp.	*Longidorus caespiticola*	*Acrobeles complexus*	*Tripyla sp.*
Acrobeloides sp.	*Meloidogyne hapla*	*Acrobeloides* sp.	
Criconema sp.	*Prionchulus punctatus*	*Globodera rostochiensis/pallida*	
Ditylenchus dipsaci	*Pristionchus lheritieri*	*Hemicycliophora wyei*	
Globodera rostochiensis	*Rhabditis* cf. *terricola*	*Longidorus macrosoma*	
Hemicycliophora conida	*Steinernema carpocapsae*	*Laimaphelenchus deconincki*	
Laimaphelenchus penardi	*Trichodorus primitivus*	*Meloidogyne hapla*	
Longidorus caespiticola	*Tripyla glomerans*	*Plectus* sp.	
Meloidogyne hapla	*Xiphinema diversicaudatum*	*Prionchulus* sp.	
Plectus andrassyi		*Rhabditis* sp.	
Prionchulus punctatus		*Pristionchus lucani*	
Pristionchus lheritieri		*Steinernema carpocapsae*	
Rhabditis cf. *terricola*		*Trichodorus primitivus*	
Steinernema carpocapsae		*Xiphinema diversicaudatum*	
Trichodorus primitivus		*Tripyla sp.*	
Tripyla glomerans			
Tylenchus arcuatus			
Xiphinema			

3.3.6 Taxonomy assignment using phylogeny

The trees produced by the four markers all had some clades bearing very low supports, to such an extent that most of them would be rendered meaningless were they to be condensed based on branch supports cut-off of even 50%. Because of this, the original topologies were kept as they were. The NF1-18Sr2b-based tree placed most of the otus together with named sequences from NCBI within monophyletic clades (Figure 3.1). The results of the tree-based assignments were very similar to the blast approach, with at least 22 out of the 23 taxa identified. *Criconema* was the only taxon that did not form a clade

with any of the otus in the tree. The otu that matched *Criconema* from the blast search clustered with *Ogma* and *Bakernema,* both of which are close phylogenetic relatives of *Criconema*. It was discovered that the failure to have assignment for the *Criconema* otus was because it had no sequence in the reference sequences. Although one of its reference sequences was present in the PR^2 reference database, it got removed during the PCR trimming step (using the command *search_pcr* in USEARCH) that was performed before alignment. From the SSUF04-SSUR22 tree, only four of the sampled taxa could be correctly identified (Figure 3.2). For the D3Af-D3Br marker, otus of sixteen of the sampled taxa were identified with the phylogeny method (Figure 3.3). With the JB3-JB5GED-based tree, four clades could be identified that were monophyletic but only three could be used to identify the otus to the genus level. The otus clustered with these three genera: *Steinernema*, *Longidorus* and *Meloidogyne* (Figure 3.4).

3.3.7 Taxonomic coverage

The calculation of taxonomic coverage of the markers was based on how many of the sampled taxa were recovered by at least one of the three replicates. This was based on a consensus of the results of the taxonomy assignment via utax, blast and the phylogenetic tree method. The NF1-18Sr2b had the highest coverage, producing 100% recovery of the sampled taxa (Table 3.9). All 23 taxa were detected in all three replicates, apart from *Acrobeles* and *Criconema* which both failed to appear in one of the replicates. In the case of the SSUF04-SSUR22 marker, eight taxa were missing from all three assignment methods. The ones that were recovered occurred in all three replicates. The 28S rDNA-based D3Af-D3Br marker recovered all taxa except *Criconema* in the consensus taxonomy. Amongst the recovered taxa, *Hemicycliophora* occurred in one of the replicates, *Acrobeles* in two, while the rest were found in all three replicates. For the COI-based JB3-JB5GED marker, even the consensus taxonomy drawn from all three assignment methods could only recover two taxa, *Meloidogyne* and *Steinernema*. Although the tree-based method included *Longidorus* in the assignment, it was discovered that Otu17 and the NCBI reference sequence KJ741245 *Longidorus* sp which were clustered together had only 81% similarity which was insufficient for assignment. In general, the consensus taxonomies for all the markers were almost exactly as what the blast search produced because all successful assignments made by *utax* against the references were also made by blast search against the nucleotide database, which also detected even more taxa that were missing in the *utax* results.

Figure 3.1. Maximum likelihood tree of the 18S rDNA-based NF1-18Sr2b otus and 18S rDNA reference sequences from NCBI

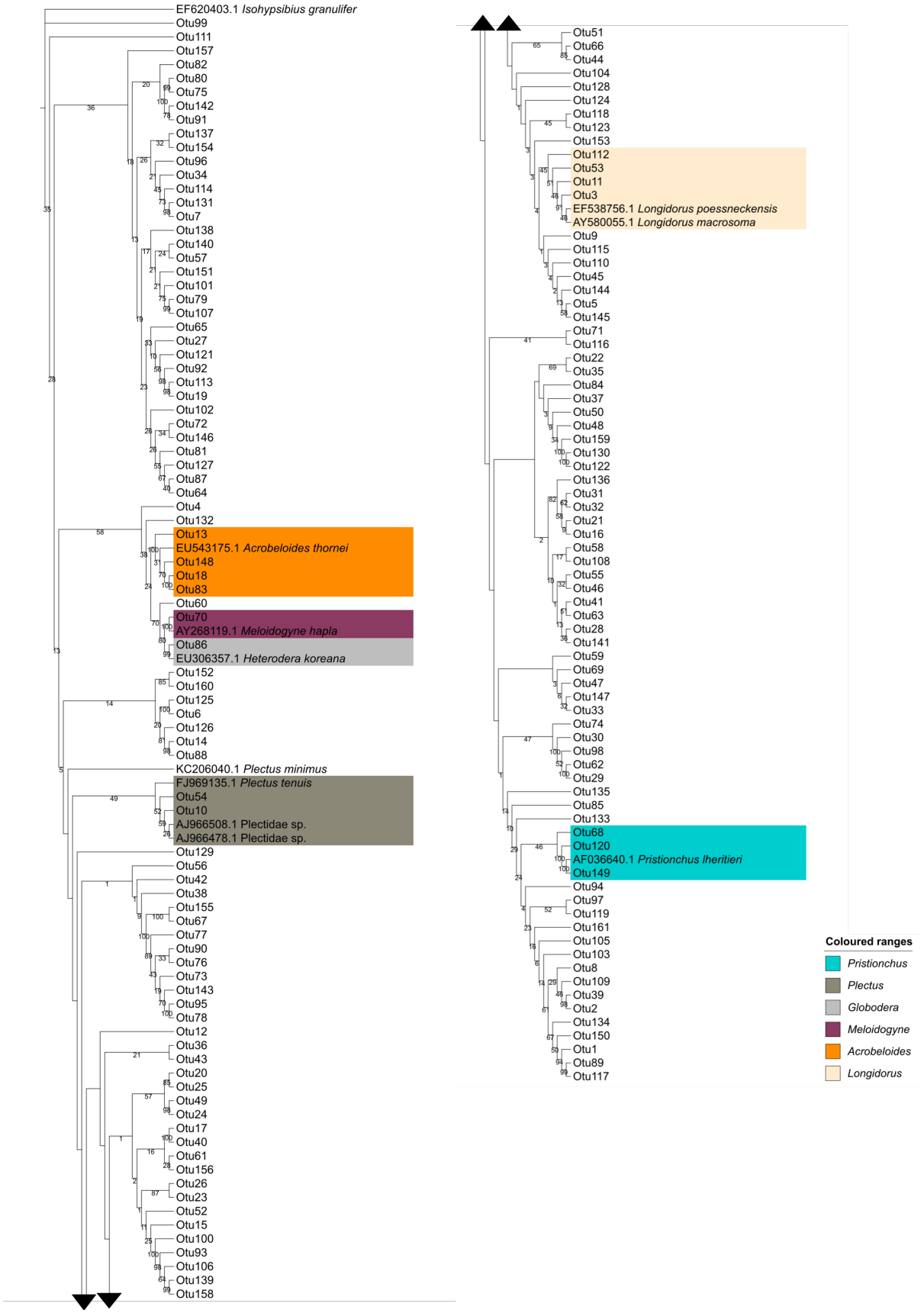

Figure 3.2. Maximum likelihood tree of the 18S rDNA-based SSUF04-SSUR22 otus and 18S rDNA reference sequences from NCBI

Figure 3.3. Maximum likelihood tree of the 28S rDNA-based D3Af-D3Br marker otus and 28S rDNA reference sequences from NCBI

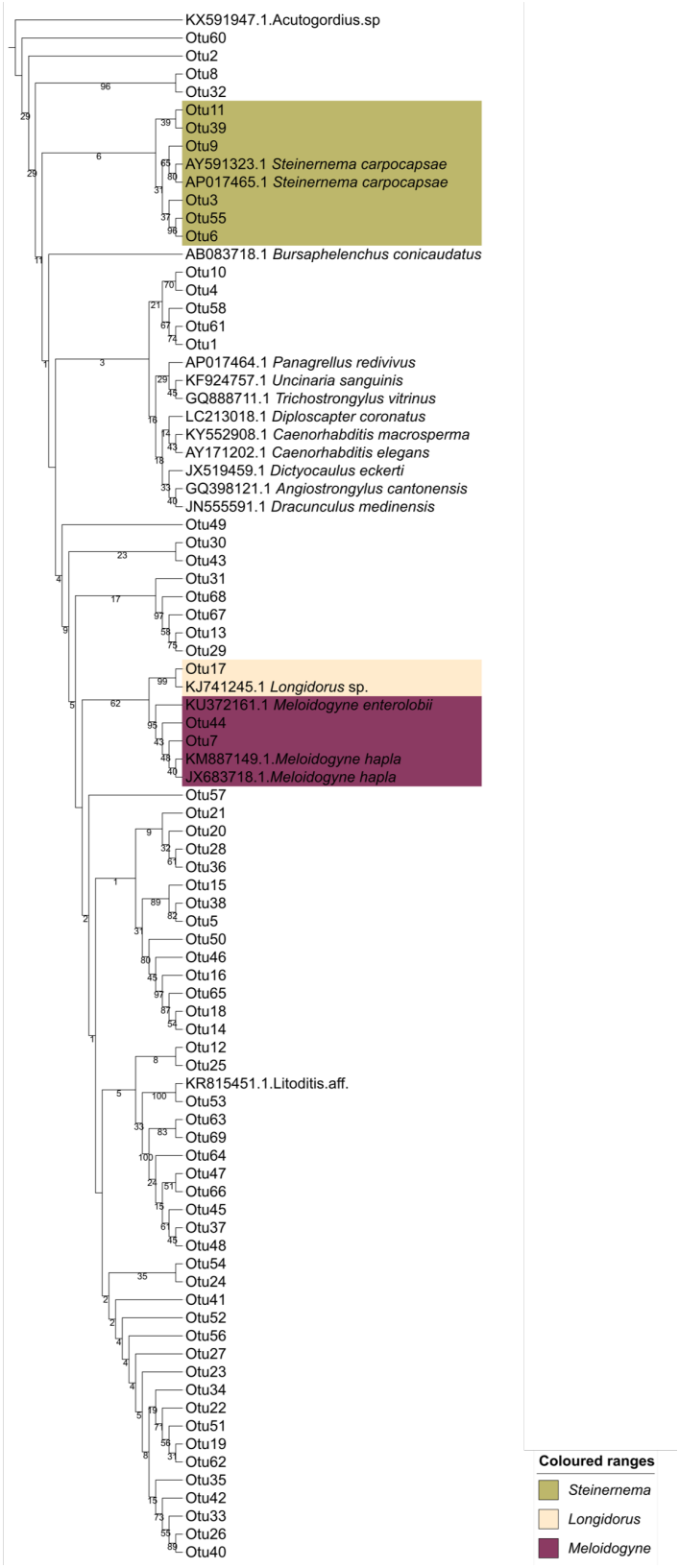

Figure 3.4. Maximum likelihood tree of the COI-based JB3-JB5GED marker otus and COI reference sequences from NCBI

Table 3.9. Taxa recovered by the barcode markers in at least one of the replicates

Taxa	NF1-18Sr2b			SSUF04-SSUR22			D3Af-D3Br			JB3-JB5GED		
Alaimus sp.	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Trichodorus primitivus	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Tripyla glomerans	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Longidorus caespiticola	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Xiphinema diversicaudatum	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Aporcelaimellus sp.	☑	☑	☑				☑	☑	☑			
Prionchulus punctatus	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Anatonchus tridentatus	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Anaplectus sp.	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Plectus sp.	☑	☑	☑				☑	☑	☑			
Pristionchus sp.	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Ditylenchus dipsaci	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Rhabditis sp.	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Steinernema carpocapsae	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Acrobeles sp.	☑	☑					☑	☑				
Acrobeloides sp.	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Tylenchus sp.	☑	☑	☑				☑	☑	☑			
Laimaphelenchus penardi	☑	☑	☑				☑	☑	☑			
Aphelenchoides sp.	☑	☑	☑				☑	☑	☑			
Hemicycliophora sp.	☑	☑	☑				☑					
Criconema sp.	☑	☑										
Globodera rostochiensis	☑	☑	☑	☑	☑	☑	☑	☑	☑			
Meloidogyne hapla	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑

3.3.8 Read frequencies and abundance

The relative number of sequence reads associated with some of the sampled taxa mostly deviated from their relative abundance in the mock community. This was true for all four markers (Figure 3.5). The relative read frequencies between the replicates, however, revealed a certain degree of correlation, indicating some level of reproducibility. For the NF1-18Sr2b marker, the relative read frequencies associated with *Xiphinema*, *Trichodorus* and *Aporcelaimellus* were similar to their relative abundances in the mock community. The most extreme deviation between relative read frequencies and relative abundance was observed in *Prionchulus*. In the case of the SSUF04-SSUR33 marker, *Prionchulus* and *Anatonchus* were the taxa for which relative read frequencies that deviated the most from the relative abundance. With this marker, the relative read frequencies of *Acrobeloides*, *Alaimus* and *Tripyla* were quite similar to their relative abundances. Relative frequencies of the D3Af-D3Br reads generated for *Tripyla*, *Rhabditis* and *Prionchulus* also deviated significantly from their respective relative abundances. The relative read frequencies of *Xiphinema* and *Acrobeloides* were similar to their respective relative abundances. And finally, of the two taxa that were successfully assigned from the JB3-JB5GED marker, *Steinernema* was the one with the closest match

between relative read frequencies and relative abundance. The reads of *Meloidogyne*, the other identified taxon deviated significantly from the relative abundance.

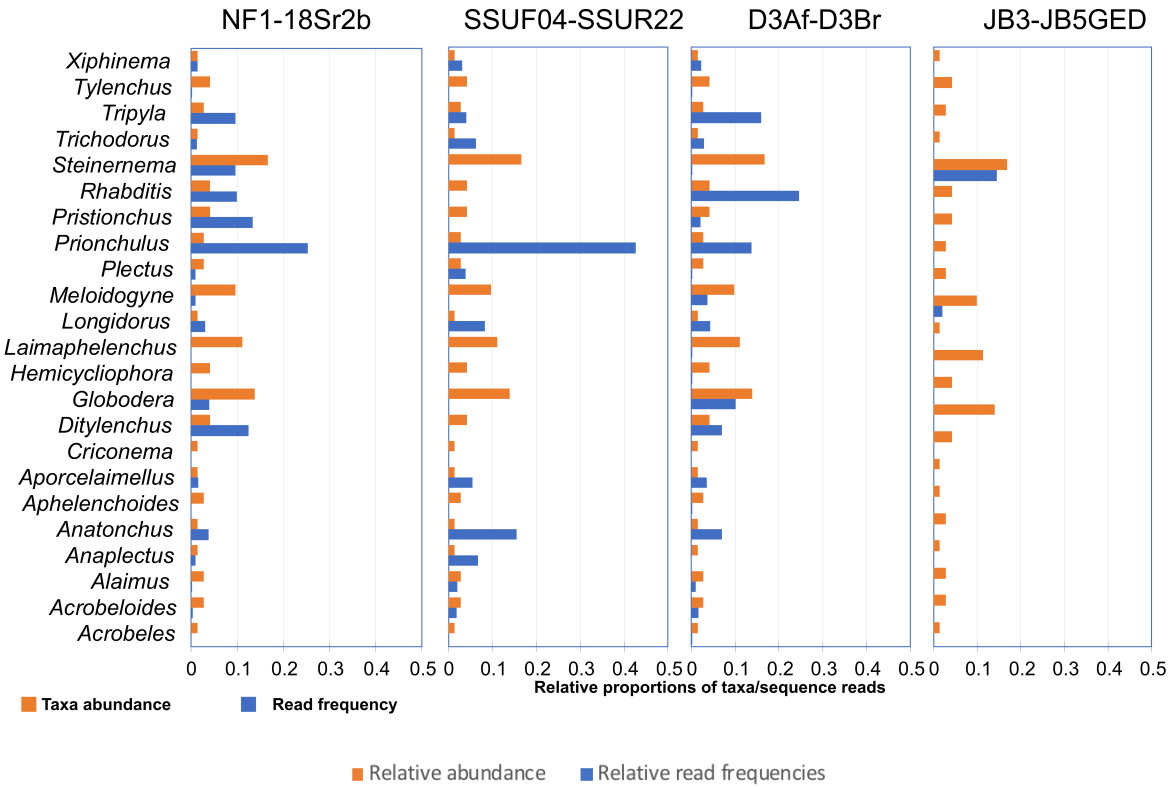

Figure 3.5. Comparison of the relative read frequencies and relative abundances of sampled taxa for each of the four markers. Relative read frequency values are averages of the three replicates.

3.3.9 Cross-talk

This phenomenon occurs when a sequence read is assigned to the wrong sample during demultiplexing (Edgar, 2016). It was observed through close examination of the blast results and the otu table with taxonomy that there were a number of wrongly placed otus (Table 3.10). The NF1-18Sr2b, SSUF04-SSUR22 and JB3-JB5 samples were all sequenced in a single run. Between the two 18S rDNA markers, the detection of these spurious otus was made possible by the differences in where they aligned with the full-length 18S reference sequences in NCBI. The SSUF04-SSUR22 marker is situated within the first 500bp of most of the full-length 18S sequences while the NF1-18Sr2b marker is located much further downstream, around the 1200th to 1600th base positions. By observing how the otus aligned to the reference sequences, these products of cross-talk could be detected. For cross-talk between JB3-JB5GED and SSUF04-SSUR22, the detection was easier because a simple search of the keyword, 18S in the output file of the blast search against the NCBI database with the JB3-JB5GED otus revealed all the otus that were not COI-based and vice versa. There was zero incidence of NF1-18Sr2b otus in

the JB3-JB5GED marker. The 28S-based marker, D3Af-D3Br was not involved in the cross-talk because it was set on a separate run of its own.

Table 3.10. Locations and sources of otus identified as products of cross-talk, their read frequencies in the replicates and their taxonomic identities. Number after 'Otu' are otu IDs after clustering. Each marker has its own set of otus. It is possible for the same out ID to be associated with different taxa as long as the taxa are from a different marker.

Location	Source	otu	MC1	MC2	MC3	ID
NF1-18Sr2b	SSUF04-SSUR22	Otu52	10	10	18	*Prionchulus punctatus*
		Otu79	5	5	4	*Anatonchus tridentatus*
		Otu115	0	1	2	*Trichodorus primitivus*
		Otu138	1	0	3	*Longidorus caespiticola*
		Otu69	9	9	6	*Bursaphelenchus* sp.
	JB3-JB5GED	Otu109	4	4	0	*Bursaphelenchus* sp.
SSUF04-SSUR22	NF1-18Sr2b	Otu67	7	18	10	*Prionchulus punctatus*
		Otu73	4	10	1	*Steinernema carpocapsae*
		Otu76	4	10	0	*Tripyla* sp.
		Otu77	3	15	2	*Pristionchus lheritieri*
		Otu78	3	7	3	*Ditylenchus dipsaci*
		Otu90	1	8	2	*Longidorus caespiticola*
		Otu95	2	5	0	*Globodera pallida*
		Otu143	2	6	1	*Rhabditis cf. terricola*
		Otu155	1	4	0	Uncultured nematodes
	JB3-JB5GED	Otu100	0	0	7	*Crossogaster* sp.
JB3-JB5GED	SSUF04-SSUR22	Otu45	4	6	4	*Anatonchus tridentatus*
		Otu47	5	3	0	*Trichodorus primitivus*
		Otu48	7	0	0	*Mylonchulus*
		Otu63	1	1	1	*Steinernema carpocapsae*
		Otu66	2	3	0	*Tripyla* sp.
		Otu69	3	2	0	*Longidorus caespiticola*

3.4 Discussion

Taxonomic coverage is very crucial to any metabarcoding study. The ability of a marker to recover as many taxa as possible could easily be one of the main benchmarks for determining its suitability for metabarcoding. With the main objective here being to evaluate four widely used markers for how suitable they are for metabarcoding of nematodes, this discussion will be focussed only on how the markers performed based on a consensus of all the assignment approaches and not on the differences in performance of the taxonomy assignment methods themselves. This subject is well covered in (Holovachov *et al.*, 2017) for some aquatic nematodes. The results of the current study have shown that despite some recommendations to adopt COI-based markers (Prosser *et al.*, 2013; Peham *et al.*, 2017), there is still significant amount of effort needed to get this region ready as a barcode marker for nematodes. Besides being a region with currently poor taxonomic coverage, the lack of a comprehensive reference database strongly hampers its adoption for nematode metabarcoding in a manner similar to other animal groups such birds, fishes and insects. Although the length of the JB3-JB5GED gives it the appropriate size for application on a wide array of NGS platforms, it has the same limitation as other COI-based markers–their primer annealing sites are not sufficiently conserved across the diversity of groups such nematodes (Deagle *et al.*, 2014). Nonetheless, the fact that only two taxa could be assigned taxonomy may not necessarily mean that these were the only taxa that were successfully recovered. It is suspected that the absence of sequences in the reference database that covered the location of this marker may have led to the failure of otus from some the sampled taxa to be correctly assigned taxonomy.

Another marker whose poor coverage could be attributed to insufficient matching reference sequences was the 18S rDNA-based SSUF04-SSUR22. Being part of a region that is supposedly well conserved and with a large collection of reference sequences, particularly for nematodes, the failure to detect eight members of the mock community was not the expected outcome. However, the issue with this marker is its location within the full-length 18S rDNA operon. As mentioned earlier, this marker is situated within the first 500 bp of the 18S rDNA region and so unless the entire length of the 18S rDNA or this specific region are covered by a reference sequence, the reference is likely to not contain the homologous region for this maker. Although this marker has been used in a number of metagenetic studies involving meiofauna (Fonseca *et al.*, 2010; Bik *et al.*, 2012b), its coverage within a diverse mock community has never been evaluated. Creer *et al*. (2010) examined the homology of their amplification primers to ~170 aligned sequences of nematodes and showed that 100% of base pair positions are conserved in the nematode sequences which supports this argument in suggesting that primer mismatch may not be the issue with this marker. Nonetheless, it is still popular for studies

involving marine nematodes. For soil nematodes, however, there is still a number of taxa whose full SSU sequences need to be made available to make this marker broadly suitable.

The D2-D3 expansion segment of 28S rDNA region may be the region besides the 18S rDNA region that has just the right amount of conservation and variability for a perfect barcode marker. The region spanning these two high variability segments (D2-D3) have also been the focus of phylogenetic studies for various groups of soil nematodes (Kaplan *et al.*, 2000; Litvaitis *et al.*, 2000; Subbotin *et al.*, 2005, 2006, 2007, 2008, 2011; Giblin-Davis *et al.*, 2006; Douda *et al.*, 2013; Gutiérrez-Gutiérrez *et al.*, 2013). It is also a common practice for phylogenetic analysis of certain groups of nematodes to be inferred based on both the commonly used 18S rDNA and the 28S rDNA region (Gutiérrez-Gutiérrez *et al.*, 2013; Ahmed *et al.*, 2013). And because the number of published studies on particular genomic regions may determine the number of reference sequences available publicly, it is quite convenient to build a reference database for a marker located within either the D2 or D3 segments. The entire length of the 28S rDNA marker used for the mock community study is situated within the latter segment and so there was no issue of it falling out of range with most published sequences in NCBI. This therefore explains the high percentage assignments obtained for D3Af-D3Br compared to the SSUF04-SSUR22 marker despite there being more 18S rDNA than 28S rDNA nematode sequences. A search using the keywords, 'Nematoda' and '18S' of the NCBI database on 6th February 2018 yielded 24,370 entries as compared to 19,674 for the keywords, 'Nematoda' and '28S'.

Unlike the SSUF04-SSUR22, the location of the NF1-18Sr2b marker within the 18S rDNA region puts it within the flanks of most sequences used for reconstructing 18S rDNA-based phylum-wide phylogeny of nematodes (Holterman *et al.*, 2006; van Megen *et al.*, 2009). The utility of this marker benefits greatly from the ample reference sequences of 18S rDNA and availability of highly conserved primer annealing site. Although Porazinska *et al*. (2009) reported a very high coverage of this marker, there were still three taxa they could not account for in the sequence reads. According to them this could have been caused by a number of factors: unmet species-specific PCR requirements, nematode-nematode template competition or nematode-non-nematode templates competition. The disparity in coverage with the current study could also be due to the fact that they sampled more taxa than was done in this study and unlike the current study, they included some genera with more than one species representatives.

There are several important community indices used in ecological studies that depend on abundance of taxa in the community. These include the maturity index (Bongers, 1990), enrichment, structure and basal indices (Ferris, Bongers, & De Goede, 2001). Abundance

has been used widely to compare processes such as community food web dynamics, stability and response to mineral and mechanical perturbations. In other words, almost all indices of community structure rely on abundance information of soil inhabiting taxa. Therefore, for metabarcoding to be accepted as an alternative to the more laborious, expensive and time-consuming classical morphology-based method of community analysis for the purposes as soil health assessment or food web analysis, it is pertinent that it is able to provide reliable abundance information based on the read frequencies of the recovered taxa. However, no mock community study to date has reported a strong correlation between actual abundance and read frequencies, implying that using read frequency data in the computation of any abundance-based index to assess various characteristics of a community may not reflect its actual state. The same is true of this study, read number showed no correlation with actual abundance. In microbial ecology, however, read frequency information continues to be used even though the read frequency issue has been shown to transcend beyond studies involving metazoans such as nematodes to microbes, particularly in mixed species samples (Amend, Seifert, & Bruns, 2010; Edgar, 2017a). This lack of association between relative abundance and read frequency, therefore remains one of the key limitations of metabarcoding. The quantitative reproducibility of the replicates, however, was shown in this study. Thus, on both the lack of correlation between read frequencies and relative abundance and reproducibility across replicates, this study agrees with (Porazinska *et al.*, 2010c), especially for the NF1-18Sr2b samples.

Cross-talk between samples as seen in this experiment has the potential of causing misrepresentation of the diversity in a sample. This phenomenon has been reported in a number of previous studies (Kircher, Sawyer, & Meyer, 2011; Carew *et al.*, 2013; Nelson *et al.*, 2014; Esling, Lejzerowicz, & Pawlowski, 2015). An instance of how this could happen was seen in the recovery of an otu of *Anatonchus tridentatus* in the NF1-18Sr2b samples when in fact it was from the SSUF04-SSUR22 samples. It was easily detected because the two markers covered different genomic regions. If all the samples were amplified by the same primer sets, then it would have been hard detecting it. Edgar (2016) investigated cross-talk from a total of twelve Illumina runs and developed an algorithm, *uncross* capable of detecting as high as 80% of these spurious reads in samples. Schnell *et al*. (2015) also described a laboratory set-up to help minimize their occurrence. The authors implicated factors such as errors during primer synthesis, PCR errors, and sequencing errors as being some of the causes of cross-talk and suggested using tags with as many base differences from one another as possible.

As observed from the different taxonomy assignment methods, it is best to always combine two or more different approaches in order to get the best assignments. Methods usually employed in analysis pipelines such QIIME (Caporaso *et al.*, 2010), MOTHUR

(Schloss *et al.*, 2009) or USEARCH (Edgar, 2010) depend on the use of curated reference databases such SILVA , UNITE, RDP or PR^2 which all depend on publicly available sequences. For that reason, at any point in time these databases may not be as up-to-date in terms of collection as the NCBI GenBank. Therefore, alternative methods such as blast search against all NCBI database may provide assignments that perhaps these pipeline-based methods might fail to provide. And despite the fact that a number of entries in GenBank may have dubious accuracies in terms of taxonomy, no blast assignment in this study contradicted any of *utax's* that had sufficient support (≥ 0.5).

In summary, for metabarcoding of soil nematodes, there are many reasons for one to favour the NF1-18Sr2b marker as the most suitable both in terms of coverage and ease of access to reference sequences. The issue of non-specificity of this marker, although a problem, can mostly be avoided by extracting nematodes from soil before DNA extraction to make sure most non-targets are excluded. According to Creer *et al*. (2010), the fact that only a very minute proportion of soil volumes are made up of nematode biomass means that nematodes should always be extracted from the sediments first (Creer *et al.*, 2010). Moreover, as stated by Peham *et al*. (2017), direct extraction kits can handle only a fraction of the recommend volumes of soil needed for analysis of nematode communities at the moment. In an ideal situation where a complete reference database of 28S rDNA is available, the D3Af-D3Br may equally be suitable alone or in combination with any of the 18S-based markers. In fact, claims have been made that suggests that the D2-D3 segment may provide a better barcoding marker than the 18S rDNA (Creer *et al.*, 2010). For the COI region, unless the approach involves a cocktail of primers covering the diversity of the entire phylum as suggested by (Prosser *et al.*, 2013), the COI markers will continue to be of very little use in nematode metabarcoding. The 'primer cocktail' approach has not been tested on soil nematodes, and even if it worked, there is still the issue of very limited nematode reference sequences in public databases. Finally, when making recommendations for appraisal and adoption of new barcode marker(s) other than the ones known and used so far, an important consideration that always has to be made is the availability of a comprehensive reference database. It will take tremendous amount of work to develop new reference databases as vast as what exists now for the 18S or 28S rDNA region.

4 A comparison of tools for analysing nematode next generation sequencing data

4.1 Introduction

The use of next generation sequencing for nematode identification has the potential to make community analysis less tedious, fast and inexpensive. This technology makes it possible to generate enormous amounts of sequence data cheaply and within a very short time (Porazinska *et al.*, 2009). With such large amounts of sequence data comes the need for bioinformatics tools that are sophisticated enough to process them. A number of programmes and online resources offer tools and services for analysing high-throughput sequence data (Bik *et al.*, 2012a). Some of these tools perform exclusive tasks such as quality filtering, and trimming, otu picking and chimera detection, while others provide complete data processing pipelines. Quantitative Insights Into Microbial Ecology (QIIME) (Caporaso *et al.*, 2010), MOTHUR (Schloss *et al.*, 2009), metagenomics rapid annotation using subsystems technology (MG-RAST) (Meyer *et al.*, 2008), Ribosomal database project (RDP) (Cole *et al.*, 2014) and USEARCH (Edgar, 2010) are among examples of those that offer complete data processing. All of these tools are fully open-source except USEARCH, which only offers a 32-bit version without charge.

The qiime.org homepage describes QIIME as a bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data generated from next generation sequencing platforms such as 454 pyrosequencers and Illumina platforms. It is developed with the capability to analyse sequences generated by these platforms through to publication-ready standard (Caporaso *et al.*, 2010). QIIME can easily be installed and run locally on the user's computer. MOTHUR, like QIIME is also targeted towards private users, and hence easily installable on a personal computer (Schloss *et al.*, 2009). RDP is mainly an online resource for sequence data analysis. It also offers tools that can be installed and run locally on user's computer. It was initially developed for processing bacterial and archaeal 16S rDNA pyrosequencing data but has since been expanded to include fungal ITS and 28S rDNA. It was also optimised for analysing Illumina data, presumably, to keep up with the then growing popularity of the Illumina sequencing method. Like the aforementioned two, USEARCH (Edgar, 2010) integrates many different algorithms into a single package with substantial amount of documentation and support. MG-RAST is only available as web application where users can upload sequences for automated analyses. Although all these pipelines to some extent appear to be biased towards microbiome analyses, based on their focus on mostly 16S rDNA and ITS, they are still supportive of metazoan analyses provided the user can obtain reference databases to train the algorithms available for microbiome analyses.

A number of previous studies have compared and reviewed some of these pipelines (D'Argenio *et al.*, 2014; Nilakanta *et al.*, 2014; Plummer *et al.*, 2015). In their comparison of QIIME and MG-RAST on 16S rDNA dataset from human gut, D'Argenio *et al*. (2014) found that QIIME performed much better than MG-RAST at assigning taxonomy, particularly at the family rank and lower. Diversity analyses also yielded more accurate results with QIIME. Analyses run time between the two was also significantly different, with QIIME taking approximately two hours to complete the analyses whiles MG-RAST took about 10 days. The comparison of the run time was a somewhat flawed one since, and as admitted by the authors, this difference could in part have had to do with the fact that the two pipelines were not run under the same computing power. Plummer *et al*. (2015), using 16S rDNA from gut microbiome, also expanded this comparison to include the MOTHUR pipeline. Between QIIME and MOTHUR, both of which were run on the same machine, the former took one hour while the latter took ten hours to complete the analyses. The authors concluded that QIIME and MOTHUR were more powerful than MG-RAST in terms of the analysis options available to the user. They also concluded that QIIME was more user friendly than MOTHUR, the latter taking more tweaking to obtain sensible results (Plummer *et al.*, 2015).

Although these studies (D'Argenio *et al.*, 2014; Plummer *et al.*, 2015) provided very helpful comparisons of two of the most widely used analysis pipelines, both were based on microbial 16S rDNA data not metazoan 18S rDNA data. Also, the recent popularity of USEARCH, which currently has over 35,000 users, may warrant its inclusion in such comparisons to assess its performance against the two most recommended pipelines, QIIME and MOTHUR. Additionally, by using nematode 18S rDNA data, this becomes probably the first of such comparisons made using metabarcoding data from metazoans. The objective here was to present a comparison of QIIME, MOTHUR and USEARCH in terms of their overall user-friendliness, accuracy and speed under the same computing power using the same mock community data as in Chapter 3.

Specifically, they were compared based on:

1. Efficacy of their read filtering methods
2. Accuracy of the read clustering algorithms they use by default
3. Reliability of their taxonomy assignments
4. Ease of use including the number of lines of commands to run in order to reach similar outcome
5. And time required to complete analysis, which is defined here as the stage where a biom table is generated and each otu has been assigned taxonomy.

Since the previous comparisons attested to the superiority of MOTHUR and QIIME over MG-RAST, it was not considered necessary to include MG-RAST in this comparison. And

also, the focus here was on pipelines that can be installed and run locally on a PC, not on a server. Although OCTUPUS (Sung *et al.* unpublished) is another such analysis pipeline that has previously been used in studying meiofaunal diversities (Creer *et al.*, 2010; Fonseca *et al.*, 2010), it has not been updated since May 2010 and is hence not considered in the comparison. Moreover, despite many attempts to run the analysis using RDP command line tools, the extremely limited documentation available on its command line tools made it impossible to obtain a meaningful output. Additionally, since the RDP command line pipeline requires aligned reads it would have taken an extremely long time to carry out the analyses with the large number of sequence reads used for this comparison. Furthermore, this chapter will not be reporting on the numerous diversity analyses such as alpha and beta diversity possible with these pipelines. Instead the focus will be on the suitability of these pipelines for metazoan metabarcoding which mainly involves determining the identities of the sequences recovered from each sample.

4.2 Materials and methods

4.2.1 Mock community

The sequences were produced from three replicates of artificially assembled nematode community consisting of 23 different species spanning the two classes of nematodes. In all, 19 different families belonging to six orders within Nematoda were represented. As described in section 3.2.2, the communities were dominated by members of the Order Rhabditida due their numerical dominance in soil environments. Table 3.1 summarises the taxonomic composition of the mock community and the abundance of each genus/species that was included. The data used were the same ones obtained from the analysis involving the NF1 and 18Sr2b primer pairs from Chapter 3.

4.2.2 Data analysis

The raw fastq sequences files were analysed using the three different pipelines, QIIME (v.1.9.1), MOTHUR (v.1.39.1) and USEARCH (v.9.2 and v.8.1). For most of the analysis parameters were set to default wherever possible. For clustering, a universal value of 97% similarity cut-off was used for all three pipelines. For taxonomy assignment, pr2_gb203_version_4.5 18S rDNA sequences (Guillou *et al.*, 2012) were used as reference. In USEARCH, the fasta reference file first had to be formatted into a udb file before being used for taxonomy assignment (Edgar, 2015).

4.2.2.1 QIIME

The analyses within the QIIME pipeline started with assembling the paired reads to generate merged reads using the fast-join method (Aronesty, 2011) (Table 4.1) with the command, *join_paired_ends.py*. The merged reads were then filtered applying a quality threshold of 29 using *split_libaries.py* command. This ensured that reads with quality scores of less than 30 were discarded. All the quality-checked merged reads were then concatenated into a single fasta file. These sequences were checked for chimeras using uchime (Edgar *et al.*, 2011) and the chimeric sequences were filtered out. Sequences were then clustered using the *pick_otus.py* command at 97% similarity cut-off using the uclust method (Edgar, 2010). From each of the otu clusters, the most abundant read was picked to represent all of the reads using the command *pick_rep_set.py*. The OTU representative sequences were assigned taxonomy based on the uclust approach using the command *assign_taxonomy.py* and the pr2_gb_203_version_4.5 as reference database (Guillou *et al.*, 2012). The taxonomy and otu text files were then used to generate an otu table wherein each otu was shown as a record with the number of reads it has under each sample and its associated taxonomy. This was performed using the

command *make_otu_table.py*. Singletons were removed from the biom table using the command *filter_otus_from_otu_table.py.*

4.2.2.2 MOTHUR

The steps used here follow the protocol described in the MOTHUR tutorial for MiSeq SOP (https://www.mothur.org/wiki/MiSeq_SOP) which was accessed on 25th August 2017. The raw fastq reads were first merged using the *make.contigs* command with default parameter settings. This step uses similar algorithm as the one used in PANDAseq (Masella *et al.*, 2012) (Table 4.1). The merged reads were then screened to remove sequences longer than 400 bp or ones with one or more ambiguous bases via the command *screen.seqs*. Reads that passed this step were then dereplicated using the command *unique.seqs*. The next step was to align the unique sequences against SILVA v.119 aligned 16/18S rDNA reference sequences (Quast *et al.*, 2013) with the command, *align.seqs*. The SILVA database was used because it performed much better compared to a custom database of sequences of nematodes downloaded from NCBI and aligned using MAFFT (Multiple Alignment using Fast Fourier Transform) (Katoh *et al.*, 2002). This alignment step in the MOTHUR analysis was used to determine the length of the region where the reads overlap the reference sequences including gaps. The sections of the sequence reads outside the overlaps were screened out. Gaps in the alignment were then removed using the *filter.seqs* command, thus reducing the number of positions to exactly the number of bases in the sequences. Using the command *unique.seqs*, the gap-free sequences were again dereplicated to ensure that redundant sequences that may have resulted after trimming are unified with their identical matches [the trimming may have caused sequences that were originally unique to become identical, if the trimmed region was where the reads differed]. The unique sequences were grouped into clusters of sequences not more than three nucleotides different from one another using the command, *pre.cluster*. Chimeric sequences were identified using the uchime algorithm (Edgar *et al.*, 2011) as implemented in the command *chimera.uchime*. The chimeric sequences were subsequently removed using the command *remove.seqs*. Using the pr2_gb203_version_4.5 18S rDNA sequences as reference database, the chimera-filtered reads were assigned taxonomy using the default method employed in MOTHUR (Wang *et al.*, 2007). This method computes the probability that a query sequence is within a reference sequence based on the k-mers that constitute it. While k can be set to a value of the user's choosing, the default is 10. The command used for this step is *classify.seqs*. Thereafter, the command *dist.seqs* was used to generate a distance matrix which was implemented in the clustering command, *cluster* which grouped reads into otus of 97% similarity cut-off. The clustering algorithm used was OptiClust, the default method

MOTHUR (Westcott & Schloss, 2017). With the command *make.shared*, the number of sequences within each otu was determined. And with this information singletons could be removed from the 'shared' file using the command, *remove.rare.* The taxonomies of the otus were determined using the command *classify.otu*. Following this, a biom file was created using the command *make.biom*. Finally, to generate a fasta file of the otu representative sequences, the command *get.oturep* was used.

4.2.2.3 USEARCH

Within this pipeline, the paired reads were first merged based on the sequence of the overlapping region. The command, *fastq_mergepairs* which performs the assembling also concatenates all merged sequences of the different samples into one fastq file (Edgar & Flyvbjerg, 2015). Mismatches of up to 10 bp were allowed within the overlapping regions of the paired reads to be aligned. The average read length after joining paired reads was 360 bp. The merged reads were then filtered using the command, *fastq_filter* which removes any read with more than one base pair expected error (EE) (Edgar & Flyvbjerg, 2015). The filtered reads were then dereplicated using the command *fastq_uniques*. Using a similarity cut-off of 97% sequence similarity, the unique sequences were clustered using the command *cluster_otus* which implements the uparse algorithm (Edgar, 2013). This clustering command also finds and removes chimeras. A standalone command for removing chimeras exists that can be used but this extra chimera removal step is strongly cautioned against (Edgar, 2017b). Taxonomy assignment was based on the *utax* algorithm using the PR^2 reference database for 18S rDNA for eukaryote version 4.5 (Guillou *et al.*, 2012). Finally, the command *usearch_global* was used to generate an otu table, similar to the one produced by QIIME. With exception of this last command which was executed using USEARCH v.8.1, all the previous commands used USEARCH v. 9.2.

Table 4.1. Summary of the different methods/algorithms used in the three pipelines, QIIME, MOTHUR and USEARCH.

Step	QIIME	MOTHUR	USEARCH
1. Merging	fast-join	PANDAseq	USEARCH fastq_mergepairs
2. Quality filtering	Minimum quality score	Minimum quality score	Maximum expected error
3. Chimera removal	uchime	uchime	uparse
4. Clustering	uclust	OptiClust	uparse
5. Taxonomy	uclust	wang	utax

4.2.3 Taxonomic coverage

Taxonomic coverage of the three pipelines was based on the proportion of the sampled taxa they each were able to recover from the data. To avoid losing any diversity, singletons were kept. For a taxon to be classified as detected, it had to be found in at least one of the three replicates. Although taxonomic coverage was based on the assignment methods employed by each of these pipelines, independent blast searches were performed on the representative sequences before singletons were removed in order to confirm the identities of those sequences whose taxonomies were inconclusive.

4.3 Results

4.3.1 Paired-read merging

The first step in all the analysis pipelines was merging the paired reads into sequences that overlap. For all three pipelines a mismatch of 15 bp within the overlapping region and a minimum overlap of 150 bp were allowed. With these two parameters, the percentage of the total reads that were successfully merged was comparable between all three pipelines: 57.4% for QIIME, 56.3% for USEARCH and 57.1% for MOTHUR. Similarly, average lengths of merged reads were similar across all the three pipelines, 364 bp for both QIIME and MOTHUR and 361 bp for USEARCH.

4.3.2 Quality filtering

This step was carried out to remove all poor-quality sequences. Within QIIME, this resulted in the removal of 1,592,033 sequences from the merged reads. For USEARCH, a total of 155,624 reads were discarded while within MOTHUR 1,422,211 sequences were removed. With both QIIME and USEARCH, the sequence filtering step involved just one command, *split_libraries.py* and *fastq_filter*, respectively. With MOTHUR, this was done in two steps using the commands, *trim.seqs* and *screen.seqs*. It is worth noting that the filtering method used with USEARCH discards reads based on a maximum expected error threshold, whereas with QIIME and MOTHUR, reads were discarded if they had average quality scores of less than 30.

4.3.3 Chimeric reads

There were some similarities between the chimera filtering approach of QIIME and MOTHUR. Two steps were involved, one that identifies chimeras and one that filters them out. Chimera removal in USEARCH is integrated in the otu clustering step. In QIIME chimeric sequences were picked from sequence reads; in MOTHUR, from preclustered sequences; and in USEARCH, from the unique sequences. Numerically, 746,671 individual reads were identified as chimeric and removed with QIIME; 74,048 from a total of 105,256 preclustered reads with MOTHUR and 6557 unique reads with USEARCH.

4.3.3.1 Clustering

The number of otus generated during clustering were 5,442 with QIIME, 2,244 with MOTHUR and 161 with USEARCH at 97% cut-off. Without excluding singletons, QIIME generated 11,513 otus, whereas MOTHUR and USEARCH generated 4,588 and 400 otus respectively. Compared to the expected diversity, the number of otus produced with the QIIME pipeline was orders of magnitude higher even without singletons. For all three

pipelines proportions of the otus that emerged as singletons were similar (Figure 4.1A). The number of otus produced by USEARCH, however, was relatively closer to the expected diversity.

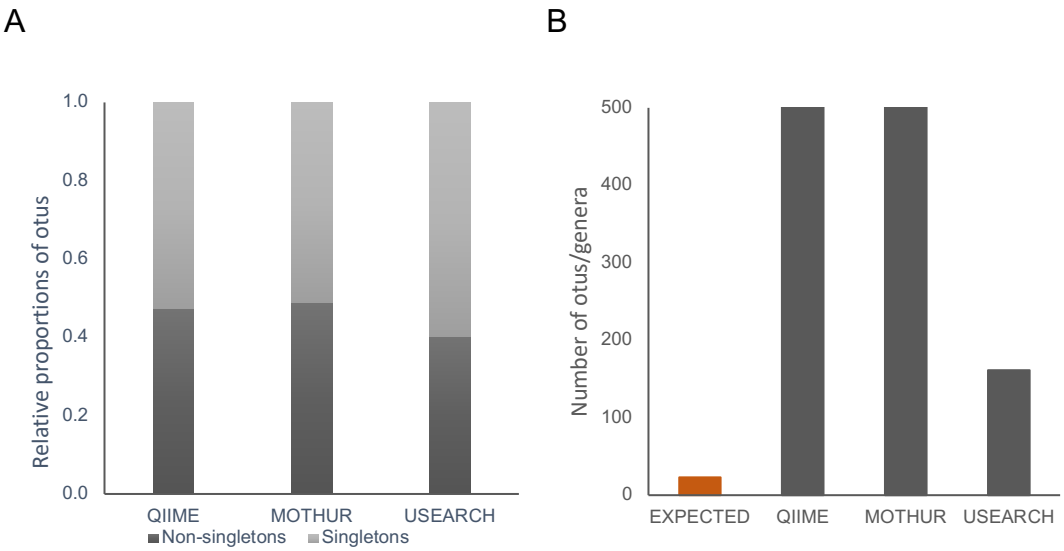

Figure 4.1. (A) Proportion of singleton and non-singleton otus for each of the three pipelines. (B) Number of otus generated by each of QIIME, MOTHUR and USEARCH compared to the actual species diversity in the mock community. Bars representing QIIME and MOTHUR are out of scale and so have been shortened to fit in the graph.

4.3.4 Taxonomy assignment

The results of the taxonomy assignments highlighted more a limitation of the reference database used than the limitations of taxonomy assignment algorithms associated with these pipelines. A number of the entries in the reference sequence files had taxonomies that were not annotated to the order level making their use for assigning sequences to genus or species level impossible. As discussed in Chapter 3, this is not peculiar to just the PR^2 database releases but was observed with some of entries of the SILVA releases that were examined as well. Because of this, the taxonomy assignment of all three pipelines had to be cross-checked to ensure they were assigned to the correct genus or species if possible. Blast search with the NCBI database provided more accurate assignments, although this was often slow.

4.3.5 Taxonomic coverage

All three pipelines recovered all of the sample taxa in the analyses. Some of the otus in both QIIME and MOTHUR matched nematode taxa that were not in the mock community. Upon closer examination of their first few blast search matches, these otus appeared to be chimeras because of the taxonomic distances between their top few blast hits.

4.3.6 Ease of use

All three pipelines offer significant documentation with comprehensive guidance on steps to perform typical microbiome data. The QIIME script page, http://qiime.org/scripts/ has detailed guides and examples on how to run each script. There are also tutorials on how to follow a standard operating procedure for data analyses. Similarly, both MOTHUR and USEARCH offer detailed guides on how to install and run their various commands. MOTHUR and USEARCH were the easier to install on a computer running linux, requiring only the user ensure that the executable programmes were in the path of the system. With both of these, the standalone executable file could be enough to run every command, although options are available in MOTHUR to use other programmes such as uchime for chimera detection. And most importantly, in addition to being available for Linux and macOS, they are both capable of running natively on Windows operation system. QIIME, however, can only be run on Windows through a third-party application. Also, third-party tools such as fast-join need to be installed alongside QIIME before some commands can be used. In terms of possibility of streamlining the analyses, USEARCH was the easiest to streamline, because all analyses involved five commands. MOTHUR presented the highest difficulty in terms of ease of streamlining the entire analyses due to the number of commands that needed to be executed and large memory space needed for some of the computations.

4.3.7 Duration of analyses

Analysis run time varied significantly between the three pipelines (Figure 4.2). It took QIIME in total 37 minutes to complete the analysis of the ~10GB of data used, all through to the biom table stage. It took MOTHUR ~7 hours while USEARCH took 13 minutes. The timing was based on how long it took for all the commands used to finish running on a Lenovo ThinkPad T430 having 8 GB memory, two-core processor with base frequency of 2.5GHz and maximum turbo frequency of 3.1 GHz. In both QIIME and USEARCH, the commands were listed in bash executables files and run as a workflow; and in MOTHUR, they were written in a batch file (https://www.mothur.org/wiki/Batch_mode) and run with a single command. With MOTHUR, the workflow failed several times. In fact, due to the number of commands that had to be run and the memory and storage requirements, there were a couple of instances where the analysis would terminate before completion. Hence, the total analysis run time was not based on the time taken to complete the workflow, but rather, the sum of durations for each command to run to completion. In fact, MOTHUR does well to report these durations after each completed command.

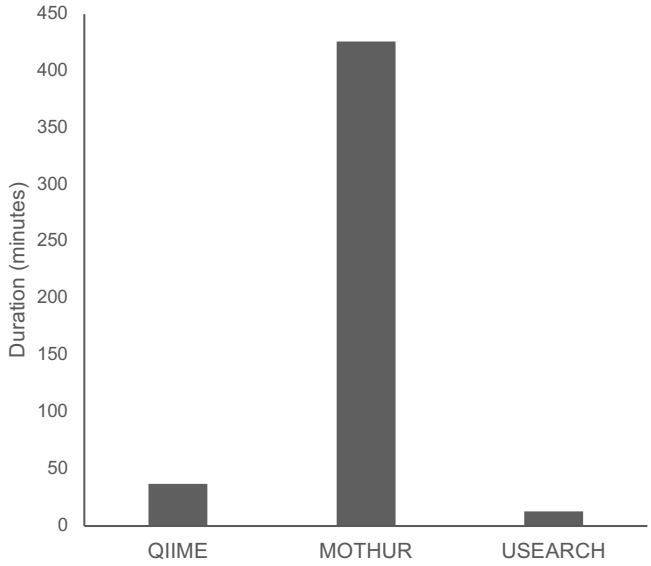

Figure 4.2. Time taken for each pipeline to complete the analysis in minutes. Analysis run on a PC with the specifications: 8 GB memory, two-core processor with base frequency of 2.5GHz and maximum turbo frequency of 3.1 GHz

4.4 Discussion

In the comparison of QIIME, MOTHUR and USEARCH pipelines on nematode NGS data, USEARCH was the best pipeline, particularly in terms of which clustering algorithm gave the most realistic diversity in the samples. In terms of ease of use and speed, it again outdid the other pipelines. And none of the steps where USEARCH emerged superior were at the expense of any other step in the analysis. The initial processing step for each pipeline was similar, except in the number of reads discarded during the filtering step; there were significantly fewer reads discarded by the USEARCH pipeline compared to QIIME and MOTHUR. One reason for this could be the filtering method the USEARCH pipeline employs which is different from the average Q-score-based approach used with the other two pipelines (Edgar & Flyvbjerg, 2015). USEARCH filtering is based on the expected error value of a read and this was set to not exceed one in this experiment (Edgar & Flyvbjerg, 2015). By this method it is possible for reads with an average Q score of less than 30 to be kept even though these would fail to pass the Q30 threshold set for QIIME and MOTHUR. This difference in approach to filtering sequences could be the reason for such an enormous difference in the number of discarded reads.

With regards to chimera removal, the observations made were similar to previous reports (Edgar, 2013), where, for both QIIME and MOTHUR, reads discarded as chimeric were significantly higher than the number detected and discarded by USEARCH. Many of these could have been due to false positive detections which has been reported to be a potential problem with the uchime algorithm (Edgar *et al.*, 2011) that was used for both QIIME and MOTHUR analysis. Other chimera detection options available within QIIME are the blast fragment method (Altschul *et al.*, 1990), chimera slayer (Haas *et al.*, 2011), and within MOTHUR, vsearch (Rognes *et al.*, 2016) and chimera slayer (Haas *et al.*, 2011) can also be used. The choice of the uchime method for both QIIME and MOTHUR was to facilitate their comparison with uparse within USEARCH.

Despite the similarity cut-off percentage being set to the same level across all three pipelines, the number of otus generated between them varied significantly. If the number of otus is to be considered as an estimation of biological diversity, QIIME and MOTHUR would be predicting diversity 500 and 200 times more than what it actually is, respectively. Similar observations were made by Bachy *et al.* (2013) on protists when they compared the diversities assessed through morphological, sanger sequencing and pyrosequencing methods. Using *Escherichia coli*, Kunin *et al.* (2010) also highlighted this issue in prokaryotes and suggested high stringency in filtering out erroneous reads as a means of minimising this diversity inflation. They found that the pyrosequencing data overestimated diversity, in some cases by several orders of magnitude compared to the morphological method. By discarding singletons, the otu numbers dropped down by approximately half.

The number of otus generated with the uparse algorithm of USEARCH corroborates the findings of (Edgar, 2013), which demonstrated that otus numbers from QIIME or MOTHUR often exceed actual diversity by hundreds to thousands. It is important to mention, however, that the closeness of otu number to actual diversity in a sample is not always the best measure of the efficacy of a clustering method (Edgar, 2013). This is because unless this 'good' diversity estimation is combined with very good taxonomic coverage, especially for metabarcoding-focussed studies such as this one, fewer otus is not necessarily the best outcome. USEARCH was able to provide otu numbers closest to the actual diversity compared to the other pipelines while at the same time achieving taxonomic coverage comparable to these other pipelines.

In metabarcoding, the main goal is to identify individuals in a sample based on the sequence of a specific region of their DNA (Taberlet *et al.*, 2012), thus making taxonomy assignment one of the most important steps in the analysis process. This step involves matching query sequences, usually otu representatives, against a reference database of sequences with curated taxonomies. Therefore, unless the query sequence matches a sequence in the reference database, the sequence cannot be assigned taxonomy. At the moment, only the SILVA (Quast *et al.*, 2013) and PR^2 (Guillou *et al.*, 2012) databases offer 18S rDNA sequences with curated taxonomies formatted to work within most analysis pipelines. As discussed earlier, a few of the database entries have incomplete taxonomies, with most classified only to ranks above order level. In a case where the query sequence matches a sequence with an incomplete taxonomy, the result is an assignment that, while correct, makes further pipeline-independent examination of the sequence [such as blast search] compulsory. Many instances of this were observed in each of the pipelines, prompting the use of blast search as a means of cross-checking the assignments. Alternatively, a local reference database can be generated with the user's own sequences and their full taxonomies, in which case assignment is constrained by the taxonomic coverage of this database. Otherwise, NCBI blast search also provides a good check for these incompletely assigned otus. The result from blast search showed that where the taxonomy in the reference database was complete, all the three pipelines assigned the otus to the correct taxa. For a long-term solution to this issue, it may be a good idea for seasoned nematode taxonomist to come together and build a well curated database.

QIIME and MOTHUR are currently the most popular tools for high throughput sequence analysis. So far, all comparisons involving these two pipelines have not provided conclusive verdict as to which is better (Nilakanta *et al.*, 2014; Plummer *et al.*, 2015). They both have comprehensive sets of functionalities for carrying out various (otu)-based methods of describing and comparing communities and documentation on these functionalities. However, the biggest difference according to Nilakanta *et al*. (2014)

between these two pipelines was the analysis run time, with QIIME taking 1 hr and MOTHUR taking 10 hrs. A similar observation was made in this study with QIIME taking 10 times less time than MOTHUR to finish. USEARCH took an even shorter time to complete, about three times less than QIIME. And like QIIME and MOTHUR, USEARCH offers similar suites of functionalities and a plethora of documentations to go with these functionalities. MOTHUR combines features from pre-existing software reconstituted and written in C and C++ programming language for speed and robustness (Schloss *et al.*, 2009). There is no denying that it provides a faster implementation of each of these programmes on their own but compared to the other pipelines evaluated in this study, it is the slowest.

The MOTHUR pipeline involves a step where reads are aligned using a template alignment supplied by the user. This is so that sequences can be trimmed to a common region in the alignment to improve both the speed and robustness of otu assignment (Schloss, 2013). The composition and the quality of alignments of the sequences in this template database has a huge impact on the clustering step further down the analysis. Initially a template file generated from 38 nematode sequences downloaded from NCBI and aligned using MAFFT (Katoh *et al.*, 2002) was used. Using this alignment caused over 80% of the reads to be assigned to one species, *Steinernema carpocapsae* leaving more than 50% of the sampled taxa missing. It was when this step was repeated using a subset of 1000 aligned 16S/18S rDNA sequences from the SILVA database as template that a much better otu assignment was obtained and all sampled taxa were recovered. A subset was used because using the complete reference database generated an alignment file of over 240GB causing the programmeto terminate with an error due to lack of space. And even without the storage space limitations, the process would have taken a very long time to complete.

In conclusion, if all three pipelines were to be scored, based on diversity approximation by number of otus, speed and ease of use, the ranking would be USEARCH, QIIME and then MOTHUR. Automation was extremely easy with USEARCH due to fact that the whole analysis involved significantly fewer command lines than the other pipelines. Taxonomy assignment was not any different between the pipelines based on the dataset used here. Although clustering accuracy remains a difficult process to assess (Edgar, 2013), it was obvious uparse was the fastest clustering algorithm compared to uclust and OptiClust used in QIIME and MOTHUR, respectively. To summarise, USEARCH may be the best of the three for nematode metabarcoding analysis, at least, based on the results of this study.

5 In-depth analysis of the relationship between sequence reads and nematode abundance

5.1 Introduction

Speed, cost-effectiveness, accuracy and the ability to simultaneously identify a wide diversity of taxa are all features of metabarcoding that appeal to taxonomists interested in biodiversity monitoring. Metabarcoding has, over the years, helped unravel great insights into microbial and metazoan communities in both aquatic and terrestrial environments (Sogin *et al.*, 2006; Buée *et al.*, 2009; Fonseca *et al.*, 2010; Porazinska *et al.*, 2010b, 2012a; Bik *et al.*, 2012b; Lindeque *et al.*, 2013). Its application in nematology may date as far back as almost a decade ago through the work of Porazinska *et al*. (2009), which showcased the tremendous potential of this method as a means of identifying nematode taxa within soil communities. The study used an artificially assembled nematode community to test the suitability of the approach for nematode identification. The authors used two pairs of primers that anneal to sites within the highly conserved 18S and 28S rDNA to amplify two regions that were highly successful in recovering and resolving a significant number of the sampled taxa.

It is a common, but rarely tested assumption that read abundance correlates with taxon abundance in amplicon sequencing (Amend *et al.*, 2010; Edgar, 2017a). While the ability of metabarcoding to detect and accurately identify species in bulk samples has been proven across most taxonomic groups (Zhan & MacIsaac, 2015), most mock community metabarcoding studies have reported poor correlation between relative abundance of taxa and their read frequencies. In fact Porazinska *et al*. (2009) acknowledged this limitation in their study. Studies that have attempted to relate abundance/proportion of taxa with their sequence read numbers/proportions found no correlation or advised that caution be exercised when attempting such assumption (Amend *et al.*, 2010; Yu *et al.*, 2012; Deagle *et al.*, 2013). This is a key concern because of the role abundance data play in community assessment of organisms, especially nematodes and their use as biological indicators of soil status.

Theoretically, any factor that influences or introduces bias in the number of DNA templates for each individual organism in a sample prior to PCR or causes bias in the number of copies of each amplicon generated for each individual after PCR can be a culprit to the poor correlation between read numbers and actual abundance. These include innate factors such as variability in cell numbers and differences in the number of copies of the marker gene between species. Nematodes were previously considered to be eutelic (Martini, 1909; van Cleave, 1932; Malakhov, 1994), meaning individuals of the same life stage have identical number of cells across all species. The subsequent disproof of this idea by Cunha *et al*. (1999), who showed that most species have variable cell

numbers in their tissues particularly their epidermis, meant nematodes can be subject to those bias associated with cell number difference depending their life stage.

Other factors responsible for failure of read numbers to correlate with actual abundance of taxa include DNA extraction and PCR biases (Amend *et al.*, 2010). In nematodes, there is evidence that DNA of some taxa may be more easily extractable than that of other taxa, mainly due to differences in their cuticle thickness (Waeyenberge *et al.*, 2017). For example, members of the super-family Criconematoidea are characterised by thick strongly annulated cuticles and were shown to yield a very low amount of DNA compared to other taxa. Universal primers employed in metabarcoding for amplifying targeted DNA regions often have positions of base mismatch against certain taxa within the group they are designed for. Such primer mismatches have been shown to lead to significant amplification bias in multi-template PCR, where the templates with perfect matches are preferentially amplified (Sipos *et al.*, 2007). The extent of this becomes even escalated at higher annealing temperatures. And as is mostly the case in NGS analysis, the additional amplification steps such as the bridge amplification on Illumina platforms further magnify this bias (Pinto & Raskin, 2012). This, in some cases, has been avoided by the use of primers with ambiguous bases at known mismatch positions.

One of the first ever attempts at resolving the read abundance bias focussed mainly on correcting the gene copy number disparity (Kembel *et al.*, 2012). According to the authors, a much better correlation between read abundance and species abundance of bacterial species could be achieved by predicting and incorporating information about the 16S rRNA gene copy numbers. Edgar (2017) created a solution through UNBIAS; an algorithm that corrects for biases attributable to gene copy number variations and primer mismatches inside the USEARCH ((Edgar, 2010) pipeline. He acknowledged that this algorithm performs with very limited success and requires a number of conditions to be met. For example, it works best when the number of mismatches within the primer annealing site is known. It also relies on prior knowledge of the number of copies of targeted region present in the genome, an undertaking possible only if the entire genome of the species involved has been sequenced.

Previous studies have suggested that in monospecific samples, read numbers, to some extent, correlate with actual abundance (Amend *et al.*, 2010). In multi-specific samples, as ones in Chapter 3 and in Porazinska *et al*. (2009) inherent differences between species [such as number of gene copies and number of cells] skew the relationships between read numbers and actual abundances of the community. The findings from Chapter 3 showed that the distribution of reads remained consistent across replicates, which means that even in multi-specific samples, each of the sampled taxa produced the same reproducible read abundance. Based on this information, it can be hypothesised that the disparity

between relative read numbers of taxa and their relative abundance in a sample increases with increasing complexity; and that identical communities, regardless of how complex they might be, are quantitatively similar in terms of read distribution. Therefore, this study is aimed at improving our understanding of how species complexities within samples and varying taxa combinations influence read numbers.

It was also set up:

1. To understand how differences in biomass translate into the read numbers between two noticeably different life stages of the same species
2. To determine how read orientation and quality filtering influence read proportions of nematode taxa in the samples at the analysis steps

To achieve these, triplicates of fifteen mock samples were set up each with specific set of taxa and number of individuals. The samples, their species composition, the number of each species included and the specific question(s) they were set up to answer are summarised in Table 5.1.

5.2 Materials and methods

5.2.1 Mock Communities

Fifteen mock community samples with various compositions and abundances of nematode species were set up, each in triplicates. Because of time and resource constraints, it was not possible to use more replicates than the three that were used. Table 5.1 summarises the taxonomic composition of the samples and what questions they were set up to address.

5.2.2 DNA extraction and amplification

For each sample, the specimens were transferred into 20 µl of distilled water and frozen at -20ºC until they were ready for DNA extraction. Using the Qiagen extraction kit and the procedure described in section 3.2.4, total DNA was eluted into 200 µl of AE buffer. The targeted barcode region, located within the 18S rDNA region was amplified using the nematode universal primer pairs, NF1 and 18Sr2b both with Nextera adapters attached to their 5' ends. The same PCR programme used in Chapter 3 was employed here.

5.2.3 Library preparation

Following a quality control step involving the use of gel electrophoresis for 5 µl of the PCR products from each sample and a clean-up step, index PCR was carried out as described in Chapter 3. The products were cleaned up again before being multiplexed according to the concentrations of each sample. A third clean-up was performed on the multiplexed sample before checking the size of the amplicons using the Agilent TapeStation (Agilent Technologies, Germany). The sample was mixed with 10% PhiX, denatured before being sequenced on the Illumina MiSeq using the paired-end approach at Fera Science Ltd.

Table 5.1. Taxa constituting each of the samples used, number of individuals (N) included, their sources and the purpose of the sample

Sample	Species	N	Purpose	Source
MC1	*Pratylenchus penetrans*	3	1. Reproducibility of read distribution 2. Quality filtering effect on read distribution	HAU
	Mylonchulus sp.	1		ES
	Oscheuis tipulae	7		CGC
	Oscheuis myriophila	2		CGC
	Trichodorus primitivus	4		Fera
	Trichodorus cylindricus	1		Fera
	Acrobeloides butschlii	4		CGC
	Panagrolaimus rigidus	1		CGC
	Panagrellus redivivus	3		CGC
	Rhabditis brassicae	2		CGC
	Rhabditis sp.	9		HAU
	Meloidogyne naasi	6		HAU
Pp2	*Pratylenchus penetrans*	3	Amplification bias and which taxa are favoured	HAU
PpPc3	*Pratylenchus penetrans*	3		HAU
	Pratylenchus crenatus	3		HAU
PpPcR4	*Pratylenchus penetrans*	3		HAU
	Pratylenchus crenatus	3		HAU
	Rhabditis sp.	1		ES
PpPcRPr5	*Pratylenchus penetrans*	3		HAU
	Pratylenchus crenatus	3		HAU
	Rhabditis brassicae	1		CGC
	Panagrellus redivivus	1		CGC
Ab6	*Acrobeloides butschlii*	1	Read abundance in mono-specific samples	CGC
Ab7	*Acrobeloides butschlii*	2		CGC
Ab8	*Acrobeloides butschlii*	4		CGC
Ab9	*Acrobeloides butschlii*	8		CGC
Ab10	*Acrobeloides butschlii*	12		CGC
Ab11	*Acrobeloides butschlii*	16		CGC
L12	*Longidorus* (female)	1	Life stage (biomass) vs read abundance	ES
L13	*Longidorus* (juvenile)	1		ES
X14	*Xiphinema* (female)	1		ES
X15	*Xiphinema* (juvenile)	1		ES

Sources of the sampled taxa included ones obtained from cultures kept in Harper Adams University (HAU), those obtained from Fera as well as strains ordered from the Caenorhabditis Genetic Center (CGC), University of Minnesota. Some were also identified from environmental samples (ES) collected from locations within Harper Adams University.

5.2.4 Bioinformatics

All bioinformatic analyses were carried out using the USEARCH pipeline (Edgar, 2010). The raw reads were first merged using the *fast_mergepairs* command allowing up to 15 mismatches within the overlap between the read pairs. The minimum length of overlap was kept at the default setting of 16 bp. The merged reads were then filtered and reads with expected error (as described in Edgar and Flyvbjerg 2015) of more than one were discarded and the remaining sequences stored in a fasta file. The filtered reads were dereplicated into unique sequences. The unique reads were then clustered into operational taxonomic units (otus) at 97% similarity cut-off using the uparse algorithm. Singleton otus were discarded. Taxonomies were assigned to the otus using the PR^2 database (Guillou *et al.*, 2012) of 18S rDNA sequences as reference. The taxonomy-assigned otus were combined with the merged sequences to create otu tables. Relative read abundance of each taxon was calculated as a proportion of the total reads in a sample.

To determine how less stringent filtering approach affects quantification bias, analysis was performed on each of the unpaired reads. This procedure allowed reads that would normally be discarded for not merging to be retained. For this analysis, data from each read direction were analysed separately. The reads were first truncated to 200 bp at the 3' end before being filtered to remove low quality reads. The reason for truncating the 3' end was to remove the usually low-quality part of the sequence, thus keeping the expected error of the reads low. The maximum expected error (EE) parameter within USEARCH was set to four different values, 1, 2, 4 and 8. The subsequent commands were executed using the same parameter settings as with the merged reads. Subsequently, otu tables were viewed and edited using MS Excel. Figures were also generated within MS Excel.

5.3 Results

5.3.1 MC1 with merged reads

About half of the paired reads successfully merged constituting about 187,243 reads. Only 22% of the merged reads were longer than 250 bp. A significant proportion of the other 78% were even shorter than 50 bp. This was to do with the fact that minimum overlap was kept at default setting of 16. The filtering step, which in addition to removing reads with EE > 1 also removed reads shorter than 250 bp, ensured the elimination of these very short reads. They could also have been removed, had the minimum overlap been set to 50 bp or more during the merging step. The remaining quality-checked reads formed 4,222 unique sequences. At 97% similarity cut-off, the reads were clustered into 45 otus including singletons (31=non-singletons; 14=singletons). After assigning taxonomy to the non-singleton otus, 13 were assigned to nine different species of nematodes while the rest showed up as fungi. Eight of the nine named nematode species were among the twelve nematode species sampled in the mock community. The remaining four of the sampled taxa were unaccounted for by the sequences. These missing taxa were *Acrobeloides*, *Meloidogyne naasi*, *Rhabditis brassicae* and *Rhabditis* sp. Apart from *Mylonchulus* sp. which was recovered from only one of three replicates and *Panagrellus redivivus* which was recovered in two of the three replicates, all the other recovered taxa were present in all triplicates.

Xiphinema sp. was the one taxon in the nine species that were named which was not present in the mock community. This was due to cross-talk between the multiplexed samples as the sequencing run included samples that had *Xiphinema* sp. in them. This phenomenon is described in section 3.3.9.

The distribution of sampled species in terms of relative read abundance across the three replicates was mostly similar (Figure 5.1). They were all strongly correlated (1 vs 2, $P < 0.01$, $rho = 0.87$; 1 vs 3 $P < 0.01$, $rho = 0.89$; 2 vs 3 $P < 0.01$, $rho = 0.87$). However, none of the replicates correlated with the relative distribution of the taxa in the mock community. In all three replicates, over 95% of the reads belonged to only two ‘dominant’ species, *Trichodorus primitivus* and *Pratylenchus penetrans* both of which, combined, constituted only 16% of the sampled taxa in the mock community in terms of abundance. *Trichodorus primitivus* was the more dominant of the two, reaching up to more 80% of reads in one of the replicates and around 70% in the other two. *Trichodorus cylindricus*, its close relative, did not receive such a proportionate amount of bias in its favour. The visual comparison of relative abundance of the replicates as shown in Figure 5.1 reveals no correlation with the actual species abundance. Correlation between the mean relative read abundance of taxa across the three replicates and their relative abundances in the mock community was not significant.

5.3.2 MC1 with only the forward reads

The first step in the analysis using just the forward reads was read truncation, where all reads were capped to 200 bp before filtering the reads by the allowable expected errors. The reads were truncated because of the higher likelihood of poor base quality towards the 3' end of the sequences. Without this, none of the reads passed the expected error (EE) threshold of one. The merging algorithm used in the *fastq_mergepairs* command is designed to take into account the quality of the reads being merged, specifically the overlapping region. Because merging was not performed in this single read analysis, reads that would have been discarded during this step were only discarded during the filtering step. At default setting, the filtering resulted in the removal of between 30-85% of the reads. As the error setting became increasingly less stringent the percent of reads discarded became lower (Table 5.2). Retaining more reads by relaxing the quality threshold did not result in the recovery of more taxa for the forward reads. However, three of the taxa that were not recovered from the assembled reads were recovered in this analysis. This means that analysing only the forward reads resulted in better coverage, while lowering quality threshold had no effect on neither the coverage nor the taxa distribution (Figure 5.2). A few reads from *Xiphinema* sp., *Longidorus* sp. and *Pratylenchus crenatus*, all of which were not sampled in the mock community, were recovered. Another observation was that most of the representative sequences obtained with only the forward reads gave a much higher 'expect value' [this represents the statistical significance of the alignment between the query sequence and the reference sequence (lower is better)] for blast matching than their counterparts from the merged data, suggesting that the merging produced better quality otus

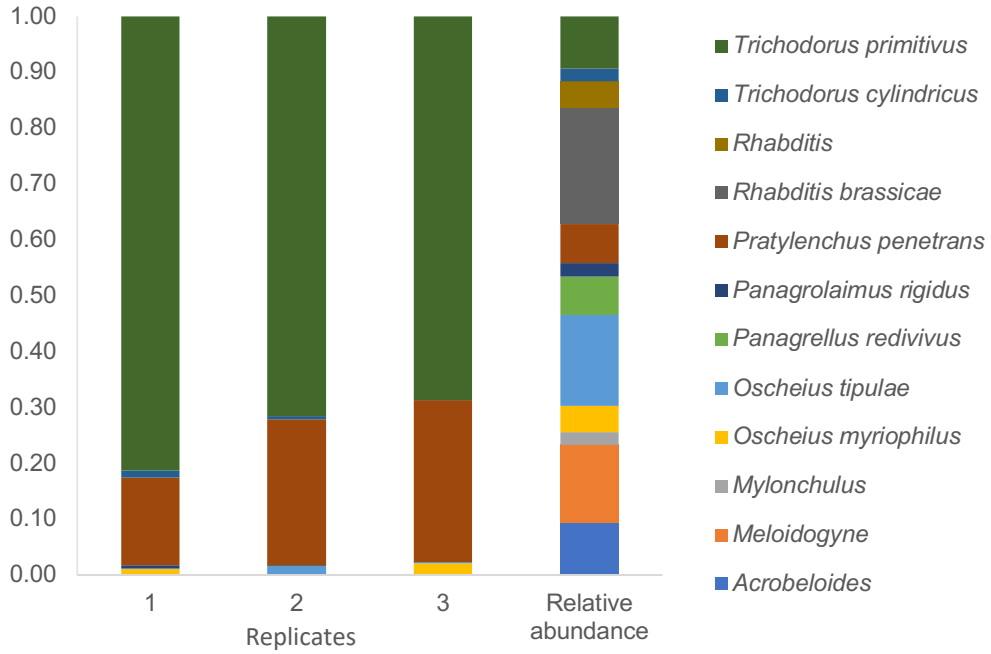

Figure 5.1. Distribution of the reads in the three replicates compared with the taxa distribution in the mock community. The first three bars represent the relative read abundances of the three replicates and the fourth bar represents abundance of taxa in the mock community.

Table 5.2. Percentage of FORWARD reads discarded at various maximum expected error (EE) settings. Max error represents the maximum expected error a read has to have to be retained after filtering. Reads with expected errors higher than max error were discarded.

Replicates	Max EE = 1	Max EE = 2	Max EE = 4	Max EE = 8
1	35.0%	25.8%	14.4%	3.0%
2	29.1%	20.0%	11.2%	2.4%
3	85.1%	82.9%	79.1%	55.0%

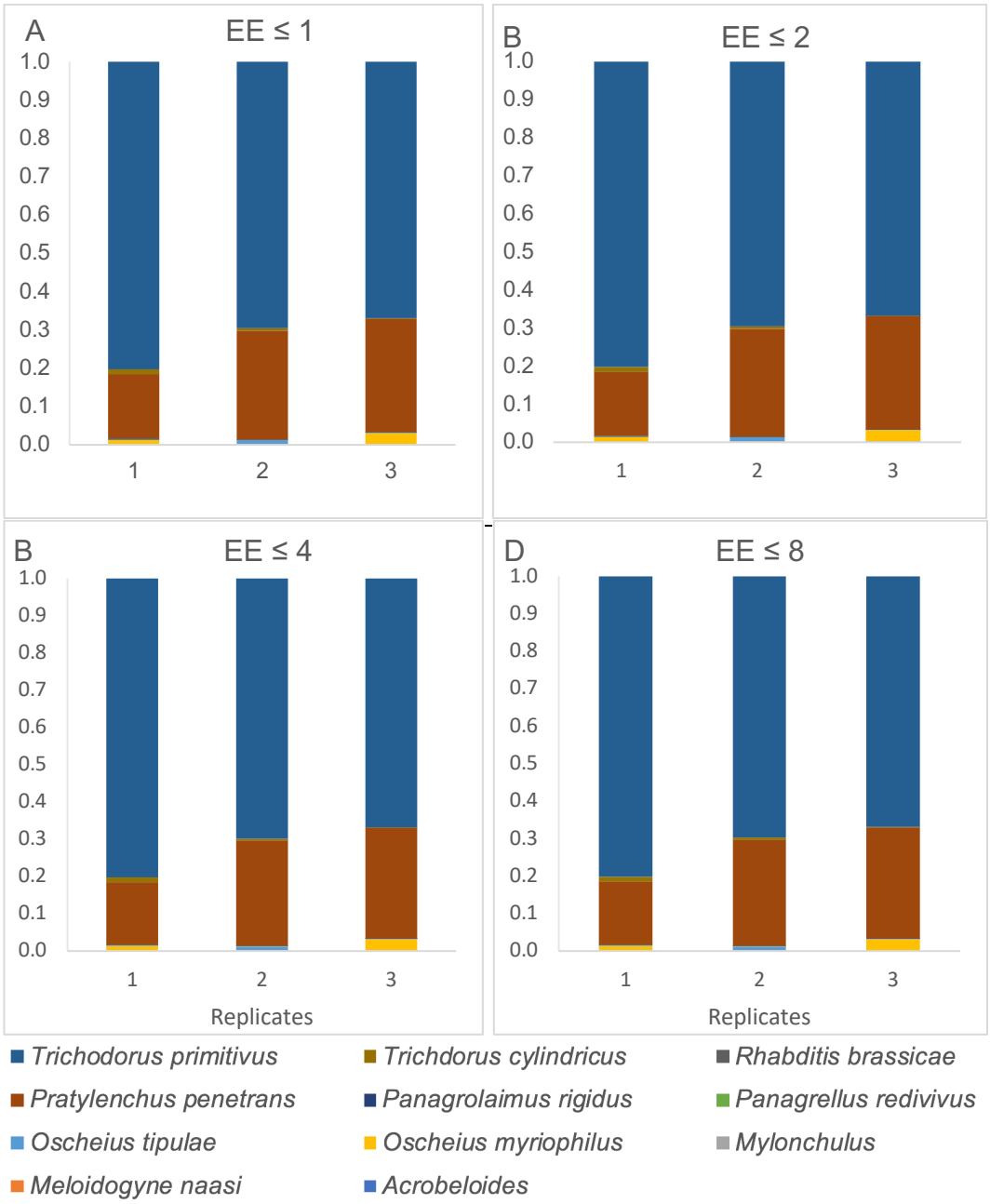

Figure 5.2. Read distribution of taxa under different quality filtering settings when only FORWARD READS were used. The bar plots show that all four maximum expected error values (1, 2, 4 and 8) resulted in the same read distributions across the three replicates.

5.3.3 MC1 with only the reverse reads

By following the same steps as with the forward reads, more reads were retained as the expected error (EE) value increased (Table 5.3). At the default EE threshold of one, up to 97.1% reads were discarded. Increasing the EE threshold to eight retained 22.8% more reads for replicate three (Figure 5.3). Unlike the forward reads, increasing EE threshold values above one resulted in the recovery of *Rhabditis brassicae*, which was missing when the EE parameter was set to one. The distribution of the reads among the species remained almost the same as was with the forward reads and the merged reads. As with the forward reads, quality of otus was lower compared to the merged reads.

Table 5.3. Percentage of REVERSE reads discarded at various maximum expected error (EE) settings. Max error represents the maximum expected error a read has to have to be retained after filtering. Reads with expected errors higher than max error were discarded.

Replicates	Max EE = 1	Max EE = 2	Max EE = 4	Max EE = 8
1	75.6%	51.9%	26.7%	5.2%
2	77.2%	53.7%	27.7%	6.0%
3	97.1%	92.2%	85.8%	74.3%

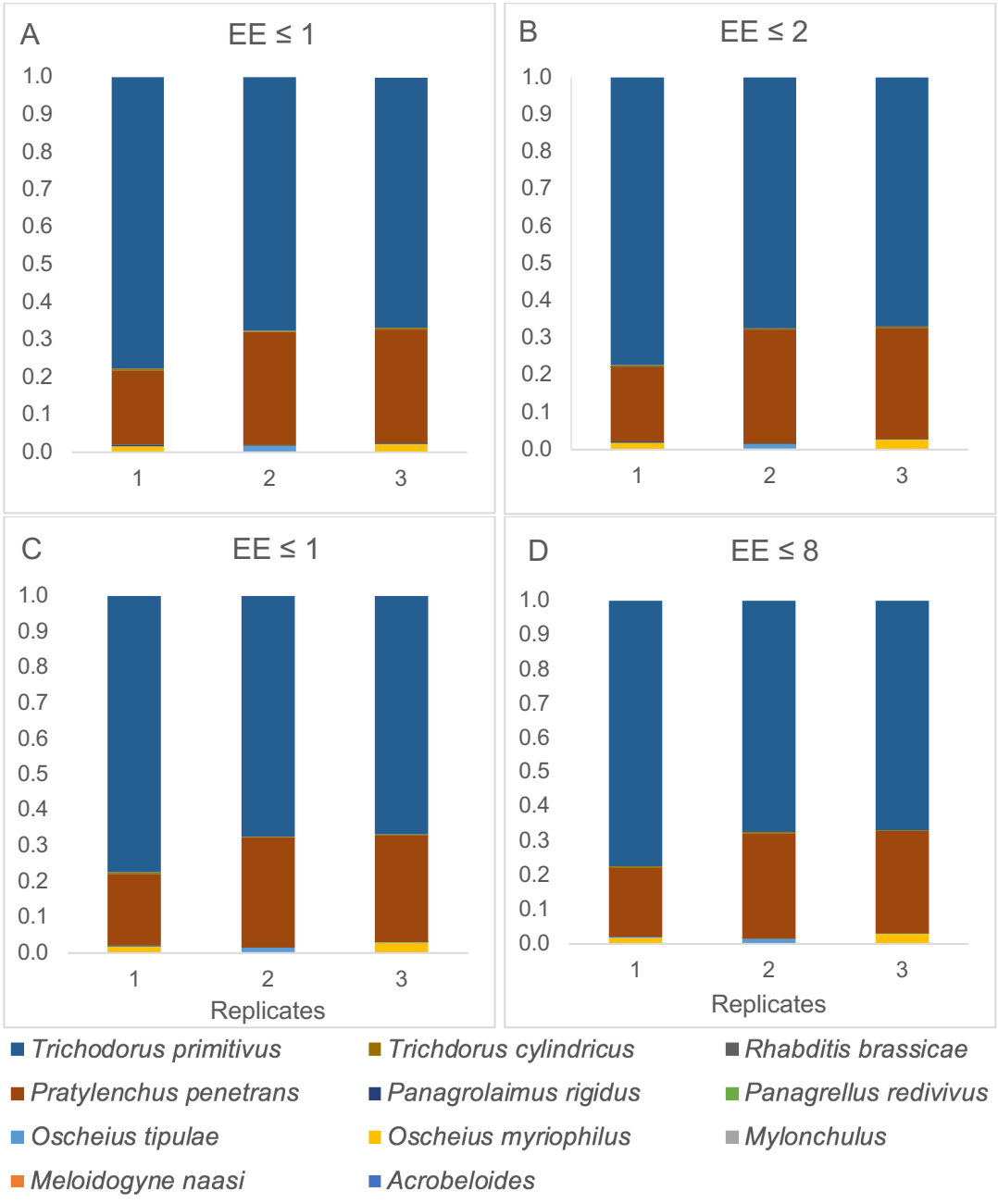

Figure 5.3. Read distribution of taxa under different quality filtering settings when only REVERSE READS were used. The bar plots show that all four maximum expected error values (1, 2, 4 and 8) resulted in the same read distributions across the three replicates. A, B, C, and D represent the results when the maximum expected error values were set to 1, 2, 4 and 8 respectively. Absence of Rhabditis brassicae in pane A not apparent due to low representation in other panes).

5.3.4 Reproducibility

The three replicates were mostly similar in terms of the relative number of reads recovered for each of the sampled taxa. This was evident in all three analysis approaches used; those involving the merged forward and reverse reads as well as forward and reverse reads only. *Trichodorus primitivus* and *Pratylenchus penetrans* remained the taxa with the highest read abundances in all of them. In a monospecific set up involving three replicates with each containing three individuals of *P. penetrans* (sample Pp2), the read numbers generated were significantly disproportionate between the replicates. However, the proportion of the reads of *P. penetrans* relative to all the reads (including the non-nematode reads) recovered in each of the samples were very similar between the replicates.

5.3.5 Preferential amplification of certain taxa

Using simple sets of mock communities with diversities ranging from two to four (PpPc3, PpPcR4 and PpPcRPr5) (Table 5.1), the nature of the bias in read abundance was examined. In PpPc3 which had three individuals each of *P. penetrans* and *P. crenatus,* reads belonging to the former constituted about two-thirds of the reads in two of the replicates and over 90% in one of the replicates (Figure 5.4A). This could suggest a possible preferential amplification of *P. penetrans* over *P. crenatus*. In PpPcR4, a slightly more complex sample with three of *P. penetrans*, three of *P. crenatus* and one of *R. brassicae*, *R. brassicae* formed a significant proportion of the reads, leaving less than 20% of reads to the other two species (Figure 5.4B). Increasing the complexity of the sample through the addition of a single specimen of *Panagrellus redivivus*, resulted in *Panagrellus* dominating the read counts in two out of the three replicates. In the third replicate, only a very few reads belonging to *Panagrellus* were recovered (Figure 5.4C).

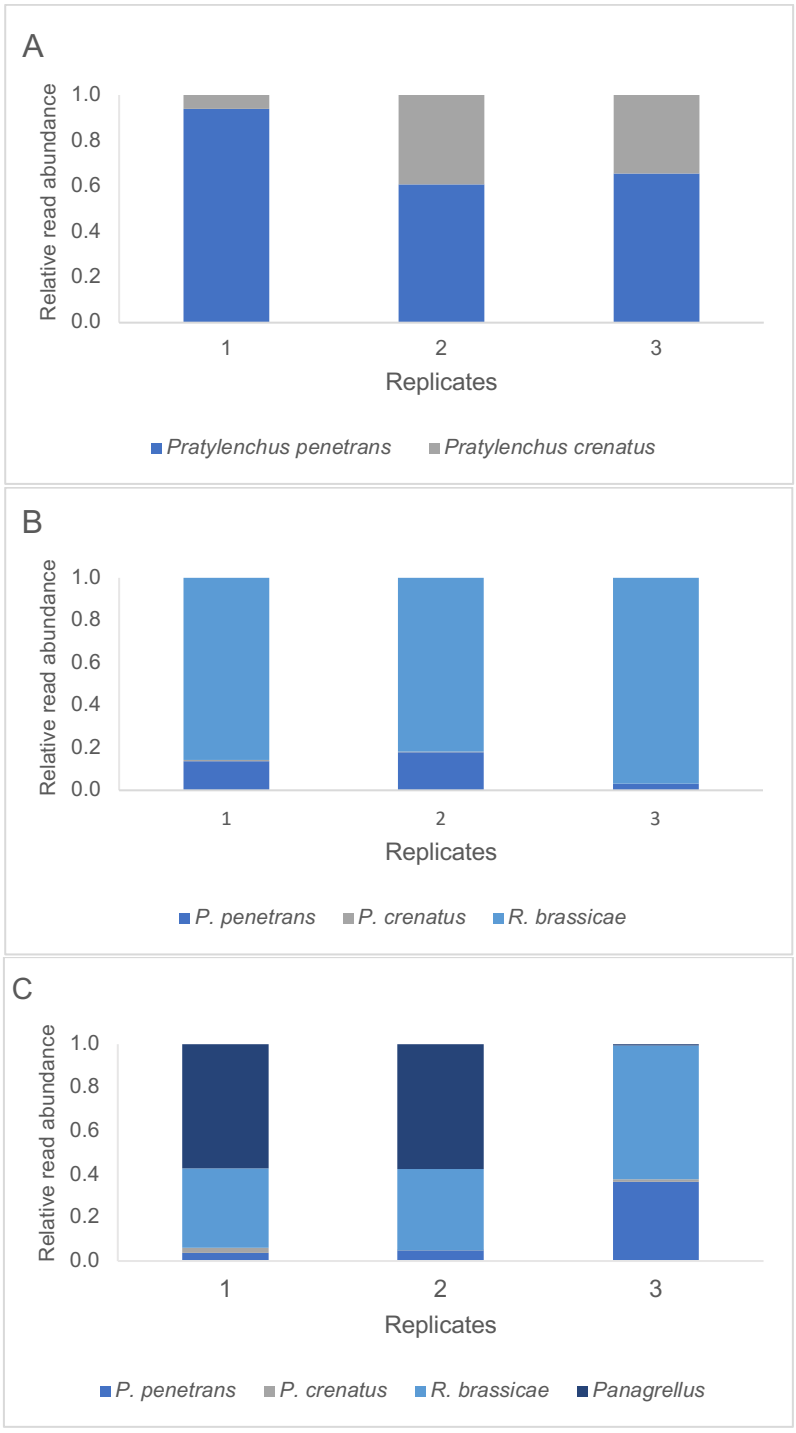

Figure 5.4. Relative read abundances of taxa depicting the bias in read distributions. (A) the distribution of the reads between Pratylenchus penetrans *and* P. crenatus *in a sample containing three individuals of each species. (B) the read distribution for three species,* P. penetrans*,* P. crenatus *and* Rhabditis brassicae *in a sample with three individuals of each of* P. penetrans *and* P. crenatus *and one individual of* R. brassicae*. (C) the distribution of reads for* P. penetrans*,* P. crenatus*,* R. brassicae *and* Panagrellus redivivus *in a sample containing three of* P. penetrans*, three of* P. crenatus*, one* R. brassicae *and one* Pangrellus redivivus*.*

5.3.6 Number of individuals and life stage (biomass) vs read abundance in monospecific samples

The read abundances of *Acrobeloides buetschlii* associated positively with the actual number of individuals (Figure 5.5) but their correlation was not significant (*P*=0.186). The effect of biomass difference tested using two different life stages of *Xiphinema* sp. and *Longidorus* sp. showed that there was no significant difference between the read numbers obtained from female adults and those from the juveniles (Figure 5.6A and B) based on Welch's t-test.

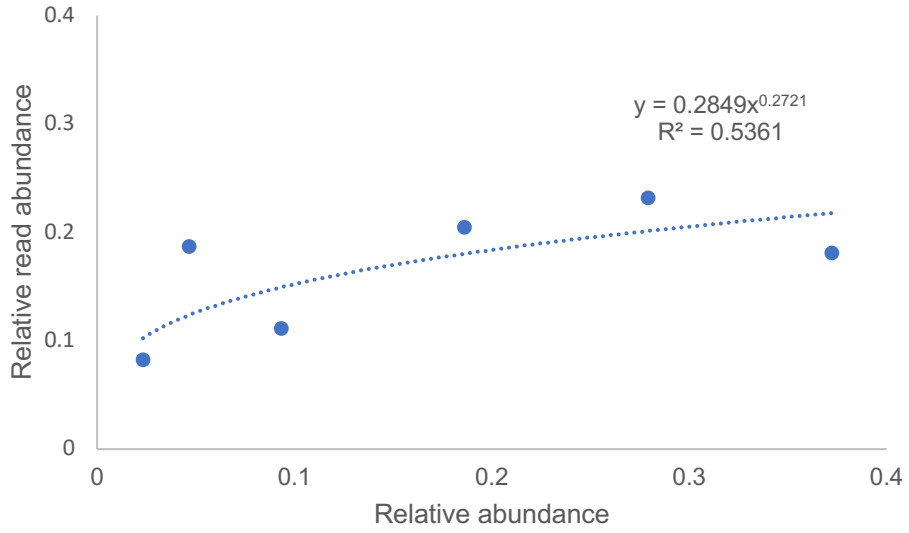

Figure 5.5. Relative read abundance plotted against relative abundance. Equation and R^2 *values are based on exponential regressions of the relative number of reads on the relative number of individuals of* Acrobeloides butschlii *added.*

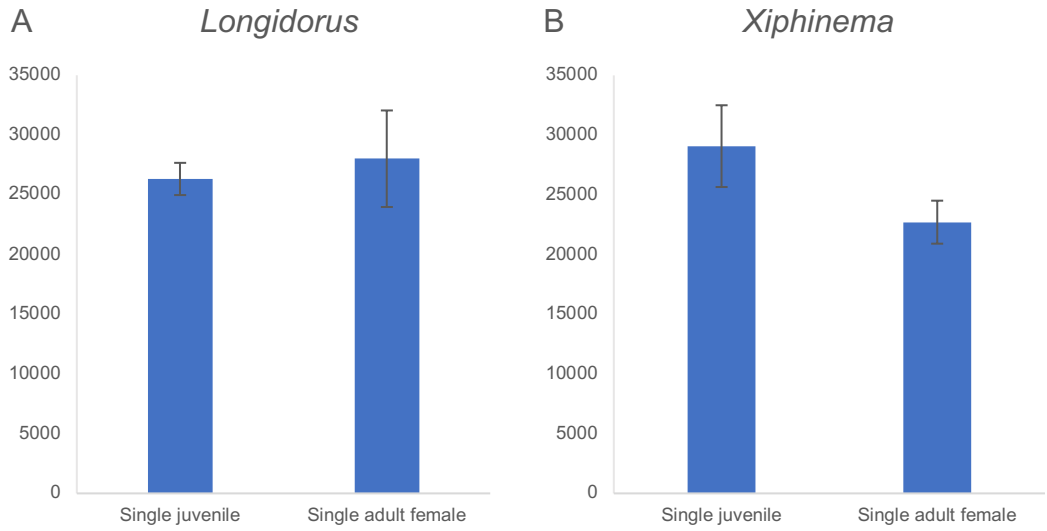

Figure 5.6. Number of reads associated with juvenile and adult specimens of Longidorus *sp. and* Xiphinema *sp. Error bars represent standard error of means (n = 3).*

5.4 Discussion

Four of the sampled taxa, *Acrobeloides*, *Meloidogyne naasi*, *Rhabditis brassicae* and *Rhabditis* sp., were not accounted for in the sequence data. The massive loss of data observed after merging the paired reads may have contributed to this. Even when the filtering step was 'relaxed' to include slightly lower quality reads, none of aforementioned taxa could be recovered. This was because the loss of the merged reads after filtering was mostly to do with their lengths being shorter than the set threshold of 250 bp and not their low average base qualities. This was confirmed when an attempt to relax the quality threshold from expected error of one to two, four or eight resulted in the retention of a very insignificant number of reads (data not shown).

Although taxonomic coverage was not identical between the three replicates, due to the absence of *Mylonchulus* sp. in two of the replicates and *Panagrellus redivivus* in one of the replicates, *Trichodorus primitivus* and *Pratylenchus penetrans* remained dominant across all three and read distribution was generally similar. This corroborates other studies that found consistency in the manner in which reads are distributed between replicates (Porazinska *et al.*, 2009; Deagle *et al.*, 2013) and it further confirms the constancy of amplification bias during PCR. The failure to recover the two missing taxa could not have been due to the primers used because both were amplified by the same set of primers used in other samples. *Panagrellus redivivus*, for example, was successfully recovered in sample PpPcRPr5 in this chapter, where the PCR bias even appeared to have favoured it over the other taxa. For *Mylonchulus*, the same sets of primers were able to amplify it in one of the replicates used in Chapter 3 where a single specimen of it was inadvertently added instead of *Prionchulus*. Absence of *P. redivivus* and *Mylonchulus* sp. did not affect the distribution most likely because they both had very low read abundance in the replicates that they were recovered from and so the effect of their absence or presence were not significant. Moreover, such minor differences in read distribution of taxa between replicates are common among artificial community studies, and their incidence can be attributed to unpredictable or unknown factors rather than any of the known sources of PCR bias (Edgar, 2017a).

It is an established fact that base qualities of amplicons sequenced using Illumina platform tend to drop towards the 3' end for both the forward and the reverse strands (Schröder *et al.*, 2010). And because the reads are assembled based on an overlapping stretch of bases of this 3' end, errors in this region can result in a huge mismatch between the paired reads. In a typical analysis, the unmerged reads are discarded and so are not included in the downstream analysis. In some cases, these unmerged reads can constitute a significant proportion of the data. By using either the forward reads or the reverse reads alone, it was shown here that some taxa could be recovered that otherwise

were missing where paired reads were merged first. In this study, the effect of using the unpaired reads on distribution of read abundance was not as evident as its effect on coverage. And between the forward and the reverse read data, the difference was only manifested at the most stringent quality threshold (EE ≤ 1), where the results showed that the reverse read data recovered one taxon (*Rhabditis brassicae*) less than the forward read data. It is often common for forward reads from Illumina sequencing to have a higher quality than their corresponding reverse reads, and some of the possible causes of this can be overclustering which affects the image analysis. Reagent depletion during the second round of sequencing when the reverse reads are generated could also be a factor.

In contrast with the merged reads, increasing the value of the expected error parameter at the filtering step of both the forward and reverse reads resulted in retention of more sequences. The three additional taxa detected from the unpaired reads showed how, by retaining more reads, a better representation of a community can be achieved. In fact, the only missing species across all analysis methods (merged or single reads) was *Rhabditis* sp. which had a congeneric relative, *R. brassicae* also present in the mock community. Based on this, at least three possible explanations can be given as to why *Rhabditis* sp. was not recovered in the unpaired reads despite using less stringent filtering methods. The first one is that the unidentified *Rhabditis* species is actually the same species as *R. brassicae* or a very close relative, which means reads from both taxa could be lumped together as a single otu at 97% similarity cut-off. Secondly, it may have been that the barcode marker failed to provide the needed resolution to discriminate between the two species. It is common knowledge that the 18S rDNA region can be relatively conserved within certain groups of nematodes (Floyd *et al.*, 2002). Finally, there is also the possibility that reads from this taxon were not present in the read data. In other words, for some reason the primers failed to amplify the reads for *Rhabditis* species. Incomplete recovery of sampled taxa, especially, in bulk samples is not a rare occurrence even with the relatively conserved 18S rDNA region (Porazinska *et al.*, 2009; Adl, Habura, & Eglit, 2014). Recoverability of a taxon by a primer pair can be examined, beforehand, using longer aligned sequences to see how well the annealing sites match the primers. Even then, this may not guarantee successful recovery in NGS data as shown in Porazinska *et al*. (2009).

In theory, factors that contribute to preferential amplification of the abundance-read number ratio distortion such as primer homology, gene copy number and cuticle thickness are constant within species (Amend *et al.*, 2010). However, this study demonstrates that depending on the species composition of a sample, the way these factors affect bias in amplification can be different. For instance, while *P. penetrans* was dominant over *R. brassicae* and *Panagrellus redivivus* in terms of read proportion in the mock community (MC1), it did not show the same dominance over these two taxa in PpPcRPr5. In fact, in

the sample PpPcR which contained all four except *P. redivivus*, reads belonging to *R. brassicae* dominated with over 80% of reads in two of the replicates and about 97% in one. The difference, of course, between the mock community and these other two is that the mock community is more diverse suggesting that amplification bias does not always occur in favour of one taxon over the other but instead can switch depending the other taxa present in the sample. It also confirms the already established fact that factors such as taxon's gene copy number and the presence or absence of mismatches at the primer annealing sites as reported by other authors (Sipos *et al.*, 2007; Kembel *et al.*, 2012; Edgar, 2017a) are only a few among some of factors driving bias in read number distribution (including ones that are not known).

Differences in the biomass of individuals in mixed taxa samples is among the key determinants of the amount of amplicon copies generated at the end of a PCR reaction for each taxon. Despite the clear biomass difference, number of reads from adult female *Longidorus* was not significantly different from those from its juveniles. Since these were monospecific samples, without the influence of other nematode taxa, the most probable explanation would be that the difference in gene copy numbers and cell numbers between these two life stages had no significant impact on the number of reads generated during PCR. Both results are in contrast with Elbrecht and Leese (2015) who found a strong positive correlation between biomass and read numbers.

In microbial studies, size of individuals in the community is not as crucial as in larger nematodes where size difference can be quite significant between species as well as life stages. This study, in addition to confirming the presence of read number bias and the similarity of its pattern across identical communities, also demonstrates how such pattern can drastically change even under a slightly different community. This study further demonstrates how limited our understanding of the factors influencing read abundance of taxa in samples puts to question studies that use taxonomic abundance to infer community structure. Carrying on from the previous chapter, it is clear from this study that specimen size or life stage probably has no effect on the read numbers; retention of low-quality reads may increase detection but has no influence on read distribution. There is still a lot to understand with regard to factors that determine how reads are distributed between taxa within a community.

6 Nematode communities under different tillage and traffic conditions

6.1 Introduction

Nematodes constitute a highly diverse and species-rich group of organisms and their numeric dominance make them a vital component of the soil ecosystem. They possess most of the requisite attributes of soil health indicators, such as being relevant for maintaining balanced communities, being sensitive to the kind of stress being investigated, being widely applicable to different pressures and environments, being measurable, interpretable and cost-effective to sample (Cairns Jr, McCormick, & Niederlehner, 1993; Ritz *et al.*, 2009). Additionally, there is more information on nematode taxonomy and trophic functions that make them better biological indicators of soil conditions than other fauna in the soil such as mites, termites, springtails and earthworms (Neher, 2001). The use of their community structure to reveal changes in the soil environment resulting from certain field management practices is an area that has received quite some attention in the past (Yeates *et al.*, 1993; Ferris, Venette, & Lau, 1996; Porazinska *et al.*, 1998, 1999).

Intensive use of farm machinery has been identified as one of the main causes of compaction of soil. Over the years, increasingly heavier machinery has been used to work land for the cultivation of crops. Such increased impact on the soil can lead to compaction of the subsoil, causing irreparable damage (Arvidsson, 2001; Horn & Fleige, 2003). Soil compaction caused by uncontrolled traffic with heavy machinery and continuous cropping is a well-recognised challenge facing modern agriculture (Håkansson, Voorhees, & Riley, 1988; Gysi, 2001). It can impact on yield through changes it causes to soil pore size (Horn & Fleige, 2003), damage to the soil structure, restricted root system development and penetration, impaired water uptake, poor growth and reduced yields (Voorhees, Evans, & Warnes, 1985; Czyż, 2004). Grain yield loss of between 2.0 – 30.5% have been reported in spring barley on fields compacted with tractor passes ranging from one to four prior to sowing (Czyż, 2004). Changes to the soil pore system caused by compaction can also reduce water holding capacity, aeration and rate of water infiltration, which does not only reduce the amount of available water for plant growth but increases the risk of flooding and erosion and has a detrimental effect on nutrient cycling (Gerrard, 2000). Limited availability of oxygen resulting from impaired aeration in the soil can be detrimental to soil biological processes such as organic matter decomposition and soil mineralization as most organisms depend on oxygen to function.

Controlled traffic farming (CTF) is a quickly advancing system intended to avert or minimise compaction caused by random traffic farming (RTF). This system is based on a fixed layout of machinery passes across a field (Kroulík *et al.*, 2011) designed to limit the

total area trafficked. By minimising the level of compaction resulting from machinery passes, it has been shown that energy savings of 25-40% could be achieved during tillage (Tullberg, 2000). Another system used to minimise compaction while allowing normal operation of machineries is the low ground pressure system (LGP). This is based on the idea that lower tyre inflation pressures with LGP-specific tyres can reduce the pressure exerted on the ground by spreading load over a wider area than a normal tyre does (Tijink, Döll, & Vermeulen, 1995).

Conventional tillage, otherwise known as deep-tillage has been practiced under several agricultural systems (Wardle, 1995) to enhance soil aeration, increase water retention and drainage, improve water infiltration, to improve incorporation of plant debris in the soil and to regulate pests, weeds and pathogens (Phillips *et al.*, 1980; Gebhardt *et al.*, 1985; Unger, 1990; Khan, 1996). Tillage can also have such adverse effects as exposing soil to erosion and loss of top soil, reduced soil moisture retention and organic matter contents as well as possible loss of soil biodiversity (Kladivko, 2001; Pretty, Morison, & Hine, 2003; Holland, 2004). Tillage has a very significant effect on the structure and abundance of soil organisms through the influence it has on soil physical and chemical properties (Kladivko, 2001). Additionally, tillage choice can also affect the community of fauna including nematodes through changes it might cause to organisms at lower trophic levels (Wardle, 1995). Agricultural systems that involve little perturbation of the soil are often regarded as resembling a natural system, which is characterised by the existence of a wider diversity of soil organisms (Altieri, 1991).

According to Wardle (1995) many authors have investigated the impact of tillage regimes on nematode communities; with a number of them reporting contrasting results. Parmelee and Alston (1986) studied the changes in the community structure of nematode trophic groups in conventional (CT) and no-tillage (NT) agroecosystems on a monthly basis for a year under two crop types; rye and clover. They found a significantly higher monthly mean total nematode abundance in conventional tillage (CT) than in no tillage (NT) plots grown with rye and clover. The seasonal data, however, showed no differences between CT and NT plots for most part of the year except for late winter and early spring. Bacterivores and fungivores were significantly higher in CT than in NT plots while the population of omnivore-predator nematodes showed no apparent differences between the two tillage methods. The abundance of herbivorous nematodes also was not significantly different. This is in contrast with an earlier study by Thomas (1978) who reported a significantly higher population of most of the sampled taxa in no-till plots than in the spring and autumn-tilled plots. Lenz and Eisenbeis (2000) found that nematode density significantly reduced following the first application of tillage treatment, although secondary tillage had little effect on the density. Bacterivores became dominant in tilled plots while herbivores dominated the non-tilled plots. Sanchez-Moreno *et al*. (2006) looked at the response of

nematode taxa to tillage and intermittent fallow treatments. Their conclusion was that high enrichment indices (EI) values were associated with a combination of conventional tillage and continuous cropping. No-tillage and fallow systems did not necessarily support more structured food webs, but instead, supported the slow enrichment of a fungal-based community.

A number of studies have investigated the environmental impacts of traffic and its effects on yield (Czyż, 2004; Tullberg, Yule, & McGarry, 2007; Gasso *et al.*, 2014). However, while the effect of tillage on nematodes in the soil has been the subject of numerous studies, traffic effect on soil faunal community has received much less attention. The aim of the current study was to determine the effect of soil tillage and traffic-induced compaction on nematode communities in the soil and how other environmental variables influence these communities in an established field-scale tillage/traffic experiment. For this reason, in addition to the sampling of nematodes, data on physical and chemical properties of soil were collected to examine how they relate with the abundance of certain nematode taxa.

6.2 Materials and methods

6.2.1 Experimental site

The experimental field site (Large Marsh field: 52°46.7899'N, 002°25.5236'W) located at Harper Adams University is an integrative project that has been running for the past five years (2012-2017). In the first year of the experiment, the area was surveyed, and areas of uniformity marked out using data on soil type, texture, crop growth and yield of winter wheat (*Triticum aestivum* cv. Duxford). The soil at the site is mostly very slightly stony sandy loam (Smith *et al.*, 2013). In the year that followed, a 3x3 factorial plot experiment was set up in 4 randomised blocks that consisted of three tillage types and three traffic types in plots 4 m wide and 80 m long. After each harvest (August of each year), the tillage and traffic treatments were reapplied. The site was under its fourth treatment application when samples for this study were taken in March 2016. The crops grown since the establishment of the experiment were winter wheat (2012), winter barley (years 2013 and 2014) and then spring oats (2015).

The traffic treatments consisted of random traffic farming (RTF), low ground pressure (LGP) and controlled traffic farming (CTF). The RTF treatment was designed to mimic how machinery operates on the farm without restriction on the tractor passes on the soil. The LGP system is essentially an RTF system where LGP-specific tyres, with reduced pressure, are used to minimise the impact on the ground. The CTF system involves the least number of vehicular passes where vehicles are limited to a strict circuit within the field using GPS. The tillage treatments included deep-tillage (250 mm deep), shallow-tillage (100 mm deep) and zero-tillage. There were four strips on each plot. The two outer strips were designated as the tractor wheel ways for operations such as drilling and tilling for the tilled plots. For this study, CTF samples were taken from the middle strips in the CTF plots that had zero passes while the LGP and RTF samples were taken from the wheel way plots with either two or three passes depending on the tillage treatment.

6.2.2 Sampling, extraction and identification of nematodes

Soil sampling was carried out over a period of three weeks between 29th February and 22nd March 2016. Fifteen cores were taken from each plot using an auger of 30 cm length and 2.5 cm width. The individual cores were then combined into one composite sample per plot. Sampling was done at three different depths: 0-5 cm, 5-15 cm and 15-30 cm. Nematodes were extracted from a sub-sample of 200 g soil using the Whitehead tray method (Whitehead & Hemming, 1965). Suspensions of nematodes produced after extraction were concentrated to 10 ml and transferred into 15 ml Falcon tubes. For identification, each sample suspension was first homogenised by gently inverting the tubes at least ten times before 1 ml subsample was pipetted into a gridded rectangular

counting dish. An additional 2 ml of tap water was added to the counting dish to ensure the suspension spread across the base of the dish. The nematodes in the suspension were then examined under a stereoscope (Meiji EMT, Meiji Techno Co. Ltd) at 40x magnification where they were all counted. Following this, all nematodes were picked and mounted onto glass slides (15 individuals at a time) and identified to family or genus under high magnification (400x) using a compound microscope (Meiji MX5300L, Meiji Techno Co. Ltd).

6.2.3 Community indices

Using the genera or families they belonged to, nematodes were classified into the five feeding types according to Yeates *et al*. (1993) and given the coloniser-persister (cp) scores according to Bongers (1990). With the cp scores of the taxa, maturity index (MI) (Bongers, 1990) -with and without the inclusion of dauer juveniles in the computation- was calculated for each sample. The MI is simply the weighted mean cp-score of an average nematode in the assemblage and it ranges from one to five. Lower MI scores are typically indicative of high abundance of enrichment opportunistic or general opportunistic taxa, associated with eutrophication events or stress conditions (Bongers, 1990). Higher values, on the other hand, indicate the abundance of large-bodied, sensitive persisters typically associated with structured and stable environments. Another dependent of MI, MI.2-5, was computed by excluding the enrichment opportunist cp-1 taxa from the calculations. Channel index (CI) (Ferris *et al.*, 2001) was also calculated; this index indicates which energy channel, either fast bacterial or slow fungal, is the most predominant in the community based on the populations of fungivores and bacterivores. Other descriptors of food web such as Enrichment (EI), Basal (BI) and Structure (SI) indices that describe the proportion of enrichment opportunistic microbivores, general opportunistic microbivores and stress-sensitive predators and omnivores respectively, were also calculated (Ferris *et al.*, 2001).

6.2.4 Soil analysis

During sampling, penetrometer resistance readings were taken at 15 different spots each close to where soil cores were taken using an Eijkelkamp penetrologger (Eijkelkamp Soil & Water, Giesbeek, The Netherlands). The instrument recorded the resistance as a unit of pressure in MPa at each centimetre down to the depth of 40 cm. A fraction of the composite samples, about 500 g, for each plot was prepared and sent to Natural Resource Management Laboratories Ltd (Berkshire, UK) for analysis (http://www.nrm.uk.com). For each sample, dry soil matter percentage, organic matter

content (loss on ignition method), total nitrogen, pH, phosphorus, potassium, magnesium, ammonium and nitrate were measured.

6.2.5 Data analysis

Direct effect of treatments on the soil food web was illustrated as the index V of the different trophic groups, based on their responses to tillage and traffic (Wardle, 1995).

$$V = \frac{2Mt}{Mb + Mt} - 1$$

Where: *Mb* and *Mt* are abundances under no treatment and treated conditions, respectively.

Zero tillage was used as the baseline (no treatment) condition for tillage treatments while CTF was used as the baseline for the traffic treatments. For tillage, zero tillage was compared with shallow tillage (Zero vs Shallow) and then compared with deep tillage (Zero vs Deep). For traffic, CTF was compared with LGP (CTF vs LGP) and then with RTF (CTF vs RTF). The index V value ranges from -1 to +1 for each trophic group, which correspond to total inhibition and total stimulation by treatment, respectively. Zero value indicates no change in abundance by the treatment.

Canonical Correspondence Analysis (CCA) was used to illustrate how nematode taxa related with environmental variables and compaction. The analyses of the effects of tillage and traffic treatments on abundance of nematode taxa and nematode trophic groups were performed separately for each depth that was sampled. This was done using generalized linear models with negative binomial distribution as recommended for count data (O'Hara & Kotze, 2010). Post-hoc tests on the univariate analysis were performed using Tukey's test. The R functions *cld* and *lsmeans* from multcomp (Hothorn *et al.*, 2008) and lsmeans (Lenth, 2016) packages , respectively were used for post-hoc tests on the interaction effects. Treatment effects on community indices were analysed using the robust two-way Analysis of Variance (ANOVA) function for trimmed means, *t2way* from the WRS2 (Mair & Wilcox, 2018) R package. All statistical analyses were carried out using RStudio (RStudio Team, 2016) development environment for R.

6.3 Results

6.3.1 Nematode taxa recovered

Across all the samples, a total of 83 different taxa including dauer juveniles were recovered and categorised into genera, families, trophic groups, cp-scores and functional guilds (Appendix 10.4). There were 35 families in total. Most of these taxa only occurred in a minority of the samples analysed. In fact, close to 80% of the taxa occurred in less than half the samples. None of the taxa occurred in all samples. *Helicotylenchus* (herbivore) was the genus most frequently recovered; present in 100 of the 108 samples. Representatives of *Wilsonema*, *Clarkus, Penjatinema, Panagrolobus, Propanagrolaimus*, *Miconchus,* Heteroderidae, Xyalidae and Linhomoeidae were only recovered once in the samples. The most frequently recovered bacterivorous nematodes in order of increasing frequencies were *Eucephalobus*, *Cephalobus* and *Acrobeloides*.

6.3.2 Response of nematodes to tillage and traffic treatments

At the top (0-5cm) depth, there was a significant tillage and traffic interaction effect on numbers of *Aphelenchus* (P<0.001), *Tylenchorhynchus* (P<0.001), *Meloidogyne* (P=0.014) and *Trichodorus* (P<0.001) (Table 6.1). For *Aphelenchus*, both LGP and RTF caused an increase in abundance compared to CTF under deep tillage (CTF = 1.03, LGP = 2.17, RTF=2.18) and shallow tillage (CTF = 1.76E-15, LGP = 4.23, RTF = 2.15). Under zero tillage, however, their numbers decreased with both LGP and RTF (CTF = 3.45, LGP = 1.78E-15, RTF = 1.68E-15).The number of *Tylenchorhynchus* was found to increase with LPG traffic under zero tillage but decreased under shallow and deep tillage (Figure 6.1). Random traffic produced the highest number of *Meloidogyne* compared to the other tillage methods under deep (CTF = 33.10, LGP = 24.99, RTF = 48.95) and zero tillage (CTF = 36.43, LGP = 85.72, RTF = 1241.67), while under shallow tillage LGP produced the highest abundance (CTF = 18.54, LGP = 64.02, RTF = 20.76). There was also a significant increase in the abundance of *Trichodorus* in response to traffic under deep tillage and a decrease under shallow tillage but no significant change under zero tillage (Figure 6.2). Tukey's multiple comparison test showed no significant difference in the interaction means for *Aphelenchus* and *Meloidogyne*.

At the (5-15cm) depth, there was a significant tillage and traffic interaction effect on the abundance of *Acrobeloides* (P=0.043), *Aphelenchus* (P=0.002), *Alaimus* (P<0.001), Dorylaimidae (P=0.006), *Meloidogyne* (P<0.001) and *Trichodorus* (P<0.001). The abundance of *Acrobeloides* increased with RTF under deep tillage (CTF = 14.52, LGP = 15.31, RTF = 25.08) but decreased under shallow tillage (CTF = 13.15, LGP = 14.40, RTF = 8.80). Under zero tillage both LGP and RTF caused a reduction in their abundance (CTF = 33.74, LGP = 8.59, RTF = 13.87). *Alaimus* increased in response to RTF and

decreased with LGP under deep tillage (CTF = 33.10, LGP = 24.99, RTF = 48.95). Under shallow tillage, their numbers increased with LGP and did not change with RTF (CTF = 18.54, LGP = 64.02, RTF = 20.76). Under zero tillage, both LGP and RTF resulted in an increase in their abundance (CTF = 36.43, LGP = 85.72, RTF = 1241.67). Under deep tillage, RTF caused a decrease in the number of Dorylaimidae (CTF = 3.69, LGP = 2.57, RTF = 2.22E-16) while under shallow tillage both LGP and RTF reduced their abundance (CTF = 18.9, LGP = 1.28, RTF = 4.33). Under zero tillage, however, RTF caused an increase in their numbers and LGP caused a decrease in their numbers (CTF = 2.56, LGP = 2.22, RTF = 7.09). With deep tillage, both LGP and RTF reduced the number of *Meloidogyne* (CTF = 29.54, LGP = 9.54, RTF = 14.91) whereas under shallow (CTF = 1.22, LGP = 17.78, RTF = 19.11) and zero (CTF = 4.30, LGP = 398.31, RTF = 391.11) tillage there was an increase in their numbers with both LGP and RTF. *Trichodorus* under deep (CTF = 11.8, LGP =1.07, RTF = 2.14) and zero (CTF = 19.7, LGP = 2.96, RTF = 1.94E-15) tillage reduced in abundance in response to both LGP and RTF treatments. Under shallow tillage, RTF increase their abundance while LGP reduced it (CTF = 5.24, LGP = 2.25, RTF = 53.7). In the case of *Aphelenchus*, traffic effect was only significant under deep tillage with their abundance within LPG being significantly higher than within RTF. However, under shallow and zero tillage, traffic effect was not significant (Figure 6.3). For all but *Aphelenchus*, Tukey's multiple comparison test showed no significant difference between interaction means.

The number of *Tylenchus* found to be significantly different between tillage treatments (P<0.001) but there was no interaction between tillage and traffic. Specifically, deep tillage caused a significant decrease in the abundance of *Tylenchus*.

At the lower depth (15-30cm), *Tylenchus* (P=0.008) and *Pratylenchus* (P<0.001) were the only taxa affected by the interaction between tillage and traffic. For *Tylenchus*, the effect of traffic was significant under shallow tillage but not under deep or zero tillage (Figure 6.4). Similarly, response of *Pratylenchus* to traffic was significant under zero tillage alone (Figure 6.5).

Total nematode abundance within the top depth was significantly affected by tillage and traffic interaction (P=0.002) (Table 6.2). There was an increase in total abundance in response to traffic under zero tillage but not under deep or shallow tillage where traffic had no effect. Within the middle depth, tillage had a significant effect on total abundance (P=0.020) with both shallow and deep tillage causing a decrease in abundance. Within the lower depth, traffic and tillage interaction effect was significant on total abundance. As in the top depth, traffic caused an increase in total abundance only under zero tillage within the lower depth.

Within the top depth, PPI (P=0.001) was the only index affected by the interaction between tillage and traffic treatments (Table 6.2). While traffic did not affect PPI under deep and shallow tillage, RTF caused a significant increase in PPI under zero tillage. Within the middle depth, tillage and traffic interaction had no effect on any of the indices. Traffic affected MI2-5, EI and SI significantly. There was a statistically significant decrease in MI2-5 between CTF and RTF. Enrichment index also decreased significantly in response to both LGP and RTF. Within the lower depth, interaction effect of tillage and traffic was not significant across all indices. Neither of the main factors had effect on the indices.

The effects of tillage and traffic treatments on nematode trophic groups illustrated by index V showed some varying responses along the three depths (Figure 6.6). The analyses of the impact of tillage and traffic were made with reference to zero tillage and CTF treatments in order to demonstrate how the various trophic groups are affected by soil perturbation and compaction, respectively. Tillage treatments caused mild to extreme inhibition of all the trophic groups except bacterivores and omnivores which were mildly stimulated by shallow-tillage within the top 5 cm (Figure 6.6A). The two traffic treatments, LGP and RTF, had contrasting effects on both herbivores and fungivores. In herbivores, LGP caused mild inhibition while RTF resulted in moderate stimulation. Similarly, LGP caused mild inhibition of fungivores. However, instead of moderate stimulation as was the case in herbivores, RTF caused mild stimulation fungivores (Figure 6.6A). Both LGP and RTF caused mild or moderate inhibition of bacterivores and predators. Omnivores were the only group that responded to the two traffic systems positively, with mild stimulation.

Table 6.1. Summary of analysis of variance (ANOVA) for nematode density – $\chi 2$ values from glm.nb function and P *values shown in parentheses. Data are from 108 samples representing 36 plots sampled at three different depths. Treatments consist of three levels of tillage (zero, shallow and deep) and three levels of traffic (random traffic, low ground pressure and controlled traffic).*

Source (degrees of freedom)	Density (individuals/100 g soil)										
	Acrobeloides	*Aphelenchoides*	*Aphelenchus*	*Tylenchus*	*Rhabditis*	*Alaimus*	*Tylenchorhynchus*	*Dorylaimidae*	*Meloidogyne*	*Pratylenchus*	*Trichodorus*
Top Depth											
Tillage (2)	0.41 (0.813)	0.72 (0.695)	7.00 (0.031)	2.26 (0.166)	3.02 (0.221)	5.69 (0.058)	22.94 (<0.001)	2.28 (0.308)	0.61 (0.738)	0.02 (0.992)	13.86 (<0.001)
Traffic (2)	1.54 (0.462)	2.83 (0.161)	0.54 (1.000)	0.05 (0.070)	3.97 (0.282)	12.70 (<0.000)	26.78 (<0.001)	0.61 (0.177)	0.51 (0.774)	6.72 (0.035)	13.86 (<0.001)
Block (3)	3.47 (0.324)	6.02 (0.111)	14.83 (0.002)	17.27 (<0.001)	6.98 (0.072)	2.34 (0.505)	176.99 (<0.001)	29.20 (<0.001)	9.22 (0.027)	3.35 (0.340)	49.24 (<0.001)
Tillage X Traffic (4)	2.77 (0.596)	7.25 (0.124)	20.08 (<0.001)	4.46 (0.348)	1.82 (0.768)	7.33 (0.119)	55.61 (<0.001)	6.19 (0.185)	12.54 (0.014)	4.63 (0.326)	37.34 (<0.001)
Middle Depth											
Tillage (2)	5.67 (0.059)	1.69 (0.429)	0.68 (0.713)	21.85 (<0.001)	0.915 (0.633)	4.47 (0.107)	2.15 (0.341)	3.88 (0.143)	11.22 (0.004)	4.39 (0.112)	1.99 (0.370)
Traffic (2)	1.84 (0.399)	0.45 (0.797)	17.53 (<0.001)	3.04 (0.218)	0.056 (0.972)	57.10 (<0.001)	0.01 (0.998)	6.89 (0.032)	1.07 (0.586)	0.50 (0.780)	6.29 (0.043)
Block (3)	6.34 (0.096)	7.99 (0.046)	9.87 (0.020)	38.14 (<0.001)	0.58 (0.901)	3.00 (0.391)	23.76 (<0.001)	5.23 (0.156)	2.81 (0.422)	1.09 (0.779)	35.51 (<0.001)
Tillage X Traffic (4)	9.84 (0.043)	2.40 (0.662)	16.06 (0.002)	9.21 (0.056)	2.20 (0.699)	30.35 (<0.001)	4.73 (0.317)	16.41 (0.006)	25.08 (<0.001)	1.19 (0.880)	21.64 (<0.001)
Lower Depth											
Tillage (2)	0.89 (0.641)	5.65 (0.059)	2.27 (0.322)	2.54 (0.280)	0.000 (1.000)	1.10 (0.577)	1.27 (0.531)	10.99 (0.004)	0.07 (0.964)	11.50 (0.003)	0.65 (0.721)
Traffic (2)	0.43 (0.807)	4.31 (0.116)	1.56 (0.458)	2.84 (0.242)	0.000 (1.000)	7.51 (0.023)	1.06 (0.589)	8.11 (0.017)	0.38 (0.828)	8.23 (0.016)	0.35 (0.839)
Block (3)	14.59 (0.002)	7.57 (0.056)	6.60 (0.086)	20.29 (<0.001)	13.862 (0.003)	6.96 (0.073)	8.74 (0.033)	13.86 (0.003)	1.53 (0.676)	11.62 (0.008)	18.39 (<0.001)
Tillage X Traffic (4)	4.52 (0.340)	6.94 (0.139)	3.90 (0.419)	13.91 (0.008)	0.000 (1.000)	6.78 (0.147)	1.73 (0.785)	0.000 (1.000)	4.38 (0.358)	18.68 (<0.001)	7.76 (0.101)

Table 6.2. Summary of analysis of variance (ANOVA) for community indices – F values from Wilcox two-way robust ANOVA with trimmed means for community indices or χ^2 values (P values) from glm.nb for total abundance (N) with P *values shown in parentheses. Data are from 108 samples representing 36 plots sampled at three different depths. Treatments consist of three levels of tillage (zero, shallow and deep) and three levels of traffic (random traffic, low ground pressure and controlled traffic).*

Source (degrees of freedom)	Community index							
	MI	MI.2-5	PPI	CI	EI	SI	BI	N
Top Depth								
Tillage (2)	3.30 (0.247)	3.75 (0.211)	15.70 (0.006)	4.68 (0.156)	2.73 (0.299)	3.17 (0.254)	3.30 (0.238)	0.21 (0.903)
Traffic (2)	0.27 (0.882)	3.80 (0.211)	1.75 (0.461)	3.48 (0.243)	2.91 (0.276)	5.08 (0.131)	3.92 (0.185)	1.38 (0.500)
Block (3)	0.47 (0.712)	0.04 (0.988)	0.99 (0.427)	1.82 (0.195)	1.62 (0.234)	0.10 (0.959)	1.15 (0.367)	32.01 (<0.001)
Tillage X Traffic (4)	3.12 (0.634)	6.99 (0.263)	45.18 (<0.001)	5.40 (0.395)	0.17 (0.997)	7.57 (0.230)	0.41 (0.985)	16.24 (0.002)
Middle Depth								
Tillage (2)	2.86 (0.286)	2.47 (0.335)	2.77 (0.294)	0.62 (0.760)	3.01 (0.287)	2.91 (0.275)	1.92 (0.445)	7.85 (0.020)
Traffic (2)	5.96 (0.085)	24.58 (<0.001)	0.34 (0.850)	2.87 (0.313)	8.35 (0.041)	23.32 (0.001)	5.49 (0.110)	5.81 (0.055)
Block (3)	0.08 (0.971)	1.07 (0.396)	4.09 (0.030)	1.93 (0.179)	0.16 (0.922)	1.25 (0.333)	0.21 (0.885)	10.75 (0.013)
Tillage X Traffic (4)	10.44 (0.109)	1.76 (0.821)	2.15 (0.763)	5.21 (0.421)	7.93 (0.210)	2.12 (0.765)	6.32 (0.309)	9.28 (0.054)
Lower Depth								
Tillage (2)	0.84 (0.678)	0.61 (0.760)	1.06 (0.614)	0.08 (0.966)	1.92 (0.441)	0.244 (0.894)	2.74 (0.320)	0.901 (0.636)
Traffic (2)	1.86 (0.431)	5.42 (0.126)	0.35 (0.851)	3.27 (0.277)	0.46 (0.813)	6.99 (0.073)	1.94 (0.446)	8.645 (.013)
Block (3)	1.33 (0.307)	1.67 (0.221)	7.57 (0.003)	3.50 (0.056)	1.16 (0.366)	2.36 (0.120)	0.78 (0.527)	14.77 (0.002)
Tillage X Traffic (4)	0.48 (0.980)	5.87 (0.351)	1.45 (0.867)	4.88 (0.449)	5.38 (0.395)	1.90 (0.808)	7.60 (0.246)	9.90 (0.042)

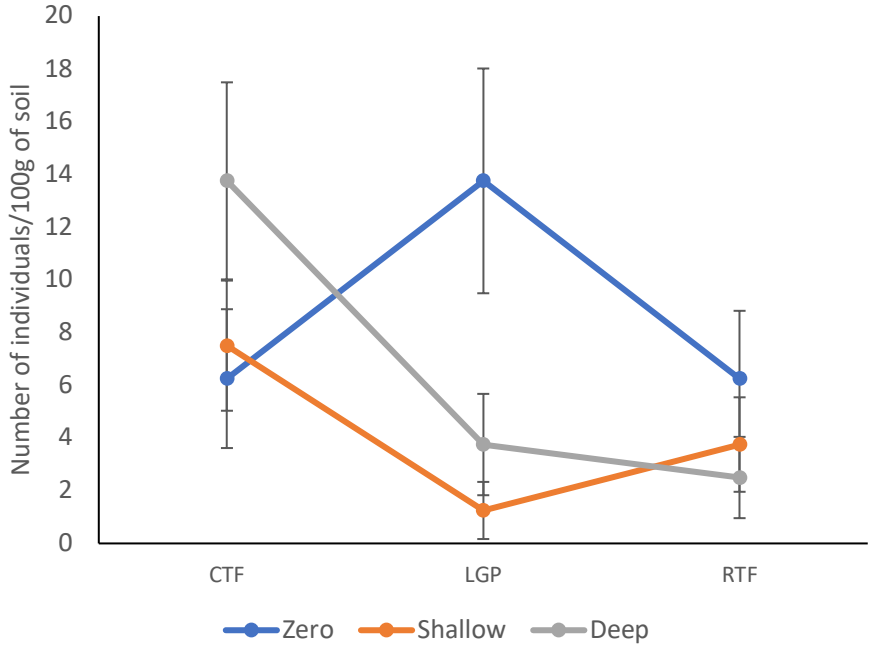

Figure 6.1. Mean abundance of Tylenchorhynchus *across the different tillage and traffic combinations within the top depth (0-5 cm). Error bars are based on the confidence intervals of the respective means.*

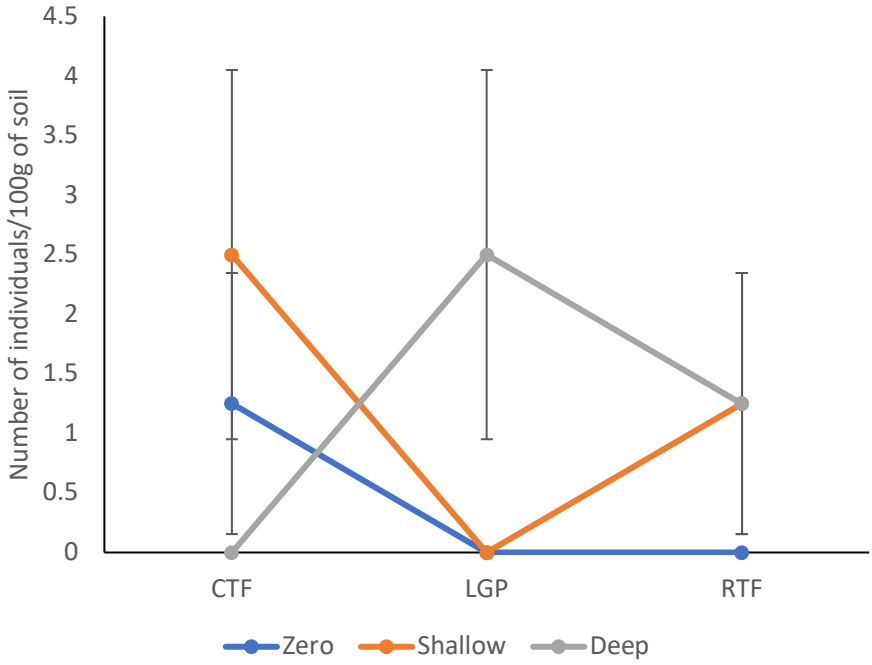

Figure 6.2. Mean abundance of Trichodorus *across the different tillage and traffic combinations within the top depth (0-5 cm). Error bars are based on the confidence intervals of the respective means.*

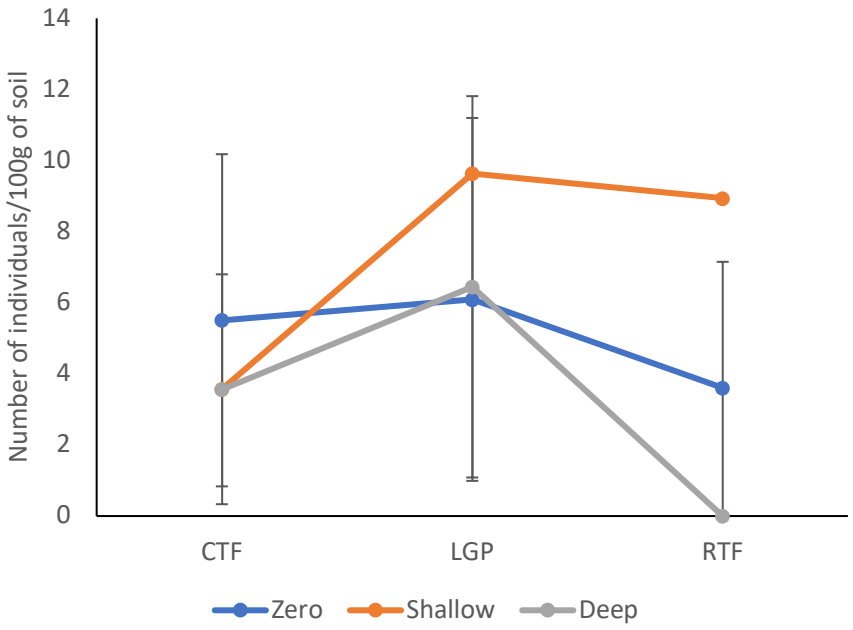

Figure 6.3. Mean abundance of Aphelenchus *across the different tillage and traffic combinations within the middle depth (5-15 cm). Error bars are based on the confidence intervals of the respective means.*

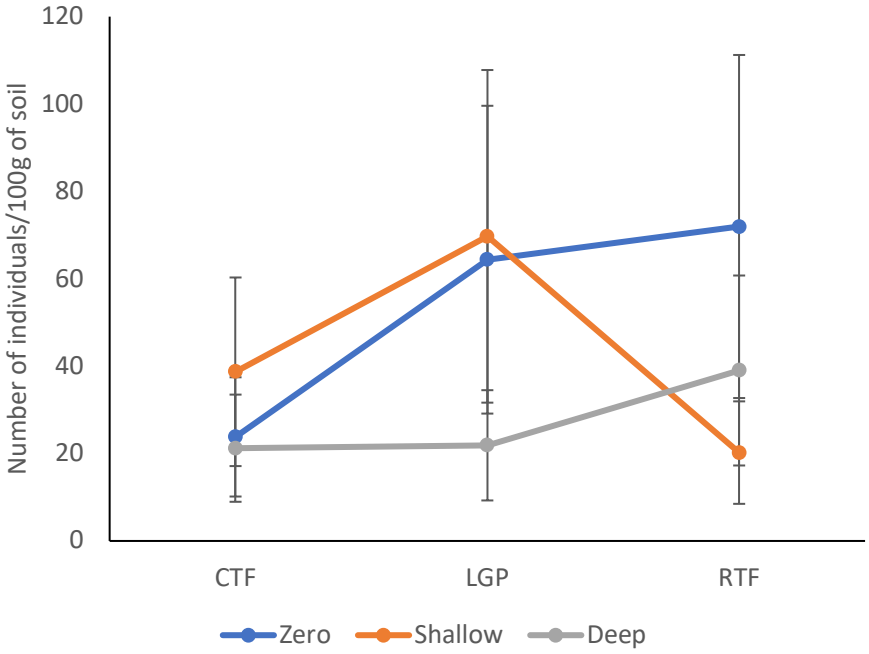

Figure 6.4. Mean abundance of Tylenchus *across the different tillage and traffic combinations within the lower depth (15-30 cm). Error bars are based on the confidence intervals of the respective means.*

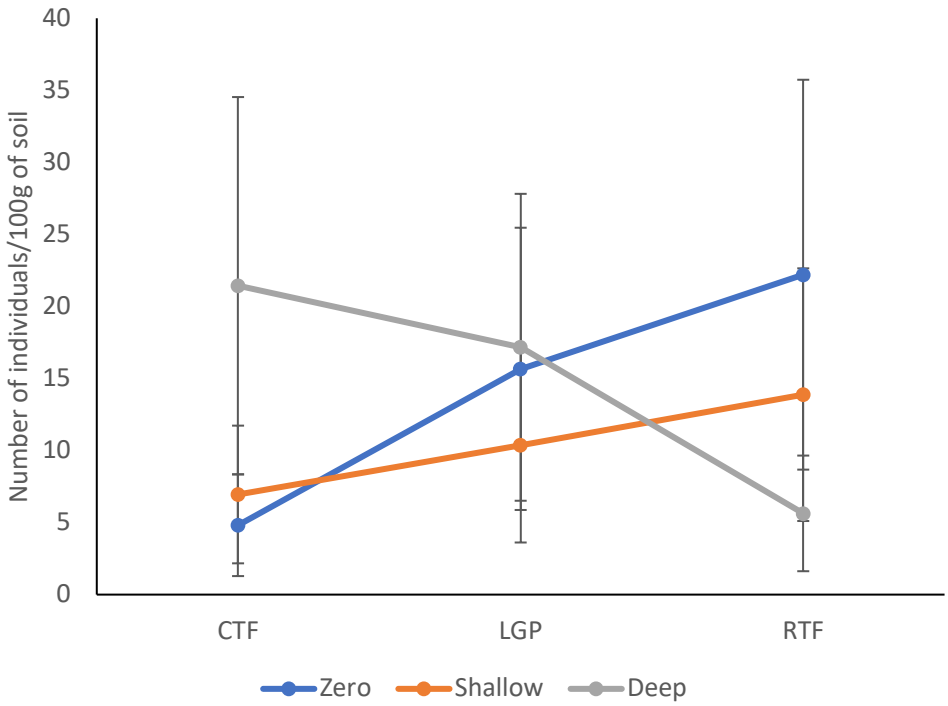

Figure 6.5. Mean abundance of Pratylenchus *across the different tillage and traffic combinations within the lower depth (15-30 cm). Error bars are based on the confidence intervals of the respective means.*

The two tillage methods caused mild to moderate inhibition of herbivores, fungivores and bacterivores in the 5-15 cm depth (Figure 6.6B). In predators, the two tillage treatments had opposite effects, with shallow-tillage causing mild inhibition and deep-tillage causing mild stimulation. Both tillage methods had the same mild stimulation effect on omnivores at this depth. Herbivores were mildly stimulated by both LGP and RTF treatments. In predators and omnivores, LGP resulted in moderate inhibition. In fungivores and bacterivores, there was a mild inhibition by LGP. RTF caused mild stimulation of bacterivores and omnivores while causing mild inhibition of fungivores and predators.

Within the 15-30 cm depth, shallow and deep tillage systems caused inhibition of herbivores and bacterivores (Figure 6.6C). Fungivores, predators and omnivores, on the other hand, were stimulated by tillage. Deep-tillage did not elicit any response from omnivores, which only responded positively to shallow-tillage. The two traffic systems at this depth elicited opposite responses from fungivores and bacterivores. In herbivores and omnivores, both traffic systems resulted in mild to moderate stimulation. Predators were the only group inhibited by both LGP and RTF within this depth.

6.3.3 Effects of environmental variables on nematode fauna

Most of the measured environmental conditions differed significantly across the treatment plots (Table 6.4). The CCA plot revealed various associations between some of selected nematode taxa and a few of the measured environmental variables (Figure 6.7). The bacterivores, *Eucephalobus* (r = -0.53, P < 0.001), *Acrobeloides (*r = -0.59, P < 0.001), *Plectus (*r = -0.39, P < 0.001), *Rhabditis* (r = -0.44, P < 0.001) *Alaimus* (r = -0.20, P < 0.05) were negatively associated with compaction. *Prionchulus* (r = -0.25, P < 0.01) and Dorylaimidae (r = -0.36, P < 0.001) were also negatively associated with compaction. *Tylenchus* (r = 0.40, P < 0.001), *Trichodorus* (r = 0.39, P < 0.001), *Tylenchorhynchus* (r = 0.26, P < 0.05) and *Aphelenchus* (r = 0.42, P < 0.001) were positively associated with compaction. *Meloidogyne* was positively associated with organic matter content (OM) (r = 0.26, P < 0.01) and tillage intensity (r = 0.35, P < 0.001). Magnesium content positively correlated with *Tylenchus* (r = 0.22, P < 0.05) and Aporcelaimidae (r = 0.20, P < 0.05), but negatively correlated with *Trichodorus* (r = -0.22, P < 0.05).

The community indices also showed varying association with some environmental variables (Figure 6.8). Maturity index correlated negatively with NO_3 (r = -0.22, P < 0.05), and NH_4 (r = -0.23, P < 0.05). Structure index (r = 0.45, P < 0.001) and enrichment index (r = 0.44, P < 0.001) were positively correlated with compaction while basal index (r = -0.54, P < 0.001) showed a negative association with compaction.

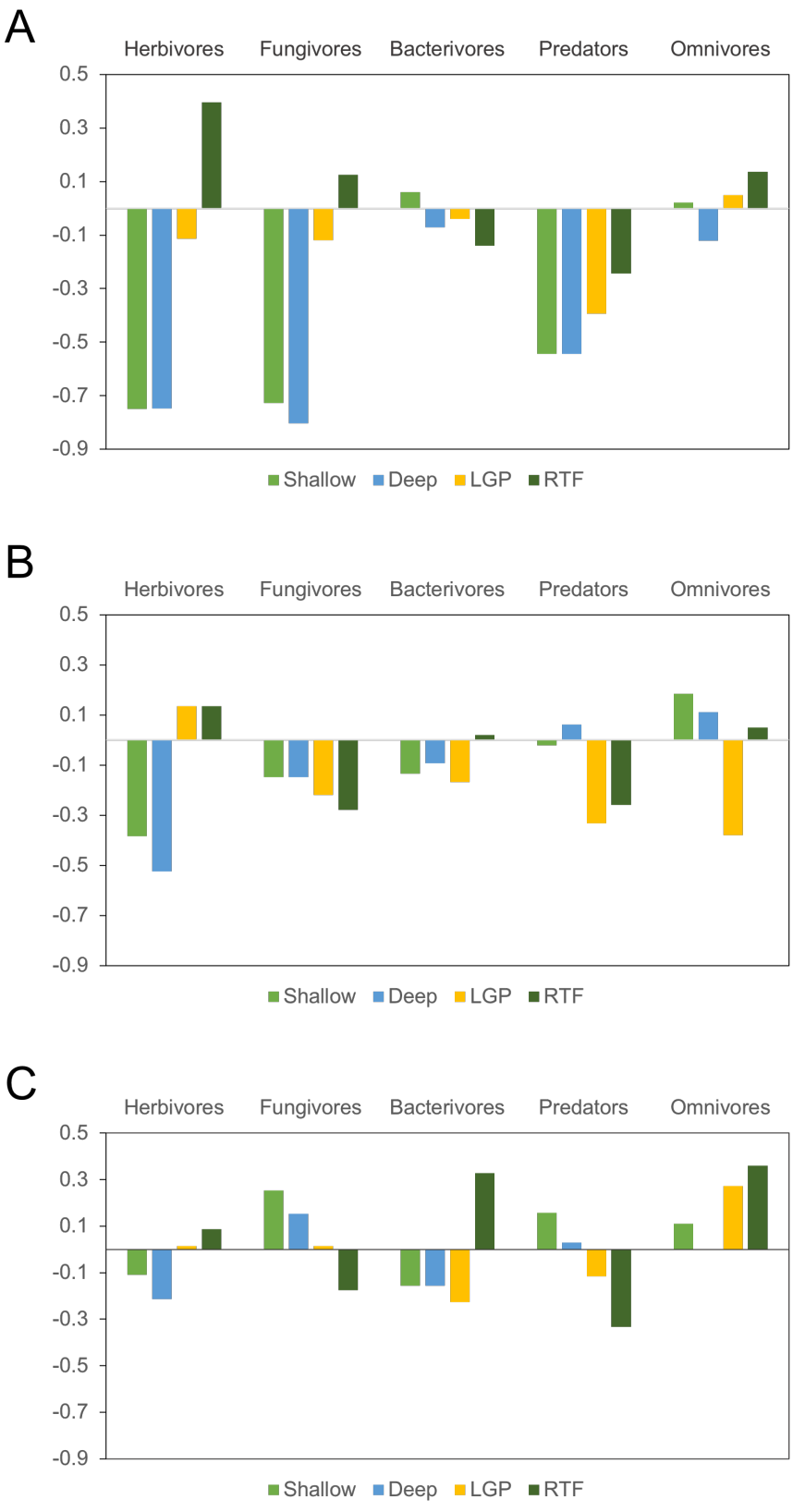

Figure 6.6. Graphical representation of V index values of various nematode trophic groups in their response to shallow and deep-tillage systems, and LGP and RTF traffic regimes. A) Top 5 cm depth B) middle 5-15 depth and C) Lower 15-30 cm. Computation of the V index for tillage and traffic effects on the nematode trophic groups was carried out with reference to values obtained from zero-tillage and the CTF system, respectively. Positive values denote stimulation and negative values, inhibition. V < -0.67: extreme inhibition; -0.33 > V > -0.67: moderate inhibition; 0 > V > -0.33: mild inhibition; 0 < V < 0.33: mild stimulation; 0. 33 < V < 0.67: moderate stimulation; V > 0.67: extreme stimulation.

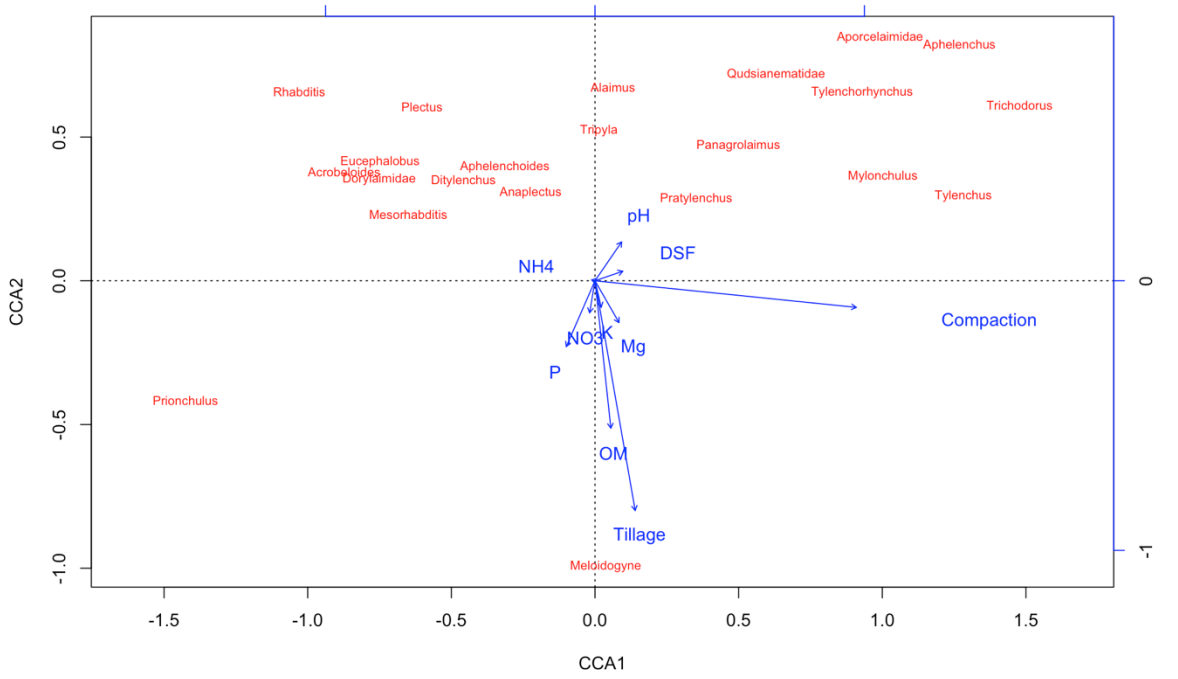

Figure 6.7. CCA bi-plots of nematode taxa and environmental variables. Eigenvalues for CCA1 and CCA2 were 0.252 and 0.141 respectively. Percentages of the variance of CCA1 and CCA2 explained were 51.3 and 28.6%, respectively. DSF = dry soil fraction.

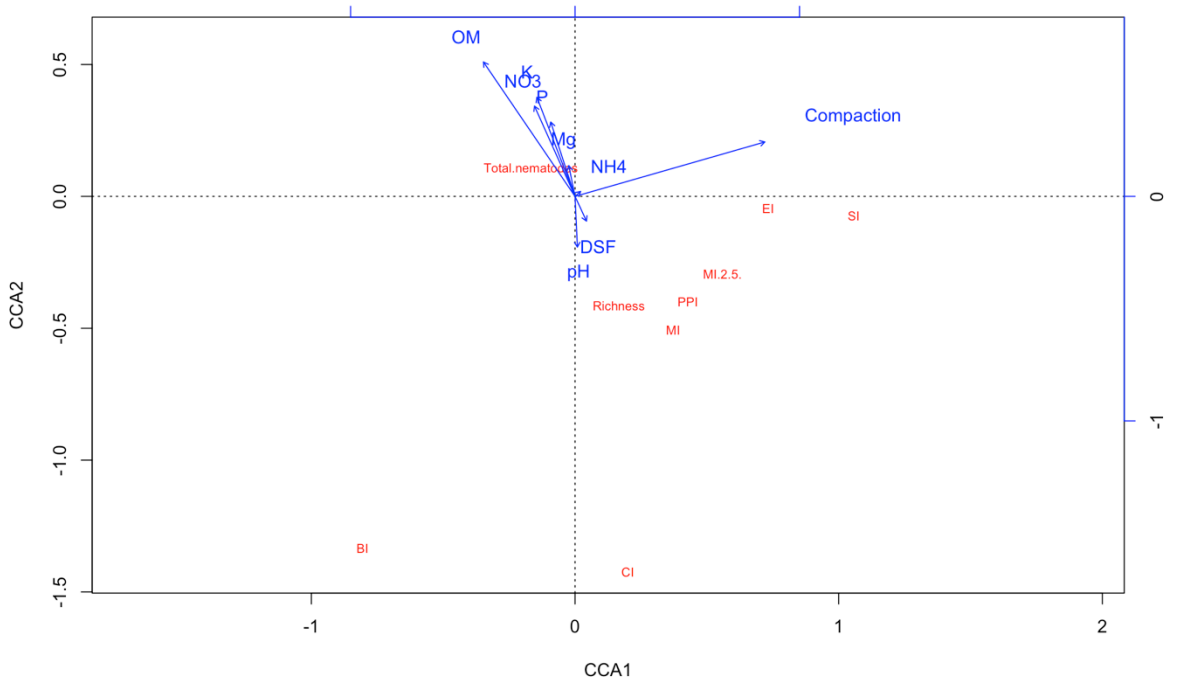

Figure 6.8. CCA bi-plots of nematode indices and environmental variables. Eigenvalues for CCA1 and CCA2 were 0.066 and 0.017 respectively. Percentages of the variance of CCA1 and CCA2 explained were 71.7 and 19.0%, respectively. MI = maturity index; MI.2.5 = maturity index 2-5; PP = plant parasitic index; CI = channel index; EI = enrichment index; SI = structure index; BI = basal index.

Table 6.3. Summary of analysis of variance (ANOVA) for trophic groups[a]. $\chi 2$ values from glm.nb *function* and P *values shown in parentheses. Data are from 108 samples representing 36 plots sampled at three different depths. Treatments consist of three levels of tillage (zero, shallow and deep) and three levels of traffic (random traffic, low ground pressure and controlled traffic).*

Source (degrees of freedom)	Density (individuals/100 g soil)				
	Herbivores	Fungivores	Bacterivores	Predators	Omnivores
Top Depth					
Tillage (2)	0.632 (0.729)	0.528 (0.768)	0.039 (0.981)	1.432 (0.489)	0.518 (0.772)
Traffic (2)	1.916 (0.384)	1.471 (0.479)	2.326 (0.313)	4.964 (0.084)	1.674 (0.433)
Block (3)	34.677 (<0.001)	7.434 (0.059)	13.524 (0.004)	6.010 (0.111)	27.014 (0.234)
Tillage X Traffic (4)	29.158 (<0.001)	4.170 (0.383)	0.696 0.696)	4.159 (0.385)	7.964 (0.093)
Middle Depth					
Tillage (2)	0.704 (0.703)	1.866 (0.393)	6.286 (0.043)	1.304 (0.521)	1.295 (0.523)
Traffic (2)	6.52 (0.038)	0.141 (0.932)	1.623 (0.444)	6.854 (0.032)	0.872 (0.647)
Block (3)	8.496 (0.037)	8.36 (0.039)	19.469 (<0.001)	13.594 (0.004)	3.511 (0.319)
Tillage X Traffic (4)	27.125 (<0.001)	1.626 (0.804)	7.562 (0.109)	3.479 (0.481)	1.895 (0.755)
Lower Depth					
Tillage (2)	6.596 (0.037)	4.182 (0.124)	1.962 (0.375)	3.118 (0.210)	1.399 (0.497)
Traffic (2)	8.614 (0.013)	0.347 (0.841)	0.277 (0.871)	11.662 (0.003)	1.308 (0.520)
Block (3)	46.109 (<0.001)	11.736 (0.008)	2.197 (0.533)	24.512 (<0.001)	0.524 (0.914)
Tillage X Traffic (4)	43.289 (<0.001)	1.507 (0.825)	10.117 (0.038)	13.783 (0.008)	3.337 (0.503)

[a] The list of taxa constituting each trophic group are in Appendix 10.4

Table 6.4. Environmental variables and how they differ across plots. Soils tested were sampled across 0-30 cm depth. Mean data ± standard error of mean.

Environmental variables	CTF-Deep	CTF-Shallow	CTF-Zero	LGP-Deep	LGP-Shallow	LGP-Zero	RTF-Deep	RTF-Shallow	RTF-Zero	*P* values (bivariate)
NO_3^{-1}	9.03±0.78	9.67±0.65	10.17±0.56	6.61±0.42	9.57±1.29	13.57±1.64	8.55±0.77	8.87±0.46	9.90±1.07	<0.001
NH_4^+	0.77±0.26	0.38±0.07	0.42±0.08	0.22±0.04	0.94±0.46	13.94±6.59	0.48±0.13	0.76±0.08	0.44±0.16	<0.001
Compaction†	1.07±0.14	1.20±0.13	1.34±0.17	1.44±0.15	1.43±0.13	1.58±0.13	1.44±0.14	1.52±0.17	1.55±0.10	0.197
Organic matter	3.15±0.04	3.35±0.06	3.47±0.05	3.20±0.02	3.37±0.02	3.45±0.07	3.37±0.02	3.42±0.04	3.45±0.07	<0.001
Total N	36.75±3.89	37.70±2.18	39.70±2.12	25.60±1.73	39.40±6.55	103.15±30.38	33.85±3.11	36.15±1.99	38.75±4.59	<0.001
K^+	166.00±10.05	161.75±3.59	189.25±9.10	168.75±2.00	159.50±3.75	189.00±11.62	150.50±9.19	164.50±12.01	166.25±3.04	<0.001
P	35.20±2.26	36.30±0.81	39.70±1.78	38.15±1.45	33.30±1.29	36.80±2.01	37.15±1.48	37.45±1.37	37.70±1.44	<0.001
Dry matter	81.35±0.32	80.75±0.09	81.62±0.22	81.60±0.24	81.62±0.15	81.82±0.25	81.17±0.14	81.22±0.27	81.27±0.29	0.042
pH	6.22±0.06	6.12±0.05	6.12±0.07	6.45±0.06	6.02±0.11	6.05±0.08	6.42±0.05	6.25±0.10	6.10±0.08	<0.001
Mg^{2+}	113.50±3.34	123.50±2.34	121.50±3.14	129.75±2.82	91.75±14.12	123.00±4.06	121.50±2.11	118.75±2.61	123.50±3.14	<0.001

† Significant in univariate analysis across tillage, traffic and depth. The only environmental variable measure across the three depths.

Within the top depth, herbivores were the only trophic group affected by the interaction between tillage and traffic (P<0.001). The number of herbivores significantly increased in response to traffic, specifically RTF, under zero tillage but not under shallow or deep tillage (Figure 6.9). Within the middle depth, the interaction between tillage and traffic had a significant effect on herbivore abundance (P<0.001). As in the top depth, the number of herbivores appeared to increase with traffic under zero tillage but not under deep or shallow tillage (Figure 6.9). Tillage and traffic had a significant main effect on bacterivores (P=0.043) and predators (P=0.032), respectively. Bacterivores decreased in response to shallow tillage while predators decreased in response to both LGP and RTF. However, Tukey's post-hoc analysis showed no significant difference for tillage in the case of bacterivores and traffic in the case of predators. Within the lower depth, herbivores (P<0.001), bacterivores (P=0.038) and predators (P=0.008) were significantly affected by tillage and traffic interactions. For predators under deep tillage, RTF caused a decrease in the number of individuals while LGP caused an increase in their number (CTF = 8.86, LGP = 11.79, RTF = 0.99). Under shallow tillage, both RTF and LGP resulted in a decrease in their number (CTF = 14.58, LGP = 8.81, RTF = 6.05). Under zero tillage, their number increased with RTF and decreased with LGP (CTF = 5, LGP = 4.94, RTF = 10.41). While the number of herbivores increased with RTF under zero tillage, they decreased with RTF under shallow tillage. They also decreased with LGP under deep tillage (Figure 6.9). Bacterivore populations responded differently to tillage across the three traffic systems. Their number increased significantly with shallow tillage under LGP but were not affected by tillage under CTF or RTF (Figure 6.10). For predators, Tukey's post-hoc analysis showed no significant difference in the interaction means.

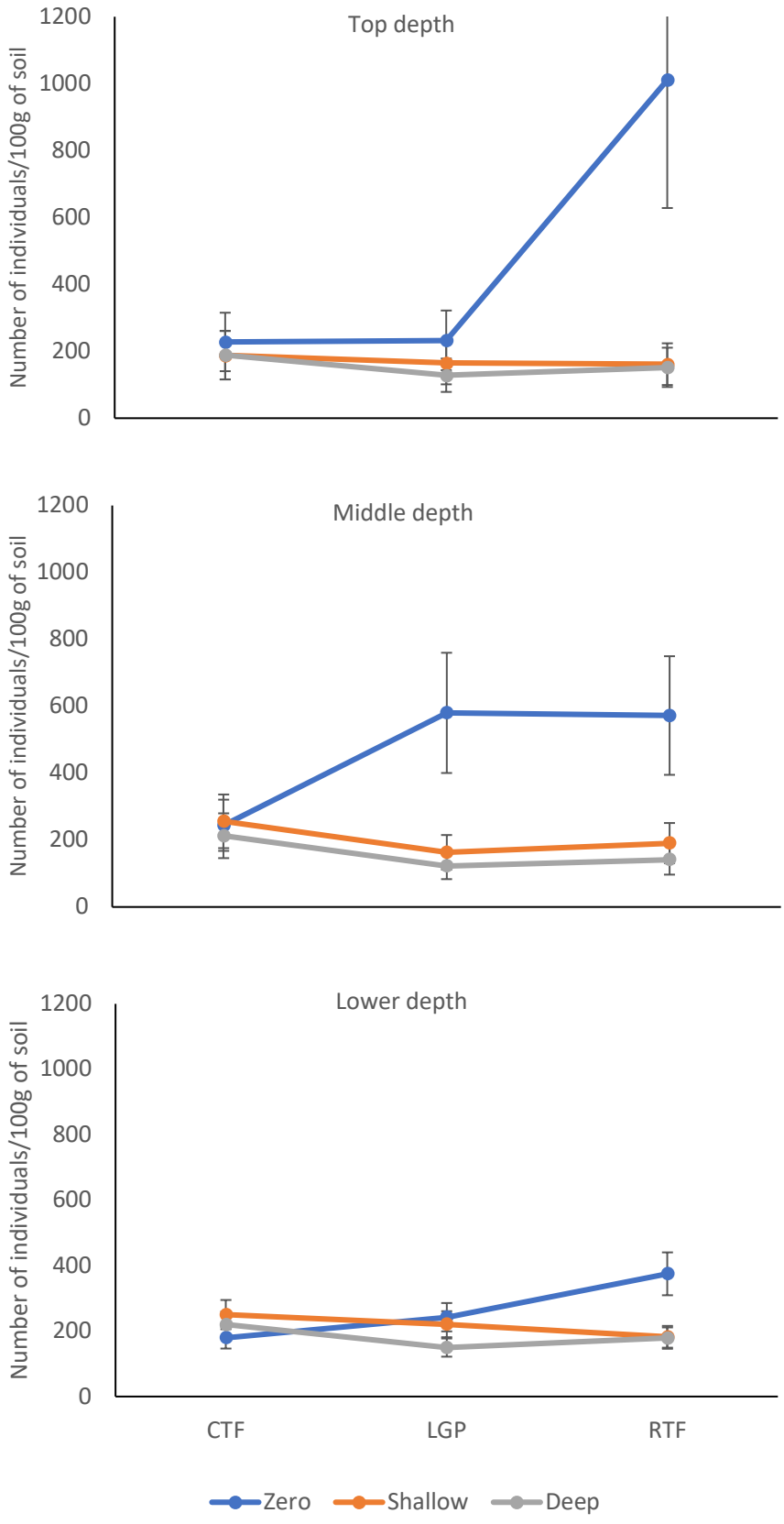

Figure 6.9. Mean abundance of herbivores *across the different tillage and traffic combinations for the top (0-5 cm), middle (5-15 cm) and lower (15-30 cm) depths. Error bars are based on the confidence intervals of the respective means.*

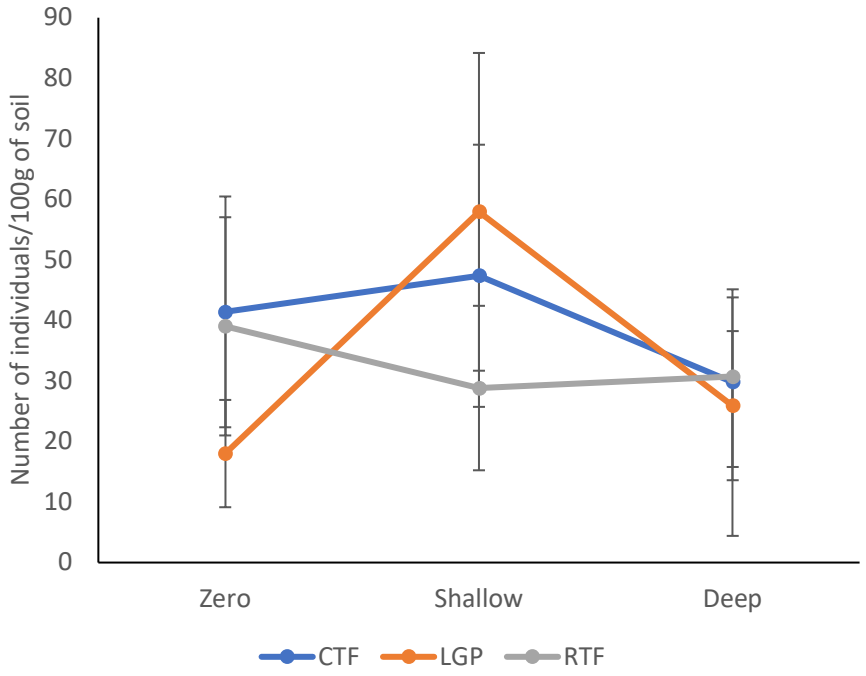

Figure 6.10. Mean abundance of bacterivores across the different tillage and traffic combinations within the lower depth (15-30 cm). Error bars are based on the confidence intervals of the respective mean

6.4 Discussion

The 83 taxa recovered across the 108 samples analysed was high in comparison with the number reported by Okada and Harada (2007) from a soybean field, who incidentally analysed the same number of samples as in this study and recovered 51 taxa. Also, with regards to incidence of specific taxa across samples, only about 20% of the recovered taxa occurred across 50% or more of the samples analysed. This low incidence of the taxa across samples was not surprising since according to Ettema (1998) species similarity across samples is usually very low. In fact, the 20% observed here was high compared to the 6% observed by Price and Siddiqi (1994) and 3% observed by Orr and Dickerson (1966) who analysed 150 and 240 individual cores, respectively. This discrepancy was probably due to the fact that while samples in this study consisted of 15 cores combined to form one sample, in the above-mentioned studies, samples consisted of single cores. Johnson *et al.* (1972) combined 50 cores into single samples and discovered that up to 50% of the taxa occurred in more than 50% of the samples, suggesting, therefore that the more the cores constituting the samples, the more similar the samples are likely to be in terms of diversity and the more representative the sample can be of the soil condition.

In this study, there was a significant interaction between tillage and traffic with respect to total abundance within the top and lower depths suggesting that the impact of traffic on total abundance within these depths depended the type of tillage practiced. And while traffic, particularly RTF, enhanced total abundance of nematodes under zero tillage system, this effect was not observed when the soil was tilled. The effect of tillage as a main factor on total abundance of nematodes was mostly evident within the middle depth. incidentally, it was within this depth that the highest total abundance was recorded. Both shallow and deep-tillage significantly reduced total number of nematodes. This finding contradicts a few previous studies that found no apparent response of total nematode numbers to tillage treatments (Wardle, 1995; Lenz & Eisenbeis, 2000; Sanchez-Moreno *et al.*, 2006; Okada & Harada, 2007). It is also in contrast with the Liphadzi *et al*. (2005) and Parmelee and Alston (1986), who found an overall increase in nematode abundance under tilled conditions. This, however, they attributed to the predominance of bacterivores and fungivores, both of which tend to thrive when the soil is disturbed. The reason for the high abundance of nematodes in the zero-tillage treatment in this study was the occurrence of *Meloidogyne* in remarkably high numbers, especially in the zero-tillage and RTF combinations. The different traffic systems affected the total abundance of nematodes within the lower depth. This finding conflicts with Bouwman and Arts (2000) who found no difference in the total abundance of nematodes between low traffic load (4.5 t) and high traffic load (14.5 t) plots after four years of repeated application of compaction

treatments on a non-grazed grassland. However, at the end of the fourth year, they did find a shift in the composition of the feeding groups towards one that is dominated by herbivores but with a reduced number of bacterivores, predators and omnivores. Although they used specific weight of the soil–which was not the done in the current study, the level of compaction the different loads caused correlated with the weights in the same way compaction increased with the intensity of traffic system used.

It has been established that different taxa exhibit varied responses to soil perturbation (Sanchez-Moreno *et al.*, 2006). In samples taken from the top 5cm of the soil, both tillage systems caused inhibition of all the trophic groups except bacterivores and omnivores. In these two groups, shallow-tillage caused a mild stimulation while deep-tillage resulted in a mild inhibition. The population of herbivores in this study was dominated by *Meloidogyne*, particularly in the zero-tillage samples. The significant inhibition of the herbivores by the deep and shallow-tillage systems observed across all three depths were largely due to the response of *Meloidogyne*, a phenomenon also observed by Okada and Harada (2007). They attributed it to the destruction of possible weed hosts of these obligate endoparasites by the tillage treatment. *Pratylenchus* was also inhibited by tillage in this study, a finding that disagrees with that reported by Minton (1986), Parmelee and Alston (1986) and Okada and Harada (2007), who observed no significant difference in the number of *Pratylenchus* between no-tillage and conventional tillage. It is, however, supported by (Thomas, 1978) who observed greater number of *Pratylenchus* in no-till plots than in cultivated ones. It must be noted that the population dynamics of these endoparasites is also affected by the availability and type of host plant, which could be the reason for these inconsistencies.

Although most previous studies have reported a significant stimulation in the abundances of bacterivores and fungivores by conventional tillage (Parmelee & Alston, 1986; Lenz & Eisenbeis, 2000; Sanchez-Moreno *et al.*, 2006), a completely opposite response was observed in this study as shown by the index V. Fungivores responded to the deep and shallow-tillage regimes with a decline in numbers within the 0-5 cm and 5-15 cm depths, while bacterivores exhibited only a mild stimulation by shallow-tillage and a similar extent of inhibition by deep-tillage in the 0-5 cm but were inhibited by both shallow and deep-tillage within 5-15 and 15-30 cm depths. It has often been assumed that the increase in bacterivores and fungivores associated with tillage had to do with the incorporation of organic matter into the soil and the subsequent availability of this to microbes that multiply on it (Sanchez-Moreno *et al.*, 2006). It is this increase in microbes that most consider as being responsible for the increase in these microbial grazers such as bacterivorous and fungivorous nematodes. This decline in microbivores, especially the fungivores could have resulted indirectly from the fungicides applied to control fungi (Records show fungicide

application each year for three years prior to the sample collection and two years following that). It is unlikely that this was the result of a direct effect because that would have affected other trophic groups as well.

A number of complex processes regulate the dynamics of food webs in the soil, rather than just one constant. There could be two theories to explain the decline in microbial grazers observed in this study. Firstly, for the fungivores, the tillage may have caused nematodes and the surface organic matter to be buried deeper in the soil beyond the top 5 cm, and as a consequence were slow to recover due to their relatively long generation times. This would partly explain why the extent of inhibition was much less in 5-15 cm depth, and why there was a stimulation in their abundance within the 15-30 cm depth. Secondly, bacterivores may have, despite being buried deeper, been able to recover quickly to show a mild stimulation in case of shallow-tillage and mild inhibition in the case of deep-tillage due to their shorter generation time (Bongers, 1990). Nonetheless, this mild change to the bacterivore abundance was not significant (Table 6.3). Within the middle and lower depths, their abundance was already low and were less likely to recover as they did in the top depth. To a very small extent, predation could be a reason for their decline within the lower depths due to the observed increase in the number of predators and omnivores. However, there is a question of how the soil properties could have manipulated the occurrence of microbial grazers across the three depths, and this could have been better explained had the soil properties been analysed for each of the depths rather than for the bulk samples.

In addition to the usual assertion that predators and omnivores are sensitive and tend to always decline in disturbed soils (Wardle, 1995), their decline within the top depths of the tilled plots could also be attributed to the reduction in the number of other trophic groups on which they prey. This trend differed within the 5-15 and 15-30 cm depths, where stimulation, rather than inhibition was the general response to tillage. Omnivores may have utilized some organic products, as well as some microbial grazers, buried as a result of tillage. Predators under deep-tillage may also have benefitted from the availability of deeply buried grazers to build their population in the middle and lower depths. That said, this change in abundance was found not to be significant (Table 6.3).

Traffic and its consequential increase in compaction resulted in varying responses from the various feeding groups across the three depths. Bacterivores and omnivores showed inhibition response to LGP treatment but were stimulated under RTF treatment. It is possible that at the depth of 5-15 cm, only RTF could introduce sufficient compaction to consolidate the soil after deep-tillage. The mild stimulation effect observed in the bacterivores and omnivores could have been as a result of the soil consolidation bringing them in close proximity to their microbial food sources and prey.

A few of the community indices responded to the tillage and traffic treatments. They also differed across the three depths of sampling. These indices are based on the structure and composition of the nematode community. They can therefore, be used to make inferences on the condition of the soil (Neher & Darby, 2006). Maturity index (MI) measures the degree of disturbance or ecological succession in the soil. Being an index that is dependent on abundance of herbivores, PPI was influenced by tillage and traffic interaction in a manner similar to how herbivore abundance was. And as with herbivore abundance, PPI increased in response to RTF only under zero tillage. The reason why traffic influence was only observed under zero tillage could be because application of perturbation, as was done with shallow and deep tillage, interrupted with the traffic treatments to even them out. Traffic caused a reduction in MI2-5, albeit only by a very small magnitude. Tillage, on the other hand, did not influence the MI2-5 value significantly. The lack of significant response in terms of community indices is perhaps to do with the pre-existing conditions of agricultural soils. The rare sensitive taxa whose detection dictate the values of indices such as MI and SI are usually absence is most cultivated fields. Their absence, therefore, even in the untilled plots means that tillage does very little to impact these indices that depend on them. A similar finding was reported by (Sanchez-Moreno *et al.*, 2006).

General opportunistic bacterivores which in this study were dominated by *Acrobeloides, Eucephalobus* and *Plectus* exhibited negative association with compaction on the CCA plot. They are known to usually thrive well under disturbed conditions and unlike the enrichment opportunists, do not depend on eutrophication of the soil (Bongers, 1990). This, perhaps, explains their lack of association with minerals in soil. The herbivorous taxa, *Tylenchorhynchus*, *Pratylenchus* and *Trichodorus* associated positively with compaction, hence negatively associated with tillage and any form of soil disturbance. This agrees with many of the previous studies that reported a decrease in herbivores' population by tillage (Thomas, 1978).

The soil factors, NO_3^-, NH_4^+, N, P, K^+ and OM were all associated positively with EI, an index which measures the proportion of the total nematode abundance composed of enrichment opportunistic taxa (Ferris *et al.*, 2001). Channel index (CI), however, was negatively associated with all these soil minerals. Higher CI value is indicative of the fact that the predominant decomposition pathway is the slow fungal-driven channel. Such conditions are usually associated with relatively lower mineral availability in the soil, hence the observed negative association between this index and soil minerals. The two disturbance-indicator indices, MI and SI, responded differently to compaction. Since both indices are higher when the community is dominated by sensitive higher cp taxa, the expectation was that they would position similarly with respect to compaction. As well as

being higher under the LGP and RTF treatments, compaction was also associated more with zero tillage than the other tillage treatments. And therefore, while compaction could be detrimental to these sensitive taxa, they may still be thriving because of zero tillage which inflicts no disturbance on the soil. Being a measure of abundance of the disturbance-tolerant general opportunists of the cp 2 taxa, BI was expectedly associated with low compaction and tillage.

In summary, the results of this study demonstrate the importance of nematodes as indicators of soil condition through the different changes in communities observed as a result of traffic and tillage treatments as well as the fact that nematode taxa with different feeding strategies respond differently to perturbation and/or compaction depending on the depth of the profile examined. For growers, the results presented here can serve as a guide to what management practice inflicts the least damage on the soil and its biodiversity, as well as how some practices can inadvertently keep in check the populations of some parasitic nematode species.

7 Using high-throughput sequence data to predict trophic-group composition and community indices of nematodes

7.1 Introduction

The role of nematodes as indicators of soil quality is a well-researched sub-discipline of nematology (Bongers, 1990; Bongers & Bongers, 1998; Porazinska *et al.*, 1998; Bongers & Ferris, 1999; Porazinska *et al.*, 1999; Ferris & Bongers, 2006; Okada & Harada, 2007). As discussed earlier, nematode community composition and structure provide an excellent tool for determining whether the soil has been disturbed or not (Bongers, 1990). And as demonstrated in chapter 6, this typically entails identifying the different taxa and recording their abundance in soil samples. This information is then used to generate community indices such as the maturity index (MI), plant-parasitic index (PPI) (Bongers, 1990), enrichment index (EI), basal index (BI), structure index (SI) and channel index (CI) (Ferris *et al.*, 2001) of a sample, which if determined correctly allows inferences about the condition of the soil the sample was taken from. The taxonomic aspect of this process has traditionally been carried out using morphological characteristics of nematodes. Except for channel index, which only requires trophic group classifications, all the above-mentioned indices require identifying the individual nematodes to the genus level or at least to the family level (Bongers, 1990; Ferris *et al.*, 2001). Although identifying most nematode taxa to the family level may not present significant difficulty to a non-expert, the sheer number of specimens in each extract and the number of extracts requiring examination may preclude carrying out such study beyond local scales.

The usual workaround to this constraint caused by nematode densities and sample size is to identify only a given number of randomly selected specimens, usually at least 100 (Yeates & Bongers, 1999), from each sample while placing a certain limit to the number of samples taken for each study (Neher & Campbell, 1996). Although this approach of identifying a stipulated number of nematodes also serves as a means of normalising the data, it is a step that should better be done at the statistical analysis stage rather than during the identification because leaving the investigator to decide what specimens to identify can introduce potential bias, which can then lead to false representation of the community. Especially in samples that are taxonomically rich, the sampling size of 100 may not capture the diversity sufficiently. The outcome of this is a study deficient both in terms of scale and ability to capture the full diversity within the community. This can also result in the exclusion of some rare but important taxa in the analysis especially in agricultural fields, where samples can be dominated by one or two plant-parasitic taxa, thus resulting in a very low representation of free-living taxa needed for the computation of indices such as MI, EI, BI, SI and CI. By reducing the resolution of the classification to

trophic groupings, it is also possible to facilitate the identification process without picking only a fixed number of individuals per sample. Despite some favourable arguments put forward to support this approach as providing ecologically and functionally relevant information (Ritz & Trudgill, 1999), adopting it completely precludes the use of most of these indices since they all require at least family level identification.

The use of metabarcoding has the potential to allow for identifying unlimited number of individuals from several samples all in just a single sequencing run. Moreover, because of the high sensitivity of some next generation sequencing (NGS) technology, even those taxa that are represented in low numbers can be detected. Identification to the genus level is routinely obtainable with this technology, assuming all the necessary bioinformatics tools and expertise are available (Porazinska *et al.*, 2009). Therefore, not only can this method increase the speed with which multiple samples are analysed but could also reduce the requirement for expert knowledge on how to identify nematode genera or species while still ensuring highly accurate identification.

Despite the promise that metabarcoding holds with regard to nematode community analysis, there is still one important aspect of it that makes its full implementation thus far highly challenging. All the aforementioned indices used as indicators of soil quality require abundance information for their computation. To fully substitute the traditional method, a new tool needs to detect the taxa present in a sample, and also correctly predict the abundances of these taxa. No study has so far been able to accurately predict relative abundance of taxa in mixed samples of any group of organisms based on their read frequencies from a metabarcoding data (Porazinska *et al.*, 2009; Amend *et al.*, 2010; Edgar, 2017a). Chapters 3 and 4 of this thesis also attest to this lack of correlation between sequence read frequency and relative abundance of the sampled taxa. Due to this, some have suggested that all PCR-dependent assessment of biodiversity should be based solely on presence/absence, and not abundance (Elbrecht & Leese, 2015).

Criticism of the ability of metabarcoding approaches to quantify taxa within mock or real communities have largely been based on specific taxa, species or genera, and how their read frequencies deviate from their relative abundance (Porazinska *et al.*, 2009; Yu *et al.*, 2012) (also Chapters 3 and 5). Although all the indices used to assess nematode communities depend on abundance data (Bongers, 1990; Yeates, 1994; Bongers *et al.*, 1995; Ferris *et al.*, 2001), the required abundance information does not necessarily have to be of the individual taxa. In the case of MI for instance, it is the frequencies of the colonizer-persister groups that are needed for its computation. Similarly, the EI, BI, SI and CI each utilises abundance information at the functional guild level. In simple terms, a functional guild is a group of taxa with the same feeding habit and life history characteristics (Neher, Bongers, & Ferris, 2004). In view of this, there is a possibility that

read numbers obtained from metabarcoding would correlate better at these functional and trophic group levels with their respective actual abundances from morphological identification more than they would at the genera or species level. Additionally, transformation of the read counts into binary (presence/absence) records have been recommended and successfully used as a means circumventing the lack of correlation between read frequencies and relative abundance (Yu *et al.*, 2012; Ji *et al.*, 2013).

This study is based on the experiment described in Chapter 6 on how tillage and traffic treatments influence the structure of nematode communities in the soil. In the current chapter, the objective was to examine to what extent indices obtained using relative read frequencies of taxa correlate with those obtained for the same samples using standard methods of analysis which involved morphology-based identification. Additionally, the assessment of the impact of different tillage and traffic regimes on nematode community according to the metabarcoding data was compared with that according the standard morphological method. Finally, metabarcoding data obtained from samples taken from the same experiment a year later were analysed to determine if there have been any changes in the nematode community after a one-year period (between year four and five of the experiment).

7.2 Materials and methods

7.2.1 Sample preparation

Soil samples were collected in February and March 2016 from the controlled traffic farming experiment at Harper Adams University (52°46.7899'N, 002°25.5236'W). The experiment was set up in four randomised complete blocks of nine treatments. Each treatment is a combination of one of three tillage and one of three traffic systems. The tillage treatments were zero, shallow or deep tillage and the traffic treatments were controlled, random and low ground pressure traffic. Details of the treatment types have been described in Chapter 6. Samples were taken at three different depths (0-5 cm, 5-15 cm and 15-30 cm) for each of the 36 plots, making a total of 108 samples. Nematodes were extracted from sub-samples of 200 g and the extracts concentrated to 10 ml as described in chapter 6.

To determine if there were changes in the nematode community a year later following the sampling carried out from February to March 2016, new samples were taken in February 2017, following the harvesting of spring oats, from the top 0-5 cm from the same plots as in 2016. Samples in 2017 were analysed using only metabarcoding. Sampling was limited to a single depth because of time constraint. The 0-5 cm depth was chosen, since prior morphology-based analysis (Chapter 6) recorded the highest diversity within this profile.

7.2.2 DNA extraction

For the DNA extraction 1 ml subsample was taken from each of the 10 ml sample extracts and stored at -20 °C. From the 1ml subsamples, 100 µl subsamples were used for DNA extraction using the Qiagen DNeasy Blood and Tissue Kit. To each sample, 800 µl of Qiagen ATL buffer and 100 µl proteinase K were added before incubation overnight at 56 °C and shaking at 109 rpm. The genomic DNA extraction followed the manufacturers guide. To maintain the same ratio of proteinase K and other reagents as described in the manufacturer's protocol, five times the recommended volumes of AL buffer and ethanol were added.

7.2.3 PCR and library preparation

Polymerase chain reaction was performed on each sample using the primer pairs NF1 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTGGTGCATGGCCGTTCTTAGTT-3') and 18Sr2b (5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTACAAAGGGCAGGGACGTAAT-3') (Porazinska *et al.*, 2009). The underlined sections of the primer sequences here represent the Illumina adapter sequence. Each reaction tube

contained a total of 25 µl reaction mix consisting of 1X Phusion HF buffer (New England Biolabs, Ipswich, MA, USA), 0.2 mM dNTPs (New England Biolabs), 0.5 µM each of adapter-ligated forward and reverse primers, 1U of Phusion DNA polymerase (New England Biolabs) and 5 µl of template DNA. The PCR programme was set at 98ºC for 2 min, 30 cycles of 98ºC for 20 secs, 66ºC for 45 secs, 72ºC for 45 secs before a final extension step at 72ºC for 5 mins.

Amplicons were purified using Ampure XP Beads (Bechman Coulter, Inc. USA) and quantified using a Thermo Scientific™ Fluoroskan Ascent™ Microplate Fluorometer (Thermo Fisher Scientific, Wilmington, DE, USA). This was then followed by the index PCR step where unique dual indexes and the sequencing adapters were attached to each amplicon using Nextera XT index primers (Illumina inc. San Diego, CA, USA) for amplification (Illumina's 16S Metagenomic Sequencing Library Preparation protocol). The PCR conditions at this stage were: 98ºC for 3 min, 8 cycles of 98ºC for 30 secs, 55ºC for 30 secs, 72ºC for 30 secs and a final extension step at 72ºC for 5 mins. The products were purified again, quantified using the Microplate Fluorometer and pooled according to their molarity. Length of the amplicons in the pooled sample was verified on the Agilent 2200 TapeStation (Agilent Technologies, Santa Clara, CA, USA). The pool was quantified and diluted to 4 nM concentration. Using the Illumina protocol, the pool was denatured by mixing with 0.2N NaOH. 10% denatured PhiX control library was added to the denatured pool to introduce diversity. The mixture was incubated for 2 min at 96ºC and immediately put on ice, before being loaded on a MiSeq machine (Illumina) for sequencing. The sample was sequenced at Fera (York, UK) in a paired-end approach using a V3 run kit and 2 x 300 cycles.

7.2.4 Bioinformatic analysis

The sequence reads were demultiplexed by the MiSeq Reporter software on the MiSeq sequencer into the individual samples based on the paired nextera indexes used, applying default settings. All subsequent analyses were carried out using the USEARCH pipeline (Edgar, 2010). Initial attempts at merging the paired reads using the command *fastq_mergepairs* in USEARCH resulted in a significant amount of the forward and reverse reads failing to merge. Therefore, the merging step was skipped, and the forward reads alone were used for the analysis due to their superior quality over the reverse reads. A similar approach was implemented in Unterseher *et al*. (2016). For each of the samples, reads were first trimmed to 200 bp from the 3' end, while at the same time removing primer sequences with the command *fastx_truncate*. The truncated reads were then used to pass the command, *fastq_filter* which filtered out any read with expected

error of more than one (Edgar & Flyvbjerg, 2015) and at the same time converted the files from fastq to fasta formats. The filtered reads of all samples were then combined into one file. The command, *fastx_uniques* was used to dereplicate the reads. The subsequent steps, ie clustering, taxonomy assignment, and creating otu table followed the method described in section 3.2.8.

The otu table was converted into a format in which the taxa names are arranged in columns and the samples names in rows. Firstly, all nematodes-only otus with their records in each sample were selected and copied onto a separate worksheet in Microsoft Excel. For otus whose assignments did not include genus names, blast search (Zhang *et al.*, 2000) was performed against the NCBI reference database on 5th February 2018 to confirm their identities at the genus level or species if possible. After this, reads of otus with the same species names were lumped so that each species name is unique.

The steps involved in obtaining the species table for the standard morphological approach are detailed in Chapter 6. In short, from 1 ml subsample of each sample extract, all individuals were identified to the genus level, although for some individuals, only family-level identification was possible. The number of each taxon (genus or family) in each sample was recorded in a species table similar to one described for the metabarcoding approach.

7.2.5 Statistical analysis

For calculating community indices, both the metabarcoding species table and the morphology species table were analysed in three different formats. First analysis involved computing all indices and trophic classifications using the unrarefied data (raw data). The second analysis involved the computation of these indices and trophic classifications on rarefied data from both the metabarcoding and morphological identification. The standard and metabarcoding datasets were rarefied to 50 individuals and 100 reads per sample, respectively. The third analysis used presence-absence data. The three different analyses were performed to determine if the format of the data had any impact on the comparison. In each of these analysis classes, difference in means of metabarcoding- and morphology-based indices were tested using the Welch's t-test (Welch, 1951) in RStudio development environment for R (RStudio Team, 2016).

All computations of nematode indices, trophic group classifications and food web diagnostics were performed using NINJA (Sieriebriennikov, Ferris, & de Goede, 2014), an online tool for nematode faunal analyses. The indices obtained for all samples were added to the species table together along with some measured environmental variables. Using NINJA, c-p triangle (De Goede, Bongers, & Ettema, 1993a), which graphically

presents nematode community structure based on coloniser-persister (c-p) classifications of the community were also obtained.

Difference in nematode community compositions across different treatments and soil depths were demonstrated using non-metric multidimensional scaling (NMDS) ordination of Jaccard dissimilarity metrics. Most of the analysis involving NMDS were performed following the workflow described in (Ji *et al.*, 2013). For this procedure alone, presence-absence data from the metabarcoding was compared with the unrarefied data from the standard morphological data in order to ensure this analysis conforms with the procedure described by Ji *et al*. (2013). An NMDS plot comparing the presence-absence metabarcoding data with the rarefied standard data instead of the unrarefied one did not change the plot. The decision to use the unrarefied standard data was, therefore, to avoid loss of data that may result from the rarefaction. The function *vegdist* of the vegan package (Oksanen *et al.*, 2015) was used to calculate distances, and *metaMDS* for performing NMDS. For the metabarcoding data alone, the parameter 'binary' in *vegdist* was set to true because the data was in a binary (presence/absence) format. Correlations between the metabarcoding and morphological data were performed using Mantel and Procrutes correlation tests in *vegan* as implemented in the functions *mantel* and *protest*, respectively.

Effects of treatments on community compositions were analysed using *mvabund* (Wang *et al.*, 2012) package. Differences in the effects of treatment levels were calculated using *summary.manyglm*, and *p.adjust* function based on the 'fdr' method (Benjamini & Hochberg, 1995) was used to correct for multiple testing. The 'fdr' method controls the expected proportion of false discoveries amongst the rejected hypotheses.

7.2.6 Comparisons between 2016 and 2017 samples

Sequence files for the 0-5 cm depth of the samples from 2016 and those collected in 2017 were combined and analysed using USEARCH. Analyses inside USEARCH were carried following the same steps, using the same procedure as described in the section, 7.2.4. Species tables for the two sampling years were separated before performing statistical analysis. A possible outcome when data are separated in this manner is the occurrence of species with zero records for all samples in one part of the data because they only occurred in the other. Therefore, an additional code was added to remove species with zero records in all samples in an individual year. Distances were calculated using *vegdist* for this comparison. Treatment effects on community compositions were also calculated using *summary.manyglm* function with *p.adjust* to correct for multiple testing as was done in the previous section.

7.3 Results

7.3.1 Recovered taxa

Standard morphological analysis recovered 83 unique taxa identified to the genus level, including a few taxa that were identified only to the family level for individuals whose genus identification could not be achieved. The metabarcoding data set yielded 194 otus at 97% sequence similarity that were assigned 55 unique species names constituting 41 different genera. Although the 97% similarity cut-off produced otus in excess of the number of species recovered by taxonomy assignment, using lower cut-offs failed to recover some of the taxa. And since taxonomic identities to at least family level was needed for nematode community indices and trophic classifications, taxonomic diversity instead of otus richness had to be used.

7.3.2 Trophic groupings

Apart from the dominance of herbivorous nematodes in the metabarcoding data set, changes in the relative abundance of the three dominant trophic groups (herbivores, bacterivores and fungivores) followed similar patterns in both standard and metabarcoding data sets (Figure 7.1). Both data sets also revealed the numerical dominance of the bacterivorous and herbivorous nematodes under all treatments. Across the traffic treatments, both data sets showed minimal to non-existent difference in the distribution of the trophic types. Although predacious and omnivorous nematodes had low representations in both data sets, this was more extreme in the metabarcoding data where they were rarely recovered from the samples.

7.3.3 Nematode community indices

Two terminologies used in this section and the subsequent ones need clarification. The term data format is used to refer to unrarefied, rarefied or presence-absence data while data type refers to the standard morphological or metabarcoding data. For both the standard and metabarcoding data types, indices were calculated from three different data formats: the unrarefied data, data rarefied to 100 and 50, respectively and binary data. Attention was focussed only on MI and MI-derived indices alone, namely, MI, MI2-5 PPI and Sigma MI. Structure, channel and basal indices between the two data types were significantly different (data not shown). Of the three data formats, the binary data revealed the most deviations between indices from the standard and metabarcoding data. By using rarefied or unrarefied data, most of the indices were not different between the two data types (Table 7.1). Plant-parasitic index was the only index that showed no significant difference between the two data types across all data formats. Maturity index 2-5 varied

significantly between the two data types for all treatments in at least one data format. There was generally more evidence to support the fact that indices computed from the three data types did not differ than there was to support otherwise.

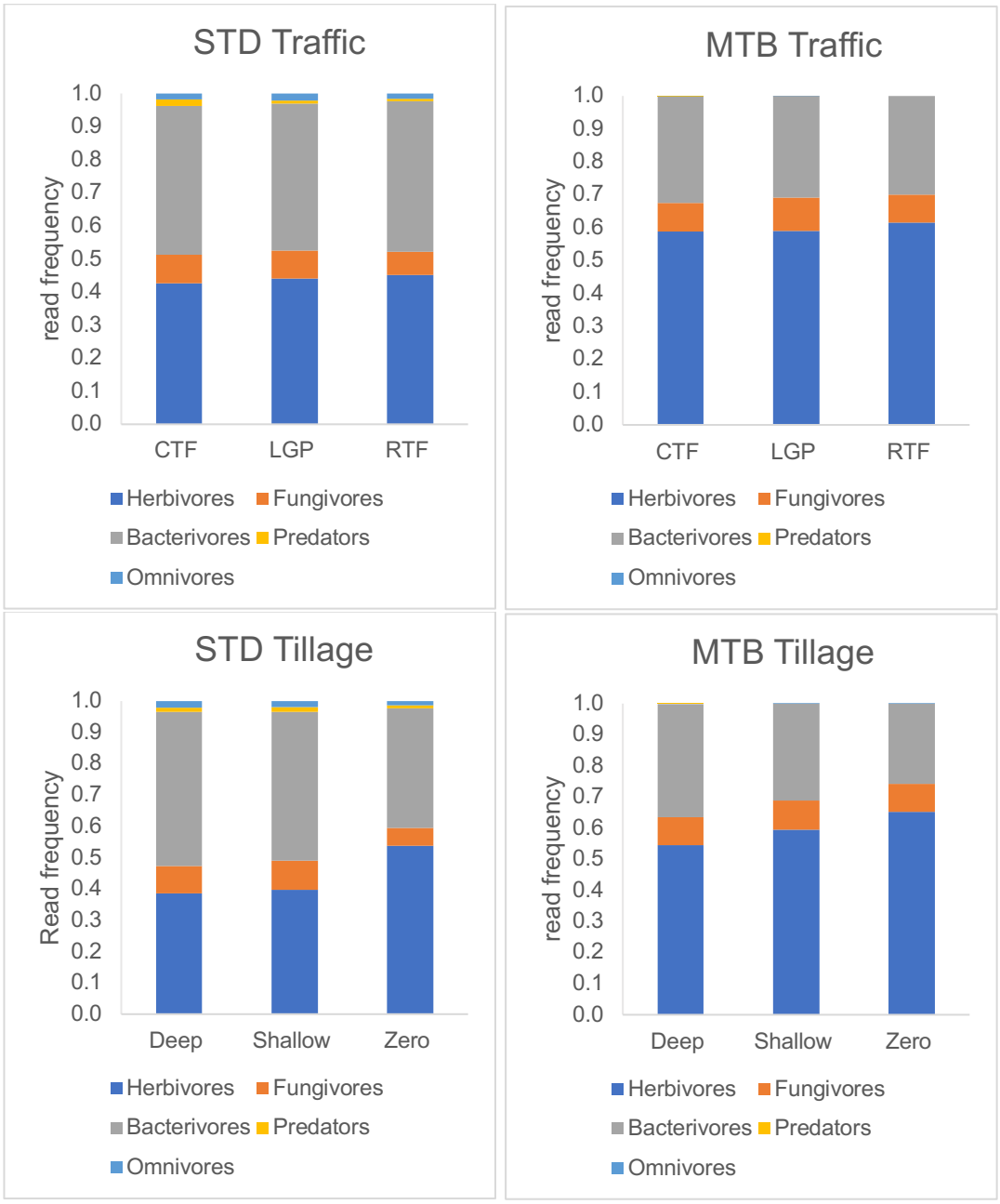

Figure 7.1. Nematode trophic groups' distribution across traffic and tillage treatments from standard and metabarcoding data sets. STD represents standard data set and MTB represents metabarcoding data sets.

Table 7.1. Comparison between MI-family indices calculated from standard (MOR) and metabarcoding (MTB) data types from raw (UNR), rarefied (RAR) and the binary data (PrAb) formats under different traffic and tillage treatments. Numbers are means calculated for each treatment group. Insignificant P *values are highlighted in grey.*

	Index	Unrarefied		Rarefied		Pres-Abs		Welch's t-test *P* value		
Traffic										
CTF		MTB	MOR	MTB	MOR	MTB	MOR	UNR	RAR	PrAb
	MI	1.75	1.99	1.79	2.03	1.77	2.43	0.100	0.163	<0.001
	MI2-5	2.05	2.62	2.00	2.62	2.06	2.73	0.003	<0.001	<0.001
	PPI	2.82	2.63	2.75	2.65	2.81	2.81	0.132	0.472	1.000
	SigMI	2.41	2.22	2.49	2.24	2.31	2.57	0.255	0.297	0.001
LGP										
	MI	1.86	1.86	1.76	1.86	1.89	2.22	0.978	0.735	0.060
	MI2-5	2.12	2.36	2.13	2.63	2.12	2.56	0.018	0.222	0.018
	PPI	2.68	2.66	2.66	2.66	2.71	2.75	0.898	0.984	0.610
	SigMI	2.32	2.19	2.33	2.19	2.36	2.46	0.172	0.159	0.153
RTF										
	MI	1.71	1.76	1.74	1.74	1.70	2.31	0.399	1.000	0.001
	MI2-5	2.00	2.49	2.00	2.46	2.00	2.66	0.005	0.002	0.005
	PPI	2.75	2.68	2.86	2.66	2.72	2.68	0.424	0.222	0.269
	SigMI	2.39	2.11	2.43	2.10	2.31	2.45	0.201	0.154	0.045
Tillage										
Deep										
	MI	1.83	1.92	1.90	1.95	1.80	2.32	0.617	0.827	0.003
	MI2-5	2.07	2.51	2.13	2.50	2.04	2.65	0.008	0.093	0.001
	PPI	2.73	2.66	2.72	2.65	2.72	2.76	0.700	0.701	0.631
	SigMI	2.42	2.14	2.50	2.15	2.32	2.48	0.184	0.159	0.049
Shallow										
	MI	1.77	1.89	1.80	1.89	1.80	2.37	0.148	0.303	0.004
	MI2-5	2.07	2.48	2.00	2.48	2.08	2.66	0.021	0.022	0.002
	PPI	2.78	2.62	2.82	2.62	2.72	2.73	0.133	0.317	0.825
	SigMI	2.31	2.12	2.32	2.12	2.34	2.51	0.107	0.242	0.048
Zero										
	MI	1.72	1.80	1.59	1.79	1.75	2.28	0.253	0.268	0.038
	MI2-5	2.04	2.47	2.00	2.46	2.06	2.64	0.038	0.041	0.021
	PPI	2.74	2.70	2.73	2.71	2.80	2.74	0.578	0.841	0.480
	SigMI	2.39	2.25	2.43	2.27	2.32	2.48	0.290	0.230	0.039

Hypothesis testing to determine the impact of tillage and traffic treatments on the indices using Kruskal-Wallis test revealed some similarities between the two data types. Traffic was the only treatment that had significant effect and it affected only the MI25. This effect was presented as significant only in the standard data set ($P = 0.015$). The other indices showed no response to any of the treatments.

7.3.4 Depiction of disturbance conditions

Using the c-p triangle to visualise a snapshot of level of stress in the soil, the two data types gave different depictions of the nematode communities (Figure 7.2). While both data types mostly indicated communities with low stability, this condition was highly exaggerated in the metabarcoding data where only very few communities were situated

away from the zero-stability border. This was even more so in the rarefied data which not only showed no stability but also depicted most of the communities as being highly stressed. The use of the binary (presence-absence) data for the c-p triangle gave the best representation of the community through metabarcoding when compared with the rarefied or unrarefied standard data but not with the binary-formatted standard data.

7.3.5 Beta diversity

Differences between nematode communities as a result of tillage and traffic treatment as well as depths of sampling across field compaction gradients were depicted similarly by the non-metric multidimensional scaling (NMDS) ordinations in both the standard and metabarcoding data types. There was no clear separation of the treatment levels in the NMDS ordinations for both data types (Figure 7.3 A and B). The ordinations involving the different depths, on the other hand, reveal clear separations again in both the standard and metabarcoding data types (Figure 7.4 A and B). The Procrustes tests of correlation between ordinations of the standard and metabarcoding data sets was highly significant (r = 0.42, P = 0.001). Between their dissimilarity matrices, the Mantel test also showed a highly significant correlation (r = 0.18, P = 0.001).

Tests of significance of the treatments using *summary.manyglm* test showed that in both data types, treatments had no effect on the nematode community. But as was evident from the NMDS ordinations, difference in the community structure across the sampling depths were highly significant in both the standard data (0-5 cm vs 5-15 cm: Wald values = 23.99; P < 0.001, 0-5 cm vs 15-30 cm: Wald values = 29.30; P < 0.001) and metabarcoding data (0-5 cm vs 5-15 cm: Wald values = 11.57; P < 0.001, 0-5 cm vs 15-30 cm: Wald values = 10.07; P < 0.001).

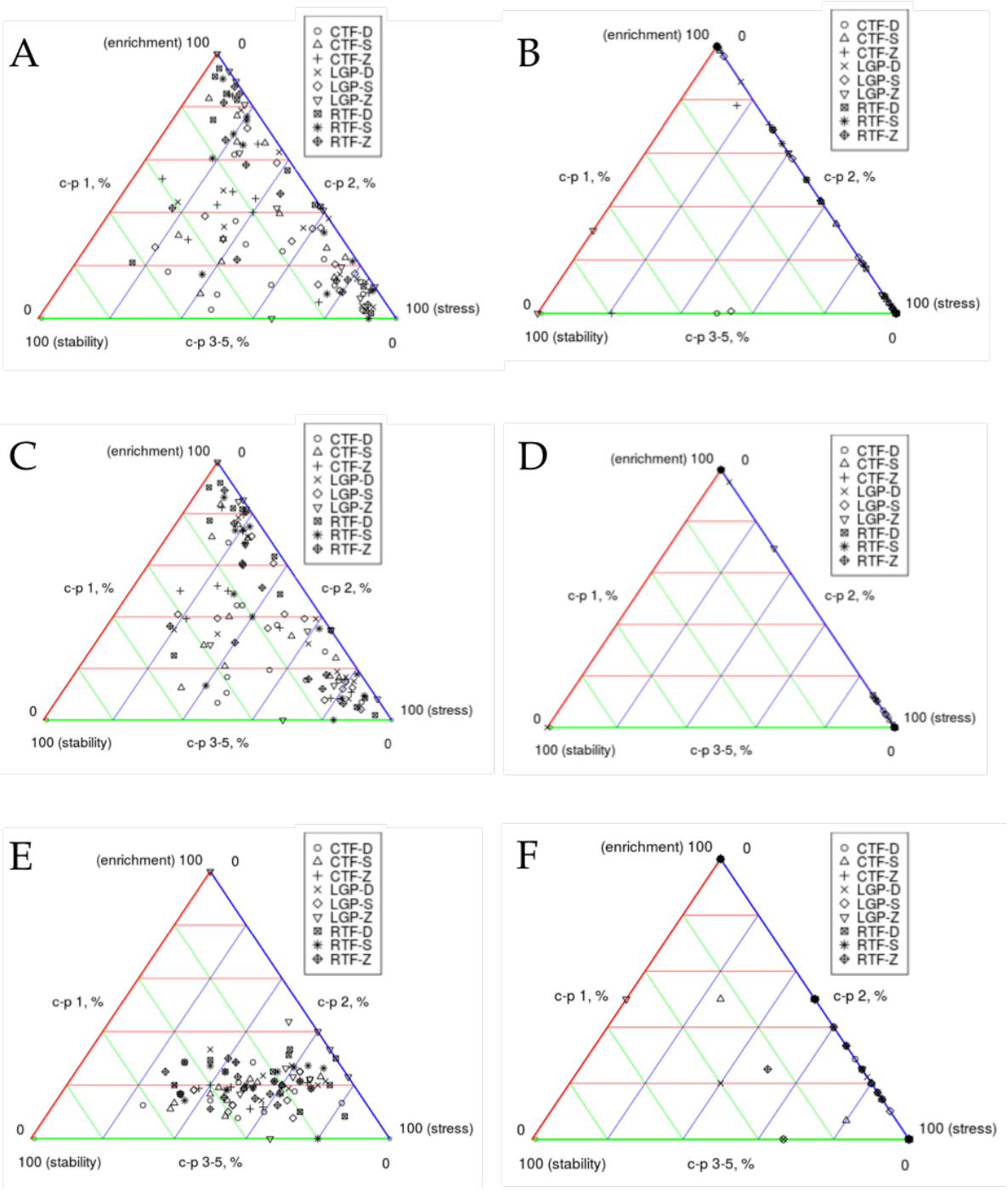

Figure 7.2. Graphical representation of nematode communities using c-p triangles. Each marker represents one of the 108 samples. Samples within the same treatment categories have identical marker shapes. A and B depict the status of the communities based on unrarefied data of the standard and metabarcoding approaches, respectively. C and D depict the status of the communities based on rarefied data of the standard and metabarcoding approaches, respectively. E and F depict the status of the communities based on presence-absence data of the standard and metabarcoding approaches, respectively

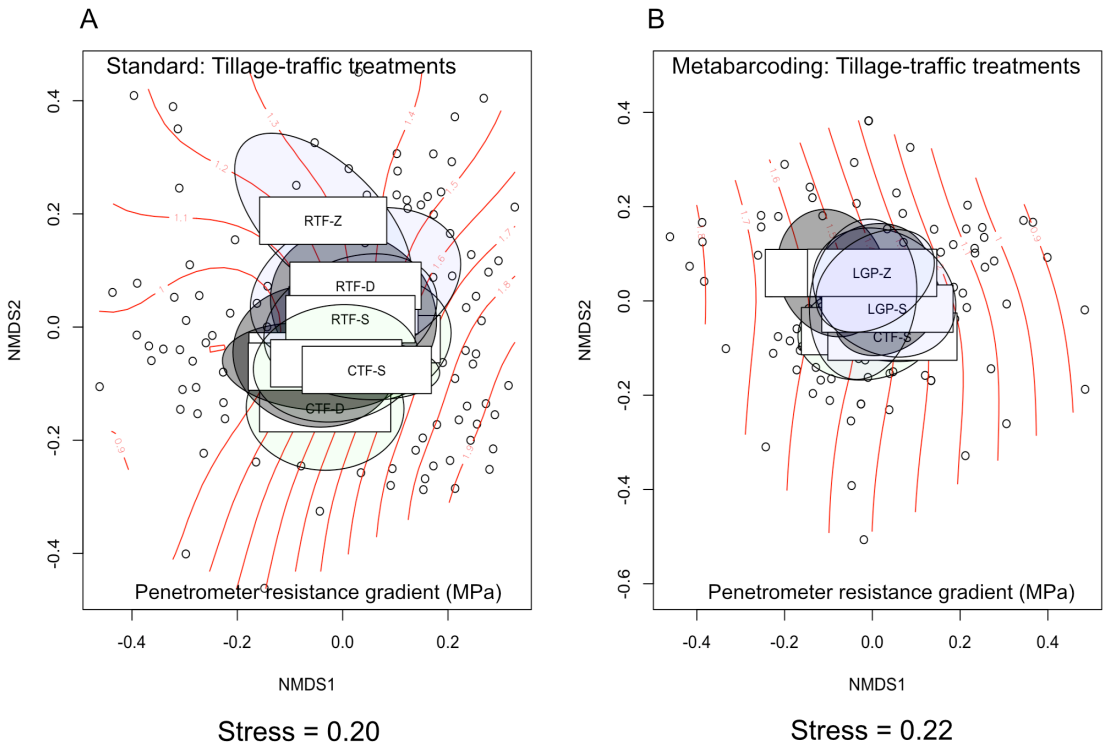

Figure 7.3. Non-metric multidimensional scaling (NMDS) ordinations based on morphology (A) and metabarcoding (B) data for samples collected in 2016. Points are experimental plots, and coloured ellipses are 95% confidence intervals of species centroids for each treatment (ellipses generated using the function, 'ordiellipses' inside vegan). Plots along the same line segments have the same level of compaction. Colour assignments to ellipses in the tillage-traffic treatments ordination is based on the different traffic systems.

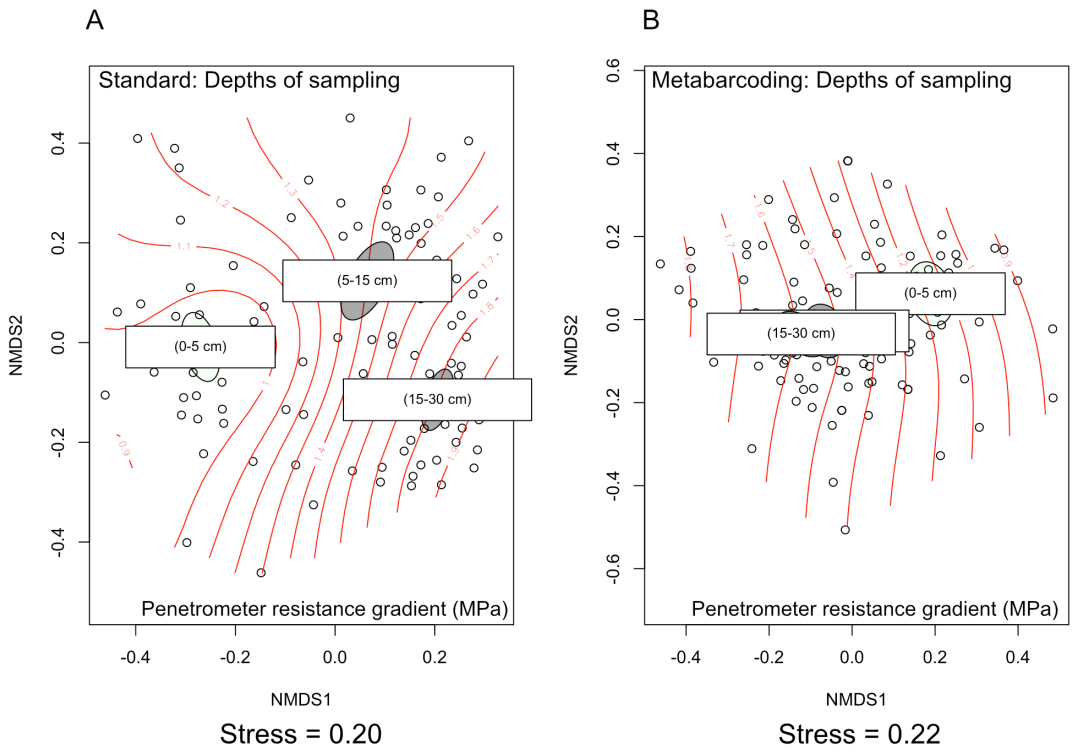

Figure 7.4. Non-metric multidimensional scaling (NMDS) ordinations based on standard morphological (A) and metabarcoding (B) data for samples collected in 2016. Points are experimental plots, and coloured ellipses are 95% confidence intervals of species centroids for each depth (ellipses generated using the function, 'ordiellipses' inside vegan). Plots along the same line segments have the same level of compaction. Colour assignments to ellipses in the sampling depths ordination is based on the different depths of sampling (A and B)

7.3.6 Seasonal difference

Between the sampling times, most indices did not change significantly under the different tillage and traffic treatment conditions (Table 7.2). Plant-parasitic index (PPI) changed significantly and it did so under most treatments. In all cases where PPI changed, it increased from years 2016 to 2017. Apart from the PPI, MI and SigMI were the only other indices that changed significantly from 2016 to 2017. While MI decreased under deep tillage in the one-year period, SigMI did so under zero tillage.

Hypothesis testing of treatment effects on the indices using Kruskal-Wallis test showed that neither tillage nor traffic had any effect on the indices in the 2016 samples. However, there was a significant difference in the median MI of the tillage treatments in the 2017 (P = 0.023). The Dunn test (Dunn 1964) for multiple comparison showed that the difference in median was between deep and shallow tillage alone, with the latter having the higher value. Sigma MI and PPI were only marginally affected in 2017 by traffic (P = 0.073) and tillage (P = 0.077), respectively.

Table 7.2. Comparison between MI-family indices of the two metabarcoding data obtained from sampling years, 2016 and 2017 under different traffic and tillage treatments. The data for the two sampling years were both rarefied to 100 reads per sample. Numbers are means calculated for each treatment group.

	Index	Sampling Year		Welch's t-test *P* value
		2016	2017	
Traffic				
CTF				
	MI	1.94	1.67	0.150
	MI2-5	2.00	2.10	0.339
	PPI*	2.62	4.25	<0.001
	SigMI	2.44	1.80	0.055
LGP				
	MI	1.69	1.77	0.630
	MI2-5	2.00	2.14	0.168
	PPI	2.85	3.55	0.074
	SigMI	2.32	2.24	0.769
RTF				
	MI	1.92	1.72	0.442
	MI2-5	2.00	2.22	0.217
	PPI*	2.25	3.38	0.012
	SigMI	2.12	1.96	0.613
Tillage				
Deep				
	MI*	1.83	1.43	0.013
	MI2-5	2.00	2.12	0.151
	PPI*	2.33	4.06	<0.001
	SigMI	2.17	1.92	0.449
Shallow				
	MI	1.80	1.98	0.451
	MI2-5	2.00	2.17	0.336
	PPI*	2.67	3.89	0.007
	SigMI	2.30	2.14	0.676
Zero				
	MI	1.92	1.76	0.421
	MI2-5	2.00	2.17	0.166
	PPI	2.73	3.23	0.138
	SigMI*	2.42	1.95	0.025

* statistically significant (*P* = 0.05)

7.3.7 Beta diversity across seasons

Non-metric multidimensional scaling ordinations on Jaccard binary distance matrices for the two sampling periods, although similar, showed no difference in community composition across the different treatment levels (not shown). Hypothesis testing using the *summary.manyglm* in the R package *mvabund* showed that treatments had no significant effect on the community composition of the top 0-5 cm profile in both 2016 and 2017 (not shown). The c-p triangle showed a slight increase in stability of some communities in the period between the two years (Figure 7.5). Surprisingly, some of these plots were ones that actually received disturbance treatments such as traffic and tillage.

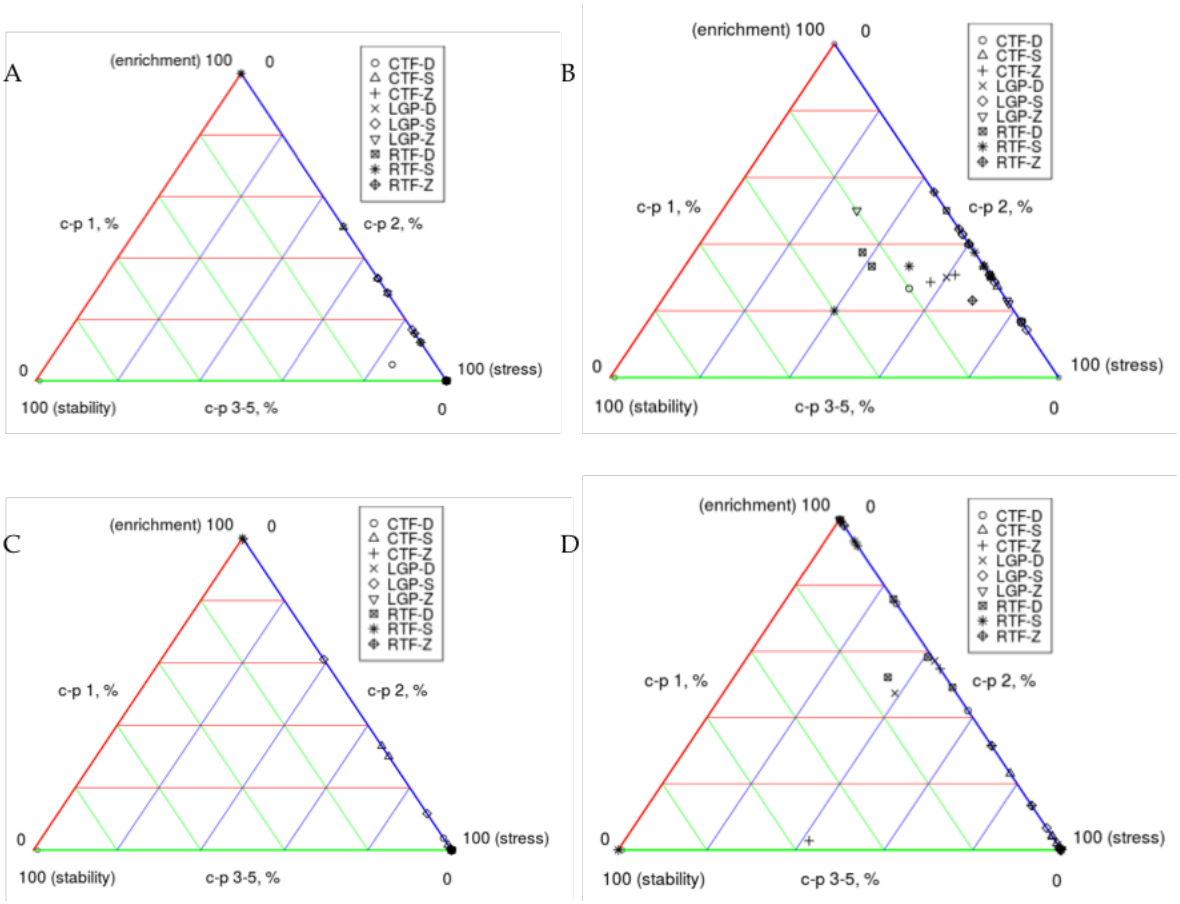

Figure 7.5. Graphical representation of the change in nematode communities in an interval of one year using c-p triangles. Each marker represents one of 36 samples collected within the top 0-5 cm depth in the years, 2016 and 2017. Samples within the same treatment categories have identical marker shapes. A and B depict the status of the communities based on presence-absence data for the 2016 and 2017 samples, respectively. C and D depict the statuses of the communities based on unrarefied data for the 2016 and 2017 approaches, respectively.

Table 7.3. Abundance of herbivorous and bacterivorous nematodes recovered from the three depths according to the unrarefied data. Values within the rows labelled herbivores and bacterivores are mean and standard errors of percentages of the total reads associated with herbivores and bacterivores, respectively. Values within rows of taxa are actual read numbers associated with each of those taxa. Test of statistical significance was performed using robust ANOVA as implemented in the R function, t1waybt *within the WRS2 package* (Wilcox, 2012)*.*

	Metabarcoding				**Standard**			
	0-5 cm	5-15 cm	15-30 cm	*P* values	0-5 cm	5-15 cm	15-30 cm	*P* values
Herbivores (%)	36.69 ± 7.11	62.98 ± 6.40	79.69 ± 4.77	<0.01	35.19 + 3.10	35.37 ± 3.15	61.39 ± 2.91	< 0.001
Pratylenchus	0 ± 0	1261.50 ± 1258.30	429.39 ± 403. 25	NA	2.17 ± 0.30	3.14 ± 0.48	2.58 ± 0.36	0.447
Tylenchus	0 ± 0	453.31 ± 452.76	0.50 ± 0.33	NA	1.22 ± 0.24	3.72 ± 0.55	8.17 ± 1.24	< 0.001
Paratylenchus	36.19 ± 35.88	48.638 ± 24.578	2568.94 ± 1207.82	NA	7.47 ± 1.52	5.77 ± 0.99	13.47 ± 1.46	< 0.001
Bacterivores (%)	52.67 ± 7.15	27.28 ± 5.83	13.42 ± 4.12	<0.01	50.62 ± 2.64	54.47 ± 3.21	29.86 ± 3.09	< 0.001
Rhabditis	5.94 ± 3.22	725.06 ± 389.89	1017.08 ± 696.66	0.427	5.64 ± 0.86	1.14 ± 0.64	0.03 ± 0.03	NA
Eucephalobus	3658.81 ± 1815.22	1151.56 ± 1016.83	1.05 ± 0.40	0.643	11.06 ± 1.14	0.53 ± 0.13	3.36 ± 0.48	< 0.001

Table 7.4. Abundance of herbivorous and bacterivorous nematodes recovered from the three depths according to the rarefied data. Values within the rows labelled herbivores and bacterivores are means and standard errors of percentages of the total reads associated with herbivores and bacterivores, respectively. Values within rows of taxa are actual read numbers associated with each of those taxa. Test of statistical significance was performed using robust ANOVA as implemented in the R function, t1waybt *within the WRS2 package* (Wilcox, 2012)*.*

	Metabarcoding				**Standard**			
	0-5 cm	5-15 cm	15-30 cm	*P* values	0-5 cm	5-15 cm	15-30 cm	*P* values
Herbivores (%)	36.91 ± 9.38	60.50 ± 9.58	81.04 ± 6.37	0.018	35.27 + 3.49	35.27 ± 3.20	60.19 ± 3.06	<0.001
Pratylenchus	0 ± 0	5.00 ± 5.00	9.04 ± 5.86	NA	1.00 ± 0.14	1.11 ± 0.23	1.74 ± 0.34	0.109
Tylenchus	0 ± 0	0.17 ± 0.17	2.40 ± 2.40	NA	0.70 ± 0.24	1.17 ± 0.26	6.13 ± 0.74	<0.01
Paratylenchus	0.35 ± 0.31	9.70 ± 6.57	23.39 ± 7.91	NA	2.78 ± 0.55	2.11 ± 0.43	8.87 ± 0.97	<0.001
Bacterivores (%)	31.65 ± 8.17	23.14 ± 8.06	37.17 ± 8.61	0.315	50.61 ± 3.06	55.22 ± 3.37	29.61 ± 3.39	<0.001
Rhabditis	0.30 ± 0.23	8.60 ± 5.17	7.30 ± 4.61	NA	3.00 ± 0.62	0.38 ± 0.16	0 ± 0	NA
Eucephalobus	15.91 ± 6.10	7.45 ± 4.90	0.04 ± 0.04	NA	5.42 ± 0.64	1.27 ± 0.41	0.26 ± 0.10	NA

7.4 Discussion

The results of this study demonstrate the possibility of determining the condition of a soil through metabarcoding of nematode communities in the soil. This was true for rarefied, unrarefied and to some extent the binary community data. It also showed that metabarcoding data can be expected to provide a reliable representation of the difference in community composition between treatments, time periods or possibly locations. Correlation between the morphological data and the binary-transformed metabarcoding data was highly significant with the correlation coefficient *r*, based on *Mantel* and *Procrustes* tests showing low and medium effect sizes, respectively.

The standard morphological data set produced a much higher diversity in terms of the number of unique taxa recovered than the metabarcoding data set did. The opposite would have been the case had diversity for the latter been based on the number of unique otus as is often done with metabarcoding data (Ji *et al.*, 2013). It was necessary to utilize the taxonomic identities of the otus in this study in order to correctly assess how the two data sets describe disturbance conditions in the nematode communities. The relatively low taxonomic coverage of the metabarcoding data set concurred with some previous studies that showed that metabarcoding is not always able to detect all resident taxa in samples (Cowart *et al.*, 2015; Zimmermann *et al.*, 2015). In Zimmermann *et al.* (2015), it was only at the genus level where fewer taxa were recovered by the metabarcoding method (28 versus 30). At the species level, over twice the number recovered by the standard approach were recovered by the metabarcoding method (265 versus 102). In Cowart *et al*. (2015), the metabarcoding approach based on both the COI and 18 rDNA could only recover 36% of the morphologically-determined diversity. This may be for a number of reasons ranging from primer mismatch to dominant taxa leading to the masking of rare taxa from the data.

The importance of nematode trophic groups to soil processes such as decomposition and primary production pathways has been well documented (Beare *et al.*, 1992; Wardle *et al.*, 1995). This study showed that metabarcoding data sets could depict how anthropogenic disturbances such as tillage affects their relative abundance in the soil based on the relative abundance of the various trophic groups. The similarity was strong, particularly for the dominant groups, the herbivores and bacterivores. Lejzerowicz *et al.* (2015), in their comparison of morphological and metabarcoding-based assessment of metazoans in benthic environments, also observed such high congruence in the abundance of the dominant taxa.

Sufficient molecular data exists that supports the phylogenetic origins of most of the current nematode classifications (Blaxter *et al.*, 1998; Holterman *et al.*, 2006; van Megen *et al.*, 2009). The same studies have demonstrated that features such as herbivory may

have arisen multiple times independently in the course of evolution. Similar evidence exists for at least fungivory and predation. One of the things these independent evolutions of feeding behaviour may point to is the discrepancies between trophic classification and taxonomic classification, at least based on the 18S rDNA region. This may therefore explain why metabarcoding data based on the same region was able to correctly quantify trophic groups even though the contrary has largely been reported when it comes to quantifying actual taxa in communities. (Porazinska *et al.*, 2009; Amend *et al.*, 2010; Edgar, 2017a).

Much like the trophic classification, the colonizer-persister groupings and functional guilds used to determine nematode indices do not depend on the phylogenetic positions of their constituent taxa, but rather the life history characteristics. And the same as it was for the trophic compositions, almost all indices obtained using metabarcoding data showed no deviation from those obtained from the standard data set. The implications of this are that most information needed to understand the soil status can reliably and fairly accurately be obtained from metabarcoding even if the same data fails to quantify the actual taxa.

The c-p triangle presents graphically three aspects of the condition of the soil, how enriched, stressed or stable it is based on the proportion of the community constituted by c-p1, c-p 2 or c-p 3-5 nematodes, respectively (De Goede *et al.*, 1993a). Although the main goal of c-p triangles is to demonstrate patterns in community change over time, they can also give a snapshot of the state of soil at any point in time. The two data sets in this study provided different depiction of the nematode communities. The format of the data used [rarefied, unrarefied or presence-absence] did not matter much. The metabarcoding data showed communities highly dominated by c-p 1 and c-p 2 nematodes, thus either highly enriched or stressed. Overrepresentation of a few dominant taxa in sequence reads can occur in metabarcoding data (Sogin *et al.*, 2006); and this may explain the exacerbated stress condition in the metabarcoding data.

The only significant factor according to the NMDS ordination of dissimilarity matrices was sampling depth and both the standard and metabarcoding data sets demonstrated this. In a previous study that looked at the difference between nematode communities within different depths, Sanchez-Moreno et al. (2006) also found that some taxa were significantly higher in some parts of the profile than they were in others. They observed a significantly higher number of bacterivores and fungivores, particularly those belonging to the c-p 1 and c-p 2 groups in the top soil layers (0-15 cm) of the profile. The lower depths had significantly higher number of herbivores such as *Pratylenchus* and Tylenchidae. Similar observations were made regarding the vertical distribution of herbivores versus bacterivores in both standard and metabarcoding data sets (Table 7.3 and Table 7.4). Bacterivores were most abundant in 0-5 or 5-15 cm in both data sets regardless of

whether the data were rarefied or unrarefied. Herbivore percentage over total nematode abundance, on the other hand, was highest in the lower 15-30 cm for both data sets. In both the metabarcoding and standard datasets, reads associated *Rhabditis* did not differ across the three depths.

Community succession is a very important aspect of environmental monitoring (De Goede *et al.*, 1993b; Ettema & Bongers, 1993; Wardle *et al.*, 1995; Ferris & Matute, 2003; Háněl, 2010). Even though this study compared only two separate time points, analysed using only metabarcoding, there were some significant changes in some of the community indices. Increased perturbation in the form of deep tillage caused a decrease in maturity index in the second sampling year. Plant-parasitic index is the MI equivalent designated exclusively for the plant-parasitic nematodes (Bongers, 1990). Previous studies have shown that its relationship with MI can be either direct (Neher & Campbell, 1994) or inverse (Bongers, van der Meulen, & Korthals, 1997). This therefore explains why the PPI in this study increased between two sampling dates while MI decreased. Another reason could be the availability of a suitable host in the interval between the two sampling since this could have resulted in the multiplication of herbivores such as *Meloidogyne* or *Pratylenchus*. If that was the case, it would certainly have increased the PPI since both of these taxa contribute highly to PPI value (both score three on plant parasite scaling according to Bongers 1990).

The c-p triangle's representation of pattern of change within the nematode communities showed some mild increases in both stability and enrichment for both rarefied and binary data formats. It has to be said that, based on the earlier comparison between c-p triangles from the standard and metabarcoding data sets, metabarcoding based c-p triangles may be presenting completely exaggerated conditions of stress. More testing may be required to increase its reliability.

In summary, this study demonstrates that metabarcoding data is capable of depicting significant aspects of a community's condition and do match in many respects the conditions presented by data obtained from the standard morphology-based approach. However, there are a few acknowledgeable limitations to this study that if addressed in future work, may improve the utility of metabarcoding for nematode community analysis.

Firstly, the DNA extraction methodology could be improved. DNA extractions were carried out using 100 µl of suspensions making it a tenth of what was used for morphological analysis and a hundredth of the whole extracts from 200 g. This could have resulted in loss of taxa that were already rare in the soil sample. Moreover, the volume of suspension containing nematodes also contained other eukaryotes which were also amplified by the universal primers used as was reported in other studies (Sapkota and Nicolaisen, 2015). In fact, a significant percentage of the otus generated for the 2016 samples were from

these non-target taxa (82.43%). To maximize the percentage of nematode reads it may be important to perform DNA extraction on either the whole suspension of nematode extracts or at least concentrate the suspension better through centrifugation (Ritz *et al.*, 2009). This may not necessarily exclude other taxa but would certainly increase the total nematode recovery. To exclude some of the non-target taxa, the nested PCR approach reported by (Sapkota & Nicolaisen, 2015) could be used to first exclude non-nematode targets before the actual amplification.

The second limitation is to do with the extent to which nematode extracellular DNA could have made it into nematode extracts and eventually, into the recovered reads. Extracellular nematode DNA has been shown to occur in bulk nematode extracts (Peham *et al.*, 2017), and the implications of this could be the detection of taxa that are no longer an active part of the community. Theoretically, this can be resolved by using RNA instead of DNA for detection, an area that may require some testing and developing in the future. For now, extracting nematodes from the soil first before DNA extraction can help reduce the amount of extracellular DNA. Finally, as has been shown in a number of tests conducted in this study, both rarefied and unrarefied metabarcoding data gave similar outcomes in terms of nematode MI-based indices and trophic composition. The exception may be for agricultural soils where some taxa, particularly the sensitive ones, are rarely recovered in the samples. In such cases, one might opt for unrarefied data to avoid losing these taxa for computing indices and trophic composition. While this study points to the right direction in terms of the potential of using metabarcoding for monitoring soil health through nematode community indices, further studies may be needed to that focus on soils from natural undisturbed environments in order to better establish the reliability of this approach.

8 General discussion

The concept of DNA barcoding is not new to nematology. In fact, the first important application of DNA barcoding was based on nematodes in which the authors, Floyd *et al*. (2002) described the molecular operational taxonomic units scheme for nematodes in soil samples collected from a Scottish upland *Agrostis-Fustuca* grassland. Their work even predated Hebert and colleagues' milestone study, "Biological identifications through DNA barcodes" which is widely credited with developing the DNA barcoding concept (Hebert *et al.*, 2003). The main idea was to find and use appropriate genetic markers to identify 'unknown' individuals by simply comparing the sequence of their marker with those of 'known' organisms. In its early years, before the advent of next generation sequencing, the Sanger method was the only means of obtaining sequences from the specimens to be identified. Sanger sequencing required the analysis of each specimen separately, from the DNA extraction step through to the sequence analysis step. As difficult as the barcoding approach appeared for community analysis, it could potentially replace the most difficult step in nematode community analysis, which is morphological identification. The concept was even more useful and certainly more convenient for tasks that only involved identifying individuals.

The introduction of next generation sequencers (NGS) capable of generating sequence reads in their millions at only a fraction of the cost opened up many exciting possibilities for nematode community analysis. Due to their massively parallel high-throughput nature, NGS platforms could potentially be used to analyse nematode communities at scales that hitherto were impossible. That is to say by applying the concept of DNA barcoding with this technology, multiple number of individuals in multiple samples could be sequenced within a single run. Consequently, it is rather fitting that the concept came to be described as metabarcoding. Amongst the first and most significant applications of this approach to nematode identification was the work published by Porazinska *et al*. in 2009. Using an artificially assembled community of nematodes and two ~400 bp long regions within the 18S and 28S rDNA, they demonstrated that up to 97% of the sampled species could be recovered with the two markers combined. Individually, however, the barcode markers could only recover ~90% of the sampled taxa. The reason for the failure of 18S rDNA-based markers to account for ~10% of the taxa was not clear, even though this was expected for the 28S rDNA-based marker. Another limitation that became evident from this study was the incongruence between abundance of taxa in the mock community and the number of reads they are represented by in the data. Specifically, it was observed that distribution of reads among taxa was biased, and the authors identified the choice of barcode marker as one of the contributing factors to this bias.

Porazinska *et al*. (2009) provided a good foundation for future exploration into how metabarcoding could be used in identifying soil free-living nematodes. It was clear, however, that the method needed more appraisal before it could reliably be implemented, particularly, in terms of exploring other widely used molecular markers, establishing a better understanding of read number and species abundance relationships as well as identifying modifications that may be required to make this approach applicable to the natural environment that is the soil. To address these questions, this study was structured into two main sections: (i) the methodology development phase where the main goal was to determine the most suitable methods and analysis tools for identifying nematodes in communities, (ii) the application phase where the field applicability of the approach was evaluated by comparing it against the standard morphology-based method.

The first experiment, Chapter 3 of this thesis, was set up to investigate further the coverage of these markers in a community that was assembled to cover a wide diversity of nematodes. The study was set up to also assess the possibility of using just a single marker instead of two or more and still be able achieve very good coverage. Two additional makers known and widely used for identifying nematodes were also included in the experiment -one was based in the 18S rDNA region and the other in the COI region. According to Porazinska *et al*. (2009) the nature of the read distribution bias can depend on the chosen barcode marker. Therefore, in addition to identifying which of the four markers was most suited for nematode metabarcoding, they were evaluated to determine which one was least prone to read distribution bias. Based on information available in literature (Floyd *et al.*, 2002; Creer *et al.*, 2010) the markers were ranked prior to the experiment in the order, 18S rDNA-based markers > 28S rDNA-based marker > COI-based marker in the terms of coverage. The results of this chapter showed that at least two of the markers gave excellent recovery of the sampled taxa. The two 18S rDNA-based markers were expected to provide the best coverage since they have been used previously in a number of metabarcoding studies and are well known for their conserved primer annealing sites. Although the 28S rDNA-based marker was expected to provide a fairly wide coverage, it was not predicted to perform better than the 18S rDNA-based SSUF04-SSUR22 marker (refer to Chapter 3). The reason for the low recovery of the SSUFO4-SSUR22 marker was attributed to the reference database bias as discussed in Chapter 3. With a complete reference database, this marker might have recovered all the sampled as observed with NF1-18Sr2b. This chapter identified the importance of having a reliable and a complete database of nematode sequences. Additionally, the chapter also highlighted the usefulness of performing taxonomy assignment with more than one method. The read distributions between replicates showed a significant correlation, which conforms with previous studies that showed the reproducibility of metabarcoding data (Porazinska *et al.*, 2010c; Edgar, 2017a). The correlation between replicates was in fact

observed for all the markers that were evaluated. Another quantification aspect of this chapter that agreed with previous studies was the lack of correlation between relative abundances of taxa and sequence reads associated with these taxa. This is a widely acknowledged and reported issue with metabarcoding (Porazinska *et al.*, 2009; Elbrecht & Leese, 2015; Edgar, 2017a; Holovachov *et al.*, 2017; Krehenwinkel *et al.*, 2017), particularly in mixed species samples. It was therefore not surprising that no such correlation was found in this experiment. Most of the reasons given for why this issue exists (discussed in Chapter 5) range from technical to innate factors that are either impossible or difficult to correct or change. For example, the difference in the number of copies of templates for each taxon in the sample prior to PCR is considered to be factor affecting read number distributions. In nematodes and other organisms, whose barcodes are located in repetitive regions of the genome, the amount of template is dependent on the number of copies of these regions within the genome. Such bias-causing factors are beyond the control of the investigator. This particular bias has, in fact, been shown by Krehenwinkel *et al*. (2017) to persist even if such recommendations as using methods devoid of PCR (Zhou *et al.*, 2013; Dowle *et al.*, 2016) are followed. By comparison, standard barcoding based on the Sanger sequencing method does not have this limitation with abundance and read numbers. This is because each individual is analysed separately, and the Sanger method produces a single sequence per individual. Thus, the number of sequences associated a taxon represents its abundance in the sample.

Being unquantifiable is certainly among the key limitations of metabarcoding that a number of studies have tried to either find solutions to or provide alternatives to. There have been attempts even at the bioinformatics step to manipulate analysis parameters or develop algorithms to lower this bias in the read abundance of taxa (Amend *et al.*, 2010; Edgar, 2017a). Some reports indicate that this bias can be minimised by relaxing some of the sequence filtering parameters such as the quality and sequence length thresholds, (Amend *et al.*, 2010; Deagle *et al.*, 2013). The main objectives of Chapters 4 and 5 were to investigate how this might work with nematode data and how the choice of bioinformatics pipeline might affect read distribution. Another objective of Chapter 4, was to compare different bioinformatics pipelines, to determine which was best in terms of accuracy, speed and convenience. Overall, the USEARCH pipeline did have the best clustering algorithm, compared to those employed in the other pipelines. For someone new to high-throughput data analysis, there is no doubt that they would find the USEARCH pipeline the easiest.

Chapter 5 sought to examine how read number bias manifests under different community complexities using different mock communities. Some of the mock samples were set up based on the knowledge from previous studies that have shown that the linear correlation between abundance and read numbers in monospecific samples is usually higher

compared to multispecies ones (Amend *et al.*, 2010). The idea was to determine how step-wise increase in the complexity of the community would reflect in the abundance-read number correlations. The results showed that it is quite common for a few taxa to consistently dominate the read numbers in multispecies samples. It also confirmed that the dominant taxa in terms of read abundance are often not the most abundant in terms of actual abundance in the samples. Furthermore, it indicated that a slight modification to a community through the introduction of a new species can have a significant effect on the overall result, with certain taxa switching from being dominant to becoming infrequent. Increasing the abundance of *Acrobeloides buetschlii* confirmed previous studies that also showed some correlation between read abundance and number of individuals for monospecific samples. This finding provides a picture of how abundance translates to read distributions under the non-typical situation where there are no other taxa to introduce bias. That said, it is obvious that this finding may not be of significant practical use for nematode samples from soil extracts where mono-specific communities do not exist. At this point in the study, it was apparent that metabarcoding-based assessment of nematode communities were not suitable for quantification regardless of the barcode marker used, analysis pipeline employed or how relaxed the sequence quality filtering parameters are set.

Artificial assemblage of nematodes, however complex or diverse, cannot replicate an actual soil nematode community. The experiment in Chapter 6 was set up to apply metabarcoding on actual nematode communities that were subjected to different soil management regimes. These tillage practices are known to have an impact on the structural and biological soil. The impact of various tillage treatments on nematode communities have been the subject of several previous community analysis studies (Lenz & Eisenbeis, 2000; Czyż, 2004; Liphadzi *et al.*, 2005; Sanchez-Moreno *et al.*, 2006; Okada & Harada, 2007; Griffiths *et al.*, 2012). Reduced or controlled traffic is a practice intended to minimise the compaction caused by vehicular passes on the field. Reducing compaction has been shown to be beneficial in terms of crop yield and the soil phyisicochemical properties. Certain soil fauna including nematodes are known to benefit from a reduced level of compaction associated with controlled traffic (Bouwman & Arts, 2000).

The experiment reported in Chapter 6 involved the analysis of the nematode communities under three different forms of tillage and traffic using the standard morphological method. Compared to a number of previous studies, the observed inhibitive effect of tillage on total abundance contradicted the expected outcome since most of them have shown significant increases in nematode abundances as a result of soil disturbance. There is often a bloom in the abundance of bacterivores after disturbance events particularly, the enrichment opportunists of the family Rhabditidae (Sanchez-Moreno *et al.*, 2006). It was this increase

in bacterivores after tillage that has been implicated as largely contributing to this increase in overall total abundance, especially in bacterivore dominated communities. All samples in this study were dominated by both bacterivorous and herbivorous taxa. In most cases, it was the latter group that had the highest presence. And since herbivores are known to decline in numbers following tillage (Sanchez-Moreno *et al.*, 2006), it made sense that this was reflected on the communities as a whole. Random traffic farming also appeared to be more favourable to total nematode abundance. Given that RTF plots were the most compacted, it can be assumed that higher compaction favoured nematode abundance particularly herbivores. A previous report by Bouwman and Arts (2000) observed that out of all the nematode trophic groups, herbivores were the only ones that responded positively to heavier compaction load. Interestingly, for most of the community indices, it was traffic, rather than tillage, that had significant effect. The maturity index (MI) including all its derivatives and the structure index all recorded their highest values under controlled traffic farming. The drop in MI values under the LGP and RTF traffic systems was expected, especially considering the fact that compaction does introduce a form of disturbance to the soil and its nematode assemblage.

Chapter 7 showed that despite the incongruence between relative abundance and read frequencies at the taxa level, the metabarcoding data correlated in a number of aspects with the standard data, specifically in terms of community indices and trophic classifications. The beta diversity computed from the data collated using the standard approach significantly correlated with those obtained using metabarcoding. This chapter also showed there was very little difference, if any, in the outcome of the analysis whether raw or rarefied abundance data from standard morphology was used. For the metabarcoding data, the binary (presence-absence) data format gave the best statistical correlation with the standard method. Ji *et al.* (2013) also reported statistically correlated alpha and beta diversities from standard and metabarcoding methods. Although it was apparent that the two methods were not exactly identical in their outcomes, the observed correlation between them laid out a method with a fair degree of reliability to follow for performing metabarcoding-based nematode community analysis.

8.1 Recommendations

An important issue of metabarcoding that remains unresolved is the abundance-read number divergence, the reason why this technique cannot currently be reliably quantifiable. It is fair therefore that a significant amount of future research subjects looks into improving the correlation between abundance and read numbers of species in a community. The current study shows that the 18S-based marker NF1-18Sr2b for most applications is sufficiently universal to recover and discriminate between most nematodes species. To the advantage of this marker, there is a huge number of sequences available in most public databases that can be used for taxonomy assignment.

A lot of the issues of read abundances bias are attributed to PCR, even though evidences and general consensus suggest it is not the only causal factor. This has been shown in insects to improve the abundance-read number correlation slightly. An area of potential future research will be to investigate to what extent metabarcoding without PCR can reduce this bias for nematode community analysis. With more whole genome sequences being published, this creates the possibility of mapping the number of copies of repeat regions present within each species. Knowledge of copy numbers of repeat marker regions together with PCR-independent analysis methods can help address two of the most important known causes of bias.

This study has also given a good insight into how far a better curated nematode sequence database can go to facilitate the taxonomy assignment step of the analyses. A well curated and annotated database of nematode sequences will be helpful in eliminating the necessity to cross-check the taxonomies assigned to otus using algorithms such as uclust or utax. As stated in Chapter 3, the main rDNA reference databases, PR2 and SILVA, have a number of entries where the taxonomies are incomplete and thus make it a necessity for the assignments to be checked using other assignment methods. This process can be time consuming especially if there is a large number of otus to be checked. Creating such database may require collaborative work by a number of nematode taxonomists and molecular biologists.

Although not immediately urgent, improving the size of soil samples that can used for direct extraction of DNA from the soil can be of immense use. This may finally make it possible to recover a wide array of taxa in the soil as well as very rare ones that could potentially be missing due to the low recovery rate of the tradition nematode extraction methods. For the Whitehead method, recovery rate has been shown to be as low as 50%. This high percentage loss could potentially be reduced if DNA can be directly obtained without nematode extraction from the soil.

8.2 Conclusions

Detection, sensitivity, speed and the throughput of metabarcoding approach to identifying organisms in bulk communities have never been questioned. It is rather the question of whether metabarcoding data is quantifiable or not that has often come under dispute. This study arrived at the same conclusion as many others preceding it by demonstrating that with the right choice of markers, metabarcoding is capable capturing diversities similar to what is possible with the standard method. Moreover, metabarcoding here has been shown to be quantifiable, specifically at such classification levels as trophic groups and functional guilds. There were, however, a few indications that such predictions using metabarcoding required some scrutiny because they were not always identical with the standard method. With a few improvements in the methodology, particularly at the DNA extraction step, the accuracy of this method can be enhanced to better predict community composition and structure at least at the trophic and functional guild levels.

9 References

Abebe E, Mekete T, Thomas WK. 2011. A critique of current methods in nematode taxonomy. *African Journal of Biotechnology* 10: 312–323.

Adl SM, Habura A, Eglit Y. 2014. Amplification primers of SSU rDNA for soil protists. *Soil Biology and Biochemistry* 69: 328–342.

Agatha S, Strüder-Kypke MC. 2007. Phylogeny of the order Choreotrichida (Ciliophora, Spirotricha, Oligotrichea) as inferred from morphology, ultrastructure, ontogenesis, and SSrRNA gene sequences. *European Journal of Protistology* 43: 37–63.

Ahmad W, Jairajpuri MS. 2010. Mononchida: the predatory soil nematodes. In: Hunt DJ,, In: Perry RN, eds. *Nematology Monographs and Perspectives*. Leiden, The Netherlands: Brill, 1–299.

Ahmed M, van de Vossenberg BTLH, Cornelisse C, Karssen G. 2013. On the species status of the root-knot nematode *Meloidogyne ulmi* Palmisano & Ambrogioni, 2000 (Nematoda, Meloidogynidae). *ZooKeys*: 1–27.

Akhtar M, Malik A. 2000. Roles of organic soil amendments and soil organisms in the biological control of plant-parasitic nematodes: a review. *Bioresource Technology* 74: 35–47.

Allen MW, Sher SA. 1967. Taxonomic problems concerning the phytoparasitic nematodes. *Annual Review of Phytopathology* 5: 247–262.

Altieri MA. 1991. How best can we use biodiversity in agroecosystems? *Outlook on Agriculture* 20: 15–23.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–10.

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

Anderson TJC, Blouin MS, Beech RN. 1998. Population biology of parasitic nematodes: applications of genetic markers. *Advances in Parasitology* 41: 219–283.

Andrássy I. 2005. *Free-living nematodes from Hungary (Nematoda errantia), I*. Budapest, Hungary: Pedozoologica Hungarica 3. Hungarian Natural History Museum and Systematic Zoology Research Group of the Hungarian Academy of Sciences.

Andrássy I. 2007. *Free-living nematodes of Hungary (Nematoda errantia), II*. Budapest, Hungary: Pedozoologica Hungarica No. 4. Hungarian Natural History Museum.

Aronesty E. 2011. ea-utils: Command-line tools for processing biological sequencing data. *Expression Analysis, Durham, NC*.

Arvidsson J. 2001. Subsoil compaction caused by heavy sugarbeet harvesters in southern Sweden: I. Soil physical properties and crop yield in six field experiments. *Soil and Tillage Research* 60: 67–78.

Bachy C, Dolan JR, López-García P, Deschamps P, Moreira D. 2013. Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study. *The ISME Journal* 7: 244–255.

Bardgett RD, Cook R, Yeates GW, Denton CS. 1999. The influence of nematodes on below-ground processes in grassland ecosystems. *Plant and Soil* 212: 23–33.

Baxter C, Rowan JS, McKenzie BM, Neilson R. 2013. Understanding soil erosion impacts in temperate agroecosystems: bridging the gap between geomorphology and soil ecology using nematodes as a model organism. *Biogeosciences* 10: 7133–7145.

Beare MH, Parmelee RW, Hendrix PF, Cheng W, Coleman DC, Crossley DA. 1992. Microbial and faunal interactions and effects on litter nitrogen and decomposition in agroecosystems. *Ecological Monographs* 62: 569–591.

Beccuti M, Carrara M, Cordero F, Donatelli S, Calogero RA. 2013. The structure of state-of-art gene fusion-finder algorithms. *Genome Bioinformatics* 1: 1–6.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*: 289–300.

Bergé JB, Dalmasso A. 1975. Caractéristiques biochimiques de quelques populations de *Meloidogyne hapla* et Melodogyne spp. *Cahiers ORSTOM. Série Biologie: Nématologie* 10: 263–271.

Besansky NJ, Severson DW, Ferdig MT. 2003. DNA barcoding of parasites and invertebrate disease vectors: what you don't know can hurt you. *Trends Parasitology* 19: 545–546.

Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK. 2012a.

Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology and Evolution* 27: 233–243.

Bik HM, Sung WAY, De Ley P, Baldwin JG, Sharma J, Rocha-Olivares A, Thomas WK. 2012b. Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Molecular Ecology* 21: 1048–1059.

Blaxter ML, De Ley P, Garey JR, Liu LX. 1998. A molecular evolutionary framework for the phylum Nematoda. *Nature* 392: 71–75.

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E. 2005. Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360: 1935–1943.

Blok VC. 2002. Molecular diagnostics for plant-parasitic nematodes. *Proceedings of the Fourth International Congress of Nematology*.195–206.

Blok VC. 2005. Achievements in and future prospects for molecular diagnostics of plant-parasitic nematodes. *Canadian Journal of Plant Pathology* 27: 176–185.

Blouin MS, Yowell CA, Courtney CH, Dame JB. 1995. Host movement and the genetic structure of populations of parasitic nematodes. *Genetics* 141: 1007–1014.

Blouin MS, Yowell CA, Courtney CH, Dame JB. 1998. Substitution bias, rapid saturation, and the use of mtDNA for nematode systematics. *Molecular Biology and Evolution* 15: 1719–1727.

Boisselier-Dubayle MC, Gofas S. 1999. Genetic relationships between marine and marginal-marine populations of Cerithium species from the Mediterranean Sea. *Marine Biology* 135: 671–682.

Bongers T. 1990. The maturity index: an ecological measure of environmental disturbance based on nematode species composition. *Oecologia* 83: 14–19.

Bongers T, Bongers M. 1998. Functional diversity of nematodes. *Applied Soil Ecology* 10: 239–251.

Bongers T, Ferris H. 1999. Nematode community structure as a bioindicator in environmental monitoring. *Trends in Ecology & Evolution* 14: 224–228.

Bongers T, van der Meulen H, Korthals G. 1997. Inverse relationship between the nematode maturity index and plant parasite index under enriched nutrient conditions. *Applied Soil Ecology* 6: 195–199.

Bongers T, De Goede RGN, Korthals GW, Yeates GW. 1995. Proposed changes of cp classification for nematodes. *Russian Journal of Nematology* 3: 61–62.

Bouwman LA, Arts WBM. 2000. Effects of soil compaction on the relationships between nematodes, grass production and soil physical properties. *Applied Soil Ecology* 14: 213–222.

Bowles J, Blair D, McManus DP. 1992. Genetic variants within the genus Echinococcus identified by mitochondrial DNA sequencing. *Molecular and Biochemical Parasitology* 54: 165–173.

Braasch H, Burgermeister W, Pastrik KH. 1995. Differentiation of three Bursaphelenchus species by means of RAPD-PCR. *Nachrichtenblatt des Deutschen Pflanzenschutzdienstes* 47: 310–314.

Bucklin A, Steinke D, Blanco-Bercial L. 2011. DNA barcoding of marine metazoa. *Annual Review of Marine Science* 3: 471–508.

Buée M, Reich M, Murat C, Morin E, Nilsson RH, Uroz S, Martin F. 2009. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist* 184: 449–456.

Cairns Jr J, McCormick P V, Niederlehner BR. 1993. A proposed framework for developing indicators of ecosystem health. *Hydrobiologia* 263: 1–44.

Cantalapiedra-Navarrete C, Navas-Cortés JA, Liébanas G, Vovlas N, Subbotin SA, Palomares-Rius JE, Castillo P. 2013. Comparative molecular and morphological characterisations in the nematode genus *Rotylenchus*: *Rotylenchus paravitis* n. sp., an example of cryptic speciation. *Zoologischer Anzeiger-A Journal of Comparative Zoology* 252: 246–268.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7: 335–336.

Carew ME, Pettigrove VJ, Metzeling L, Hoffmann AA. 2013. Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology* 10: 1–15.

Carneiro RMDG, Almeida MRA, Quénéhervé P. 2000. Enzyme phenotypes of *Meloidogyne* spp. populations. *Nematology* 2: 645–654.

Carpenter AS, Hiatt EE, Lewis SA, Abbott AG. 1992. Genomic RFLP analysis of Meloidogyne arenaria race 2 populations. *Journal of Nematology* 24: 23–28.

Carrasco-Ballesteros S, Castillo P, Adams BJ, Pérez-Artés E. 2007. Identification of *Pratylenchus thornei*, the cereal and legume root-lesion nematode, based on SCAR-PCR and satellite DNA. *European Journal of Plant Pathology* 118: 115–125.

Carvalho GR, Creer S, Allen MJ, Costa FO, Tsigenopoulos CS, Le Goff-Vitry M, Magoulas A, Medlin L, Metfies K. 2010. Genomics in the discovery and monitoring of marine biodiversity. *Introduction to Marine Genomics*. Springer, 1–32.

Castagnone-Sereno P, Vanlerberghe-Masutti F, Leroy F. 1994. Genetic polymorphism between and within *Meloidogyne* species detected with RAPD markers. *Genome* 37: 904–909.

Cenis JL. 1993. Identification of four major *Meloidogyne* spp. by Random Amplified Polymorphic DNA (RAPD-PCR). *Phytopathology* 83: 76–80.

Charchar JM, Eisenback JD. 2000. An improved technique to prepare perineal patterns of root-knot nematodes for SEM. *Nematologia Brasileira* 24: 245–247.

Chilton NB, Gasser RB, Beveridge I. 1995. Differences in a ribosomal DNA sequence of morphologically indistinguishable species within the *Hypodontus macropi* complex (Nematoda: Strongyloidea). *International Journal for Parasitology* 25: 647–651.

van Cleave HJ. 1932. Eutely or cell constancy in its relation to body size. *The Quarterly Review of Biology* 7: 59–67.

Cobb NA. 1914. Nematodes and their relationships. *USDA Yearbook of Agriculture, 1914*. US Government Printing Office, 457–490.

Coissac E, Hollingsworth PM, Lavergne S, Taberlet P. 2016. From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology* 25: 1423–1428.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research* 42: D633–D642.

Coomans A. 2002. Present status and future of nematode systematics. *Nematology* 4: 573–582.

Cowart DA, Pinheiro M, Mouchel O, Maguer M, Grall J, Miné J, Arnaud-Haond S. 2015. Metabarcoding is powerful yet still blind: a comparative analysis of morphological and molecular surveys of seagrass communities. *PLoS One* 10: e0117562.

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

Cunha A, Azevedo RBR, Emmons SW, Leroi AM. 1999. Developmental biology: Variable cell number in nematodes. *Nature* 402: 253.

Curran J, McClure MA, Webster JM. 1986. Genotypic differentiation of *Meloidogyne* populations by detection of restriction fragment length difference in total DNA. *Journal of Nematology* 18: 83–86.

Czyż EA. 2004. Effects of traffic on soil aeration, bulk density and growth of spring barley. *Soil and Tillage Research* 79: 153–166.

D'Argenio V, Casaburi G, Precone V, Salvatore F. 2014. Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. *BioMed Research International* 2014: 1–10.

Dayrat B. 2005. Towards integrative taxonomy. *Biological Journal of the Linnean Society* 85: 407–415.

Deagle BE, Thomas AC, Shaffer AK, Trites AW, Jarman SN. 2013. Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Molecular Ecology Resources* 13: 620–633.

Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P. 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters* 10: 20140562.

Decraemer W, Hunt DJ. 2006. Taxonomy and principal genera. In: Perry RN,, In: Moens M, eds. *Plant Nematology. London: CABI International*. Wallingford: CABI Publishing, 3–32.

Derycke S, Vanaverbeke J, Rigaux A, Backeljau T, Moens T. 2010. Exploring the use of cytochrome oxidase c subunit 1 (COI) for DNA barcoding of free-living marine nematodes. *PLoS One* 5: e13716.

Derycke S, Remerie T, Vierstraete A, Backeljau T, Vanfleteren J, Vincx M, Moens T.

2005. Mitochondrial DNA variation and cryptic speciation within the free-living marine nematode *Pellioditis marina*. *Marine Ecology Progress Series* 300: 91–103.

Donn S, Neilson R, Griffiths BS, Daniell TJ. 2012. A novel molecular approach for rapid assessment of soil nematode assemblages–variation, validation and potential applications. *Methods in Ecology and Evolution* 3: 12–23.

Dorris M, De Ley P, Blaxter ML. 1999. Molecular analysis of nematode diversity and the evolution of parasitism. *Parasitology Today* 15: 188–193.

Douda O, Marek M, Zouhar M, Ryšánek P. 2013. Insights into the structure and phylogeny of the 28S rRNA expansion segments D2 and D3 of the plant-infecting nematodes from the genus *Ditylenchus* (Nematoda: Anguinidae). *Phytopathologia Mediterranea*: 84–97.

Dowle EJ, Pochon X, C Banks J, Shearer K, Wood SA. 2016. Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: a case study using freshwater macroinvertebrates. *Molecular Ecology Resources* 16: 1240–1254.

Drotz MK, Saura A, Nilsson AN. 2001. The species delimitation problem applied to the *Agabus bipustulatus* complex (Coleoptera, Dytiscidae) in north Scandinavia. *Biological Journal of the Linnean Society* 73: 11–22.

Dunn OJ. 1964. Multiple comparisons using rank sums. *Technometrics* 6: 241–252.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

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

Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* 10: 996–8.

Edgar RC. 2015. makeudb_utax command.

Edgar RC. 2016. UNCROSS: Filtering of high-frequency cross-talk in 16S amplicon reads. *bioRxiv*: 88666.

Edgar RC. 2017a. UNBIAS: An attempt to correct abundance bias in 16S sequencing, with limited success. *bioRxiv*: 124149.

Edgar RC. 2017b. cluster_otus command.

Edgar RC, Flyvbjerg H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* 31: 3476–3482.

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

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323: 133–138.

Eisenback JD, Hirschmann H. 1981. Identification of *Meloidogyne* Species on the Basis of Head Shape and, Stylet Morphology of the Male. *Journal of Nematology* 13: 513–521.

Eisenback JD, Hunt DJ. 2009. General morphology. In: Perry RN,, In: Moens M,, In: Starr JL, eds. *Root-knot nematodes*. CAB International, Wallingford, UK, 18–54.

Ekblom R, Galindo J. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107: 1–15.

Elbrecht V, Leese F. 2015. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. *PLoS ONE* 10: 1–16.

Ellis RE, Sulston JE, Coulson AR. 1986. The rDNA of C. elegans: sequence and structure. *Nucleic Acids Research* 14: 2345–2364.

Esbenshade PR, Triantaphyllou AC. 1985. Use of enzyme phenotypes for identification of Meloidogyne species. *Journal of Nematology* 17: 6–20.

Esbenshade PR, Triantaphyllou AC. 1990. Isozyme phenotypes for the identification of Meloidogyne species. *Journal of Nematology* 22: 10–15.

Esling P, Lejzerowicz F, Pawlowski J. 2015. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research* 43: 2513–2524.

Ettema CH. 1998. Soil nematode diversity: species coexistence and ecosystem function. *Journal of Nematology* 30: 159–169.

Ettema CH, Bongers T. 1993. Characterization of nematode colonization and succession in disturbed soil using the Maturity Index. *Biology and Fertility of Soils* 16: 79–85.

Eves-Van Den Akker S, Lilley CJ, Reid A, Pickup J, Anderson E, Cock PJA, Blaxter M, Urwin PE, Jones JT, Blok VC. 2015. A metagenetic approach to determine the diversity

and distribution of cyst nematodes at the level of the country, the field and the individual. *Molecular Ecology* 24: 5842–5851.

Ferri E, Barbuto M, Bain O, Galimberti A, Uni S, Guerrero R, Ferté H, Bandi C, Martin C, Casiraghi M. 2009. Integrated taxonomy: traditional approach and DNA barcoding for the identification of filarioid worms and related parasites (Nematoda). *Frontiers in Zoology* 6: 1–12.

Ferris H, Matute MM. 2003. Structural and functional succession in the nematode fauna of a soil food web. *Applied Soil Ecology* 23: 93–110.

Ferris H, Bongers T. 2006. Nematode indicators of organic enrichment. *Journal of Nematology* 38: 3–12.

Ferris H, Venette RC, Lau SS. 1996. Dynamics of nematode communities in tomatoes grown in conventional and organic farming systems, and their impact on soil fertility. *Applied Soil Ecology* 3: 161–175.

Ferris H, Bongers T, De Goede RGM. 2001. A framework for soil food web diagnostics: extension of the nematode faunal analysis concept. *Applied Soil Ecology* 18: 13–29.

Ferris VR, Ferris JM, Faghihi J, Ireholm A. 1994. Comparisons of isolates of Heterodera avenae using 2-D PAGE protein patterns and ribosomal DNA. *Journal of Nematology* 26: 144–151.

Floyd R, Abebe E, Papert A, Blaxter M. 2002. Molecular barcodes for soil nematode identification. *Molecular Ecology* 11: 839–850.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek 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.

Fonseca G, Derycke S, Moens T. 2008. Integrative taxonomy in two free-living nematode species complexes. *Biological Journal of the Linnean Society* 94: 737–753.

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

Fonseca VG, Carvalho GR, Sung W, Johnson HF, Power DM, Neill SP, Packer M, Blaxter ML, Lambshead PJD, Thomas WK, Creer S. 2010. Second-generation environmental

sequencing unmasks marine metazoan biodiversity. *Nature Communications* 1: 1–8.

Fourie H, Zijlstra C, McDonald A. 2001. Identification of root-knot nematode species occurring in South Africa using the SCAR-PCR technique. *Nematology* 3: 675–680.

Gasso V, Oudshoorn FW, Sørensen CAG, Pedersen HH. 2014. An environmental life cycle assessment of controlled traffic farming. *Journal of Cleaner Production* 73: 175–182.

Gebhardt MR, Daniel TC, Schweizer EE, Allmaras RR. 1985. Conservation tillage. *Science* 230: 625–630.

Gerrard J. 2000. *Fundamentals of Soils*. Routledge Fundamentals of Physical Geography, Taylor & Francis.

Giblin-Davis RM, Ye W, Kanzaki N, Williams D, Morris K, Thomas WK. 2006. Stomatal ultrastructure, molecular phylogeny, and description of Parasitodiplogaster laevigata n. sp.(Nematoda: Diplogastridae), a parasite of fig wasps. *Journal of Nematology* 38: 137–149.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* 11: 759–769.

De Goede RGM, Bongers T, Ettema C. 1993a. Graphical presentation and interpretation of nematode community structure: cp triangles. *Mededelingen van de Faculteit Landbouwwetenschappen, Rijksuniversiteit Gent* 58: 743–750.

De Goede RGM, Georgieva SS, Verschoor BC, Kamerman JW. 1993b. Changes in nematode community structure in a primary succession of blown-out areas in a drift sand landscape. *Fundamental and Applied Nematology* 16: 501–513.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17: 333–351.

Griffiths BS, Daniell TJ, Donn S, Neilson R. 2012. Bioindication potential of using molecular characterisation of the nematode community: Response to soil tillage. *European Journal of Soil Biology* 49: 92–97.

Groombridge B. 1992. *Global biodiversity: status of the earth's living resources*. London, UK: Chapman & Hall.

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, De Vargas C, Decelle J, Del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Logares R, Mahé F, Massana R, Montresor M,

Morard R, Not F, Pawlowski J, Probert I, Sauvadet AL, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R. 2012. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research* 41: D597–D604.

Gururani MA, Venkatesh J, Upadhyaya CP, Nookaraju A, Pandey SK, Park SW. 2012. Plant disease resistance genes: current status and future directions. *Physiological and Molecular Plant Pathology* 78: 51–65.

Gutiérrez-Gutiérrez C, Cantalapiedra-Navarrete C, Montes-Borrego M, Palomares-Rius JE, Castillo P. 2013. Molecular phylogeny of the nematode genus *Longidorus* (Nematoda: Longidoridae) with description of three new species. *Zoological Journal of the Linnean Society* 167: 473–500.

Gysi M. 2001. Compaction of a Eutric Cambisol under heavy wheel traffic in Switzerland: *Soil and Tillage Research* 61: 133–142.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward D V, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* 21: 494–504.

Hague NGM, Gowen SR. 1987. Chemical control of nematodes. In: Brown RH,, In: Kerry BR, eds. *Principles and practice of nematode control in crops*. Marrickville, Australia: Academic Press Australia, 131–178.

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

Håkansson I, Voorhees WB, Riley H. 1988. Vehicle and wheel factors influencing soil compaction and crop response in different traffic regimes. *Soil and Tillage Research* 11: 239–282.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium series* 41: 95–98.

Handoo ZA, Nyczepir AP, Esmenjaud D, Van der Beek JG, Castagnone-Sereno P, Carta LK, Skantar AM, Higgins JA. 2004. Morphological, molecular, and differential-host characterization of Meloidogyne floridensis n. sp.(Nematoda: Meloidogynidae), a root-knot nematode parasitizing peach in Florida. *Journal of Nematology* 36: 20–35.

Háněl L. 2010. An outline of soil nematode succession on abandoned fields in South Bohemia. *Applied Soil Ecology* 46: 355–371.

Haque F, Li J, Wu HC, Liang XJ, Guo P. 2013. Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. *Nano Today* 8: 56–74.

Harris TS, Sandall LJ, Powers TO. 1990. Identification of single *Meloidogyne* juveniles by polymerase chain reaction amplification of mitochondrial DNA. *Journal of Nematology* 22: 518–524.

Hebert PDN, Gregory TR. 2005. The promise of DNA barcoding for taxonomy. *Systematic Biology* 54: 852–859.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Soceity of London B Biological Science* 270: 313–321.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM. 2004. Identification of birds through DNA barcodes. *PLoS Biology* 2: e312.

Hebert PDN, Ratnasingham S, Zakharov E V, Telfer AC, Levesque-Beaudin V, Milton MA, Pedersen S, Jannetta P. 2016. Counting animal species with DNA barcodes: Canadian insects. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371: 20150333.

Hillis DM, Dixon MT. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. *The Quarterly Review of Biology* 66: 411–453.

Holland JM. 2004. The environmental consequences of adopting conservation tillage in Europe: reviewing the evidence. *Agriculture, Ecosystems & Environment* 103: 1–25.

Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* 106: 12794–12797.

Holovachov O, Boström S. 2010. *Identitification of Plectida (Nematoda)*. Gent: Eumaine, Gent and Nematology, UC Riverside.

Holovachov O, De Ley IT, Mundo-Ocampo M, De Ley P. 2009. *Identification of Cephalobidae (Nematoda)*. Gent: Eumaine, Gent and Nematology, UC Riverside.

Holovachov O, Haenel Q, Bourlat SJ, Jondelius U. 2017. Taxonomy assignment approach determines the efficiency of identification of OTUs in marine nematodes. *Royal Society*

Open Science 4: 170315.

Holterman MHM, Oggenfuss M, Frey JE, Kiewnick S. 2012. Evaluation of High-resolution Melting Curve Analysis as a New Tool for Root-knot Nematode Diagnostics. *Journal of Phytopathology* 160: 59–66.

Holterman M, Karssen G, Van Den Elsen S, Van Megen H, Bakker J, Helder J. 2009. Small subunit rDNA-based phylogeny of the Tylenchida sheds light on relationships among some high-impact plant-parasitic nematodes and the evolution of plant feeding. *Phytopathology* 99: 227–235.

Holterman M, van der Wurff A, van den Elsen S, van Megen H, Bongers T, Holovachov O, Bakker J, Helder J. 2006. Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. *Molecular Biology and Evolution* 23: 1792–1800.

Horn R, Fleige H. 2003. A method for assessing the impact of load on mechanical stability and on physical properties of soils. *Soil and Tillage Research* 73: 89–99.

Höss S, Claus E, Von der Ohe PC, Brinke M, Güde H, Heininger P, Traunspurger W. 2011. Nematode species at risk—a metric to assess pollution in soft sediments of freshwaters. *Environment International* 37: 940–949.

Hothorn T, Bretz F, Westfall P, Heiberger R. 2008. Multcomp: simultaneous inference for general linear hypotheses. R Package Version 1.0-3.

Hu M, Jex AR, Campbell BE, Gasser RB. 2007. Long PCR amplification of the entire mitochondrial genome from individual helminths for direct sequencing. *Nature Protocols* 2: 2339–2344.

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

Hübschen J, Kling L, Ipach U, Zinkernagel V, Brown D, Neilson R. 2004b. Development and validation of species-specific primers that provide a molecular diagnostic for virus-vector longidorid nematodes and related species in German viticulture. *European Journal of Plant Pathology* 110: 883–891.

Hübschen J, Kling L, Ipach U, Zinkernagel V, Bosselut N, Esmenjaud D, Brown DJF, Neilson R. 2004a. Validation of the specificity and sensitivity of species-specific primers that provide a reliable molecular diagnostic for *Xiphinema diversicaudatum*, *X. index* and *X. vuittenezi*. *European Journal of Plant Pathology* 110: 779–788.

Hyman ED. 1988. A new method of sequencing DNA. *Analytical Biochemistry* 174: 423–436.

Hyman BC. 1990. Molecular diagnosis of Meloidogyne species. *Journal of Nematology* 22: 24–30.

Ibrahim SK, Davies KG, Perry RN. 1996. Identification of the root-knot nematode, Meloidogyne incognita, using monoclonal antibodies raised to non-specific esterases. *Physiological and Molecular Plant Pathology* 49: 79–88.

Jain M, Olsen HE, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* 17: 1–11.

Janssen T, Karssen G, Verhaeven M, Coyne D, Bert W. 2016. Mitochondrial coding genome analysis of tropical root-knot nematodes (Meloidogyne) supports haplotype based diagnostics and reveals evidence of recent reticulate evolution. *Scientific Reports* 6: 1–13.

Jex AR, Littlewood DTJ, Gasser RB. 2010. Toward next-generation sequencing of mitochondrial genomes—focus on parasitic worms of animals and biotechnological implications. *Biotechnology Advances* 28: 151–159.

Jex AR, Hu M, Littlewood DTJ, Waeschenbach A, Gasser RB. 2008a. Using 454 technology for long-PCR based sequencing of the complete mitochondrial genome from single Haemonchus contortus (Nematoda). *BMC Benomics* 9: 1–14.

Jex AR, Waeschenbach A, Littlewood DTJ, Hu M, Gasser RB. 2008b. The mitochondrial genome of Toxocara canis. *PLoS Neglected Tropical Diseases* 2: e273.

Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters* 16: 1245–1257.

Johnson SR, Ferris VR, Ferris JM. 1972. Nematode community structure of forest woodlots. I. Relationships based on similarity coefficients of nematode species. *Journal of Nematology* 4: 175–183.

Johnson SR, Ferris JM, Ferris VR. 1974. Nematode community structure of forest woodlots: III. Ordinations of taxonomic groups and biomass. *Journal of Nematology* 6: 118–126.

Jones P, Ambler DJ, Robinson MP. 1988. Application of monoclonal antibodies to the diagnosis of plant pathogens and pests. *Proceedings of the British Crop Protection*

Conference-Pests and Diseases. Surrey, U.K.: BCPC Registered Office, Thornton Heath, 767–776.

Kaplan DT, Thomas WK, Frisse LM, Sarah JL, Stanton JM, Speijer PR, Marin DH, Opperman CH. 2000. Phylogenetic analysis of geographically diverse Radopholus similis via rDNA sequence reveals a monomorphic motif. *Journal of Nematology* 32: 134–142.

Karssen G, Van Hoenselaar T, Verkerk-Bakker B, Janssen R. 1995. Species identification of cyst and root-knot nematodes from potato by electrophoresis of individual females. *Electrophoresis* 16: 105–109.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059–3066.

Kembel SW, Wu M, Eisen JA, Green JL. 2012. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Computational Biology* 8: e1002743.

Kerry BR. 2000. Rhizosphere interactions and the exploitation of microbial agents for the biological control of plant-parasitic nematodes. *Annual Review of Phytopathology* 38: 423–441.

Khan AR. 1996. Influence of tillage on soil aeration. *Journal of Agronomy and Crop Science* 177: 253–259.

Kiewnick S, Holterman M, van den Elsen S, van Megen H, Frey JE, Helder J. 2014. Comparison of two short DNA barcoding loci (COI and COII) and two longer ribosomal DNA genes (SSU & LSU rRNA) for specimen identification among quarantine root-knot nematodes (Meloidogyne spp.) and their close relatives. *European Journal of Plant Pathology* 140: 97–110.

Kircher M, Sawyer S, Meyer M. 2011. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research* 40: 1–8.

Kladivko EJ. 2001. Tillage systems and soil ecology. *Soil and Tillage Research* 61: 61–76.

Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG. 2017. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports* 7: 1–12.

Kroulík M, Kvíz Z, Kumhála F, Hůla J, Loch T. 2011. Procedures of soil farming allowing

reduction of compaction. *Precision Agriculture* 12: 317–333.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33: 1870–1874.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* 12: 118–123.

Lahaye R, Van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V. 2008. DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences* 105: 2923–2928.

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

Lambshead PJD. 2004. Marine nematode diversity. In: Chen ZX,, In: Chen WY,, In: Chen SY,, In: Dickson SW, eds. *Advances and Perspectives: Nematode Morphology, Physiology and Ecology*. Wallingford: CABI Publishing, 438–468.

Lejzerowicz F, Esling P, Pillet L, Wilding TA, Black KD, Pawlowski J. 2015. High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports* 5: 1–10.

Lenth R V. 2016. Least-squares means: the R package lsmeans. *Journal of Statistical Software* 69: 1–33.

Lenz R, Eisenbeis G. 2000. Short-term effects of different tillage in a sustainable farming system on nematode community structure. *Biology and Fertility of Soils* 31: 237–244.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* 44: W242–W245.

De Ley P. 2000. Lost in worm space: phylogeny and morphology as road maps to nematode diversity. *Nematology* 2: 9–16.

De Ley P. 2006. A quick tour of nematode diversity and the backbone of nematode phylogeny. *WormBook: the online review of C. elegans biology*: 1–8.

De Ley P, Blaxter ML. 2002. Systematic position and phylogeny. In: Lee DL, ed. *The biology of nematodes*. London, UK: Taylor & Francis London, 1–30.

De Ley P, De Ley IT, Morris K, Abebe E, Mundo-Ocampo M, Yoder M, Heras J, Waumann D, Rocha-Olivares A, Jay Burr AH, Baldwin JG, Thomas WK. 2005. An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 360: 1945–1958.

Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, Chen ZD, Zhou SL, Chen SL. 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences* 108: 19641–19646.

Lindeque PK, Parry HE, Harmer RA, Somerfield PJ, Atkinson A. 2013. Next Generation Sequencing Reveals the Hidden Diversity of Zooplankton Assemblages. *PLOS ONE* 8: e81327.

Liphadzi KB, Al-Khatib K, Bensch CN, Stahlman PW, Dille JA, Todd T, Rice CW, Horak MJ, Head G. 2005. Soil microbial and nematode communities as affected by glyphosate and tillage practices in a glyphosate-resistant cropping system. *Weed Science* 53: 536–545.

Litvaitis MK, Bates JW, Hope WD, Moens T. 2000. Inferring a classification of the Adenophorea (Nematoda) from nucleotide sequences of the D3 expansion segment (26/28S rDNA). *Canadian Journal of Zoology* 78: 911–922.

Mair P, Wilcox R. 2018. WRS2: Wilcox robust estimation and testing. : 1–33. https://cran.r–project.org/web/packages/WRS2.

Malakhov V V. 1994. *Nematodes: Structure, Development, Classification and Phylogeny*. Washington DC, USA: Smithsonian Institution Press.

Marché L, Valette S, Grenier E, Mugniéry D. 2001. Intra-species DNA polymorphism in the tobacco cyst nematode complex (*Globodera tabacum*) using AFLP. *Genome* 44: 941–946.

Marcussen T. 2003. Evolution, phylogeography, and taxonomy within the *Viola alba* complex (Violaceae). *Plant Systematics and Evolution* 237: 51–74.

Mardis ER. 2008. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*. 9: 387–402.

Mardis ER. 2013. Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* 6: 287–303.

Markmann M, Tautz D. 2005. Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360: 1917–1924.

Martini E. 1909. Vber Eutelie und Neotenic. *Verhandlungen der Zoologisch-Botanischen Gesellschaft* 19: 292–299.

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. 2012. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* 13: 1–7.

Mayr E, Ashlock PD. 1991. The Science of Taxonomy. *Principles of Systematic Zoology*.1–14.

McCarter JP. 2009. Molecular approaches toward resistance to plant-parasitic nematodes. In: Berg RH,, In: Taylor CG, eds. *Cell Biology of Plant Nematode Parasitism*. Springer, 1–29.

van Megen H, van den Elsen S, Holterman M, Karssen G, Mooyman P, Bongers T, Holovachov O, Bakker J, Helder J. 2009. A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. *Nematology* 11: 927–950.

Metzker ML. 2005. Emerging technologies in DNA sequencing. *Genome Research* 15: 1767–1776.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A. 2008. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 1–8.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Gateway Computing Environments Workshop (GCE), 2010*. Ieee, 1–8.

Minton NA. 1986. Impact of conservation tillage on nematode populations. *Journal of Nematology* 18: 135–140.

Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R. 2008a. Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology* 57: 750–757.

Munch K, Boomsma W, Willerslev E, Nielsen R. 2008b. Fast phylogenetic DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363:

3997–4002.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research* 39: 1–13.

NCBI Resource Coordinators. 2016. Database resources of the national center for biotechnology information. *Nucleic Acids Research* 44: D7–D19.

Neher DA. 2001. Role of nematodes in soil health and their use as indicators. *Journal of Nematology* 33: 161–168.

Neher DA, Campbell CL. 1994. Nematode communities and microbial biomass in soils with annual and perennial crops. *Applied Soil Ecology* 1: 17–28.

Neher DA, Campbell CL. 1996. Sampling for regional monitoring of nematode communities in agricultural soils. *Journal of Nematology* 28: 196–208.

Neher DA, Darby BJ. 2006. Computation and application of nematode community indices: General guidelines. In: Abebe E, ed. *Freshwater Nematodes: Ecology and Taxonomy*. CABI, 211–222.

Neher DA, Bongers T, Ferris H. 2004. Computation of nematode community indices. *Society of nematologists workshop*.1–33.

Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J. 2014. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PloS one* 9: e94249.

den Nijs L, van den Berg W. 2013. The added value of proficiency tests: choosing the proper method for extracting *Meloidogyne* second-stage juveniles from soil. *Nematology* 15: 143–151.

Nilakanta H, Drews KL, Firrell S, Foulkes MA, Jablonski KA. 2014. A review of software for analyzing molecular sequences. *BMC Research Notes* 7: 1–9.

Ntalli NG, Menkissoglu-Spiroudi U. 2011. Pesticides of botanical origin: a promising tool in plant protection. *Pesticides-Formulations, Effects, Fate*. InTech, 1–24.

Nunn GB. 1992. Nematode molecular evolution: an investigation of evolutionary patterns among nematodes based upon DNA sequences. Unpublished thesis, University of Nottingham, UK.

O'Hara RB, Kotze DJ. 2010. Do not log-transform count data. *Methods in Ecology and*

Evolution 1: 118–122.

Okada H, Harada H. 2007. Effects of tillage and fertilizer on nematode communities in a Japanese soybean field. *Applied Soil Ecology* 35: 582–598.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2015. vegan: community ecology package. R package version 2.2–1. 2015. : 1–280.

Oliveira CMG, Ferraz LCCB, Neilson R. 2006. Xiphinema krugi, Species Complex or Complex of Cryptic Species? *Journal of Nematology* 38: 418–428.

Orgiazzi A, Dunbar MB, Panagos P, de Groot GA, Lemanceau P. 2015. Soil biodiversity and DNA barcodes: opportunities and challenges. *Soil Biology and Biochemistry* 80: 244–250.

Orr CC, Dickerson OJ. 1966. Nematodes in true prairie soils of Kansas. *Transactions of the Kansas Academy of Science (1903-)* 69: 317–334.

Parmelee RW, Alston DG. 1986. Nematode trophic structure in conventional and no-tillage agroecosystems. *Journal of Nematology* 18: 403–407.

Payan LA, Dickson DW. 1990. Comparison of populations of *Pratylenchus brachyurus* based on isozyme phenotypes. *Journal of Nematology* 22: 538–545.

Peham T, Steiner FM, Schlick-Steiner BC, Arthofer W. 2017. Are we ready to detect nematode diversity by next generation sequencing? *Ecology and Evolution* 7: 4147–4151.

Perera MR, Taylor SP, Vanstone VA, Jones MGK. 2009. Protein biomarkers to distinguish oat and lucerne races of the stem nematode, *Ditylenchus dipsaci*, with quarantine significance for Western Australia. *Nematology* 11: 555–563.

Petersen DJ, Vrain TC. 1996. Rapid identification of Meloidogyne chitwoodi, M. hapla, and M. fallax using PCR primers to amplify their ribosomal intergenic spacer. *Fundamental and Applied Nematology* 19: 601–605.

Phillips RE, Thomas GW, Blevins RL, Frye WW, Phillips SH. 1980. No-tillage agriculture. *Science* 208: 1108–1113.

Pinto AJ, Raskin L. 2012. PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets (S Bertilsson, Ed.). *PLoS ONE* 7: e43093.

Plummer E, Twin J, Bulach DM, Garland SM, Tabrizi SN. 2015. A comparison of three

bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *Journal of Proteomics & Bioinformatics* 8: 283–291.

Porazinska DL, Duncan L., McSorley R, Graham J. 1999. Nematode communities as indicators of status and processes of a soil ecosystem influenced by agricultural management practices. *Applied Soil Ecology* 13: 69–86.

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

Porazinska DL, Sung WAY, Giblin-Davis RM, Thomas WK. 2010c. Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. *Molecular Ecology Resources* 10: 666–676.

Porazinska DL, Giblin-Davis RM, Powers TO, Thomas WK. 2012a. Nematode spatial and ecological patterns from tropical and temperate rainforests. *PLoS One* 7: e44641.

Porazinska DL, Giblin-Davis RM, Sung W, Thomas WK. 2012b. The nature and frequency of chimeras in eukaryotic metagenetic samples. *Journal of Nematology* 44: 18–25.

Porazinska DL, McSorley R, Duncan LW, Graham JH, Wheaton TA, Parsons LR. 1998. Nematode Community Composition under Various Irrigation Schemes in a Citrus Soil Ecosystem. *Journal of Nematology* 30: 170–178.

Porazinska DL, Giblin-Davis RM, Esquivel A, Powers TO, Sung WAY, Thomas WK. 2010b. Ecometagenetics confirm high tropical rainforest nematode diversity. *Molecular Ecology* 19: 5521–5530.

Porazinska DL, Giblin-Davis RM, Faller L, Farmerie W, Kanzaki N, Morris K, Powers TO, Tucker AE, Sung W, Thomas WK. 2009. Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Molecular Ecology Resources* 9: 1439–1450.

Powers TO. 2004. Nematode molecular diagnostics: from bands to barcodes. *Annual Review of Phytopathology* 42: 367–383.

Powers TO, Fleming CC. 1998. Biochemical and molecular characterization. *The physiology and biochemistry of free-living and plant-parasitic nematodes*. CABI Publishing, Wallingford, UK, 355–380.

Powers TO, Mullin PG, Harris TS, Sutton LA, Higgins RS. 2005. Incorporating molecular

identification of *Meloidogyne* spp. into a large-scale regional nematode survey. *Journal of Nematology* 37: 226–235.

Pretty JN, Morison JIL, Hine RE. 2003. Reducing food poverty by increasing agricultural sustainability in developing countries. *Agriculture, Ecosystems & Environment* 95: 217–234.

Price NS, Siddiqi MR. 1994. Rainforest nematodes with particular reference to the Korup National Park, Cameroon. *Afro-Asian Journal of Nematology* 4: 117–128.

Prosser SWJ, Velarde-Aguilar MG, León-Règagnon V, Hebert PDN. 2013. Advancing nematode barcoding: a primer cocktail for the cytochrome c oxidase subunit I gene from vertebrate parasitic nematodes. *Molecular Ecology Resources* 13: 1108–1115.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13: 1–13.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research* 41: D590–D596.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. 2011. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12: 1–18.

Ramage T, Martins-Simoes P, Mialdea G, Allemand R, Duplouy A, Rousse P, Davies N, Roderick GK, Charlat S. 2017. A DNA barcode-based survey of terrestrial arthropods in the Society Islands of French Polynesia: host diversity within the SymbioCode Project. *European Journal of Taxonomy*: 1–13.

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

Ristau K, Steinfartz S, Traunspurger W. 2013. First evidence of cryptic species diversity and significant population structure in a widespread freshwater nematode morphospecies (*Tobrilus gracilis*). *Molecular Ecology* 22: 4562–4575.

Ritz K, Trudgill DL. 1999. Utility of nematode community analysis as an integrated measure of the functional state of soils: perspectives and challenges. *Plant and Soil* 212: 1–11.

Ritz K, Black HIJ, Campbell CD, Harris JA, Wood C. 2009. Selecting biological indicators

for monitoring soils: a framework for balancing scientific and technical opinion to assist policy development. *Ecological Indicators* 9: 1212–1221.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4: e2584.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475: 348–352.

RStudio Team. 2016. RStudio: Integrated Development Environment for R.

Sanchez-Moreno S, Minoshima H, Ferris H, Jackson LE. 2006. Linking soil properties and nematode community composition: effects of soil management on soil food webs. *Nematology* 8: 703–715.

Sapkota R, Nicolaisen M. 2015. High-throughput sequencing of nematode communities from total soil DNA extractions. *BMC Ecology* 15: 1–8.

Sasser JN, Carter CC, Hartman KM. 1984. Standardization of host suitability studies and reporting of resistance to root-knot nematodes. *Crop Nematode Research and Control Project*: 1–7.

Schloss PD. 2013. Secondary structure improves OTU assignments of 16S rRNA gene sequences. *The ISME Journal* 7: 457–460.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75: 7537–7541.

Schneider GF, Dekker C. 2012. DNA sequencing with nanopores. *Nature Biotechnology* 30: 326–328.

Schnell IB, Bohmann K, Gilbert MTP. 2015. Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources* 15: 1289–1303.

Schots A, Gommers FJ, Bakker J, Egberts E. 1990. Serological differentiation of plant-parasitic nematode species with polyclonal and monoclonal antibodies. *Journal of Nematology* 22: 16–23.

Schröder J, Bailey J, Conway T, Zobel J. 2010. Reference-free validation of short read data. *PLoS One* 5: e12681.

Seinhorst JW. 1955. Een eenvoudige methode voor het afscheiden van aaltjes uit grond. *European Journal of Plant Pathology* 61: 188–190.

Semblat JP, Wajnberg E, Dalmasso A, Abad P, Castagnone-Sereno P. 1998. High-resolution DNA fingerprinting of parthenogenetic root-knot nematodes using AFLP analysis. *Molecular Ecology* 7: 119–125.

Shaw AJ, Allen B. 2000. Phylogenetic relationships, morphological incongruence, and geographic speciation in the Fontinalaceae (Bryophyta). *Molecular Phylogenetics and Evolution* 16: 225–237.

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. 2017. DNA sequencing at 40: past, present and future. *Nature* 550: 345–353.

Siddiqi MR. 2000. *Tylenchida: parasites of plants and insects*. Wallingford, UK: CABI Publishing, Wallingford, UK.

Sieriebriennikov B, Ferris H, de Goede RGM. 2014. NINJA: an automated calculation system for nematode-based biological monitoring. *European Journal of Soil Biology* 61: 90–93.

Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K, Nikolausz M. 2007. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiology Ecology* 60: 341–350.

Sites Jr JW, Marshall JC. 2004. Operational criteria for delimiting species. *Annual Review of Ecology, Evolution, and Systematics*: 199–227.

Smith E, Misiewicz P, Chaney K, White D, Godwin R. 2013. An investigation into the effect of traffic and tillage on soil properties and crop yields. *2013 Kansas City, Missouri, July 21 - July 24, 2013*: 2–14.

Sochová I, Hofman J, Holoubek I. 2006. Using nematodes in soil ecotoxicology. *Environment International* 32: 374–383.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences* 103: 12115–12120.

Subbotin SA, Moens M. 2006. Molecular diagnostics of plant-parasitic nematodes. *Plant nematology*. CABI Wallingford, UK, 33–58.

Subbotin SA, Sturhan D, Chizhov VN, Vovlas N, Baldwin JG. 2006. Phylogenetic analysis of Tylenchida Thorne, 1949 as inferred from D2 and D3 expansion fragments of the 28S rRNA gene sequences. *Nematology* 8: 455–474.

Subbotin SA, Ragsdale EJ, Mullens T, Roberts PA, Mundo-Ocampo M, Baldwin JG. 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.

Subbotin SA, Vovlas N, Crozzoli R, Sturhan D, Lamberti F, Moens M, Baldwin JG. 2005. Phylogeny of Criconematina Siddiqi, 1980 (Nematoda: Tylenchida) based on morphology and D2-D3 expansion segments of the 28S-rRNA gene sequences with application of a secondary structure model. *Nematology* 7: 927–944.

Subbotin SA, Sturhan D, Vovlas N, Castillo P, Tambe JT, Moens M, Baldwin JG. 2007. Application of the secondary structure model of rRNA for phylogeny: D2–D3 expansion segments of the LSU gene of plant-parasitic nematodes from the family Hoplolaimidae Filipjev, 1934. *Molecular Phylogenetics and Evolution* 43: 881–890.

Subbotin SA, Inserra RN, Marais M, Mullin P, Powers TO, Roberts PA, Van Den Berg E, Yeates GW, Baldwin JG. 2011. Diversity and phylogenetic relationships within the spiral nematodes of *Helicotylenchus* Steiner, 1945 (Tylenchida: Hoplolaimidae) as inferred from analysis of the D2-D3 expansion segments of 28S rRNA gene sequences. *Nematology* 13: 333–345.

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

Tastet C, Val F, Lesage M, Renault L, Marché L, Bossis M, Mugniéry D. 2001. Application of a putative fatty-acid binding protein to discriminate serologically the two European quarantine root-knot nematodes, *Meloidogyne chitwoodi* and *M. fallax*, from other *Meloidogyne* species. *European Journal of Plant Pathology* 107: 821–832.

Thomas SH. 1978. Population densities of nematodes under seven tillage regimes. *Journal of Nematology* 10: 24–27.

Tietjen JH. 1989. Ecology of deep-sea nematodes from the Puerto Rico Trench area and

Hatteras Abyssal Plain. *Deep Sea Research Part A. Oceanographic Research Papers* 36: 1579–1594.

Tijink FGJ, Döll H, Vermeulen GD. 1995. Technical and economic feasibility of low ground pressure running gear. *Soil and Tillage Research* 35: 99–110.

Treonis AM, Unangst SK, Kepler RM, Buyer JS, Cavigelli MA, Mirsky SB, Maul JE. 2018. Characterization of soil nematode communities in three cropping systems through morphological and DNA metabarcoding approaches. *Scientific Reports* 8: 1–12.

Tripathi AM, Tyagi A, Kumar A, Singh A, Singh S, Chaudhary LB, Roy S. 2013. The internal transcribed spacer (ITS) region and trnhH-psbA are suitable candidate loci for DNA barcoding of tropical tree species of India. *PloS one* 8: e57934.

Tullberg JN. 2000. Wheel traffic effects on tillage draught. *Journal of Agricultural Engineering Research* 75: 375–382.

Tullberg JN, Yule DF, McGarry D. 2007. Controlled traffic farming—From research to adoption in Australia. *Soil and Tillage Research* 97: 272–281.

Unger PW. 1990. Conservation tillage systems. *Advances in Soil Science*. Springer, 27–68.

Unterseher M, Siddique AB, Brachmann A, Peršoh D. 2016. Diversity and composition of the leaf mycobiome of beech (*Fagus sylvatica*) are affected by local habitat conditions and leaf biochemistry. *PloS one* 11: e0152878.

Valentini A, Pompanon F, Taberlet P. 2009. DNA barcoding for ecologists. *Trends in Ecology & Evolution* 24: 110–117.

Voorhees WB, Evans SD, Warnes DD. 1985. Effect of preplant wheel traffic on soil compaction, water use, and growth of spring wheat. *Soil Science Society of America Journal* 49: 215–220.

Vovlas N, Troccoli A, Palomares-Rius JE, De Luca F, Liébanas G, Landa BB, Subbotin SA, Castillo P. 2011. *Ditylenchus gigas* n. sp. parasitizing broad bean: a new stem nematode singled out from the *Ditylenchus dipsaci* species complex using a polyphasic approach with molecular phylogeny. *Plant Pathology* 60: 762–775.

Waeyenberge L, Haegeman A, Sutter N de, Viaene N, Maes M. 2017. DNA metabarcoding to characterize soil nematode communities. *3rd International Symposium on Nematodes as Environmental Bioindicators. June 28-June 29, 2017, Institute of*

Technology, Carlow, Ireland. Carlow, Ireland, 7.

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* 73: 5261–5267.

Wang YI, Naumann U, Wright ST, Warton DI. 2012. mvabund–an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* 3: 471–474.

Wardle DA. 1995. Impacts of disturbance on detritus food webs in agro-ecosystems of contrasting tillage and weed management practices. *Advances in Ecological Research* 26: 105–185.

Wardle DA, Yeates GW, Watson RN, Nicholson KS. 1995. Development of the decomposer food-web, trophic relationships, and ecosystem properties during a three-year primary succession in sawdust. *Oikos*: 155–166.

Wardle DA, Yeates G, Barker G, Bonner K. 2006. The influence of plant litter diversity on decomposer abundance and diversity. *Soil Biology and Biochemistry* 38: 1052–1062.

Welch BL. 1951. On the comparison of several mean values: an alternative approach. *Biometrika* 38: 330–336.

Westcott SL, Schloss PD. 2017. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. *mSphere* 2: e00073-17.

Whitehead AG, Hemming JR. 1965. A comparison of some quantitative methods of extracting small vermiform nematodes from soil. *Annals of Applied Biology* 55: 25–38.

Wiesel L, Daniell TJ, King D, Neilson R. 2015. Determination of the optimal soil sample size to accurately characterise nematode communities in soil. *Soil Biology and Biochemistry* 80: 89–91.

Wilcox R. 2012. Chapter 7 - One-Way and Higher Designs for Independent Groups BT - Introduction to Robust Estimation and Hypothesis Testing (Third Edition). *Statistical Modeling and Decision Science*. Boston: Academic Press, 291–377.

Wilkinson MJ, Szabo C, Ford CS, Yarom Y, Croxford AE, Camp A, Gooding P. 2017. Replacing Sanger with Next Generation Sequencing to improve coverage and quality of reference DNA barcodes for plants. *Scientific Reports* 7: 1–11.

Williams ST. 2000. Species boundaries in the starfish genus *Linckia*. *Marine Biology* 136:

137–148.

Wilson MJ, Khakouli-Duarte T. 2009. *Nematodes as environmental indicators*. Wallingford, UK: CABI Publishing, Wallingford, UK.

Yeates GW. 1984. Variation in soil nematode diversity under pasture with soil and year. *Soil Biology and Biochemistry* 16: 95–102.

Yeates GW. 1994. Modification and qualification of the nematode maturity index. *Pedobiologia (Germany)* 38: 97–101.

Yeates GW. 2003. Nematodes as soil indicators: functional and biodiversity aspects. *Biology and Fertility of Soils* 37: 199–210.

Yeates GW, Bongers T. 1999. Nematode diversity in agroecosystems. *Agriculture, Ecosystems & Environment* 74: 113–135.

Yeates GW, Bongers T, De Goede RG, Freckman DW, Georgieva SS. 1993. Feeding habits in soil nematode families and genera-an outline for soil ecologists. *Journal of Nematology* 25: 315–331.

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z. 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution* 3: 613–623.

Zhan A, MacIsaac HJ. 2015. Rare biosphere exploration using high-throughput sequencing: research progress and perspectives. *Conservation Genetics* 16: 513–522.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* 7: 203–214.

Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q. 2013. Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience* 2: 1–12.

Zijlstra C. 1997. A fast PCR assay to identify Meloidogyne hapla, M. chitwoodi, and M. fallax, and to sensitively differentiate them from each other and from M. incognita in mixtures. *Fundamental and Applied Nematology* 20: 505–512.

Zijlstra C. 2000. Identification of *Meloidogyne chitwoodi*, *M. fallax* and *M. hapla* based on SCAR-PCR: a powerful way of enabling reliable identification of populations or individuals that share common traits. *European Journal of Plant Pathology* 106: 283–290.

Zijlstra C, Donkers-Venne DTHM, Fargette M. 2000. Identification of *Meloidogyne incognita*, *M. javanica* and *M. arenaria* using sequence characterised amplified region (SCAR) based PCR assays. *Nematology* 2: 847–853.

Zimmermann J, Glöckner G, Jahn R, Enke N, Gemeinholzer B. 2015. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources* 15: 526–542.

10 Appendices

Appendix 10.1. Utax taxonomy assignments for all four barcode markers. ☑ used to indicate how many of the replicates the otus were found in. Each marker has its own sets of otus, so the same otu identities can be shared by different taxa as long as they belong to different markers.

Taxa	Otus	Incidence		
NF1-18Sr2b				
Alaimus sp.	Otu83	☑		
Anaplectus sp.	Otu94,12	☑	☑	
Aphelenchoides ritzemabosi	Otu18	☑	☑	☑
Globodera	Otu7,131	☑	☑	☑
Hemicycliophora conida	Otu24	☑	☑	☑
Laimaphelenchus penardi	Otu30,63	☑	☑	☑
Longidorus	Otu9	☑	☑	☑
Meloidogyne hapla	Otu13,116	☑	☑	☑
Pristionchus	Otu1,38,40,85,72,111	☑	☑	☑
Rhabditis	Otu6	☑	☑	☑
Steinernema	Otu3,64,73,133	☑	☑	☑
Tripyla sp.	Otu5	☑	☑	☑
Tylenchus arcuatus	Otu21	☑	☑	☑
Xiphinema	Otu91	☑	☑	☑
SSUF04-SSUR22				
Globodera	Otu95	☑	☑	
Longidorus	Otu90	☑	☑	☑
Prionchulus	Otu77	☑	☑	☑
Rhabditis	Otu143	☑	☑	☑
Steinernema	Otu73	☑	☑	☑
D3Af-D3Br				
Aphelenchoides gorganensis	Otu26	☑	☑	☑
Ditylenchus dipsaci	Otu6,144	☑	☑	☑
Globodera ellingtonae	Otu4,59,101,104,133	☑	☑	☑
Hemicycliophora wyei	Otu21,82	☑		
Longidorus macrosoma	Otu7	☑	☑	☑
Meloidogyne hapla	Otu8	☑	☑	☑
Rhabditis sp.	Otu1,78,126,128	☑	☑	☑
Trichodorus primitivus	Otu9,39,144	☑	☑	☑
JB3-JB5GED				
Rhabditida	Otu4,10	☑	☑	
Tylenchida	Otu7	☑	☑	☑

Appendix 10.2. Blast taxonomy assignments for all four barcode markers. ☑ used to indicate how many of the replicates the otus were found in. Each marker has its own sets of otus, so the same otu identities can be shared by different taxa as long as they belong to different markers.

Taxa	Otus	Incidence		
NF1-18Sr2b				
Alaimus sp.	Otu20,50,83	☑		
Anaplectus sp.	Otu12,94,	☑	☑	
Anatonchus tridentatus	Otu8,79	☑	☑	☑
Aphelenchoides ritzemabosi	Otu18	☑	☑	☑
Aporcelaimellus obtusicaudatus	Otu14	☑	☑	☑
Acrobeles sp.	Otu29	☑	☑	
Acrobeloides sp.	Otu17,48	☑	☑	☑
Criconema sp.	Otu45	☑	☑	
Ditylenchus dipsaci	Otu4,70,100,110,113,122,126,136	☑	☑	☑
Globodera rostochiensis	Otu7,131	☑	☑	☑
Hemicycliophora conida	Otu24	☑	☑	☑
Laimaphelenchus penardi	Otu30,63	☑	☑	☑
Longidorus caespiticola	Otu9,62,138	☑	☑	☑
Meloidogyne hapla	Otu13	☑	☑	☑
Plectus andrassyi	Otu15,98,117	☑	☑	☑
Prionchulus punctatus	Otu2,52,64,76,106	☑	☑	☑
Pristionchus lheritieri	Otu1,38,40,72,85	☑	☑	☑
Rhabditis cf. *terricola*	Otu6,46,55	☑	☑	☑
Steinernema carpocapsae	Otu3,54,73,97,133	☑	☑	☑
Trichodorus primitivus	Otu10,26,28,115	☑	☑	☑
Tripyla glomerans	Otu5,32,39,112,125	☑	☑	☑
Tylenchus arcuatus	Otu21	☑	☑	☑
Xiphinema	Otu53	☑	☑	☑
SSUF04-SSUR22				
Alaimus sp.	Otu14,	☑	☑	☑
Anaplectus sp.	Otu21,24,25,26…	☑	☑	☑
Anatonchus tridentatus	Otu2,39,122	☑	☑	☑
Ditylenchus dipsaci	Otu60,78	☑	☑	☑
Globodera rostochiensis	Otu86,95	☑	☑	☑
Longidorus caespiticola	Otu3,90	☑	☑	☑
Meloidogyne hapla	Otu70	☑	☑	☑
Prionchulus punctatus	Otu1,67,89,117,119,130,161	☑	☑	☑
Pristionchus lheritieri	Otu77,120,149	☑	☑	☑
Rhabditis cf. *terricola*	Otu143	☑	☑	☑
Steinernema carpocapsae	Otu3,54,73,97,133	☑	☑	☑
Trichodorus primitivus	Otu5,105,159	☑	☑	☑
Tripyla glomerans	Otu6,76,125	☑	☑	☑
Xiphinema diversicaudatum	Otu11	☑	☑	☑
D3Af-D3Br				
Alaimus sp.	Otu19,22,32,47,53	☑	☑	☑
Aphelenchoides ritzemabosi	Otu26	☑	☑	☑
Anatonchus tridentatus	Otu5	☑	☑	☑
Ditylenchus dipsaci	Otu6,84,144	☑	☑	☑
Aporcelaimellus obtusicaudatus	Otu14,17,45	☑	☑	☑
Acrobeles complexus	Otu29	☑	☑	
Acrobeloides sp.	Otu18	☑	☑	☑
Globodera rostochiensis/pallida	Otu4,59,101,104,111,119,124,130,133	☑	☑	☑
Hemicycliophora wyei	Otu21,82	☑		
Longidorus macrosoma	Otu7,79	☑	☑	☑
Laimaphelenchus deconincki	Otu39	☑	☑	☑
Meloidogyne hapla	Otu8,72,142	☑	☑	☑
Plectus sp.	Otu15,20,116	☑	☑	☑
Prionchulus sp.	Otu23,99	☑	☑	☑
Rhabditis sp.	Otu1,78,100,121,126,128,136	☑	☑	☑
Pristionchus lucani	Otu13,113	☑	☑	☑
Steinernema carpocapsae	Otu61	☑	☑	☑
Trichodorus primitivus	Otu9,39,144	☑	☑	☑
Xiphinema diversicaudatum	Otu11,79	☑	☑	☑
Tripyla sp.	Otu108,109	☑	☑	☑
Tylenchus sp.	Otu53,83			
JB3-JB5GED				
Anatonchus tridentatus	Otu45	☑	☑	☑
Meloidogyne hapla	Otu7	☑	☑	☑
Longidorus caespiticola	Otu69	☑	☑	☑
Steinernema carpocapsae	Otu3,9,55,59,63	☑	☑	☑
Trichodorus primitivus	Otu47	☑	☑	☑
Tripyla sp.	Otu66	☑	☑	☑

Appendix 10.3. Index combinations used for the mock community samples

Sample ID	I7 Index ID	Index 1	I5 Index ID	index2
NF1-18Sr2b 1	N714	GCTCATGA	S513	TCGACTAG
NF1-18Sr2b 2	N714	GCTCATGA	S515	TTCTAGCT
NF1-18Sr2b 3	N714	GCTCATGA	S516	CCTAGAGT
SSUF04-R22 1	N714	GCTCATGA	S518	CTATTAAG
SSUF04-R22 2	N714	GCTCATGA	S520	AAGGCTAT
SSUF04-R22 3	N714	GCTCATGA	S521	GAGCCTTA
JB3-JB5GED 1	N715	ATCTCAGG	S518	CTATTAAG
JB3-JB5GED 2	N715	ATCTCAGG	S520	AAGGCTAT
JB3-JB5GED 3	N715	ATCTCAGG	S521	GAGCCTTA
D3Af-D3Br 1	N701	TAAGGCGA	S513	TCGACTAG
D3Af-D3Br 2	N702	CGTACTAG	S513	TCGACTAG
D3Af-D3Br 3	N703	AGGCAGAA	S513	TCGACTAG

Appendix 10.4. Feeding types, families, cp scores, functional guild and mean densities of nematode taxa recovered across the 108 samples analysed. Mean densities represent the number of individuals that, on average, were recovered per sample. Ba = bacterivores, Fu = fungivores, Omnivores, Pf = herbivores, and Ca = Predators.

Feeding type	Taxon	Family	cp score	Functional guild	Mean density
Bacterivores	*Acrobeles*	Cephalobidae	2	Ba2	0.14
	Acrobeloides	Cephalobidae	2	Ba2	29.35
	Alaimidae	Alaimidae	4	Ba4	1.53
	Alaimus	Alaimidae	4	Ba4	5.69
	Amphidelus	Alaimidae	4	Ba4	2.82
	Anaplectus	Plectidae	2	Ba2	4.31
	Cephalobidae	Cephalobidae	2	Ba2	0.56
	Cephalobus	Cephalobidae	2	Ba2	13.24
	Cervidellus	Cephalobidae	2	Ba2	2.41
	Chiloplacus	Rhabditidae	1	Ba1	1.11
	Cruznema	Rhabditidae	1	Ba1	0.60
	Dauer		1	Ba1	144.26
	Diplogasteridae	Diplogasteridae	1	Ba1	2.96
	Eucephalobus	Cephalobidae	2	Ba2	24.91
	Linhomoeidae	Linhomoeidae	2	Ba2	0.05
	Mesorhabditis	Rhabditidae	1	Ba1	2.96
	Monhysteridae	Monhysteridae	2	Ba2	0.65
	Nothacrobeles	Cephalobidae	2	Ba2	0.46
	Panagrolaimidae	Panagrolaimidae	1	Ba1	0.60
	Panagrolaimus	Panagrolaimidae	1	Ba1	3.15
	Panagrolobus	Cephalobidae	2	Ba2	0.05
	Paramphidelus	Alaimidae	4	Ba4	2.18
	Pellioditis	Rhabditidae	1	Ba1	0.56
	Penjatinema	Cephalobidae	2	Ba2	0.05
	Plectidae	Plectidae	2	Ba2	1.90
	Plectus	Plectidae	2	Ba2	5.23
	Propanagrolaimus	Panagrolaimidae	1	Ba1	0.19
	Pseudacrobeles	Cephalobidae	2	Ba2	12.08
	Rhabditidae	Rhabditidae	1	Ba1	4.63
	Rhabditis	Rhabditidae	1	Ba1	11.34
	Stegelletina	Cephalobidae	2	Ba2	0.79
	Wilsonema	Plectidae	2	Ba2	0.05
	Xyalidae	Xyalidae	2	Ba2	0.24
	Zeldia	Cephalobidae	2	Ba2	0.23
Fungivores	*Aphelenchoides*	Aphelenchoididae	2	Fu2	21.02
	Aphelenchoididae	Aphelenchoididae	2	Fu2	0.83
	Aphelenchus	Aphelenchidae	2	Fu2	5.14
	Diphtherophora	Diphtherophoridae	3	Fu3	0.23
	Tylalaimophorus	Diphtherophoridae	3	Fu3	0.79
Omnivores	Dorylaimidae	Dorylaimidae	4	Om4	3.89

	Nordiidae	Nordiidae	4	Om4	1.25
	Qudsianematidae	Qudsianematidae	4	Om4	3.10
Herbivores	*Amplimerlinius*	Telotylenchidae	3	Pf3	4.68
	Anguinidae	Anguinidae	2	Pf2	1.02
	Criconematidae	Criconematidae	3	Pf3	0.14
	Ditylenchus	Anguinidae	2	Pf2	9.68
	Nagelus	Telotylenchidae	3	Pf3	0.42
	Helicotylenchus	Hoplolaimidae	3	Pf3	21.25
	Hemicycliophora	Hemicycliophoridae	3	Pf3	0.09
	Heteroderidae	Heteroderidae	3	Pf3	0.14
	Hoplolaimidae	Hoplolaimidae	3	Pf3	0.14
	Longidorus	Longidoridae	4	Pf4	4.67
	Meloidogyne	Meloidogynidae	3	Pf3	87.08
	Merlinius	Telotylenchidae	3	Pf3	1.44
	Paratylenchus	Tylenchulidae	2	Pf2	44.54
	Pratylenchus	Pratylenchidae	3	Pf3	13.15
	Rotylenchus	Hoplolaimidae	3	Pf3	0.14
	Telotylenchidae	Telotylenchidae	3	Pf3	0.23
	Telotylenchus	Telotylenchidae	3	Pf3	0.56
	Trichodorus	Trichodoridae	4	Pf4	6.94
	Tylenchorhynchus	Telotylenchidae	3	Pf3	16.62
	Xiphinema	Longidoridae	5	Pf5	0.32
Predators	Anatonchidae	Anatonchidae	4	Ca4	0.46
	Aporcelaimidae	Aporcelaimidae	5	Ca5	1.48
	Clarkus	Mononchidae	4	Ca4	0.05
	Iotonchus	Iotonchidae	4	Ca4	0.60
	Miconchus	Anatonchidae	4	Ca4	0.05
	Mononchidae	Mononchidae	4	Ca4	0.46
	Mononchus	Mononchidae	4	Ca4	0.19
	Mylonchulus	Mylonchulidae	4	Ca4	2.13
	Nygolaimidae	Nygolaimidae	5	Ca5	0.14
	Prionchulinae	Prionchulidae	4	Ca4	0.14
	Prionchulus	Prionchulidae	4	Ca4	0.28
	Tripyla	Tripylidae	3	Ca3	2.04
Root hair feeders	*Aglenchus*	Tylenchidae	2	Pf2	0.83
	Basiria	Tylenchidae	2	Pf2	6.11
	Boleodorus	Tylenchidae	2	Pf2	2.08
	Coslenchus	Tylenchidae	2	Pf2	0.37
	Filenchus	Tylenchidae	2	Pf2	5.51
	Lelenchus	Tylenchidae	2	Pf2	0.32
	Psilenchus	Tylenchidae	2	Pf2	2.87
	Tylenchidae	Tylenchidae	2	Pf2	4.40
	Tylenchus	Tylenchidae	2	Pf2	21.85

Appendix 10.5. Samples and their respective indices

Sample ID	I7 Index ID	index	I5 Index ID	index2
1-RTF-D5	N701-A	TAAGGCGA	S502-A	CTCTCTAT
1-LGP-D5	N701-A	TAAGGCGA	S503-A	TATCCTCT
1-CTF-Z5	N701-A	TAAGGCGA	S505-A	GTAAGGAG
1-CTF-D5	N701-A	TAAGGCGA	S506-A	ACTGCATA
1-LGP-S5	N701-A	TAAGGCGA	S507-A	AAGGAGTA
1-RTF-Z5	N701-A	TAAGGCGA	S508-A	CTAAGCCT
1-LGP-Z5	N701-A	TAAGGCGA	S510-A	CGTCTAAT
1-CTF-S5	N701-A	TAAGGCGA	S511-A	TCTCTCCG
1-RTF-S5	N702-A	CGTACTAG	S502-A	CTCTCTAT
2-LGP-D5	N702-A	CGTACTAG	S503-A	TATCCTCT
2-CTF-S5	N702-A	CGTACTAG	S505-A	GTAAGGAG
2-LGP-S5	N702-A	CGTACTAG	S506-A	ACTGCATA
2-RTF-S5	N702-A	CGTACTAG	S507-A	AAGGAGTA
2-CTF-Z5	N702-A	CGTACTAG	S508-A	CTAAGCCT
2-RTF-Z5	N702-A	CGTACTAG	S510-A	CGTCTAAT
2-LGP-Z5	N702-A	CGTACTAG	S511-A	TCTCTCCG
2-CTF-D5	N703-A	AGGCAGAA	S502-A	CTCTCTAT
2-RTF-D5	N703-A	AGGCAGAA	S503-A	TATCCTCT
3-CTF-S5	N703-A	AGGCAGAA	S505-A	GTAAGGAG
3-RTF-Z5	N703-A	AGGCAGAA	S506-A	ACTGCATA
3-LGP-D5	N703-A	AGGCAGAA	S507-A	AAGGAGTA
3-CTF-D5	N703-A	AGGCAGAA	S508-A	CTAAGCCT
3-RTF-S5	N703-A	AGGCAGAA	S510-A	CGTCTAAT
3-CTF-Z5	N703-A	AGGCAGAA	S511-A	TCTCTCCG
3-RTF-D5	N704-A	TCCTGAGC	S502-A	CTCTCTAT
3-LGP-Z5	N704-A	TCCTGAGC	S503-A	TATCCTCT
3-LGP-S5	N704-A	TCCTGAGC	S505-A	GTAAGGAG
4-CTF-S5	N704-A	TCCTGAGC	S506-A	ACTGCATA
4-RTF-Z5	N704-A	TCCTGAGC	S507-A	AAGGAGTA
4-LGP-Z5	N704-A	TCCTGAGC	S508-A	CTAAGCCT
4-RTF-D5	N704-A	TCCTGAGC	S510-A	CGTCTAAT
4-LGP-D5	N704-A	TCCTGAGC	S511-A	TCTCTCCG
4-RTF-S5	N705-A	GGACTCCT	S502-A	CTCTCTAT
4-CTF-Z5	N705-A	GGACTCCT	S503-A	TATCCTCT
4-LGP-S5	N705-A	GGACTCCT	S505-A	GTAAGGAG
4-CTF-D5	N705-A	GGACTCCT	S506-A	ACTGCATA

1-RTF-D15	N716-D	ACTCGCTA	S513-D	TCGACTAG
1-LGP-D15	N716-D	ACTCGCTA	S515-D	TTCTAGCT
1-CTF-Z15	N716-D	ACTCGCTA	S516-D	CCTAGAGT
1-CTF-D15	N716-D	ACTCGCTA	S517-D	GCGTAAGA
1-LGP-S15	N716-D	ACTCGCTA	S518-D	CTATTAAG
1-RTF-Z15	N716-D	ACTCGCTA	S520-D	AAGGCTAT
1-LGP-Z15	N716-D	ACTCGCTA	S521-D	GAGCCTTA
1-CTF-S15	N716-D	ACTCGCTA	S522-D	TTATGCGA
1-RTF-S15	N718-D	GGAGCTAC	S513-D	TCGACTAG
2-LGP-D15	N718-D	GGAGCTAC	S515-D	TTCTAGCT
2-CTF-S15	N718-D	GGAGCTAC	S516-D	CCTAGAGT
2-LGP-S15	N718-D	GGAGCTAC	S517-D	GCGTAAGA
2-RTF-S15	N718-D	GGAGCTAC	S518-D	CTATTAAG
2-CTF-Z15	N718-D	GGAGCTAC	S520-D	AAGGCTAT
2-RTF-Z15	N718-D	GGAGCTAC	S521-D	GAGCCTTA
2-LGP-Z15	N718-D	GGAGCTAC	S522-D	TTATGCGA
2-CTF-D15	N719-D	GCGTAGTA	S513-D	TCGACTAG
2-RTF-D15	N719-D	GCGTAGTA	S515-D	TTCTAGCT
3-CTF-S15	N719-D	GCGTAGTA	S516-D	CCTAGAGT
3-RTF-Z15	N719-D	GCGTAGTA	S517-D	GCGTAAGA
3-LGP-D15	N719-D	GCGTAGTA	S518-D	CTATTAAG
3-CTF-D15	N719-D	GCGTAGTA	S520-D	AAGGCTAT
3-RTF-S15	N719-D	GCGTAGTA	S521-D	GAGCCTTA
3-CTF-Z15	N719-D	GCGTAGTA	S522-D	TTATGCGA
3-RTF-D15	N720-D	CGGAGCCT	S513-D	TCGACTAG
3-LGP-Z15	N720-D	CGGAGCCT	S515-D	TTCTAGCT
3-LGP-S15	N720-D	CGGAGCCT	S516-D	CCTAGAGT
4-CTF-S15	N720-D	CGGAGCCT	S517-D	GCGTAAGA
4-RTF-Z15	N720-D	CGGAGCCT	S518-D	CTATTAAG
4-LGP-Z15	N720-D	CGGAGCCT	S520-D	AAGGCTAT
4-RTF-D15	N720-D	CGGAGCCT	S521-D	GAGCCTTA
4-LGP-D15	N720-D	CGGAGCCT	S522-D	TTATGCGA
4-RTF-S15	N721-D	TACGCTGC	S513-D	TCGACTAG
4-CTF-Z15	N721-D	TACGCTGC	S515-D	TTCTAGCT
4-LGP-S15	N721-D	TACGCTGC	S516-D	CCTAGAGT
4-CTF-D15	N721-D	TACGCTGC	S517-D	GCGTAAGA
1-RTF-D30	N721-D	TACGCTGC	S518-D	CTATTAAG
1-LGP-D30	N721-D	TACGCTGC	S520-D	AAGGCTAT

1-CTF-Z30	N721-D	TACGCTGC	S521-D	GAGCCTTA
1-CTF-D30	N721-D	TACGCTGC	S522-D	TTATGCGA
1-LGP-S30	N722-D	ATGCGCAG	S513-D	TCGACTAG
1-RTF-Z30	N722-D	ATGCGCAG	S515-D	TTCTAGCT
1-LGP-Z30	N722-D	ATGCGCAG	S516-D	CCTAGAGT
1-CTF-S30	N722-D	ATGCGCAG	S517-D	GCGTAAGA
1-RTF-S30	N722-D	ATGCGCAG	S518-D	CTATTAAG
2-LGP-D30	N722-D	ATGCGCAG	S520-D	AAGGCTAT
2-CTF-S30	N722-D	ATGCGCAG	S521-D	GAGCCTTA
2-LGP-S30	N722-D	ATGCGCAG	S522-D	TTATGCGA
2-RTF-S30	N723-D	TAGCGCTC	S513-D	TCGACTAG
2-CTF-Z30	N723-D	TAGCGCTC	S515-D	TTCTAGCT
2-RTF-Z30	N723-D	TAGCGCTC	S516-D	CCTAGAGT
2-LGP-Z30	N723-D	TAGCGCTC	S517-D	GCGTAAGA
2-CTF-D30	N723-D	TAGCGCTC	S518-D	CTATTAAG
2-RTF-D30	N723-D	TAGCGCTC	S520-D	AAGGCTAT
3-CTF-S30	N723-D	TAGCGCTC	S521-D	GAGCCTTA
3-RTF-Z30	N723-D	TAGCGCTC	S522-D	TTATGCGA
3-LGP-D30	N724-D	ACTGAGCG	S513-D	TCGACTAG
3-CTF-D30	N724-D	ACTGAGCG	S515-D	TTCTAGCT
3-RTF-S30	N724-D	ACTGAGCG	S516-D	CCTAGAGT
3-CTF-Z30	N724-D	ACTGAGCG	S517-D	GCGTAAGA
3-RTF-D30	N724-D	ACTGAGCG	S518-D	CTATTAAG
3-LGP-Z30	N724-D	ACTGAGCG	S520-D	AAGGCTAT
3-LGP-S30	N724-D	ACTGAGCG	S521-D	GAGCCTTA
4-CTF-S30	N724-D	ACTGAGCG	S522-D	TTATGCGA
4-RTF-Z30	N726-D	CCTAAGAC	S513-D	TCGACTAG
4-LGP-Z30	N726-D	CCTAAGAC	S515-D	TTCTAGCT
4-RTF-D30	N726-D	CCTAAGAC	S516-D	CCTAGAGT
4-LGP-D30	N726-D	CCTAAGAC	S517-D	GCGTAAGA
4-RTF-S30	N726-D	CCTAAGAC	S518-D	CTATTAAG
4-CTF-Z30	N726-D	CCTAAGAC	S520-D	AAGGCTAT
4-LGP-S30	N726-D	CCTAAGAC	S521-D	GAGCCTTA
4-CTF-D30	N726-D	CCTAAGAC	S522-D	TTATGCGA