Soil Nematode Community Structure in Southern California Deserts: Insights from 28S D1-D2 Metabarcoding

By

CHRISTOPHER AARON PAGAN

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Entomology and Nematology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Steven A. Nadler, Chair

Amanda K. Hodson

Geoffrey M. Attardo

Committee in Charge

2024

Acknowledgements

First, I would like to thank my major professor, Steven Nadler. For two decades, Steve has been my teacher, employer, role model, and friend, investing in my future more than anyone save my parents. While many students refer to their major professors as 'mentors,' the word carries extra weight when I use it to describe Steve. I don't think there will ever come a time when I stop trying to emulate him or follow the examples he has set.

I am also grateful to my dissertation committee members, Amanda Hodson and Geoffrey Attardo, for their continued support. Amanda's guidance kept me motivated and focused during the writing process, and Geoff, who also served on my qualifying exam committee, has been in my corner throughout my graduate career.

My sincere thanks to Jim Baldwin, Ed Caswell-Chen, and Dan Potter for serving on my qualifying exam committee and for the thorough preparation process that contributed so much to my understanding of nematology, evolutionary biology, and ecology.

I would also like to express my gratitude to Valerie Williamson for giving me my first independent project and the opportunity to apply my training to a meaningful cause. Her passion for scientific research was contagious, and it was during this time that I realized I wanted to pursue a career in nematology.

To my lab siblings, Lauren Camp and Corwin Parker: thank you for the movie nights, camping trips, Friendsgiving dinners, and countless hours of D&D.

Thank you to Rangaswamy Meganathan for microbiological mentorship, companionship through the COVID years, fishing trips, crunchy snacks, and laughter.

I am grateful to Ziad Khouri for his moral support, help with alignments, trees, phylogenetic parlance, and for his ceaseless and unfailing kindness.

Thank you to Curtis Carlson for enduring bouts of complaining when my experiments or analyses weren't working and, on more than one occasion, finding solutions to my problems with his smartphone… while eating a sandwich.

Thank you to Nicole Tam for her help with the amplification and sequencing of Sanger sequences and for leaving behind lab books filled with the coolest drawings.

Special thanks to undergraduate researchers Victoria Kaml and Gavin Fong for their tireless work on the morphological documentation of nematode specimens and curation of Sanger sequence data. Though they have long since graduated and moved on to great things, my gratitude remains. I must also thank the additional undergraduate researchers who have contributed to this work over the years: Jonathan Woodbury, Alex Nguyen, Perlyn Du, and Cody Hussla.

I am eternally grateful to Deanna Jackson for being one of the best students in Steve's Spring 2002 ENT156 class, securing a position as a laboratory technician in the Nadler Lab, and then recommending me as her replacement.

Thank you to Dr. Sergio Álvarez-Ortega at Universidad Rey Juan Carlos and Mirayana Marcelino Barros of the Baldwin Lab at UC Riverside for their work in identifying nematode specimens.

Thank you to Matt Settles and the UCD Bioinformatics core for introducing me to my first bioinformatics pipeline and helping me gain access to the FARM computing cluster through the pilot grant program.

I am also grateful to the many graduate students, postdocs, and visiting scientists with whom I have shared lab space and friendship over the years: Ashleigh Smythe, Andy Bieberich, Jörgen Ulberg, Hugo Mejia Madrid, Joong-Ki Park, Gerardo Pérez-Ponce de León, Anindo Choudry, Soraya Naem, Hamed Seifi, Aurelie Castinel, Liu Fan, Stefano Catalano, Brenda Solórzano, Jacinta Gimeno, Sylwia Fudali, Jinling Gao, Sonchai Kwankuae, Limin Zhang, Alex Van Damme, and Neetha Iyer.

Finally, I would like to thank my family. I am deeply grateful to my parents for their unconditional love, their implicit faith in me, and their acceptance of my unconventional career path. I also want to thank Frank and Trina, the first college graduates in our family, for setting an example for me to follow. And to Stacy Hishinuma, thank you for being my primary source of inspiration and strength, for sleeping so peacefully that I can rest vicariously through you, for tolerating seven years of a long-distance relationship, and for being the best dog mom imaginable.

This work was supported by the National Science Foundation Graduate Research Fellowship Program (2015), the Merlin Allen Travel Award (2018, 2024), and NSF grant 1257331 awarded to Steve Nadler.

For Clara “Ted” Schofield.

Abstract

Nematodes are among the most abundant and diverse organisms on Earth, occurring nearly everywhere life can exist. With over a million estimated species, their diversity remains underexplored, largely due to the challenges of accurate identification. Traditional taxonomy requires specialized expertise and is time-consuming, especially for large environmental samples. DNA metabarcoding offers a faster, more scalable approach through the sequencing of bulk DNA, though it faces challenges in accurately linking sequences to species, particularly for multicellular organisms. High-throughput sequencing (HTS) technologies have transformed biodiversity studies by increasing throughput and sensitivity, but issues such as taxonomic misassignments, sequencing artifacts, and unreliable abundance estimates complicate their use. This dissertation focuses on characterization of nematode communities using multiple sequencing technologies with the purpose of enabling future ecological studies in desert soils and other environments.

Chapter 1 is a biodiversity inventory of the Boyd Deep Canyon Reserve (BDCR) in Palm Desert, California, using a DNA barcoding strategy based on the D1-D2 region of 28S ribosomal RNA. A reverse taxonomy approach was used to document the morphological and molecular characteristics of 160 nematode specimens isolated from 9 soil samples. A total of 26 distinct lineages were identified, including 14 putative new species, while 9 lineages await further morphological study. A custom reference sequence database was then constructed,

incorporating 28S sequences generated from the reverse taxonomy study, sequences downloaded from NCBI, and additional sequences produced by our lab in previous studies. The reference database was tested using a 96-well plate format with nematodes isolated from sub-fractions of the 9 soil samples. Classification of 2480 Sanger sequences at 70% confidence yielded 79 OTUs in 44 taxonomic categories, with 32 classified to species level, including 15 putative new species. Nematode communities were primarily composed of microbivores, with smaller proportions of omnivore-predators, fungivores, and plant parasites. Beta diversity analyses revealed that nematode community composition was more strongly associated with plant species than collection site. PERMANOVA results showed that plant species accounted for 40.1% of the variation in nematode communities ($R^2 = 0.401$, $F = 2.01$, $p = 0.061$) based on Bray-Curtis distances, and 35.0% of the variation ($R^2 = 0.350$, $F = 1.61$, $p = 0.071$) using Jaccard distances. Neither metric showed a significant effect of collection site on community composition.

Chapter 2 compares three sequencing technologies (Illumina MiSeq, PacBio Sequel II, and Sanger sequencing) in their ability to detect nematode diversity from the same soil samples. All reads were trimmed to 270 base pairs to enable direct comparisons across technologies. Illumina detected 70 species-level taxa, PacBio detected 40, and Sanger detected 28. However, 37 of the taxa identified by Illumina were in extremely low abundance, accounting for less than 0.127% of total reads, are likely to be artifactual sequence variants (RSVs). PacBio identified only two taxa that were not detected by the other two sequencing technologies. The inflated richness observed in the Illumina dataset is likely due to RSVs, which had a larger effect on the

Shannon alpha diversity metric because it is more sensitive to rare taxa. In contrast, the Simpson alpha diversity metric, which gives more weight to abundant taxa, was less affected by RSVs, making it a more stable measure when these artifacts are present. While Illumina consistently overestimated the number of observed taxa compared to PacBio, both technologies showed strong agreement in relative abundance estimates for the most common species. Despite the higher abundance of RSVs observed in Illumina data, its sequencing depth and lower cost per sample make it a practical option, particularly when downstream analyses are less sensitive to richness estimates.

In Chapter 3, nematode communities were characterized from the rhizospheres of creosote bushes across 11 locations in the Sonoran and Mojave Deserts, using PacBio-based metabarcoding of the 28S rDNA D1-D2 region. Soil samples were analyzed for physicochemical properties, including sand, organic carbon, water retention, and micronutrient concentrations. We detected 2319 amplified sequence variants (ASVs), classified into 92 taxonomic categories, 83% of which were successfully classified to species level, and identified 62 putative species. Over half of the ASVs (52.5%) were restricted to single locations, while others showed broader distributions across multiple sites. Microbivores dominated the trophic structure across all locations, and geographic patterns in nematode community composition were observed, particularly in Death Valley, where *Panagrolaimus* species were unusually dominant. Phylogenetic analysis revealed that the constituent ASVs of some putative species formed monophyletic groups that were sympatric at multiple locations, indicating potential overclassification, where ASVs classified as one species are actually multiple distinct species.

Principal component analysis (PCA) identified sand content as a major contributor to variation in soil properties, followed by organic carbon, water-holding capacity, and micronutrient concentrations. Beta regression GLMM models tested the influence of soil physicochemical properties and geographic location on nematode beta diversity, revealing that soil properties, particularly those related to soil texture, had a stronger effect on nematode community composition than geographic location.

This dissertation contributes new information on nematode diversity in Southern California desert ecosystems and provides insights into the effect of environmental factors on nematode community composition. It also evaluates the relative benefits of different sequencing technologies for biodiversity assessment. Together, these findings offer a basis for future research focused on improving reference databases, addressing data quality concerns, and advancing our understanding of nematode ecology.

Table of Contents

Chapter 1: Characterization of Soil Nematode Communities in a Sonoran Desert Study System

Christopher A. Pagan[1], Steven A. Nadler[1], Victoria K Kaml[1], Gavin Fong[1], and James G. Baldwin[2].

[1]Dept. of Entomology and Nematology, University of California, Davis, CA, U.S.A.; [2]Dept. of Nematology, University of California, Riverside, CA, U.S.A.

1.1.0 Introduction

The vast majority of nematodes on earth are free-living. They are found in marine sediments, freshwater sediments, and every type of terrestrial soil (Van den Hoogen et al., 2019). They can thrive in extreme environments, such as hyperarid deserts (Edgington et al., 2011; Wharton and Brown, 1989), deep sea geothermal vents (Gollner et al., 2013), arctic ice floes (Tchesunov and Riemann, 1995), the deep subsurface biosphere (Borgonie et al., 2011), and heavily polluted soils and sediments (Park et al., 2011; Shih et al., 2019). It is a common claim that 4 out of every 5 multicellular animals on earth are nematodes (Platt, 1994). This claim is difficult to substantiate, but free-living nematodes are obviously widespread, and can achieve extremely high population densities in organically rich substrates. Measured densities range from 120 thousand nematodes per square meter in the rhizospheres of perennial desert plants

(Freckman, 1979) to 80 million nematodes per square meter in marine sublittoral sediments (Boucher & Lambshead, 1995).

As members of a highly diverse phylum, nematodes participate in a wide variety of trophic interactions and operate at multiple trophic levels (Bongers & Bongers, 1998). There are nematodes that function as primary consumers, predators, parasites, decomposers, and apex predators (Mucci et al. 2022). There are even species that, with the help of endosymbiotic chemoautotrophic bacteria, function much like producers in marine sediments near methane seeps (Austen et al., 1993). In terrestrial soils, free-living nematodes aid in the decomposition of organic matter (Freckman, 1988), mineralization and redistribution of nutrients, soil structuring, and many other ecosystem processes (Ferris, 2010). Diversity of nematodes is often linked with soil productivity and resilience (Chen et al., 2020). Their small average adult size (<1mm) makes them unable to escape habitat disturbance (Ahmed et al., 2015) and they exhibit rapid numerical and behavioral responses to environmental changes (Santos et al., 1981). Accordingly, nematode community composition has been found to correlate with a multitude of biotic and abiotic variables such as soil productivity (Yeates, 1984), sediment composition and soil pore size (Tietjen, 1989; Wallis, 1968), physio-chemical properties and pollutants (Neher, 2001; Salamun et al., 2014; Gutierrez et. al, 2016), organic matter (Marais et al., 2020), and water retention (Jones et al., 1969). The preference of some free-living nematode species for a particular soil type or geographic area suggests a degree of endemism (Porazinska et al., 2010), though this aspect of their natural history is not well-understood (Zullini, 2018).

Despite estimates of over 1 million nematode species, only about 27,000 have been described (Hugot et al., 2001; gbif.org, 2022). This is not because of a lack of species diversity in Phylum Nematoda, but rather a lack of sufficient time, funding, and personnel devoted to taxonomic description of nematode species (Baldwin et al., 2000; Coomans, 2002). Due to their small size and subtly differentiated morphological characteristics, high resolution microscopes and years of specialized training are required for taxonomic study. Moreover, phylum Nematoda is rife with instances of convergent morphological evolution (Blaxter et al., 1998; Quist et al., 2015), which further complicates morphological species delimitation.

On the molecular front, nematode taxonomy is impeded by a scarcity of genomic data. Less than 150 genomes have been sequenced, and species representation in single-locus (e.g., 18S rDNA) reference sequence databases is biased toward medically and agriculturally important taxa. Also, a significant portion of sequences in available databases are not comparable because they are from non-overlapping gene regions, and taxonomic misassignments are a sporadic yet persistent nuisance (Kozlov et al., 2016).

Our current inability to close the gap between known and unknown biodiversity is commonly referred to as “the taxonomic impediment” (Hoagland, 1996; Godfray, 2002), but it is also an impediment to the study of nematode ecology (Giangrande, 2003). Nematode ecologists have tried to circumvent these limitations by focusing on functional diversity over taxonomic diversity. Specimens are often categorized into trophic groups or functional guilds based on the

morphology of their mouthparts (Yeates, 1993). This approach does not account for the multiplicity of feeding strategies a single morphotype can adopt (Moens and Vincx, 1997), or the ability of some taxa to assume multiple morphotypes, which change depending on life stage or environmental conditions (Yeates, 1987; Kiontke and Fitch, 2010). Furthermore, there may be species with similar feeding morphology and habits that are otherwise adapted to significantly different habitats.

Model systems make ecological research more convenient when they display properties and processes of interest in a way that makes them simpler and more accessible (Vitousek, 2002). Desert soil habitats have many idiosyncrasies that make them appealing as ecological model systems. Desert landscapes are expanses of nutrient poor bulk soil dotted with localized resource patches, which are created by perennial plants and animal burrows or nests. These islands of fertility (Schlesinger, 1996) create spatial heterogeneity of soil characteristics and nutrient availability (Wright et al., 2006). As with vernal pool model systems (Carrino-Kyker et al, 2008), organisms in desert soil habitats are exposed to additional stresses related to seasonal and diurnal variations in moisture and temperature. These factors can increase the functional diversity (e.g., survival adaptations) and the species richness of below-ground communities (Procter, 1990).

Because nematodes are integral to soil health (Wilson et al, 2009), are sensitive so disturbance, and rapidly respond to environmental variables, they are appealing as model organisms for the study of spatial and community ecology. They occur in community compositions that are

particular to and representative of their habitat (de Goede and Bongers, 1994). Additionally, the species richness of desert soil nematodes is high enough to allow for the study of complex trophic interactions, yet low enough for species inventories to be tractable.

Another useful trait of desert soil nematodes is their ability to enter anhydrobiosis, a quiescent metabolic state in which their tissues contain less than 0.1% water (Crowe, et al., 1992). In this state, they can persist for extended periods (months to years) while soils remain dry (Freckman et al, 1977b). They reactivate during short wet periods following seasonal rains, when soils are saturated from rain directly or by flood waters from storms in nearby mountains. This punctuated existence, characterized by periods of dormancy and activity, allows nematodes to thrive in arid soils by synchronizing their life cycles with the availability of water. In addition to anhydrobiosis, their small size, and numerical abundance make it convenient to sample and preserve specimens for later study.

Historically, nematode systematics has tended to be compartmentalized, with taxonomic specialists focusing on their preferred taxa within disparate habitats (De Ley and Blaxter, 2002). This siloed approach (Bik, 2019) has led to a fragmented understanding of nematode diversity and ecology, hindering comprehensive insight into the role of nematodes in various ecosystems. A better strategy may be to reorient our taxonomic endeavors toward the characterization of potential ecological model systems. This would necessarily involve a more collaborative and integrative approach, drawing on the expertise of multiple taxonomists and using the best morphological and molecular techniques available. A more strategic focus on

potential model systems will not only make taxonomic efforts more efficient, but also establish a vital foundation for comprehensive ecological studies. Traditional studies typically begin with painstaking and time-consuming morphological characterization of specimens, often done by taxonomists with expertise in specific clades, followed by molecular characterization to confirm their findings. This approach has been driven by the importance of type specimens in taxonomy (Tautz, 2003) and the perception that morphological analyses were more efficient in recognizing species-level diversity (Dayrat, 2005).

In so-called reverse taxonomy (DeLey et al., 2005; Markmann and Tautz, 2005), capture of detailed morphological information from nematode specimens is done using a computer-assisted high-resolution microscope. Images of the nematodes' body regions are captured in multiple focal planes (z-series), and the data are saved for later use by specialized taxonomists to delimit, describe, and/or identify species. This method of morphological documentation, referred to as virtual microscopy (De Ley and Bert, 2002), can be done relatively quickly. Next, the specimen is subjected to DNA extraction, followed by PCR amplification and DNA sequencing of one or more loci. The data are then evaluated to determine if the specimen can be identified by sequence, or if morphological data from this and other sequenced specimens require additional study by a taxonomic specialist. This approach reduces time and resources spent on already discovered taxa, and it allows taxonomists to prioritize the specimens of greatest interest or rarity.

DNA barcoding is another useful tool for increasing the rate at which new species are discovered. While many characters (molecular and morphological) are required to delimit

species, single-gene DNA barcoding can be used detect new genotypes and thereby prospect for new species (Blouin, 2002). The main challenge is finding a barcoding locus that is evolutionarily conserved enough for PCR primers to amplify favorably in a wide range of microscopic taxa, while maintaining high enough sequence variation to resolve relationships at lower taxonomic levels.

The 18S ribosomal DNA (rDNA) gene has become the favored locus for barcoding studies of nematode diversity and ecology (Creer et al., 2010; Schenck et al., 2020). This is because it is a high copy number gene (Bik, et al., 2013) with multiple highly-conserved regions that are amenable to the design of primers. Additionally, 18S rDNA is the most represented locus for nematodes in publicly available databases, with 2174 nematode sequences in the SILVA 138.1 18S database (1446 class *Chromadorea*, 728 class *Enoplea*) and ~32,490 18S nematode sequences in the NCBI database (27,272 *Chromaorea*, 5218 *Enoplea*). A major drawback of the 18S locus is that it does not exhibit enough variation among many nematode taxa to effectively resolve diversity below the family level.

The 28S rDNA gene, despite also having conserved primer-binding sites and high copy number, is not as well represented in publicly available databases, with 371 nematode sequences in the SILVA (138.1) 28S database (274 *Chromadorea*, 97 *Enoplea*) and 7991 28S sequences in the NCBI database (6556 *Chromadorea*, 1435 *Enoplea*). Nevertheless, the 28S rDNA locus has more average variation than 18S across nematode taxa (Hillis and Dixon, 1991), making it more effective for resolving diversity below the family level in many groups (Nadler, 1992).

In this study, we used a reverse taxonomy approach to characterize soil nematode communities within Boyd Deep Canyon Reserve (BDCR), a desert habitat and potential ecological model system. We documented morphological features of nematode specimens (using virtual microscopy) taken from soil beneath perennial desert plants. We then constructed a BDCR-specific 28S reference sequence database, wherein all DNA sequences are linked to a documented morphotype. Then, to evaluate the reference sequence database, we used a 96-well plate-based DNA barcoding method to classify 2,480 nematode specimens collected from nine soil samples in the same habitat. We hypothesized that nematode community composition would correlate with the plant species or the collection locality from which the samples were taken. Our goals were to 1) estimate the number of nematode species in BDCR alluvial fans using reverse taxonomy and 96-well plate sequencing strategies, 2) detect new species for future generation of descriptions, and 3) establish a useful model system for the study of nematode ecology, and potentially for broader soil ecology studies.

1.2 Materials and methods: reverse taxonomy

This work was done by members of the Nadler Lab at The University of California, Davis in collaboration with members of the Baldwin Lab at The University of California, Riverside. Morphological documentation of specimens and PCR were done with the help of undergraduate researchers in the Baldwin Lab and Nadler Labs. Morphological identification was done by Jim Baldwin, Steven Nadler, and Mirayana Marcelino Barros. Additional identification of dorylaimidan nematodes was done by Sergio Ãlvarez-Ortega at Rey Juan Carlos University, Madrid.

1.2.1 Study site and sample collection

In May 2014, nine soil samples were collected within Boyd Deep Canyon Reserve (BDCR) in Palm Desert, California, which is located on the Northwestern edge of the Sonoran Desert (**Table 1.1**). In 2014, the year that our soil samples were collected, BDCR received approximately 2.48 inches of rain. Three collection localities were chosen within BDCR on a northward-facing alluvial fan created by stormwaters from Sheep, Deep, and Coyote canyons (Zabriskie, 1979). The soil is predominantly sandy, with numerous small rocks and plant debris. The uneven terrain is dotted by sandy hummocks, most of which are topped by perennial plants. Previous research has demonstrated that nematode abundance is greatly increased in the rhizospheres of perennial desert plants relative to the surrounding bulk soil (Freckman and Mankau, 1977). Three abundant perennial plant species at BDCR are: Creosote (*Larrea tridentata*), Palo Verde (*Parkinsonia florida*), and Chuparosa (*Justica californica*). At each of our three collection localities, one soil sample was taken from beneath a plant from each of the above species. The three selected plants were within 30m of each other to increase the likelihood of similar environmental factors. The elevation of each sample location was inferred from GPS coordinates using Google Earth Pro (version 7.3.6.9345).

Approximately 2.3kg of soil was taken from the rhizosphere of each plant to a depth of ~30cm, and within 40cm from the center of the plant. Samples were bagged, GPS coordinates and site photos taken, and a labeled aluminum tag with galvanized stake was placed below the soil

surface so that each site could be located and resampled in the future. Samples were transported back to UC Davis and kept dry at 10C. Soil samples were gently mixed and divided into three sub-fractions for use in three experiments, each employing a different strategy for assessing nematode diversity. Reverse taxonomy methods were used on the first set of 9 subfractions to construct a 5' 28S reference sequence database wherein every BDCR-derived DNA sequence is linked to morphological data in the form of microscope image captures. The newly created reference sequence database was then used to classify sequences representing 2480 individual nematodes that were isolated from the second set of 9 soil subfractions, PCR-amplified in a 96-well plate format, and sequenced using the Sanger dye-terminator method by the DBS UCDNA Sequencing Center. The third set of 9 soil subfractions was saved for a future experiment (**Chapter 2**).

1.2.2 Hydration of soil samples and isolation of nematodes

After the removal of rocks (>2cm diameter) and plant debris, soil subfractions (~285cm^3) were placed in four 3oz paper cups with drainage holes added. The soil was then fully saturated with water by first standing the cups in tap water and then removing them to drain. The cups were then incubated for 72 hours at RT. Nematodes were isolated from the hydrated soil samples using a 6" diameter plastic Baermann funnel over a 14-hour period at 28.5C in an incubator. Baermann funnel catches were decanted into a plastic petri dish, and an inverted microscope (400x) was used to select nematodes based on morphology, with the goal of sampling as many

morphotypes as possible. Selected nematodes were picked onto a microscope slide with an agarose pad for further examination and image capture under higher magnification.

1.2.3 Virtual microscopy

A Nikon Eclipse E600 with DIC optics and equipped with a computer-driven Z-axis controller and digital camera was used to obtain a series of through-focus image stack that vary incrementally in depth (Z-step varying with magnification). Morphological features of each specimen were documented via software-assisted (Nikon Elements) through-focus imagery. The result is a progressive series of images (Z-series capture), saved as an MP4 file, that includes multiple depths of field for morphological identification/characterization of specimens. Image captures were obtained at 100X-1000X magnification, depending on the specimen. In total, 160 specimens were characterized in this way. These data were used later to classify specimens on the basis of morphological characteristics.

1.2.4 DNA extraction

After documentation of their morphological features, nematode specimens were removed from the specialized microscope slides for DNA extraction. Individual worms were cut into 2 pieces using a tuberculin needle or a microscalpel, and the pieces picked into 24uL of prepGEM Gold digestion buffer in a 0.5ml microcentrifuge tube. The tubes were subjected to three freeze-thaw cycles in dry ice, and then 0.25uL of prepGEM tissue digestion enzyme was added to each tube.

The tubes were incubated at 75C for 1 hour, vortexed for 10 sec, centrifuged and incubated at 75C for another hour, then the enzyme was heat-inactivated at 95C for 10 minutes. DNA extracts were then stored at -20C for future use.

1.2.5 Amplification of 28S rDNA and sanger sequencing

A ~750bp piece of 28S rDNA (D1-D3) was amplified for each of the sampled nematodes using the KOD XL polymerase kit (MilliporeSigma). PCR reactions were done in a 25ul volume and included: 1.5ul of DNA extract, 3mM $MgCl_2$, 0.2mM of each dNTP, 0.5 units of KOD XL Polymerase, and 0.6 uM each of forward primer and reverse primer. Thermocycling conditions consisted of an initial denaturation step at 95C for 3 minutes, followed by 35 cycles of 95C for 30 seconds, 54C for 30 seconds, and an extension step at 72C for 1 minute. After 35 cycles, there was a final extension step for 6 minutes at 72C. Any DNA extracts that failed to amplify with primer pair 787F/792R were amplified with primer pair 391F/501R. PCR and sequencing primers are detailed in **Table 1.2**. Agarose gel electrophoresis was used to compare PCR products to the predicted sizes for each primer pair.

Prior to sequencing, 10ul of each PCR product was incubated at 37C with 1uL of Shrimp Alkaline Phosphatase and 1uL of Exonuclease I (USB). PCR products from 160 individual nematodes were submitted for automated sequencing (CBS UCDNA Sequencing Center) with the respective PCR primers and additional internal primers (**Table 1.2**). Sequence traces were edited and assembled

into contigs using CodonCode Aligner version 7.1.2. PCR primer sequences were removed before further analyses.

1.2.6 Construction of reference sequence database

A reference database consisting of 4292 5’ (D1-D3) 28S sequences was created for the purpose of identifying nematode specimens. The *rescript* plugin (Robeson et al., 2021) for the Qiime2 bioinformatic pipeline (Boylen et al., 2019) was used to download 35798 28S rDNA sequences from the NCBI DNA sequence database. Nematomorph (sister taxa to Nematoda) and Rotifer sequences were included in order to sharpen the distinction between nematode and non-nematode taxa during classification. The Boolean search terms were '(Nematoda[Nucleotide] OR Nematomorpha[Nucleotide] OR Rotifera[Nucleotide]) AND (28S[Title] OR large subunit ribosomal RNA[Title]) NOT environmental NOT Environmental NOT Unidentified NOT Unverified'.

The Qiime2 *extract-seq-segments* function, with the sequences of PCR primers 391F and 799R (**Table 1.2**), reduced our initial pool of 35798 reference sequences to 8801 that contained the first three domains of 28S rDNA. Redundant reference sequences were then eliminated using the *rescript dereplicate* function at a threshold of 99%. The remaining 3978 sequences were aligned using MAFFT with 5 cycles of iterative refinement. Then, the SATIVA phylogeny-aware database analysis tool (Kozlov, et al., 2016) was used to identify and change 44 reference

taxonomies that had incorrect assignments above the family level, as well as taxonomies that were inconsistent with the revised order names proposed by Blaxter and DeLey in 2002.

After elimination of redundant reference sequences and correction of taxonomical mislabeled sequences, 314 hand-picked reference sequences were manually added into the database. These included:

1) 9 additional nematode sequences from the SILVA 28s eukaryotic database (release 138.1), which did not have corresponding Genbank accession numbers.
2) 32 miscellaneous eukaryotic and archaeal sequences were selected based on sequencing of previous desert soil samples and classification of non-nematode contaminants using the *sklearn* classifier and the SILVA 28S eukaryotic database (**Table 1.3**).
3) 67 sequences that were generated in the BDRC reverse taxonomy experiment (**section 1.2**).
4) 206 sequences that were selected from two publications (Smythe and Nadler, 2006; Nadler et al., 2006), and unpublished sequences from the Nadler Lab sequence database.

1.3 Materials and methods: *96-well plate experiment*

This experiment was done using the second set of 9 soil subfractions collected in May of 2014**.**

1.3.1 Hydration of soil samples, isolation of nematodes, and counting

Soil subfractions (~712cm^3) were distributed into ten 3oz paper cups with drainage holes, fully saturated with water, then allowed to drain. They were then incubated for 72 hours at RT. Nematodes were isolated from the hydrated soil samples using three 6" diameter Baermann funnels over a 14-hour period at 28.5C using an incubator. Funnel catches (~40ml) from each sample were combined in a 250ml beaker.

Soil samples are rife with PCR-inhibitory substances, such as humic acids and polyphenols, and their presence is often indicated by the amber to brown coloration of the water in each funnel catch. To reduce the concentration of inhibitors in the bodies of the nematodes, each funnel catch was subjected to a series of water exchanges. The volume of liquid was adjusted to ~150 ml. The beakers were then placed on ice to reduce activity of the nematodes, thereby increasing the rate at which they settle to the bottom of the beaker. After 30 min on ice, the top 100 ml of liquid was aspirated off. This process was repeated twice more, or until the liquid in the beaker was colorless.

Funnel catches were decanted in a 100mm square petri dish with a 6x6 grid. Using a stereoscope, all nematodes were counted in 6 randomly selected cells of the grid. Cells were selected (X and Y coordinates) using a random number generator. Counts were averaged over 6 cells and worm count was estimated over the total area of the petri dish.

1.3.2 DNA extraction

All nematodes from a randomly selected cell were cut with a microscalpel and picked individually into 96-well plates, which had 20uL of 1X prepGEM gold digestion buffer in each well. This process was repeated until four 96-well plates were filled per soil sample. For one of the soil samples, A122, four additional plates were filled in order to provide more specimens for the optimization of DNA extraction and PCR. The plates were then subjected to three freeze-thaw cycles using dry ice, and 0.25ul of prepGEM tissue digestion enzyme was added to each well. The plates were incubated at 75C for 1 hour, vortexed for 10 seconds, then incubated at 75C for another hour. Then the prepGEM enzyme was heat-inactivated at 95C for 10 minutes. Plates were then stored at -20C for future use.

1.3.3 PCR and sequencing

A ~750bp fragment of 28S rDNA was amplified using a two-step PCR reaction in a 96-well plate format. PCR reactions were done in a 25ul volume and included: 1.5ul of DNA extract, 3mM $MgCl_2$, 0.2mM of each dNTP, 0.5 units of KOD XL Polymerase, and 0.6 uM each of forward primer and reverse primer. Reaction conditions included an initial denaturation step at 95C for 3min, followed by 35 cycles of 95C for 30sec and 67C for 1min, and a final elongation step at 72C for 7min. Products were then electrophoresed in 1.3% agarose gels, and DNA visualized with ethidium bromide.

Some sequencing reactions were re-run with alternative and/or supplemental sequencing primers. This was mainly due to a mismatch of primer 391F in some Cephalobomorphs and Aphelenchoids, prompting the use of primers 813F, 814F, and 817F (**Table 1.2**). Stutter artifacts, resulting from low-complexity sequence at the 5' amplicon end, were observed in some Dorylaimomorphs (eg. *Microdorylaimus sp.*). To address this, we used internal sequencing primers 811F and 812F.

In another 96-well plate, 10ul of each PCR product was incubated for 15 min at 37C with 1uL of Shrimp Alkaline Phosphatase and 1uL of Exonuclease I (Affymetrix). The samples were then incubated at 80C for 20 min to inactivate the enzymes. Samples were sequenced (ABI 3730) at the DBS DNA Sequencing Center at University of California, Davis. Initial sequencing reactions were done with primer 799R. Depending on the performance of these reactions, alternative and/or additional sequencing primers (**Table 1.2**) were used to obtain supplemental traces. Sequence traces were edited, and primer sequences were removed using CodonCode Aligner version 7.1.2.

1.3.4 Bioinformatics

Sequences with less than 400 bases with Phred20 quality scores over 80 (Ewing et al., 1998) were not used for further analyses. Some products required re-sequencing with internal primers (**Table 1.2**) because of low signal strength or sequence stutter. Sufficiently high-quality sequences were exported from CodonCode Aligner as FASTA-formatted text files, sorted by soil

sample, and concatenated into nine FASTA files, which were then reformatted for analysis using the Qiime 2 bioinformatics platform (Caporaso et al., 2010). A de-novo OTU-picking method was used to cluster nematode sequences at a 99% identity threshold (Creer et al., 2010; Bik et al., 2012), resulting in 79 representative OTUs. A custom reference sequence database (**Section 1.2.7**) and the *sklearn* naïve Bayesian classifier (Bokulich et al., 2018) were used to assign taxonomy to representative OTUs at a 70% confidence level. This default confidence level was implemented in the *sklearn* classifier because it offered the best balance between over-classification and under-classification of mock bacterial 16S and fungal ITS sequences. After testing the *sklearn* classifier on our BDCR Sanger sequence dataset using various confidence settings (detailed in **section 1.4.3**, shown in **Table 1.4**), we also opted to use the default setting of 70% confidence. Following classification of sequences, the resulting OTU table and metadata were exported as a BIOM formatted file (McDonald et al., 2012) for statistical analyses in R.

1.3.5 Data analyses

Statistical analyses were performed, and accompanying figures were produced in R (4.2.3). OTU accumulation curves were made in order to assess the sufficiency of sampling. An upset plot and scaled Venn plots were generated to analyze prevalence and distribution of OTUs among samples.

A heatmap was generated to provide abundance information at the lowest taxonomic level possible, and a stacked bar plot, organized by genus, was generated to visualize taxonomic

diversity among samples. In the heatmap figure, rows representing taxonomic categories were organized using a maximum likelihood tree, which was generated using the IQ-Tree java application (http://www.iqtree.org).

In order to assess functional diversity of the nematode communities in our samples, trophic groups were assigned to OTUs based on their genus-level classifications (Bonger and Bongers, 1998). Stacked barplots were generated to visualize relative abundance of trophic groups among individual soil samples and among collection sites and plant species.

Alpha diversity was estimated using Chao1, Simpson, and Shannon diversity indices. The Chao1 index was used to estimate overall richness of samples, the Shannon index was included to assess diversity with species evenness taken into account, and the Simpson index was included to provide a diversity estimate that is less affected by low-abundance OTUs, and therefore less influenced by the presence of spurious or possibly chimeric sequences.

Both plant species and collection site were explored as factors influencing nematode community composition. We performed a non-metric multidimensional scaling (NMDS) analysis using Bray-Curtis distances to visualize dissimilarity of nematode communities among our samples with respect to collection sites and plant species. Similarly, an NMDS plot was constructed using Jaccard distances in order to assess dissimilarity among samples without the influence of relative abundance.

Following the NMDS analyses, permutational multivariate analyses of variance (PERMANOVA) were done using Bray-Curtis distance to test for significant differences in species abundances and distributions among collection sites and plant species. A relatively high number of permutations were needed (9999) to achieve stable p-values. This is likely due to the low sample number (n= 9) and the lack of replication by plant species within each collection site. A similar PERMANOVA was also done using Jaccard distances to evaluate the statistical significance of differences in species presence-absence patterns within and between collection sites and plant species. To differentiate among location, plant species, and dispersion effects, we used a beta dispersion test via the *betadisper* function in the *vegan* package in R.

1.4 Results

1.4.1 Reverse taxonomy results

Of the 160 documented nematode specimens that were morphologically documented and sequenced for 5' 28S rDNA, 26 morphotypes/genotypes are hypothesized to be distinct lineages (**Table 1.4**). Of these, 3 lineages are described species collected previously in BDCR, 14 are thought to be new species and awaiting description, and nine are awaiting further morphological study by taxonomic experts to determine whether they are a new or previously described species.

1.4.2 96-well plate DNA extraction and PCR performance:

Of the 3840 nematodes that were subjected to DNA extraction, 3099 specimens (~80.7%) yielded positive PCR products, which in turn yielded 2480 (~80%) DNA sequences that met sequence quality standards (**section 1.3.5**). Sample A130 had a particularly low rate of PCR success (~46%), most likely due to a greater concentration of inhibitors in that soil sample. Sequence totals for each soil sample are detailed in **Table 1.5**. Sequence counts were highly variable among samples ($\mu \approx 275.6$, $\sigma \approx 84.9$).

1.4.3 Testing the sklearn *classifier:*

We tested the *sklearn* naïve Bayesian classifier on our Sanger sequences at 70%, 80%, 90%, 95%, and 99% confidence levels. The number of OTUs classified to species level decreases from 32 at 70% confidence to 25 at 99% confidence (**Table 1.6**). With respect to higher taxonomic levels, regardless of confidence setting, a total of three family-level and two order-level taxonomic categories were found.

1.4.4 Classification results

Classification of sequences at 70% confidence yielded 44 taxonomic categories, 32 of which were classified to species level (**Table 1.7**). Of those 32 species, 15 were unnamed putative species found during our reverse taxonomy experiment at BDCR (**section 1.2.3**). Seven taxonomic categories could not be classified below genus, three categories could not be

classified below family, and 2 categories could not be classified below order. All taxonomic categories, the number of OTUs per category, and the counts of OTUs in each soil sample are detailed in **Figure 1.1**. Of the 79 OTUs that resulted from the dereplication process (**section 1.3.5**), 53 OTUs were found to be members of Class *Chromadorea*, and 26 OTUs were of Class *Enoplea*.

Of the 2480 final Sanger sequences, 85.4% were classified to species level, 3.5% to genus, 9.8% to family, and 1.3% to order. Chromadorean nematodes were numerically dominant, making up ~94.0% of sequences, while the remaining ~6.0% consisted of enopleans. All chromadorean nematodes were members of Order *Rhabditida* and all enoplean nematodes were members of Order *Dorylaimida* (**Figure 1.2**). Of the 11 families that were successfully classified, two most abundant chromadorean families were *Cephalobidae* (~76.5%) and *Panagrolaimidae* (~16.1%). The most abundant enoplean families were Qudsianematidae (2.4%) and Aporcelaimidae (~1.1%). The remaining ~3.9% of sequences were successfully classified to seven different families, and there were a small proportion of sequences (~1.57% of total reads) that could not be classified below the order level (unclassified *Dorylamida*).

Within Family *Cephalobidae (Rhabditida)*, seven genera were found, of which *Acrobeles* and *Acrobeloides* were the most abundant (**Figure 1.3**). Within Family *Panagrolamidae*, only one genus was found, *Panagrolaimus*. The remaining panagrolaimid OTUs could not be classified below family level at 70% confidence. Within Order *Dorylamida*, eight genera were found. The

most abundant of these were *Aporcelaimus* and *Ecumenicus*, though they accounted for just 1.0% and 1.4% of total BDCR sequences, respectively.

In comparison with the 26 putative species found using reverse taxonomy methods, the Sanger 96-well plate sequence data revealed 32 species-equivalent OTUs, as well as 11 OTUs that were classified to higher taxonomic levels. Nearly half (~47%) of the species-equivalent OTUs in our Sanger 96-well sequencing experiment came from reference sequences obtained via the reverse taxonomy experiment, which in turn have linked morphological data from specimens collected in the same habitat.

1.4.5 OTU accumulation curves

OTU Accumulation curves for all soil samples approach but do not reach saturation (**Figure 1.4A**), suggesting that additional sampling would be needed to capture total diversity. The curve representing sample A122 had the steepest slope and highest OTU richness at all sampling depths. This is likely a consequence of the high DNA sequence count (n=490) of A122 compared to the other samples ($\mu \approx 275.6$, $\sigma \approx 84.9$) (**Table 1.5**). Sample A130 had the lowest DNA sequence representation (n= 176) because of a low PCR success rate, likely caused by PCR inhibitors in that particular soil sample. Despite this, the OTU accumulation curve for sample A130 achieved the highest saturation among all samples. As might be expected, the disproportionately high sample size of A122 inflates Site 1 and Palo Verde OTU richness

estimates when accumulation curves are organized by collection site (**Figure 1.4B**) and plant species (**Figure 1.4C**).

1.4.6 Distribution of OTUs

Over half (43 OTUs, 62%) of the 79 observed OTUs across DBCR Sanger sequences were unique to just one of the soil samples (**Figure 1.5**). There were 15 OTUs (~19% of all) that were unique to Palo Verde sample A122, and there were eight OTUs that were unique to sample A123. Other soil samples did not have more than four unique OTUs, and there were no unique OTUs found in samples A126 and A130. Scaled Venn plots show the distribution of OTUs among collection sites (**Figure 1.6A**), and plant species (**Figure 1.6B**). Of the 79 observed OTUs, ~73% were found in Site 1, while ~47% could be found in Site 2, and ~37% could be found in Site 3. With respect to plant species, ~72% of OTUs could be found in Palo Verde samples, ~47% in creosote samples, and ~39% in chuparosa samples. As expected, sample A122 has a significantly higher OTU count when compared with other samples (**Figure 1.7**).

1.4.7 Alpha Diversity

Sample A122 was an outlier with an OTU richness estimate (n=40) that far exceeds all other samples ($\mu \approx 19.89$, $\sigma \approx 9.44$). Because of sample A122, the mean observed OTU richness, mean Chao1 diversity, mean Shannon diversity, and mean Simpson diversity are all inflated for Site 1 samples (**Figure 1.8A**) and for Palo Verde samples (**Figure 1.8B**). Samples from Site 3 tended to

have the lowest mean OTU richness, mean Chao1 diversity, mean Shannon diversity, and mean Simpson diversity. Sample A123, from a chuparosa bush at site 1, had relatively low observed OTU richness and Chao1 diversity, yet it had relatively high Shannon and Simpson diversity, suggesting that the few species present therein were equally abundant.

The 79 observed OTUs found at BDCR were categorized into 44 taxonomic categories. Of these, 7 taxonomic categories (20 sequences) were fungivores, 24 taxonomic categories (2308 sequences) were microbivores, 12 taxonomic categories (139 sequences) were omnivore-predators, and 4 taxonomic categories (13 sequences) were plant parasites. The relative abundance of trophic groups among BDCR 96-well plate sequences was organized by soil samples, collection sites, and plant species, and is shown in **Figures 1.9** and **1.10**.

Site 1 had a greater number/proportion of fungivores than the other collection sites. All three plant species had fungivores at Site 1. Fungivores were nearly absent at other sites, except the chuparosa sample A130 at Site 3. This is also why, on average, chuparosa plants had a higher proportion of fungivores (**Figure 1.10**).

The 2308 observed microbivores consisted mainly of cephalobids (~82.1%), with a smaller proportion of panagrolaimids (~17.2%) and trace numbers of rhabditids (<1%). There are three soil samples with significant numbers of panagrolaimids: the paloverde sample at collection site 1 (A122), and the chuparosa and creosote samples at collection site 2 (A126 and A127; **Figure 1.3**). Sample 126 is unusual in that over half of the microbivores therein are panagrolaimids.

Creosote samples tended to have more omnivore-predators at all three collection sites, with sample A127 having the greatest relative abundance of omnivore-predators overall. Chuparosa samples tended to have the lowest proportion of omnivore-predators at all three collection sites (**Figure 1.10**). Palo Verde samples, on average, had the greatest number of plant parasites.

1.4.8 Beta Diversity

As seen in both the Bray-Curtis (**Figure 1.11**) and the Jaccard distance (**Figure 1.12**) based NMDS plots, nematode communities clustered more by plant species than they did by collection site. In both plots, chuparosa samples cluster at the right, while creosote samples cluster at the lower left. Palo Verde samples are more scattered, indicating relatively lower similarity among Palo Verde trees from different collection sites. Stress values 0.075 indicate that the NMDS plots captured the main patterns of dissimilarity among nematode communities in BDCR soil samples.

The PERMANOVA using Bray-Curtis distance (**Table 1.8**) found no significant effect of collection site on nematode community composition (R^2= 0.153, $F_{2,6}$= 0.544, p= 0.925), while plant species showed a marginally significant effect (R^2= 0.401, $F_{2,6}$=2.009, p= 0.061). Similarly, the Jaccard index-based PERMANOVA indicated no significant influence of collection site on nematode community composition (R^2= 0.190, $F_{2,6}$= 0.701, p= 0.871) and a borderline significant effect of plant species (R^2= 0.350, $F_{2,6}$= 1.614, p= 0.071).

The Bray-Curtis dispersion analysis indicated no significant differences in the variability of nematode community composition among collection sites ($F_{2,6}$= 0.135, p= 0.876) and plant species ($F_{2,6}$= 0.535, p= 0.611). The Jaccard dispersion analysis also did not indicate any significant differences in the variability of nematode community composition among collection sites ($F_{2,6}$= 0.165, p= 0.876).

Sample A122 is the most divergent sample. This is expected because the sample size of A122 is disproportionately large (n= 490) in comparison with our other samples (μ= 275), and the Bray-Curtis dissimilarity metric is sensitive to differences in total abundance among samples (Ricotta and Podani, 2017). Sequence data were not rarified prior to the generation of the NMDS plots shown below (**Figures 1.11 and 1.12**). In test analyses, samples were rarified at the minimum sample depth (Sample 130, n= 176), and as a result, ~36% of the data were lost. Furthermore, the stochasticity introduced by random subsampling resulted in NDMS plots that were not reproducible.

1.5 Discussion

1.5.1 Regarding the sklearn classifier

The *sklearn* classifier is essentially a Python implementation of the RDP classifier (Wang, et al., 2007). Both are naïve Bayesian classifiers, and they perform similarly at default settings, but while the RDP classifier uses a bootstrap confidence score at each level of taxonomic

classification, the *sklearn* classifier uses a confidence cutoff that is based on the posterior probability of classifications at each taxonomic level. Our test of the *sklearn* classifier (**Table 1.6**) was consistent with the results of Bokulich et al. (2018) in that there was a sharp decline in species-level classification at confidence settings over 95%. With respect to our data set, we cannot estimate recall or precision because we did not include mock or simulated nematode communities in our test. Moving forward, it will be important to establish benchmarks for classifier settings that are specific to nematode 28S rDNA sequence data.

1.5.2 Correlation of nematode community composition with collection site and plant species

Both collection site and plant species were explored as factors influencing nematode community composition using Bray-Curtis and Jaccard distance metrics. Variation in nematode communities could not be attributed to collection site with any statistical significance, whether based on Bray-Curtis distance (R^2= 0.153, $F_{2,6}$= 0.544, p= 0.925) or Jaccard distance (R^2= 0.190, $F_{2,6}$= 0.701, p= 0.871). In contrast, plant species seemed to have more substantial effects. Using Bray-Curtis distance, ~40.1% of the variation in nematode communities (R^2= 0.401) was attributed to plant species with an F-value of 2.01 and a p-value of 0.061. Using Jaccard distance, the association between plant species and nematode community composition is less pronounced (R^2= 0.350, $F_{2,6}$= 1.614, p= 0.071). Our results suggest a trend where plant species plays a role in shaping soil nematode communities while collection site has minimal effects. Future studies with more extensive sampling and increased replication of plant species within collection site will be necessary to substantiate these findings.

1.5.3 Species richness estimates

The reverse taxonomy approach produced 26 distinct nematode morphotypes/genotypes within BDCR alluvial fans, 14 of which are hypothesized to be putative new species. The increased sampling offered by the 96-well plate approach yielded 44 taxonomic categories, which included 32 species-level classifications. These findings highlight the richness of a BDCR soil habitat, while the non-saturation of OTU accumulation curves suggests that a significant portion of the nematode diversity in our BDCR study system remains undiscovered.

Of the 26 putative species documented via reverse taxonomy, 15 were among the 32 species-level classifications obtained in the 96-well plate study: *Ecumenicus monohystera* and 14 undescribed species. Two described cephalobid species found via reverse taxonomy were not among the species-level classifications in the 96-well plate dataset: *Zelida spannata* and *Stegelleta incisa* (**Table 1.4**). There were, however 49 sequences that were classified to genus *Zeldia*, and 205 cephaloid sequences that could not be classified to genus level. There were also three putative species from genus *Quinisulcius* (*Rhabiditida*; *Telotylenchidae*) that were not found among the 96-well plate sequences. It is unclear if these discrepancies are due to inadequate sampling during the 96-well plate experiment, DNA amplification failures, or conflicts between reverse taxonomy results and *sklearn* classifier results. Our custom 28S reference sequence database is unlikely to be the point of failure, given that it includes 4 *Quinisulcius* species and 79 sequences from Family *Telotylenchidae*. None of the 2480 96-well

plate sequences were classified to Family *Telotylenchidae*, and there were just 2 Rhabditidan sequences that could not be classified to family level.

Our trophic diversity estimates are consistent with the observations of Freckman and Mankau (1986), who found microbivores to be the most abundant trophic group, followed by omnivore-predators, then fungivores, and lastly plant parasites. Other surveys of nematode communities in desert perennial plant rhizospheres (Pen-Mouratov, et al. 2003; Gao at al., 2022; Nielsen et al., 2014) have tended to find more plant parasites than omnivore-predators and fungivores.

1.5.4 Detection of new species

The reverse taxonomy study yielded 14 putatively undiscovered species and nine species awaiting further study to determine whether they have been described (**Table 1.4**). Furthermore, a significant proportion (~15%) of the 96-well plate sequences were not classified to a species-equivalent level, and these data may be used to prospect for additional undescribed species.

1.5.5 Establishment of a useful model system

One of the most effective ways to facilitate future studies of nematode ecology in BDCR is to provide a reference sequence database that is useful within the study system and can be continuously built upon to increase its value. A suitably useful database will include sufficient

depth of representation (number of unique reference sequences within taxa) to distinguish interspecific variation from interspecific variation. It will also include a sufficient breadth of representation (variety of taxa) to distinguish novel taxa from described taxa.

Over 85% of 96-well plate Sanger sequences could be classified to a species-equivalent level with the current iteration of our reference sequence database (**Table 1.4**), but our results revealed some shortcomings that must be addressed. Our database had insufficient representation of three nematode families that were dominant in our study system: Family *Cephalobidae*, Family *Panagrolaimidae*, and order *Dorylamida*.

Family *Cephalobidae* accounted for 1922 (~77.5%) of the 2480 total sequences in our Sanger 96-well plate dataset, but ~15% of those sequences (spanning 6 OTUs) could not be classified to species level. Only 9.2% of our reference database consists of cephalobid sequences. Because they are the most dominant family in BDCR, the diversity of cephalobid sequences in the database must be increased in order to better represent genus and species-level diversity. In the NCBI sequence database, 28S rDNA currently has the highest representation of cephalobid genera among commonly used molecular barcoding loci (**Table 1.9**) with 22 genera, but only 14 genera are represented for the 28S D1-D2 regions. We detected six cephalobid genera among 1896 sequences in our dataset, and there were an additional 205 cephalobid sequences that could not be classified below the family level. Further analyses with a more developed database will determine whether these sequences are indeed from *Cephalobidae* or from another closely related family.

Family *Panagrolaimidae* was also one of the highest abundance families in our dataset. Of the 399 observed panagrolaimid sequences, 340 (91%) match a putative new species that was discovered via reverse taxonomy, 23 match an unnamed reference sequence from Genbank (to a species-equivalent level), and the remaining 36 sequences could not be classified to the species level. Currently, there are only 37 panagrolaimid sequences in our reference sequence database (<1.5%). This is particularly troublesome because rDNA sequence diversity tends to be high in Family *Panagrolaimidae* (Blaxter et al., 1998; Lee et al., 2002). Many additional reference sequences will be needed to better represent panagrolaimid diversity at the genus and species levels.

Order *Dorylaimida* accounted for less than 5% of 28S nematode sequences in BDCR soil samples, and ~44% of dorylamidan sequences could not be classified to species level, yet they account for 33% of OTUs overall. Currently, 780 sequences in our reference database (~18.2%) represent dorylaimidan taxa, with a disproportionately high representation of Family *Longidoridae* (359 sequences). Sampling captured much of the diversity of the nematode communities in each soil sample (**Figure 1.4A**), but there are still rare taxa that were not discovered. In most samples, these undiscovered species are likely to include dorylaimidan nematodes, which tend to have low numerical abundance but high species richness (Andrássy, 2009).

Future efforts to develop BDCR as a study system for soil nematode ecology should include:

1) Sampling of more habitats within BDCR, with greater replication and at multiple time points, to capture seasonal and interannual variability.
2) Characterization of abiotic environmental factors within BDCR habitats, such as geomorphic characteristics and soil physicochemical properties, to explore their effects on nematode distribution and community composition.
3) Characterization of soil flora and fauna in BDCR habitats, including functional diversity and relative abundance distributions, with an emphasis on revealing trophic and/or competitive interactions with nematodes.

Chapter 1 References:

Ahmed, M., et al. (2015). "Nematode taxonomy: From morphology to metabarcoding." Soil Discussions **2**(2): 1175-1220.

Andrássy, I. (2009). *Free-living nematodes of Hungary, II (Nematoda errantia)*. Budapest, Hungary: Hungarian Natural History Museum and Systematic Research Group of the Hungarian Academy of Sciences.

Austen, M., et al. (1993). "Astomonema southwardorum sp. nov., a gutless nematode dominant in a methane seep area in the North Sea." Journal of the Marine Biological Association of the United Kingdom **73**(3): 627-634.

Baldwin, J., et al. (2000). "Nematodes: pervading the earth and linking all life." Nature and human society, the quest for a sustainable world: 176-191.

Bik, H. M., et al. (2012). "Sequencing our way towards understanding global eukaryotic biodiversity." Trends in Ecology & Evolution **27**(4): 233-243.

Bik, H. M., et al. (2013). "Intra-genomic variation in the ribosomal repeats of nematodes." PloS one **8**(10): e78230.

Bik, H. M. (2019). "Microbial metazoa are microbes too." Msystems **4**(3): 10.1128/msystems. 00109-00119.

Blaxter, M. L., et al. (1998). "A molecular evolutionary framework for the phylum Nematoda." Nature **392**(6671): 71-75.

Blouin, M. S. (2002). "Molecular prospecting for cryptic species of nematodes: mitochondrial DNA versus internal transcribed spacer." International journal for parasitology **32**(5): 527-531.

Bokulich, N. A., et al. (2018). "Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin." Microbiome **6**(1): 1-17.

Bongers, T. and M. Bongers (1998). "Functional diversity of nematodes." Applied soil ecology **10**(3): 239-251.

Borgonie, G., et al. (2011). "Nematoda from the terrestrial deep subsurface of South Africa." Nature **474**(7349): 79-82.

Boucher, G. and P. J. D. Lambshead (1995). "Ecological biodiversity of marine nematodes in samples from temperate, tropical, and deep-sea regions." Conservation Biology **9**(6): 1594-1604.

Bolyen, E., et al. (2019). "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2." Nature biotechnology **37**(8): 852-857.

Carrino-Kyker, S. R. and A. K. Swanson (2008). "Temporal and spatial patterns of eukaryotic and bacterial communities found in vernal pools." Applied and Environmental Microbiology **74**(8): 2554-2557.

Chen, X., et al. (2020). "Contribution of bacterivorous nematodes to soil resistance and resilience under copper or heat stress." Soil Ecology Letters **2**: 220-229.

Coomans, A. (2002). "Present status and future of nematode systematics." Nematology **4**(5): 573-582.

Creer, S., et al. (2010). "Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises." Molecular Ecology **19**: 4-20.

Crowe, J. H., et al. (1992). "Anhydrobiosis." Annual review of physiology **54**(1): 579-599.

de Goede, R. G. and T. Bongers (1994). "Nematode community structure in relation to soil and vegetation characteristics." Applied soil ecology **1**(1): 29-44.

De Ley, P. and W. Bert (2002). "Video capture and editing as a tool for the storage, distribution, and illustration of morphological characters of nematodes." Journal of Nematology **34**(4): 296.

De Ley, P. and M. Blaxter (2002). Systematic position and phylogeny. The biology of nematodes, CRC Press**:** 1-30.

Edgington, S., et al. (2011). "Heterorhabditis atacamensis n. sp.(Nematoda: Heterorhabditidae), a new entomopathogenic nematode from the Atacama Desert, Chile." Journal of Helminthology **85**(4): 381-394.

Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome research **8**(3): 186-194.

Ferris, H. (2010). "Contribution of nematodes to the structure and function of the soil food web." Journal of Nematology **42**(1): 63.

Freckman, D. and R. Mankau (1977). "Distribution and trophic structure of nematodes in desert soils." Ecological Bulletins: 511-514.

Freckman, D., et al. (1977). "A comparison of techniques for extraction and study of anhydrobiotic nematodes from dry soils." Journal of Nematology **9**(2): 176.

Freckman, D. (1979). "Cryptobiosis and Its Effect on Metabolism and Production Estimates of Desert Nematodes." Final Progress Reports (Research Memorandum 77-19).

Freckman, D. W. (1988). "Bacterivorous nematodes and organic-matter decomposition." Agriculture, Ecosystems & Environment **24**(1-3): 195-217.

Giangrande, A. (2003). "Biodiversity, conservation, and the 'Taxonomic impediment'." Aquatic Conservation: Marine and Freshwater Ecosystems **13**(5): 451-459.

Gao, Z., et al. (2022). "Exploring the optimal grazing intensity in desert steppe based on soil nematode community and function." Land Degradation & Development **33**(14): 2512-2527.

Godfray, H. C. J. (2002). "Challenges for taxonomy." Nature **417**(6884): 17-19.

Gollner, S., et al. (2013). "Nematode succession at deep-sea hydrothermal vents after a recent volcanic eruption with the description of two dominant species." Organisms Diversity & Evolution **13**(3): 349-371.

Hillis, D. M. and M. T. Dixon (1991). "Ribosomal DNA: molecular evolution and phylogenetic inference." The Quarterly review of biology **66**(4): 411-453.

Hoagland, K. E. (1996). "The taxonomic impediment and the convention on biodiversity." Association of Systematics Collections Newsletter **24**(5): 61-62.

Hugot, J.-P., et al. (2001). "Biodiversity in helminths and nematodes as a field of study: an overview." Nematology **3**(3): 199-208.

Jones, F., et al. (1969). "The influence of soil structure and moisture on nematodes, especially xiphinema, longidorus, trichodorus and heterodera spp." Soil Biology and Biochemistry **1**(2): 153-165.

Kiontke, K. and D. H. Fitch (2010). "Phenotypic plasticity: different teeth for different feasts." Current Biology **20**(17): R710-R712.

Kozlov, A. M., et al. (2016). "Phylogeny-aware identification and correction of taxonomically mislabeled sequences." Nucleic acids research **44**(11): 5022-5033.

Lee, D. L. (2002). The biology of nematodes, CRC Press.

Marais, E., et al. (2020). "Profiling soil free-living nematodes in the Namib Desert, Namibia." Journal of Arid Land **12**: 130-143.

Markmann, M. and D. Tautz (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**(1462): 1917-1924.

Moens, T. and M. Vincx (1997). "Observations on the feeding ecology of estuarine nematodes." Journal of the Marine Biological Association of the United Kingdom **77**(1): 211-227.

Mucci, N. C., et al. (2022). "Apex predator nematodes and meso-predator bacteria consume their basal insect prey through discrete stages of chemical transformations." Msystems **7**(3): e00312-00322.

Nadler, S. A. (1992). "Phylogeny of some ascaridoid nematodes, inferred from comparison of 18S and 28S rRNA sequences." Molecular Biology and Evolution **9**(5): 932-944.

Neher, D. A. (2001). "Role of nematodes in soil health and their use as indicators." Journal of Nematology **33**(4): 161.

Nielsen, U. N., et al. (2014). "Global-scale patterns of assemblage structure of soil nematodes in relation to climate and ecosystem properties." Global Ecology and Biogeography **23**(9): 968-978.

Park, B.-Y., et al. (2011). "Effects of heavy metal contamination from an abandoned mine on nematode community structure as an indicator of soil ecosystem health." Applied soil ecology **51**: 17-24.

Pen-Mouratov, S., et al. (2003). "Seasonal and spatial variation in nematode communities in a Negev desert ecosystem." Journal of Nematology **35**(2): 157.

Porazinska, D. L., et al. (2010). "Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure." Molecular Ecology Resources **10**(4): 666-676.

Procter, D. (1984). "Towards a biogeography of free-living soil nematodes. I. Changing species richness, diversity and densities with changing latitude." Journal of Biogeography: 103-117.

Quist, C. W., et al. (2015). "Evolution of plant parasitism in the phylum Nematoda." Annual Review of Phytopathology **53**: 289-310.

Ricotta, C. and J. Podani (2017). "On some properties of the Bray-Curtis dissimilarity and their ecological meaning." Ecological Complexity **31**: 201-205.

Santos, P. F., et al. (1981). "The role of mites and nematodes in early stages of buried litter decomposition in a desert." Ecology **62**(3): 664-669.

Šalamún, P., et al. (2014). "Diversity and food web structure of nematode communities under high soil salinity and alkaline pH." Ecotoxicology **23**: 1367-1376.

Schenk, J., et al. (2020). "Comparison of morphological, DNA barcoding, and metabarcoding characterizations of freshwater nematode communities." Ecology and evolution **10**(6): 2885-2899.

Schlesinger, W. H., et al. (1996). "On the spatial pattern of soil nutrients in desert ecosystems: ecological archives E077-002." Ecology **77**(2): 364-374.

Shih, P.-Y., et al. (2019). "Newly identified nematodes from Mono Lake exhibit extreme arsenic resistance." Current Biology **29**(19): 3339-3344. e3334.

Smythe, A. B., et al. (2006). "Nematode small subunit phylogeny correlates with alignment parameters." Systematic Biology **55**(6): 972-992.

Tchesunov, A. V. and F. Riemann (1995). "Arctic sea ice nematodes (Monhysteroidea), with descriptions of Cryonema crassum gen. n., sp. n. and C. tenue sp. n." Nematologica **41**(1-4): 35-50.

Tietjen, J. H. (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**(10): 1579-1594.

Van Den Hoogen, J., et al. (2019). "Soil nematode abundance and functional group composition at a global scale." Nature **572**(7768): 194-198.

Vitousek, P. M. (2002). "Oceanic islands as model systems for ecological studies." Journal of Biogeography **29**(5-6): 573-582.

Wang, Q., et al. (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." Applied and Environmental Microbiology **73**(16): 5261-5267.

Wallace, H. (1968). "The dynamics of nematode movement." Annual Review of Phytopathology **6**(1): 91-114.

Wharton, D. and I. Brown (1989). "A survey of terrestrial nematodes from the McMurdo Sound region, Antarctica." New Zealand Journal of Zoology **16**(3): 467-470.

Wilson, M. J. and T. Kakouli-Duarte (2009). Nematodes as environmental indicators, CABI.

Wright, J. P., et al. (2006). "Predictability of ecosystem engineering effects on species richness across environmental variability and spatial scales." Journal of Ecology: 815-824.

Yeates, G. W., et al. (1993). "Feeding habits in soil nematode families and genera—an outline for soil ecologists." Journal of Nematology **25**(3): 315.

Zabriskie, J. G. (1979). Plants of Deep Canyon. - University of California Press, Riverside, 174 p.

Zullini, A. (2018). "Cosmopolitism and endemism in free-living nematodes." Biogeographia–The Journal of Integrative Biogeography **33**.

Chapter 1 Tables and Figures:

Collection Locality	Soil Sample	Coordinates	Plant Species	Elevation
Site 1	A122	N33°39.9'67, W116°22.4'75	Palo Verde	224m
	A123	N33°39.959', W116°22.467'	Chuparosa	224m
	A124	N33°39.965', W116°22.490'	Creosote	225m
Site 2	A125	N33°39.66'0, W116°22.368'	Palo Verde	242m
	A126	N33°39.654', W116°22.372'	Chuparosa	242m
	A127	N33°39.654', W116°22.365'	Creosote	242m
Site 3	A128	N33°39.477', W116°22.200'	Palo Verde	255m
	A129	N33°39.473', W116°22.192'	Creosote	255m
	A130	N33°39.471', W116°22.186'	Chuparosa	255m

Table 1.1. Collection information for 9 soil samples from Boyd Deep Canyon Reserve that were analyzed for nematode diversity using reverse taxonomic, 96-well plate Sanger sequencing, and Illumina plus PacBio sequencing-based metabarcoding (**Chapter 2**) strategies.

Primer Name	Use	Sequence (5'-3')	T_m
391F	PCR & sequencing	AACGAAGATTTCCTTAGTAACGG	52.7C
501R	PCR & sequencing	GAAGATTTCCTTAGTAACGGCGAG	55.4C
504F	internal sequencing	CAAGTACCGTGAGGGAAAGTTG	55.7C
787F	PCR & sequencing	AGCGGAGGAAAAGAAACTAACGA	56.4C
792R	PCR & sequencing	CAGCTATCCTGGGGGAAAC	55.2C
799R	PCR & sequencing	TCCGTGTTTCAAGACGG	52.2C
811F†	internal sequencing	AACGAAGATTTCCTTAGTAACGG	52.7C
812F†	internal sequencing	GAAGATTTCCTTAGTAACGGCGAG	55.4C
813F‡	internal sequencing	AAGAAACTAACGAGGATTCCC	52.1C
814F‡	internal sequencing	AGTAACGGCGAGTGAACGGGGAA	62.5C
817F§	internal sequencing	AAGAAACTAACGAGGATTCC	49.6C

Table 1.2. PCR and sequencing primers used for construction and testing of a BDCR-specific reference sequence database. Primer melting temperatures were calculated using the IDT Oligo Analyzer tool (https://www.idtdna.com/calc/analyzer).

† Primers 811F and 812F were designed based on sequences from Infraorder *Dorylaimomorpha*.

‡ Primers 813F and 814F were designed based on sequences from Infraorder *Cephalobomorpha*.

§ Primer 817F was designed based on sequences from Superfamily *Aphelenchoide*

Phylum	# of Sequences
Nematoda	4097
Rotifera	116
Nematomorpha	21
Ciliophora	12
Ascomycota	9
Basidiomycota	6
Cercozoa	4
Platyhelminthes	4
Gyrista	4
Oomycota	3
Chlorophyta	2
Streptophyta	1
Chytridiomycota	1
Apicomplexa	1
Tardigrada	1
Discosea	1
Evosea	1
Heterolobosea	1
Blastocladiomycota	1
Myzozoa	1
Haptista	1
Bigyra	1
Colponemidia	1
Choanozoa	1
Nitrosphaerota	1

Table 1.3. Number and phylum-level classification of the D1-D3 28s rDNA sequences included in a custom database that was constructed for this study. Nematomorph and Rotiferan sequences were added to improve resolution of non-nematode classifications. All other taxa were chosen based on previous identification of non-nematode sequences in desert soil samples using the *sklearn* classifier and SILVA (138.1) database.

Described Species **3**	***Chromadorea, Rhabiditida, Cepahlobidae:*** *Zeldia spannata* *Stegelleta incisa* *Ecumenicus monohystera*
Awaiting Description **14**	***Enoplea, Dorylaimida, Aporcelaimidae:*** 1 *Aporcelaimellus* species ***Chromadorea, Rhabiditida, Cepahlobidae:*** 2 *Acrobeles* species 4 *Acrobeloides* species 2 *Cervidellus* species 3 *Chiloplacus* species 1 *Stegelleta* species 1 *Stegelletina* species
Awaiting Further Study **9**	***Enoplea, Dorylaimida, Qudsianematidae:*** 1 *Microdorylaimus* species ***Enoplea, Dorylaimida, Leptonchidae:*** 1 *Utahnema* species ***Chromadorea, Rhabiditida, Panagrolaimidae:*** 1 *Panagrolaimus* species ***Chromadorea, Rhabiditida, Telotylenchidae:*** 3 *Quinisulcius* species ***Chromadorea, Rhabiditida, Aphelenchidae:*** 3 *Aphelenchus* species

Table 1.4. Summary of 26 described and putative species that were documented via reverse taxonomy.

Collection Location	Soil Sample	Plant Species	#Sequences
Site1	A122	Palo Verde	490
	A123	Chuparosa	262
	A124	Creosote	221
Site 2	A125	Palo Verde	289
	A126	Chuparosa	266
	A127	Creosote	284
Site 3	A128	Palo Verde	234
	A129	Creosote	258
	A130	Chuparosa	176
			2480

Table 1.5. Summary of Sanger sequences generated in a 96-well plate format. A total of 2840 sequences met quality standards.

Confidence Level	OTUs to Species	OTUs to Genus	OTUs to Family	OTUs to Order	Reads to Species	Reads to Genus	Reads to Family	Reads to Order	Total Reads Classified
70%	32	7	3	2	2118	87	242	33	2480
80%	31	8	3	2	2100	104	242	34	2480
90%	31	9	3	2	2070	114	262	34	2480
95%	29	8	3	2	2057	119	270	34	2480
99%	25	9	3	2	2041	125	277	37	2480

Table 1.6. Total number of OTUs and reads successfully classified to each taxonomic level by the *sklearn* classifier at 70%, 80%, 90%, 95%, and 99% confidence settings.

Class	Order	Family	Genus	Species	Count
Chromadorea	*Rhabditida*	*Cephalobidae*	*Acrobeles*	DC sp. 1	683
Chromadorea	*Rhabditida*	*Cephalobidae*	*Acrobeloides*	DC sp. 1	411
Chromadorea	*Rhabditida*	*Cephalobidae*	*Acrobeles*	DC sp. 2	349
Chromadorea	*Rhabditida*	*Panagrolaimidae*	*Panagrolaimus*	DC sp. 1	340
Chromadorea	*Rhabditida*	*Cephalobidae*	unclassified	unclassified	205
Chromadorea	*Rhabditida*	*Cephalobidae*	*Acrobeloides*	DC sp. 4	78
Chromadorea	*Rhabditida*	*Cephalobidae*	*Zeldia*	*punctata*	49
Chromadorea	*Rhabditida*	*Cephalobidae*	*Cervidellus*	unclassified	40
Chromadorea	*Rhabditida*	*Panagrolaimidae*	unclassified	unclassified	36
Enoplea	*Dorylaimida*	unclassified	unclassified	unclassified	31
Enoplea	*Dorylaimida*	*Qudsianematidae*	*Ecumenicus*	DC sp. 7	28
Enoplea	*Dorylaimida*	*Aporcelaimidae*	*Aporcelaimellus*	DC sp. 6	26
Chromadorea	*Rhabditida*	*Panagrolaimidae*	unknown	unnamed sp.	23
Chromadorea	*Rhabditida*	*Cephalobidae*	*Chiloplacus*	unclassified	21
Enoplea	*Dorylaimida*	*Qudsianematidae*	*Discolaimoides*	*symmetricus*	17
Chromadorea	*Rhabditida*	*Cephalobidae*	*Acrobeloides*	DC sp. 6	14
Chromadorea	*Rhabditida*	*Rhabditidae*	*Mesorhabditis*	*monhystera*	11
Enoplea	*Dorylaimida*	*Nordiidae*	*Longidorella*	unclassified	10
Enoplea	*Dorylaimida*	*Leptonchidae*	*Utahnema*	DC sp. 4	10
Chromadorea	*Rhabditida*	*Cephalobidae*	*Stegelletina*	unclassified	9
Chromadorea	*Rhabditida*	*Aphelenchidae*	*Aphelenchus*	unclassified	8
Chromadorea	*Rhabditida*	*Cephalobidae*	*Acrobeles*	unnamed sp. JB-132	8
Chromadorea	*Rhabditida*	*Cephalobidae*	*Acrobeles*	unclassified	7
Enoplea	*Dorylaimida*	*Qudsianematidae*	*Ecumenicus*	unnamed sp.	7
Chromadorea	*Rhabditida*	*Aphelenchidae*	*Aphelenchus*	DC sp. 2	6
Enoplea	*Dorylaimida*	unknown	unknown	DC sp. 2	6
Enoplea	*Dorylaimida*	*Qudsianematidae*	*Microdorylaimus*	DC sp. 1	6
Chromadorea	*Rhabditida*	*Cephalobidae*	*Nothacrobeles*	*spatulatus*	5
Chromadorea	*Rhabditida*	*Cephalobidae*	*Acrobeloides*	unclassified	5
Chromadorea	*Rhabditida*	*Aphelenchidae*	*Aphelenchus*	DC sp. 1	5
Chromadorea	*Rhabditida*	*Cephalobidae*	*Acrobeloides*	DC sp. 5	4
Chromadorea	*Rhabditida*	*Cephalobidae*	*Zeldia*	unnamed sp. JB-140	4
Chromadorea	*Rhabditida*	*Cephalobidae*	*Acrobeles*	*singulus*	3
Chromadorea	*Rhabditida*	*Anguinidae*	*Ditylenchus*	*persicus*	2
Chromadorea	*Rhabditida*	unclassified	unclassified	unclassified	2

Enoplea	*Dorylaimida*	unknown	unknown	unnamed sp.	2
Enoplea	*Dorylaimida*	*Qudsianematidae*	*Microdorylaimus*	unclassified	2
Enoplea	*Dorylaimida*	*Leptonchidae*	*Utahnema*	DC sp. 3	2
Chromadorea	*Rhabditida*	*Aphelenchoididae*	unclassified	unclassified	1
Chromadorea	*Rhabditida*	*Anguinidae*	*Ditylenchus*	unnamed sp. 85C1	1
Chromadorea	*Rhabditida*	*Cephalobidae*	*Cervidellus*	unnamed sp. 1 HMM2018	1
Enoplea	*Dorylaimida*	*Dorylaimida*	*Carcharodiscus*	*banaticus*	1
Enoplea	*Dorylaimida*	*Belondiridae*	*Axonchium*	*propinquum*	1
			total sequences		2480

Table 1.7. Taxonomic categories of Illumina, PacBio, and Sanger sequence reads at 70% confidence using the sklearn classifier, as well as read counts for each. The "unclassified" designator was used when the *sklearn* naïve Bayesian classifier was not able to identify OTUs to a particular taxonomic rank at 70% confidence. Reference taxa that did not include species names are designated as "unnamed sp.". Two dorylaimidan reference taxa and one panagrolaimid reference taxon were designated as "unknown" with respect to family and/or genus level classification. Taxonomic categories that are classified to genus and higher levels may include sequence reads from more than one species.

Distance Metric	**Bray-Curtis**			**Jaccard**		
Predictor	**R2**	**F**	**Pr(>F)**	**R2**	**F**	**Pr(>F)**
Plant species	0.40105	2.0088	0.06071	0.34973	1.6135	0.07143
Collection site	0.1534	0.5436	0.925	0.18949	0.7014	0.8714

Table 1.8. Results of PERMANOVA evaluating effects of plant species and collection site on the composition of nematode communities. Bray-Curtis dissimilarity and Jaccard distance were used to estimate beta diversity.

Gene	18S	28S	ITS	COx1
Acrobeles	✓	✓	✓	✓
Acrobeloides	✓	✓	✓	✓
Acromoldavicus		✓		
Cephalobus	✓	✓		
Cervidellus	✓	✓	✓	
Chiloplacus	✓	✓	✓	
Eucephalobus	✓	✓	✓	
Heterocephalobellus	✓	✓	✓	
Heterocephalobus	✓			
Macrolaimellus		✓		
Metacrobeles	✓	✓	✓	
Nothacrobeles	✓	✓	✓	
Panagrolobus		✓		
Paracrobeles		✓		
Penjatinema		✓		
Placodira		✓		
Pseudacrobeles	✓	✓	✓	
Scottnema	✓	✓	✓	✓
Seleborca	✓			
Spinocephalus	✓	✓		
Stegelleta	✓	✓	✓	
Stegelletina	✓	✓	✓	
Teratolobus		✓		
Zeldia	✓	✓	✓	
Total Non-environmental Sequences:	989	352	56	255
Total Genera:	17	22	13	3

Table 1.9. Representation of Family *Cephalobidae* genera in the NCBI sequence database for 4 commonly used molecular barcoding loci: 18S, 28S, and ITS nuclear rDNA genes and Cytochrome Oxidase I(COx1). Blackened cells indicate a lack of genus-level representation for a given locus.

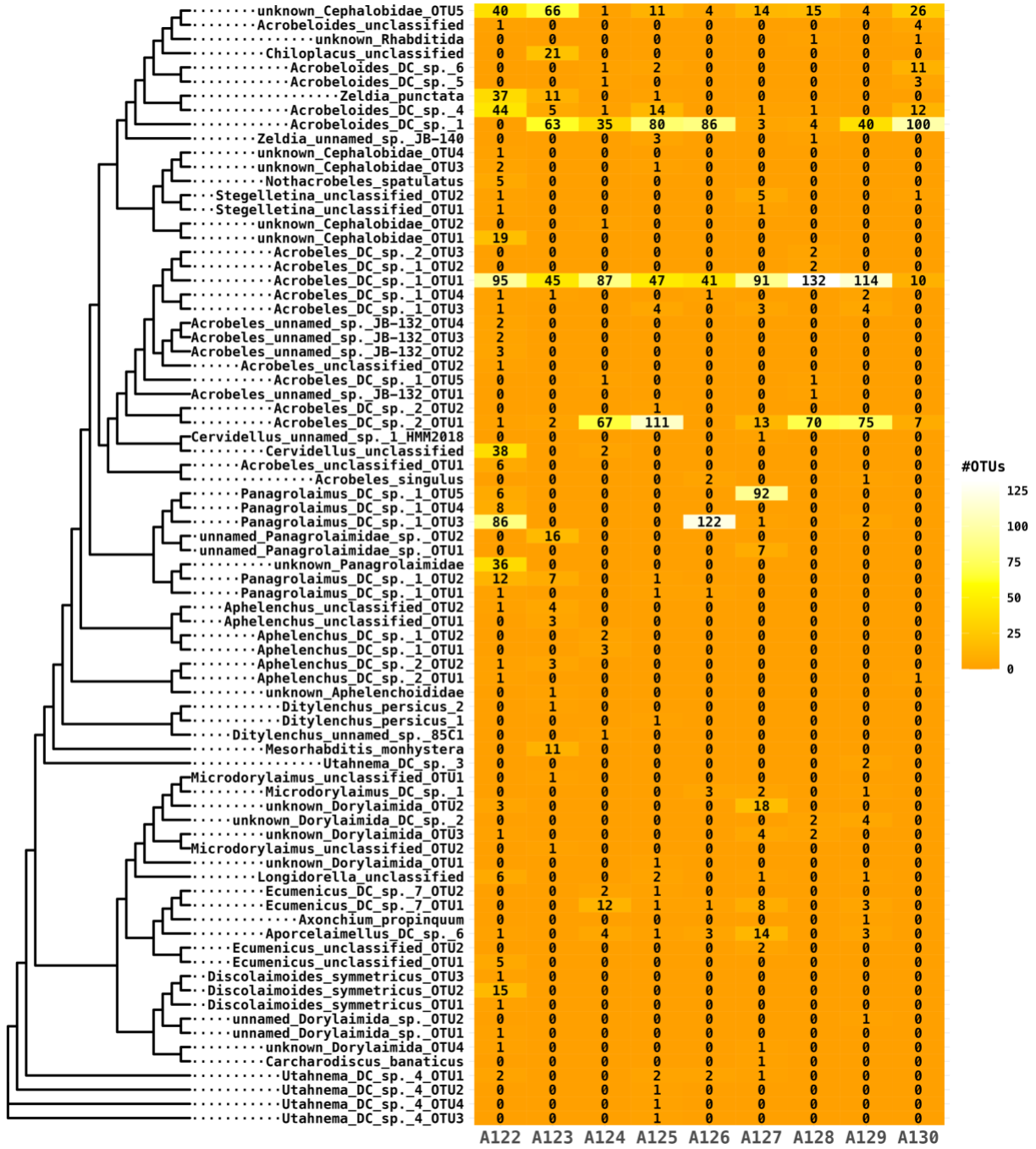

Figure 1.1. Distribution of OTU counts among BDCR soil samples. Taxonomic categories consisted of 1 to 5 OTUs. Heat map colors correlate with abundances of OTUs in each soil sample.

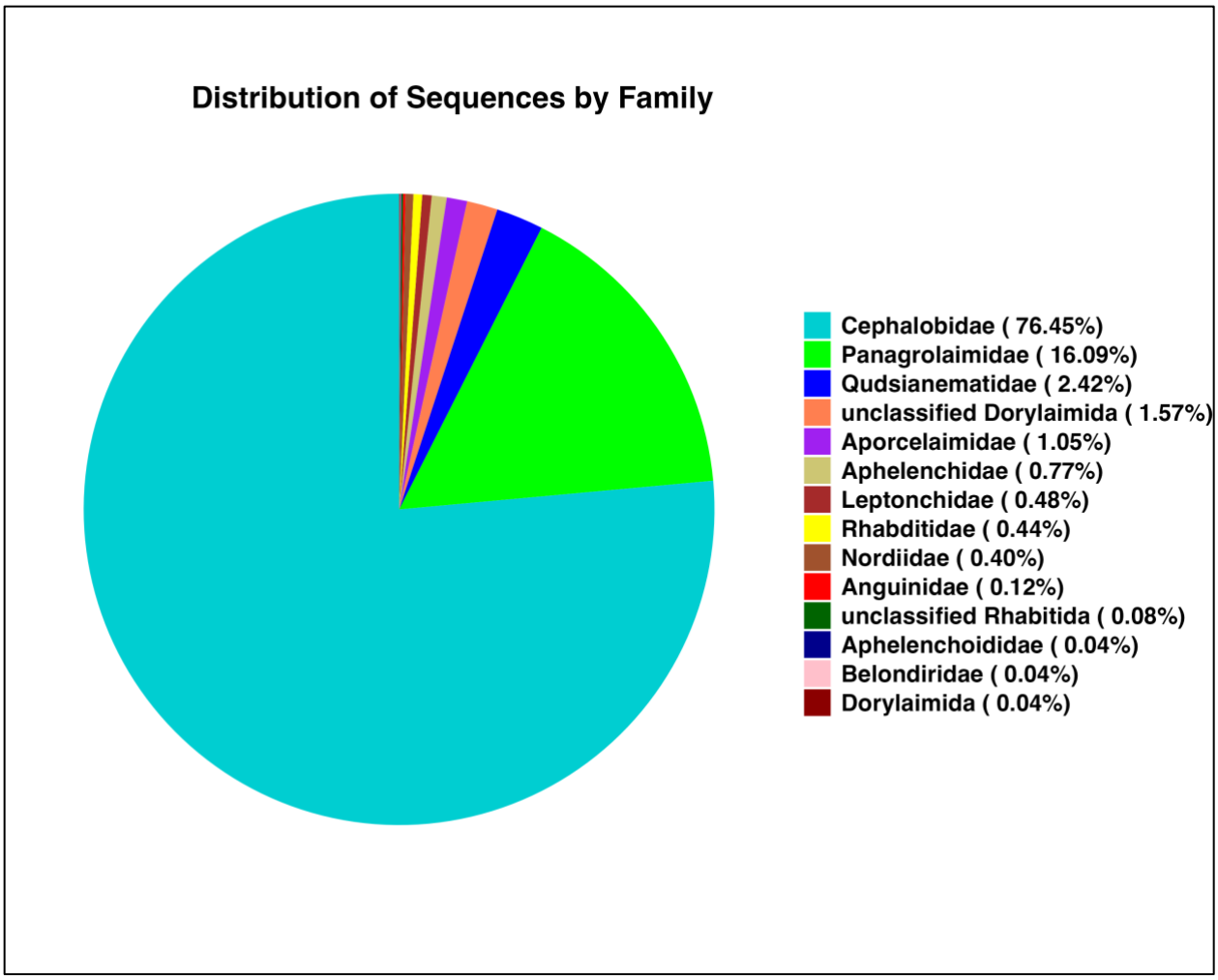

Figure 1.2. Pie chart showing the percentage of nematode families found among all BDCR 96-well plate Sanger sequences. Classifications were done with the *sklearn* naive Bayesian classifier as implemented in the Qiime2 bioinformatic pipeline. The category, "unclassified Dorylaimida" includes Dorylaimidan sequences that could be classified no lower than order level, as well as sequences that matched undescribed Dorylaimidan reference taxa (with 70% confidence) at lower taxonomic levels.

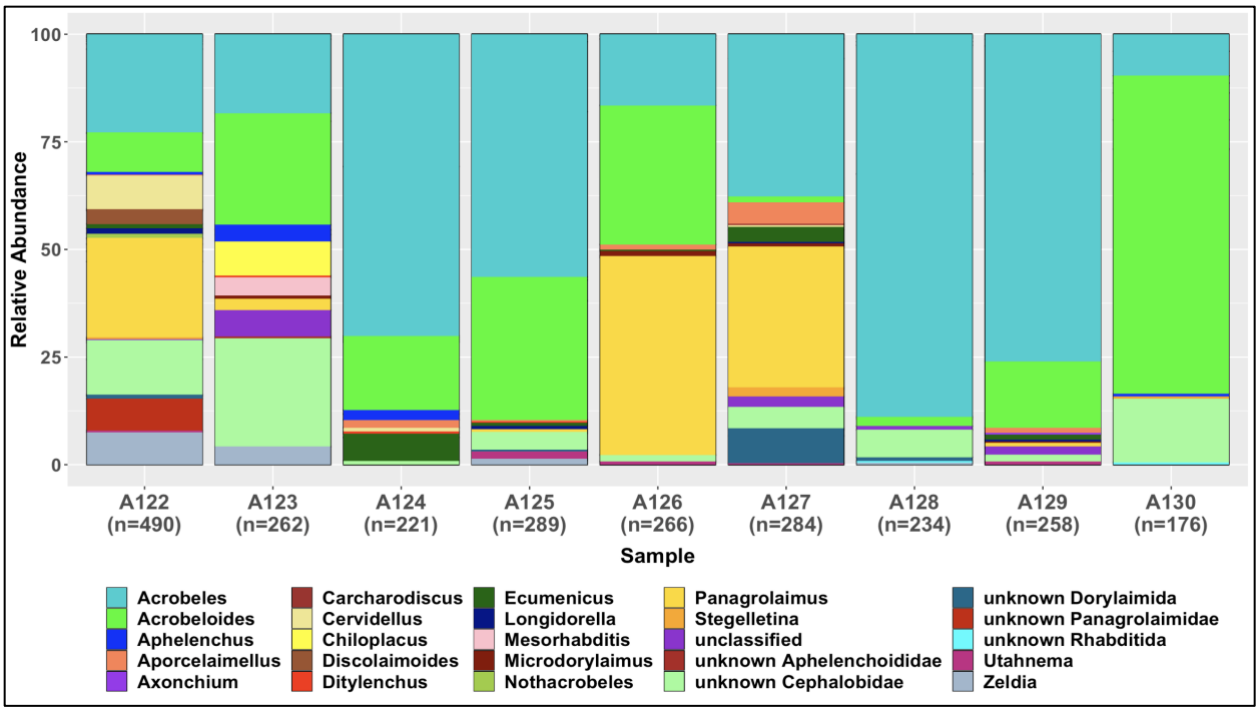

Figure 1.3. Stacked bar plot of sequence classifications to genus level at 70% confidence, by soil sample.

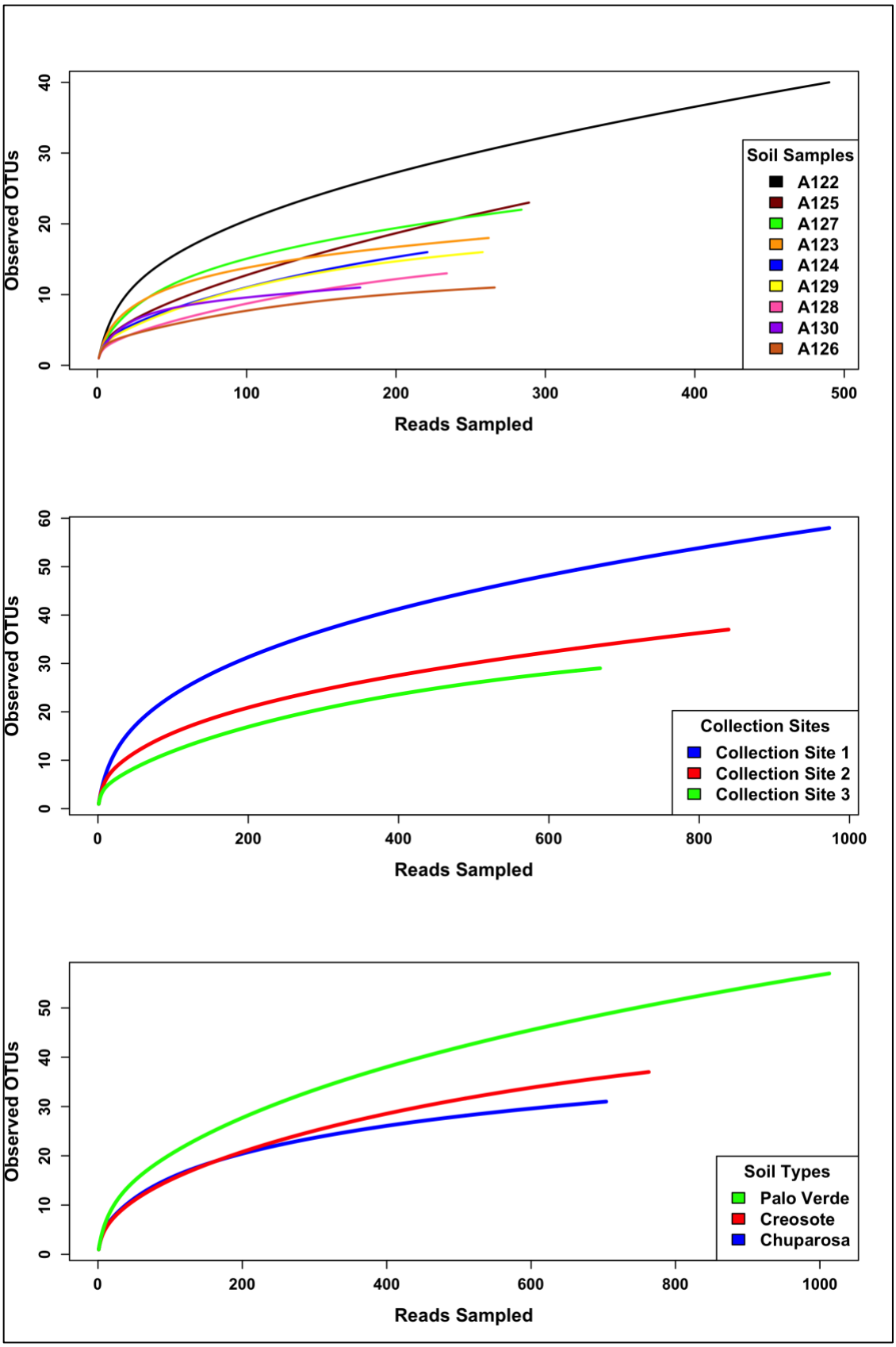

Figure 1.4. OTU accumulation curves by Soil Sample(A), Collection Site(B), and Plant Species(C).

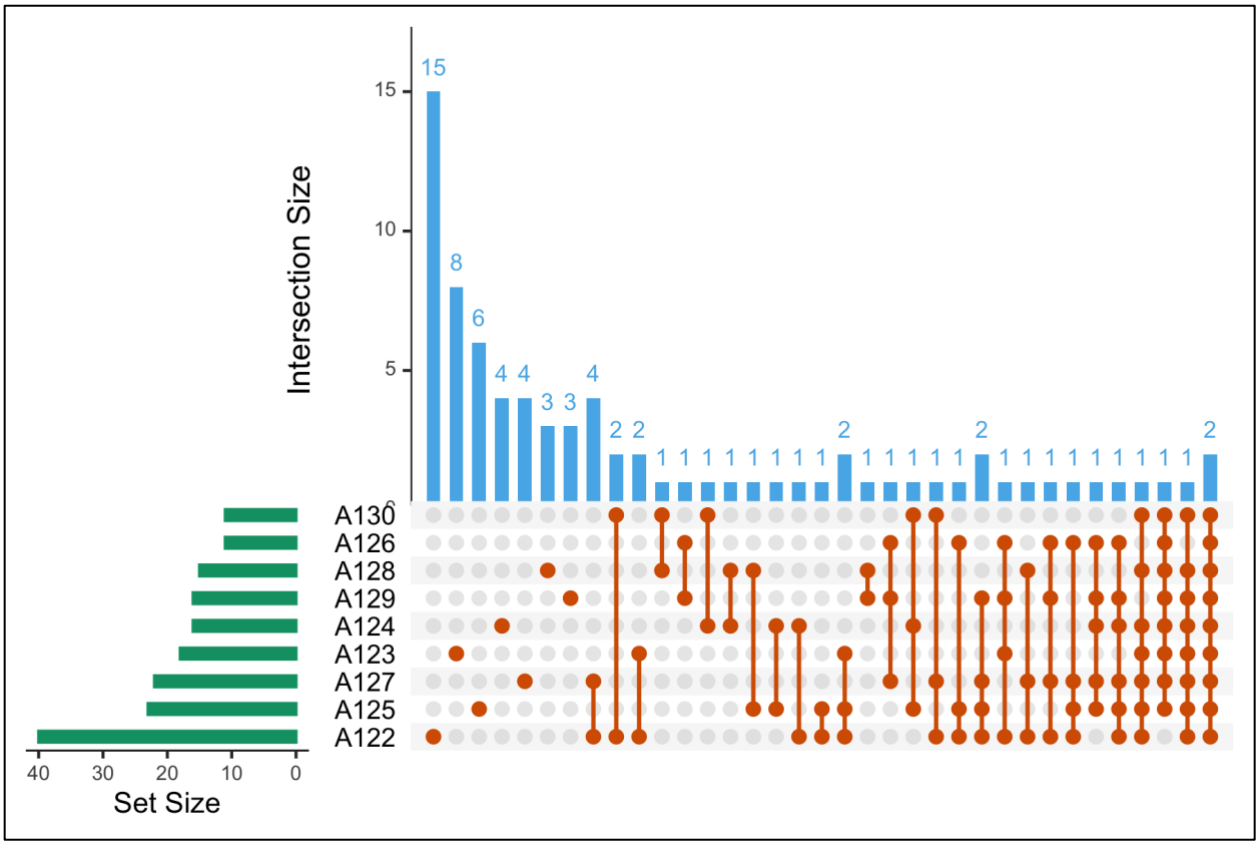

Figure 1.5. Upset plot representing prevalence of OTUs among BDCR soil samples. Intersection sizes correspond to the number of OTUs that originate from a given set of soil samples. Set sizes are the number of OTUs that originate from each soil sample.

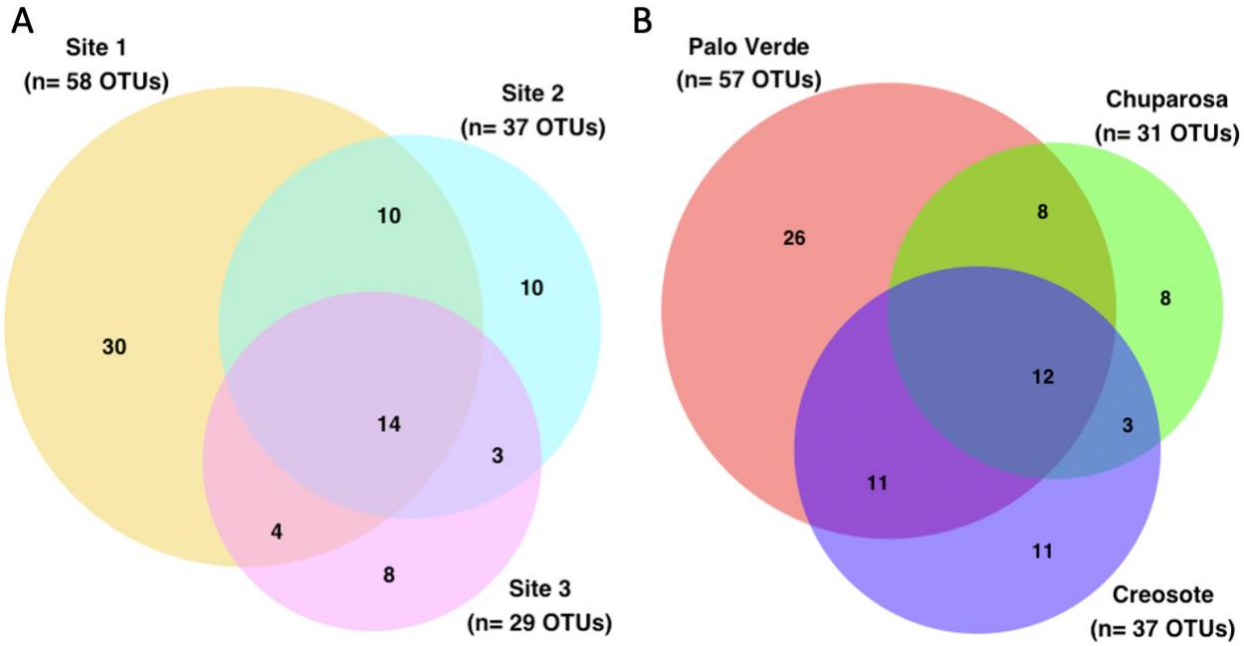

Figure 1.6. Scaled Venn diagram showing distribution of 28S OTUs among collection sites (A) and plant species (B).

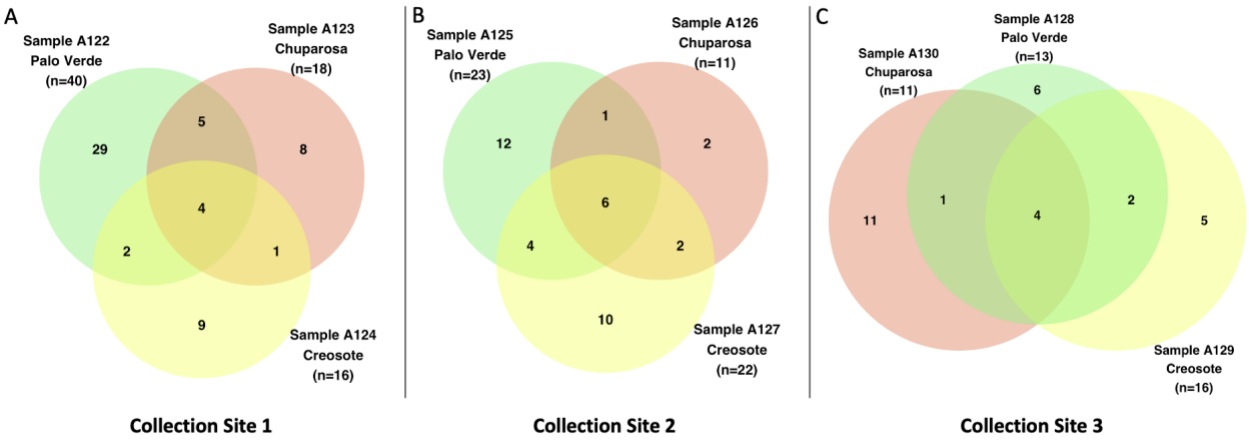

Figure 1.7. Scaled Venn diagram showing distribution of 28S OTUs among the three plant species at Collection Site 1 (A), Site 2 (B), and Site 3 (C).

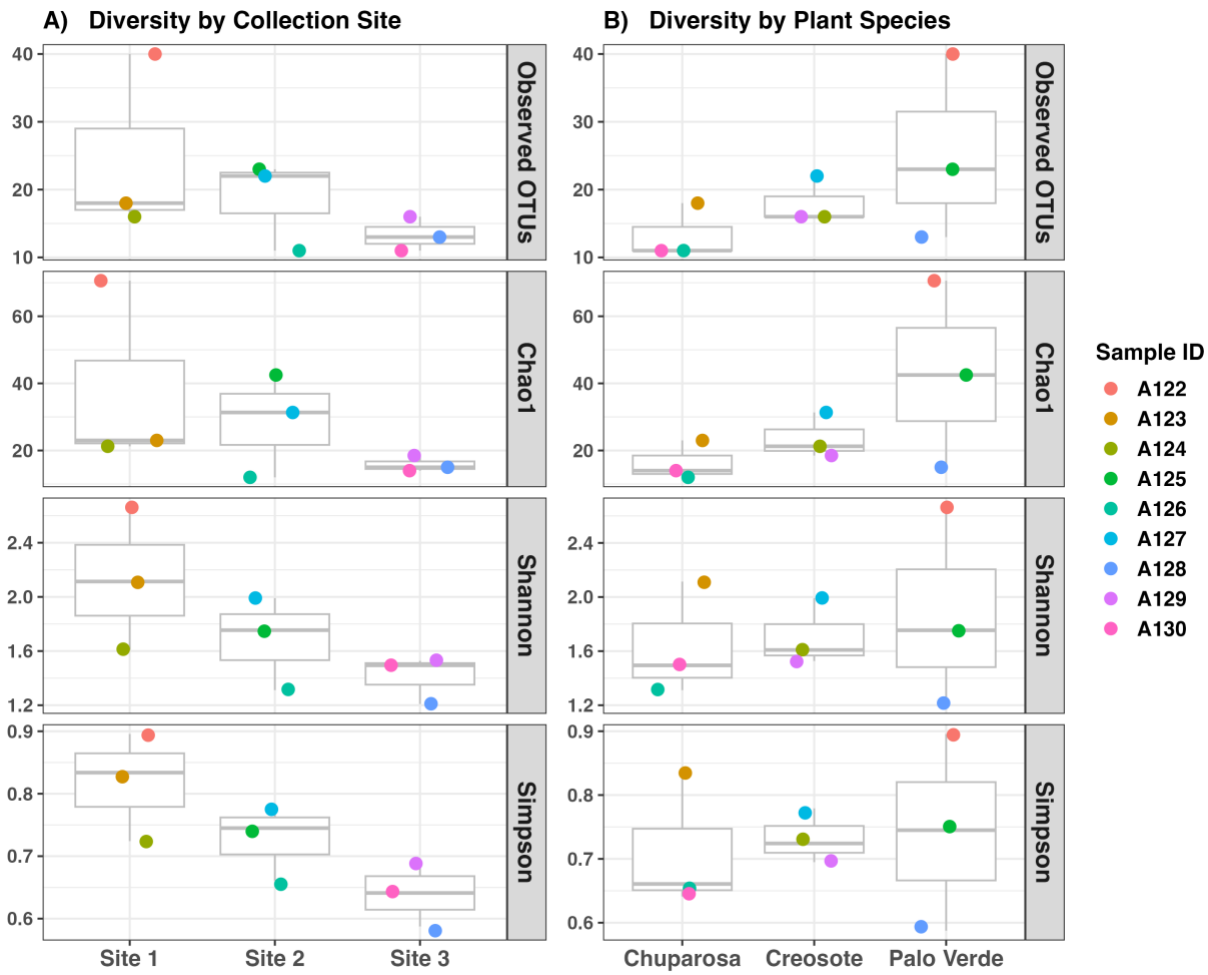

Figure 1.8. Boxplots showing number of observed OTUs, as well as Chao1, Shannon, and Simpson diversity indices by collection location (**A**) and plant species (**B**).

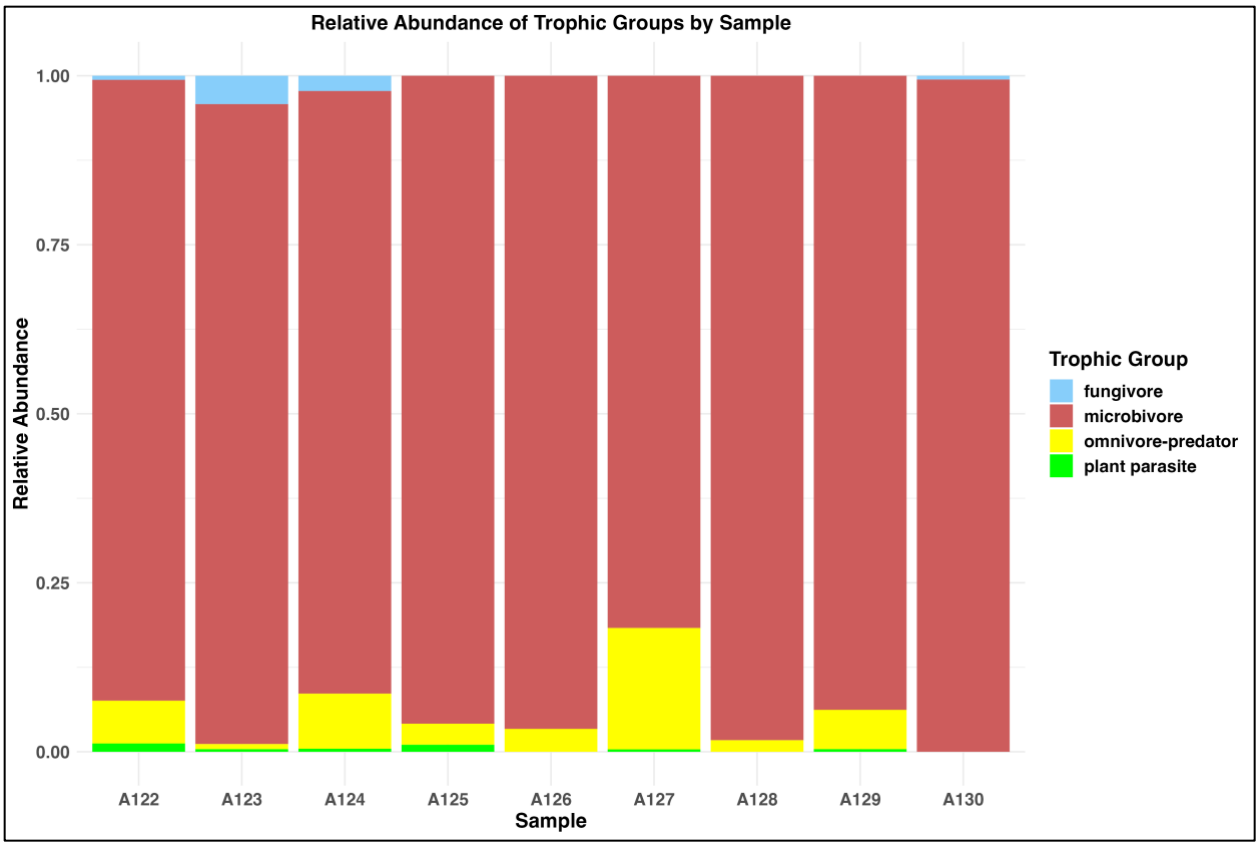

Figure 1.9. Stacked bar plot of trophic group relative abundance by soil sample.

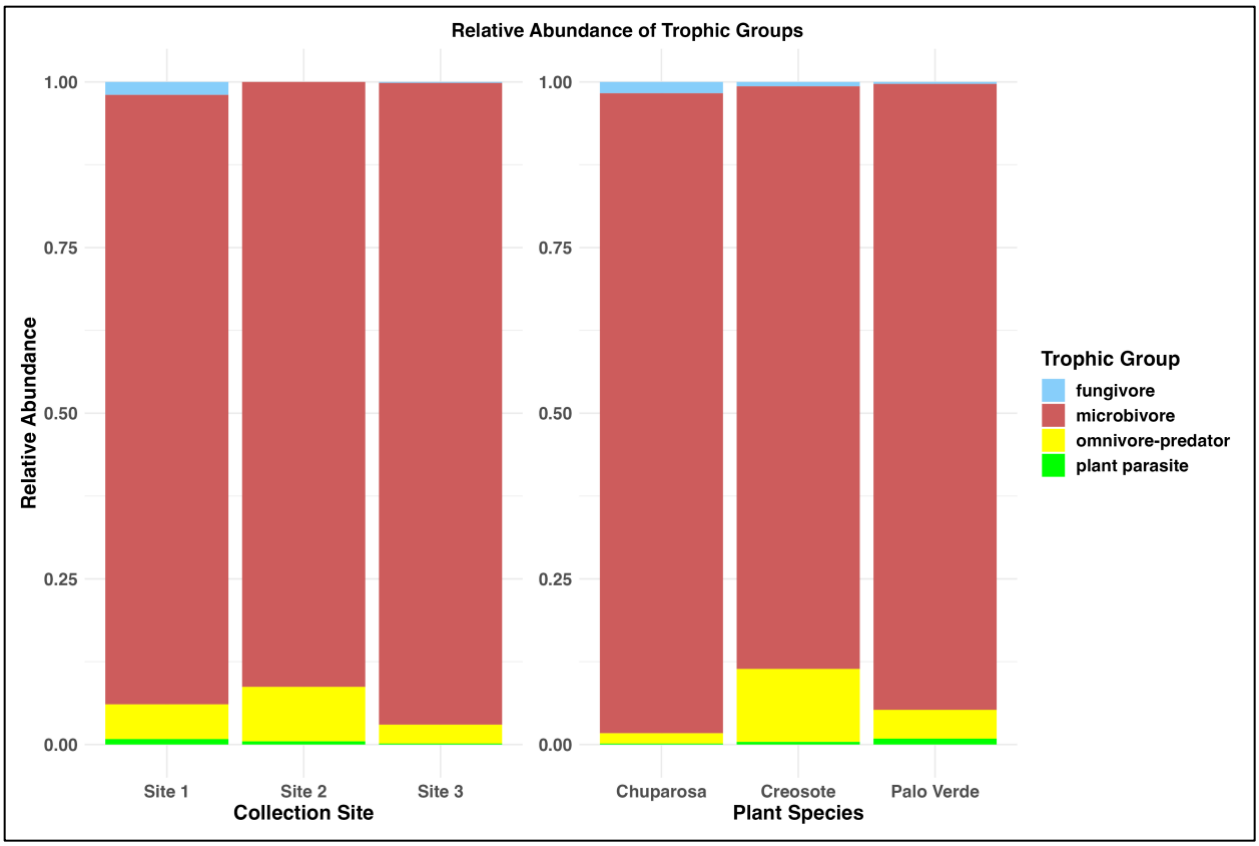

Figure 1.10. Stacked bar plot of trophic group relative abundance by collection site and plant species.

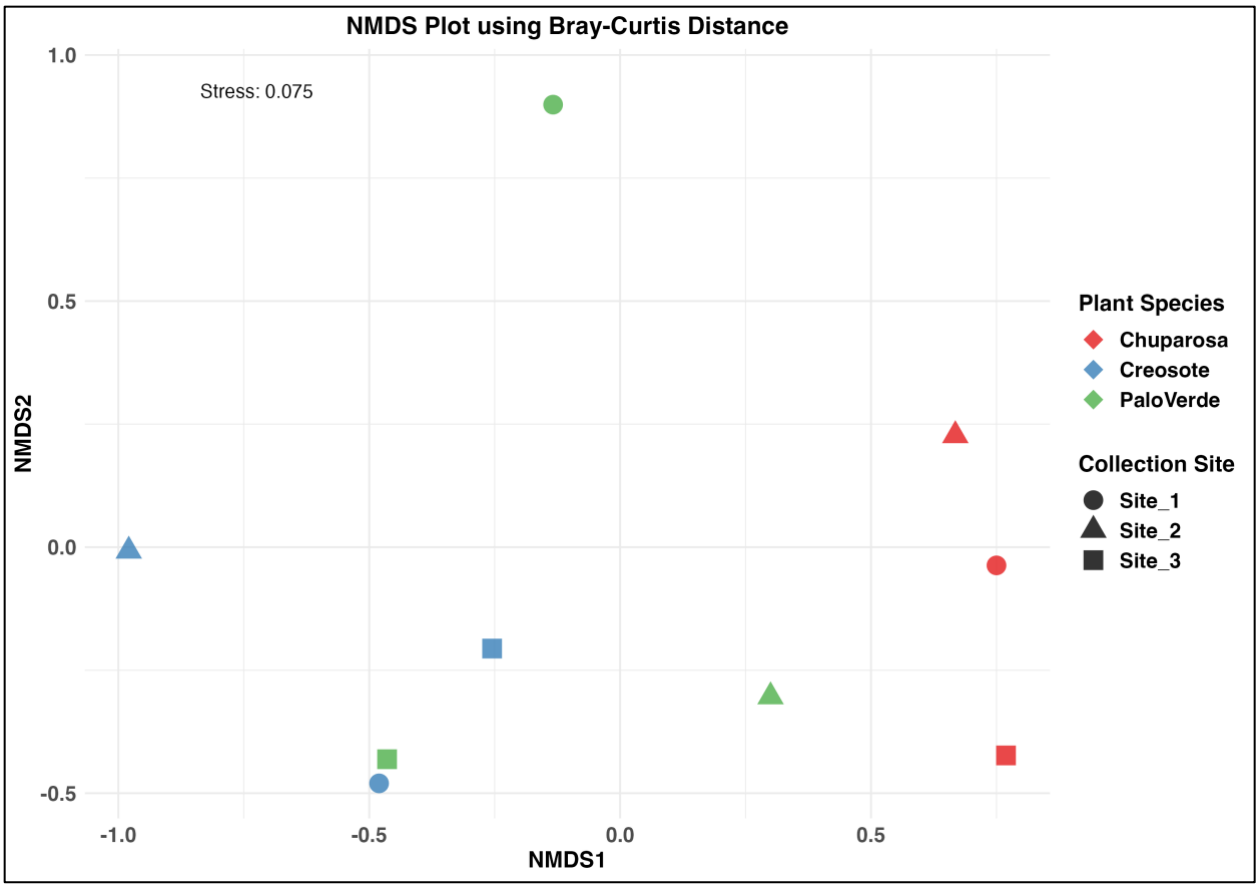

Figure 1.11. Non-metric multidimensional scaling (NMDS) plot of BDCR soil samples using Bray-Curtis distance.

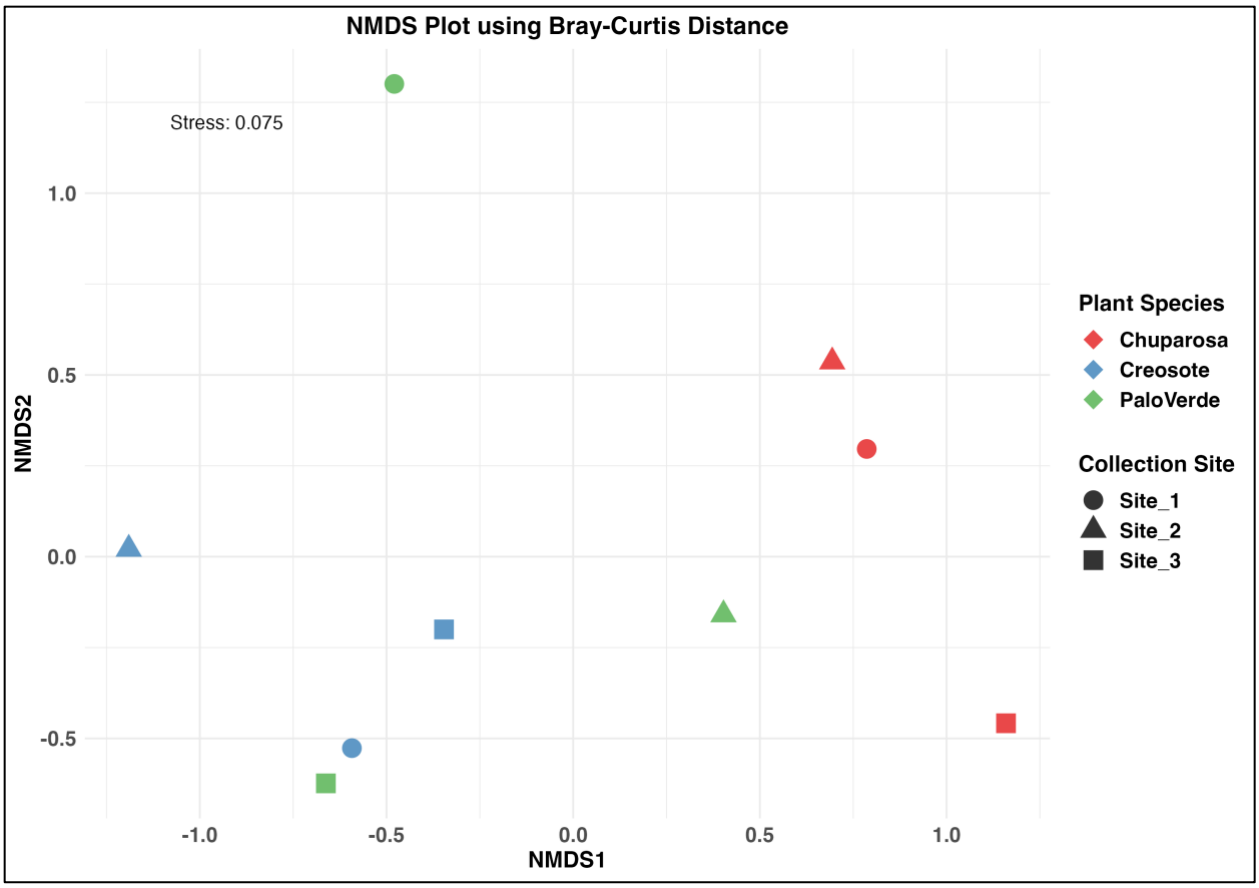

Figure 1.12. Non-metric multidimensional scaling (NMDS) plot of BDCR soil samples using Jaccard distance.

Chapter 2: Evaluating Sequencing Technologies for Nematode Community Analysis: A Comparative Study of Illumina, PacBio, and Sanger Methods

Christopher A. Pagan[1] and Steven A. Nadler[1].

[1]Dept. of Entomology and Nematology, University of California, Davis, CA, U.S.A

2.1 Introduction:

As discussed in Chapter 1, a severe lack of taxonomic knowledge is currently preventing a comprehensive study of the community and spatial ecology of soil nematodes. Even when species can be identified, the process is time-consuming and prohibitively difficult. This is especially true when working with large and diverse environmental samples. Only a fraction of nematodes from any given community can be identified to species level using morphology, and usually this can be done with only adult worms of one sex. At the current rate of species discovery, very few habitats can be fully (or even mostly) characterized with respect to nematode community composition (Maslen, 1980), and there is a clear need for more rapid methods of discovery, characterization, and identification.

DNA metabarcoding is a useful technique that significantly accelerates the identification of communities of organisms (Taberlet et al., 2012a). It begins with the extraction of bulk DNA from organisms in environmental samples. This mixed DNA is PCR-amplified and processed using high throughput sequencing technologies (HTS) to generate representative sequences

from the sample. The generated sequences are then classified by cross-referencing them with reference sequence databases. This method has greatly enhanced biodiversity and ecological studies, allowing researchers to characterize communities of organisms rapidly and inexpensively at large scales. An added benefit of DNA metabarcoding is that it expedites the discovery of potential new species by detecting taxa absent from reference sequence databases. These unrepresented taxa can then be targeted for individual sampling and further taxonomic characterization.

Metabarcoding has been used to greatest effect in the community ecology of microbes (Liu et al., 2007; Knight et al., 2018), but also in studies of fungi (Öpik et al., 2013; Taylor, et al., 2014) and other eukaryotes (Taberlet et al., 2012b), including nematodes (Porazinska et al., 2009). Genes encoding for ribosomal RNA (rDNA) are the most used loci for metabarcoding of nematodes (Blaxter et al., 2003a; Blaxter et al, 2003b). It is relatively easy to design PCR primers for nuclear rDNA loci (18S, ITS, 28S) because they contain interspersed stretches of highly conserved sequence. They are also easy to amplify because dozens to thousands of copies of these genes can be found in the genomes of most eukaryotic organisms (Prokopowich, 2003).

Illumina was, until recently, the default sequencing technology used for community metabarcoding (Kozich et al., 2013) because of its affordability and high sequencing capacity. Illumina sequencing chemistry, however, has some limitations that impose significant constraints on molecular barcoding. The Illumina MiSeq600 (2x300) platform, despite its popularity, is restricted by its ~590 base pair limit for amplicon sizes. This limitation is crucial

because PCR primer pairs that have good taxonomic coverage and span informative sequence regions are not always within this size range. This often results in the selection of primers that produce a desirable amplicon size, and not necessarily ones with the greatest utility (Fuks et al., 2018).

Recent advancements in PacBio sequencing technology have established it as a viable alternative to Illumina MiSeq for amplicon sequencing. The PacBio Sequel II platform now has sufficient sequencing capacity (~500Gb) for the characterization of diverse communities of organisms, with the added benefit of much larger read sizes (≤20,000bp). While still more expensive per base than Illumina, Pacbio's sequencing capacity is steadily improving, and the technology is increasingly used in community metabarcoding studies. Pacbio's larger read size greatly expands the range of barcoding loci that might be considered for biodiversity studies, promising greater taxonomic resolution in the future.

In ecological metabarcoding studies, distinguishing between true biological diversity and sequencing artifacts is crucial. Artifacts hinder the interpretation of metabarcoding data by inflating feature (amplified sequence variants (ASVs) or operational taxonomic units (OTUs)) counts and skewing ecological metrics such as richness, diversity, and relative abundance. The three main causes of artifactual sequence are chimeras, sequencing errors, and amplification bias.

Chimeric PCR products, or chimeras, are hybrid sequences created during a PCR reaction from fragments of two or more differing DNA templates. Chimeric sequences can be easily mistaken for true biological variation, and they are particularly common in HTS community analyses, where DNA templates from diverse organisms are amplified together in the same tube. There are many bioinformatic tools designed to detect and remove chimeras from HTS sequence data (Edgar et al., 2011; Haas et al., 2011; Callahan et al., 2016). This is usually done during data cleaning and quality control steps.

Sequencing errors are another important cause of artifactual diversity. Different sequence technologies have characteristic error profiles or the number and distribution of errors across the lengths of reads. Illumina reads tend to have concentrations of errors primarily at the 3' ends, largely because of phasing artifacts and pre-phasing artifacts (Kircher, et al., 2009), which introduce increasing numbers of frameshifted template molecules to a given cluster as cycles progress. Signal decay, deterioration of reagent quality, and context-specific errors may also contribute to the accumulation of 3’ errors (Cacho, et al., 2016; Alhoff et al., 2013).

Merging of forward and reverse reads can mitigate error by creating a consensus sequence where the forward and reverse reads overlap, but this is not helpful when amplicons are too large for reads to overlap. A common workaround is to trim reads at a predetermined length based on quality scores, and then concatenate them with one or more ambiguous bases (N's) inserted between. This is a valid solution in theory because kmer-based sequence classifiers (Wang, et al. 2007; Bokulich, et al. 2018) can ignore large deletions and ambiguous bases.

Quality scores, however, are not always strong predictors of true error density (Schirmer, et al., 2015) because some proportion of these errors were introduced prior to sequencing, during PCR and library preparation. Concatenation of forward and reverse reads has been shown to increase recovery of control taxa in simulated data experiments (Liu et al., 2020) but with the caveat that sequence diversity might be artificially inflated. Ultimately, the challenge of read-trimming is to find a balance between elimination of error-rich sequence and preservation of valuable data (Haider, et al, 2024), and single-end (R1-only) data have yielded very similar results to paired-end data with respect to alpha and beta diversity, and OTU profiles (Werner et al., 2012). Whether joined reads (with poly(N)'s) are used, or forward and reverse reads are analyzed separately, information gaps in non-overlapping reads are not well accounted for by existing classifiers (Jeraldo et al, 2014).

In contrast to Illumina's error-prone end regions, Pacbio sequencing errors are evenly distributed throughout subreads. Where Illumina reads receive some correction via consensus at the junction of merged reads, the Pacbio platform generates full-length consensus sequences from up to thousands of subreads (Eid et al., 2009), producing an overall higher level of accuracy. This is an improvement over Illumina in terms of sequence quality, but Pacbio-based experiments are equally affected by error and bias generated during PCR and library preparation.

Amplification biases can occur during initial PCR amplification, during subsequent library preparation steps, or during the process of sequencing, and they cause an unequal distribution

of sequences among the taxa in an amplicon library. Biases occur when certain amplicons are amplified more efficiently than others because of factors like variation in primer affinity, amplicon size, GC content (Aird et al., 2011), and because of the presence of sequencing indexes (O'Donnell, et al., 2016) or secondary structures. Sequencing biases introduced by the Illumina MiSeq platform mainly result from its clustering phase, which is a resource-competitive process akin to PCR. This clonal amplification method (a.k.a. bridge amplification; Mayer, 2011) also preferentially amplifies smaller DNA fragments. Because of this, a sequencing library with significant size variation typically results in an overrepresentation of sequence reads from the smallest amplicons present. Pacbio sequence runs do not have an enrichment step comparable to the Illumina clustering phase, which may reduce the severity of such sequencing biases.

In this study, we assess and compare the performance of Illumina (MiSeq 600) sequence technology with PacBio (Sequel II) and Sanger sequencing technologies in estimating the richness, diversity, and relative abundance of nematode species from desert soil samples. To do this, we use the location-specific 28S reference sequence database and subfractions of the same soil samples that were used in **Chapter 1**. We PCR-amplified the same 750bp fragment, spanning the first through third domains of 28s rDNA. Our chosen PCR primers have excellent coverage within Nematoda and among other distantly related taxa. The first three domains of 28S rDNA have sufficient sequence diversity to resolve species-level differences in many nematode clades, which might be attributed to the amplicon's larger-than-average size or to its information density, particularly in the first domain. Due to the inherent difficulties of using joined reads with inserted poly(N)'s, we restrict our analyses in this chapter to Illumina forward

reads that are trimmed to 270bp. In the interest of making direct comparisons across sequencing technologies, we use PacBio and Sanger reads that are trimmed to the same length.

Using Illumina forward reads, PacBio reads, and Sanger reads that are generated from the same soil samples and trimmed to the same length (270bp), we compare:

1) Differences in recovered taxa between Illumina and PacBio sequences from the same amplicon library, including comparison with Sanger sequencing results from subfractions of the same soil samples (**Chapter 1**).
2) Effects of Illumina read 3' error profiles and chimeric sequences on the presence of artifactual sequence diversity.
3) Agreement of species richness, relative abundance, and alpha diversity measures across 10 PCR replicates and the Illumina and PacBio sequencing technologies.

We hypothesize that Illumina will show the highest number of observed features (ASVs, OTUs, observed taxa, species), followed by PacBio, with Sanger sequencing showing the least. This study aims to provide insights into the effectiveness of different sequencing technologies in characterizing nematode communities, as well as the challenges posed by the limitations of current molecular barcoding approaches.

2.2 Materials and Methods: Illumina Metabarcoding Study

This study was conducted on the third set of 9 soil subfractions collected in May of 2014.

2.2.1 Hydration of soil samples and isolation of nematodes

Approximately 850 cm^3 of soil was placed in 12 3oz paper cups that were punctured with a metal probe to make drainage holes. Eight rows of four holes were evenly spaced around the sides of the cups, and seven holes were punched in the bottom of each cup. Each of the 12 cups was filled to ~80% capacity with soil. The soil was then fully saturated with tap water, allowed to drain, and then incubated at 25C for 72 hours. After hydration and incubation, soil samples were placed on four 6" Baermann funnels (3 cups per funnel, 32x47cm one-ply Kimwipes holding the soil) at 28.5C in an incubator. Nematodes were collected from the funnels after 24 and 48 hours.

To reduce the concentration of inhibitors in or on the bodies of the nematodes, each funnel catch was subjected to a series of water exchanges. The volume of liquid was adjusted to ~150 ml. The beakers were then placed on ice to reduce the activity of the nematodes, thereby increasing the rate at which they settle to the bottom of the beaker. After 30 min on ice, the top 100 ml of liquid was aspirated off. This process was repeated twice more, or until the liquid in the beaker was colorless.

2.2.2 Nematode counting

After the water exchange procedure, samples were decanted in a 100mm square petri dish with a 6x6 grid. Using a stereoscope, all nematodes were counted in 6 randomly selected grid cells. Cells (X and Y coordinates) were selected using a random number generator. Counts were averaged over six cells and a total worm count was estimated over the total area of the petri dish. After counting, nematodes from each sample were decanted into a 50ml polystyrene tube (Falcon) and spun at X180g in a clinical centrifuge. Centrifugation and aspiration were used to pellet worms and reduce the liquid volume each sample until they could be transferred to a 2.0 ml microcentrifuge tube. The tubes were then centrifuged, and the supernatant was reduced again until the pellet of worms was barely submerged. The tubes were stored at -20C for future use. The number of nematodes per DNA extract ranges from 4006 to 6854 with an average of 5077 (**Table 2.1**).

2.2.3 DNA extraction

Tubes with frozen nematode pellets were thawed and resuspended in 600ul of Epicentre Masterpure (Madison, WI) Cell & Tissue Lysis Solution, then transferred to 3.0ml tubes that were pre-filled with 1mm zirconium beads (Benchmark Scientific, Sayreville, NJ). A Beadbug microtube homogenizer (Millipore Sigma) was used to homogenize the samples at maximum speed for two minutes. Two microliters of 50ug/ul proteinase K were added, the tubes were incubated at 65C for 1 hour, and vortexed every 20 minutes. A portion of the supernatant from the tissue digestions (~300ul) was transferred to 1.5ml tubes, and the Masterpure kit instructions were followed thereafter. DNA pellets were dried at RT in a desiccator and resuspended in 35ul of TE pH 8.0.

2.2.4 PCR, library preparation, and sequencing

Amplicon PCRs: The KOD XL polymerase kit was used to amplify a fragment (~750bp) of 5' 28S rDNA from the DNA extracts. PCR reactions were done in a 25ul volume and included 1.5ul of DNA extract, 3mM $MgCl_2$, 0.2mM of each dNTP, 0.5 units of KOD XL Polymerase, and 0.6uM each of primers 391F and 799R. Cycling conditions consisted of an initial melting step of 95C for 3 min, followed by 35 cycles of 95C for 30 seconds and 67C for 1 min, and a final extension step at 72C for 6 min. For each of the nine soil samples, three replicate PCRs were done, and the products were pooled to account for any amplification variability among individual reactions. To test the consistency of individual PCR reactions, ten replicate reactions were amplified from sample A123 and barcoded individually (below).

A series of primer pairs were designed and ordered from Eurofins Genomics, each with a unique forward or reverse barcode. Barcodes were created using a Hanning barcode generator (Cold Spring Harbor, NY), and added to primers 391F and 799R. The IDT Oligo Analyzer Tool (https://www.idtdna.com/calc/analyzer) was used to pair forward and reverse barcoded primers that had minimal self and cross complementarity (Δg > -6kcal/mol). The KAPA Hifi Hotstart polymerase kit was used to generate barcoded PCR products. Reactions were done in 50ul, 75ul, or 100ul volumes, depending on levels of PCR inhibition. The reactions included 3mM $MgCl_2$, 0.2mM of each dNTP, 1 unit of KAPA Hifi Hotstart polymerase, 1.2uM of forward and reverse barcoded primers, and the templates for these reactions were 1.5ul of 1/1000th dilution PCR products from each soil nematode sample. Cycling conditions consisted of an initial

melting step at 95C for 5 min, followed by 24 cycles of 98C for 20 sec, 69C for 15 sec, 72C for 45 sec, and a final extension step at 72C for 1 min.

Barcoded PCR products were cleaned using a custom magnetic bead protocol (Faircloth & Glenn, 2016). Depending on the measured concentration (Qubit fluorometer) of amplicon PCRs, either 20ul or 40ul of the product was incubated with a 50ul solution containing 10mM Tris pH 8.0, 1mM EDTA, and a 1:1 ratio of Seramag Speedbeads (Sigma-Aldrich) to polyethylene glycol (PEG-8000). Cleaned products were resuspended in 12ul of Tris-HCl pH 8.0, and the DNA concentrations were measured using a Qubit fluorometer. The products were then made equimolar and pooled. The sequencing library was submitted to the UC Davis DNA Technology Core, where Illumina adapters were added via a ligation reaction. The library, consisting of nine soil samples (each consisting of three PCRs combined) and ten replicates of A123 (single PCRs), was sequenced on the MiSeq600 platform. In a previous Miseq600 run, these samples were sequenced along with control reactions intended to represent laboratory contaminants. The control samples were greatly overrepresented (>65% of sequence capacity) in the resulting sequence files because of their smaller average fragment size (~400bp vs ~750bp). For this reason, they were omitted from this sequencing run. A subfraction of the final amplicon library was set aside and stored at -80C for PacBio sequencing.

2.2.5 Pacbio and Sanger sequencing

An aliquot of the amplicon library previously used for Illumina sequencing (**section 2.2.5**) was submitted to the UB3 sequencing center at The University of California, Berkeley. The library was subjected to SMRTbell adapter ligation and sequenced on the Sequel II platform. Sanger sequences were generated and preprocessed as detailed in **Chapter 1** (**section 1.3.3**).

2.2.6 DNA Sequence preprocessing

Because our targeted amplicon was ~750bp, there was no overlap between our Illumina forward and reverse reads. For the aforementioned reasons (see introduction), we decided to use only the Illumina forward (R1) reads. PacBio and Sanger reads were trimmed to the same length so data sets by the three sequencing technologies could be more directly compared. Illumina sequence fastq files were demultiplexed using the Sabre bioinformatic tool (Joshi, 2010). Demultiplexed Illumina sequences were imported into the Qiime2 bioinformatic pipeline, where they were subsequently denoised, stripped of primers, trimmed (to various lengths), and subjected to a chimera removal step (at varying levels of stringency) using the dada2 plugin *denoise-single* function.

Raw Pacbio sequencing data, in the form of subreads.bam files were converted into circular consensus sequences using the PacBio proprietary software, ccs. A --min-passes value of 20 was used. The resulting fastq files were demultiplexed using PacBio's lima tool, with a --min-score requirement of 80, a --min-end-score of 50, and a --min-ref-span of 0.75. Demultiplexed Illumina and PacBio sequences were imported into the Qiime2 bioinformatic pipeline, where

they were subsequently stripped of primers, denoised, trimmed to 270bp, and subjected to a chimera removal step using the dada2 plugin, with a --min-fold-parent-over-abundance (mfp) setting of 3.5, the default for the *denoise-ccs* function.

Sanger sequences from **Chapter 1** were truncated to 270bp using the Trimmomatic bioinformatic tool, then imported into Qiime2, where they were de-novo clustered at a similarity threshold of 99%.

All sequences were classified with the sklearn naïve Bayesian classifier (Bokulich et al., 2018) as implemented in Qiime2, in tandem with the custom 28S reference sequence database generated in **Chapter 1**. Classifications were assigned to representative ASVs and OTUs with a 70% confidence level.

2.2.7 Illumina read trimming test

Illumina reads (R1) and PacBio reads of the same length were compared to assess the effects of 3' end-specific errors on overall sequence variation. After primer sequences were removed, four datasets were generated for each sequencing technology by trimming reads to 270bp, 260bp, 250bp, and 240bp. Chimeras were removed from the Illumina data set using the 'consensus' method with an mfp setting of 1.0, the Qiime2 default setting for the *denoise-single* function. Chimeras were removed from the PacBio data using the *denoise-ccs* function with the default mfp setting of 3.5.

2.2.8 Illumina chimera removal test

To assess the effects of chimeras on overall Illumina sequence diversity, we compared the number of features generated at varying levels of stringency (mfp setting) in the chimera removal step employed by the dada2 plugin of Qiime2. The dada2 plugin in Qiime2 incorporates a chimera removal step in its denoising process. The mfp setting controls the filtering of putative chimeric sequences based on their abundance relative to their presumed "parent" sequences (Callahan, et al., 2016). During denoising, Illumina reads were trimmed at 270bp with mfp settings of 0.5, 1.00, 2.0, 3.0, 3.5, 4.0, 5.0, 6.0, 7.0, and 8.0.

2.2.9 Sequence diversity analyses

To search for the presence of and assess the effects of artifactual read diversity, we compared counts of taxonomic categories in our Illumina data set to those of the PacBio and Sanger (**Chapter 1**) data sets. To better facilitate comparisons, reads from all three technologies were trimmed to 270 bp. Illumina and Pacbio reads were denoised using mfp settings of 1.0 and 3.5, respectively. Sanger reads were de-novo clustered at a threshold of 99% similarity. All reads were classified at a confidence setting of 70%.

Statistical analyses and visualizations were created in R (4.2.3). Figures were generated to examine the effects that incremental read trimming (240bp-270bp) and stringency of chimera

removal (mfp = 0.5-8.0) have on the number of reads passing quality control filters during bioinformatic preprocessing and the effect those factors have on multiple measures of Illumina sequence diversity: number of ASVs, total number of taxonomic categories classified at 70% confidence, and number of species-level classifications made at 70% confidence. We define 'taxonomic category' as the lowest level the *sklearn* classifier can classify any given ASV or OTU at 70% confidence. It is important to note that any taxonomic category above species level may include ASVs or OTUs from lower taxonomic levels. By 'species-level classifications', we mean taxonomic categories successfully classified to the 7th taxonomic rank (species) at 70% confidence.

The vegan package in R was used to calculate observed taxa, Shannon Diversity, and Simpson Diversity for all samples in the Illumina, PacBio, and Sanger data sets. Shapiro-Wilk tests were done to assess the normality of calculated alpha diversity metrics. Observed taxa counts were used to estimate the taxonomic richness (real or artifactual) of samples, Shannon Diversity was used to assess sequence diversity with evenness taken into account, and the Simpson index was included to provide a diversity estimate that is less affected by low-abundance taxa, and therefore less influenced by spurious or artifactual sequence variants (RSVs). Analysis of variance (ANOVA) tests were performed to measure the significance of the effects that sequencing technologies have on each of the alpha diversity metrics, and then post hoc pairwise comparisons of significance were done using a Tukey's honestly significant difference (HSD) test.

Stacked bar plots were generated to visualize the consistency of species relative abundance within and among Illumina, PacBio, and Sanger sequencing technologies. In these plots, species-level taxonomic categories were sorted in order of descending relative abundance. Species over 0.05% relative abundance across the whole data set were assigned colors, while species below this abundance were colored grey, and any taxonomic categories not classified to species level were colored black.

Two sets of Bland-Altman plots were generated using the ggplot2 package in R. The first set of Bland-Altman plots were made to visually assess the agreement of the nine soil samples between technologies with respect to a number of observed taxa, Shannon Diversity, and Simpson Diversity. The second set of Bland-Altman plots was made to compare ten pairs of corresponding (sample A123) PCR replicates between the Illumina and PacBio datasets, thereby inferring the consistency of relative abundance measurements between Illumina and PacBio sequencing technologies. One Bland-Altman plot was made for each of the six most abundant species-level classifications among all A123 PCR replicates: *Acrobeloides* DC sp2, *Panagrolaimus* DC sp1, *Chiloplacus* DC sp1, *Acrobeles* DC sp1, *Aphelenchus* DC sp2, and *Acrobeloides* DC sp4.

To compare the degree of variation among Illumina A123 PCR replicates with the degree of variation among PacBio A123 PCR replicates, coefficients of variation (CVs) were calculated for the number of observed taxa, Shannon Diversity, and Simpson diversity for PCR replicates within each technology.

2.3 Results:

2.3.1 Classification results

Total counts of ASVs, taxonomic categories, and species-level classification across all three sequencing technologies are shown in **Table 2.2**. Illumina produced over 10 times more reads than PacBio and identified 69% more taxonomic categories, 75% more species-level classifications, and nearly 5 times more ASVs. When comparing PacBio to Sanger, PacBio generated over 10 times more reads and identified 53% more taxonomic categories, 43% more species-level classifications, and about one-third as many ASVs.

Taxonomic categories and their counts in Illumina, PacBio, and Sanger data sets are detailed in **Table 2.3.** Classification of 270bp reads yields 70 species-level classifications in the Illumina data set, 40 species-level classifications in the PacBio data set, and 28 species-level classifications in the Sanger data set. Several taxa were detected by PacBio and Sanger but not by Illumina. PacBio identified *Microdorylaimus DC sp1* (381 reads), *unclassified Microdorylaimus* (3 reads), and *unclassified Chromadorea* (319 reads), which were not found by Illumina. Sanger detected *Mesorhabditis monhystera* (11 reads) and *unknown Longidorella sp.* (10 reads), both of which were missed by Illumina. Additionally, *Stegelletina salinaria* was found by both PacBio (18 reads) and Sanger (5 reads), but not by Illumina.

2.3.2 Read trimming test results

There was a significant increase in the percentage of Illumina reads passing dada2 quality filters as bases were incrementally trimmed from 3' ends of reads (**Figure 2.1**). Incremental trimming of reads also resulted in a mild increase in the number of ASVs produced by denoising the Illumina data set (**Figure 2.2**). However, as more bases were trimmed from the 3' end of reads, the total number of taxonomic categories remained constant, as did the number of taxonomic categories that were successfully classified into species. The PacBio data set showed even less response to the loss of 3' bases, with virtually no change in ASV, taxonomic category, or species counts as reads were incrementally trimmed.

2.3.3 Chimera test results

As the stringency (mfp setting) of the chimera removal step incrementally increases, there was a gradually increasing percentage of Illumina reads that passed quality filters during the dada2 denoising process (**Figure 2.3**). There was also a gradual increase in the estimated number of ASVs as mfp settings increased. Despite these increases in sequence diversity, the number of taxonomic categories and species-level classifications remained constant (**Figure 2.4**).

2.3.4 *Calculated Alpha diversity metrics*

Number of observed taxa, Shannon index, and Simpson Index were calculated for the nine BDCR soil samples and the ten PCR replicates from sample A123, and they are shown in **Table**

2.4. Boxplots of observed taxa counts and Shannon and Simpson diversity indices for the nine BDCR soil samples are shown in **Figure 2.5**. ANOVA and subsequent Tukey's HSD test results are shown in **Table 2.5**. Observed taxa counts significantly differ in pairwise comparisons of all three sequencing technologies ($p < 0.001$). The Shannon Diversity Indices are not significantly different between Illumina and PacBio datasets ($p = 0.999$). Still, significant differences are found when Sanger is compared with Illumina ($p = 0.011$) and when Sanger is compared with PacBio ($p = 0.011$). Concerning Simpson Diversity, Pacbio samples are not significantly different from Illumina samples ($p = 0.995$), and there are marginally significant differences found in Sanger versus Illumina samples ($p = 0.063$) and Sanger versus PacBio samples (0.052).

2.3.5 Agreement between sequence technologies

As seen in the stacked bar plots of species from the three sequencing technologies (**Figures 2.6, 2.7, and 2.8**), the Illumina and PacBio reads yield visually similar patterns of species abundance. The Illumina dataset has 30 more species-level classifications than the PacBio data set, but all these additional taxa (colored grey) are below 0.05% relative abundance among all samples. Patterns of relative abundance also appear to have a high level of similarity among 10 PCR replicates included in the Illumina and PacBio data sets.

The Bland-Altman plot for observed taxa among the nine BDCR soil samples (**Figure 2.9A**) shows a mean difference (bias) of +15, indicating that the Illumina dataset overestimates the number of observed taxa relative to the PacBio data set by an average of 15 taxa across all soil samples.

The slope of the line of best fit indicates a positive proportional bias, meaning that Illumina's overestimation of observed taxa increases in severity as the number of observed taxa increases.

The Bland-Altman plot for Shannon and Simpson Indices among the nine BDCR soil samples (**Figure 2.9B and 2.9C**) shows mean differences close to zero, indicating very little systematic bias between Illumina and the PacBio measurements across all soil samples. Lines of best fit indicate a slightly negative proportional bias for Shannon diversity and a slightly positive proportional bias for Simpson diversity. These alpha diversity metrics have a high degree of consistency between the two sequence technologies.

Bland-Altman plots for the six most common species in A123 PCR replicate samples (**Figure 2.10**) vary by species with respect to the amount of systematic bias and proportional bias present and the consistency of relative abundance measurements in Illumina samples relative to PacBio samples.

As seen in the Bland-Altman plot for *Acrobeloides* DC sp2 (**Figure 2.10A**), the bias (0.03275) is nearly four times larger than the standard deviation (SD= 0.00832), suggesting that relative abundance measurements in the Illumina PCR replicates are significantly overestimated (by an average of +3.28%) in comparison with the PacBio PCR replicates and to a relatively high degree of precision. The line of best fit has a slope close to zero, indicating very little proportional bias.

The Bland-Altman plot for *Panagrolaimus* DC sp1 (**Figure 2.10B**) reveals a modest underestimation (Bias = -0.00782) of relative abundance in Illumina replicates, but the bias is still significant given the range of relative abundance measurements among the 10 PCR replicates (0.14 - 0.18). There is a low spread of variation, as indicated by the standard deviation (SD= 0.00409). The slope of the line of best fit is nearly zero, indicating minimal proportional bias.

For *Chiloplacus* DC sp1 (**Figure 2.10C**), the Bland-Altman plot shows relatively little bias (-0.00172) compared to the SD of 0.00425. In this case, the variability in measurements is more significant than the systematic bias between Illumina and Pacbio sequencing technologies, but both are still quite low. As with *Panagrolaimus* DC sp1, the slope of the line of best fit for *Chiloplacus* DC sp1 is close to zero, indicating minimal proportional bias.

The bias of *Acrobeles* DC sp1 (**Figure 2.10D**) relative abundance measurements (0.00949) is much larger than the SD (0.00198), indicating a significant (~38% of the total range) and consistent overestimation of relative abundance measurement in the Illumina replicates when compared to PacBio. The line of best fit suggests no proportional bias.

Aphelenchus DC sp2 (**Figure 2.10E**) relative abundance is consistently and relatively significantly underestimated in Illumina PCR replicates, with a moderate spread in the differences. A negative bias of -0.02354 is several times larger than the SD (0.00554). The slope of the line of

best fit indicates a negative proportional bias, so there is a higher agreement between the two technologies as relative abundance values increase.

Measurements of *Acrobeloides* DC sp4 (**Figure 2.10F**) relative abundance are consistently overestimated in Illumina PCR replicates relative to PacBio replicates. The bias (0.02166) is substantially larger than the SD (0.00176), indicating a significant systematic difference between the methods with high precision and very little proportional bias.

Coefficients of variation (CV) for the number of observed taxa between Illumina (8.5211) and PacBio (8.3066) PCR replicates are very similar (**Table 2.6**), suggesting the two technologies have a similar degree of variation relative to their means. The slightly higher CV for Illumina suggests marginally higher variability in the number of observed taxa across PCR replicates relative to PacBio. For Shannon Diversity, the two technologies have more similar CV values (Illumina= 2.1300, PacBio= 2.0885). The CVs for Simpson diversity show the highest difference in variability among all measurements between the two technologies, with 2.695 for Illumina and 2.2460 for PacBio. Within technologies, coefficients of variation (CV) for Shannon and Simpson diversity are low (all < 3), suggesting a high level of agreement of measurement within Illumina and PacBio PCR replicates. The CVs for observed taxa are higher within Illumina and PacBio replicates (both > 8), suggesting less agreement within technology for that metric.

2.4 Discussion

2.4.1 Differences in recovered taxa between Illumina and PacBio datasets.

There were 57 taxonomic categories (43 species-level) in the Illumina and Pacbio datasets that were not present in the Sanger data set (**Table 2.3**). Unsurprisingly, the two HTS data sets have higher species richness because they represent ~46,0000 nematodes across nine soil samples, as opposed to the 2340 individuals represented in the Sanger data. However, 37 taxonomic categories (30 species-level) are unique to the Illumina dataset. These Illumina-unique taxa are ~39% of all taxonomic categories but less than 0.127% of all Illumina reads. The most abundant of these Illumina-unique taxonomic categories, *Acrobeloides* sp. FHD001, consists of 2415 reads or just 0.045% of all Illumina reads. Their extremely low abundance, coupled with the fact that they do not occur in the PacBio or Sanger datasets, suggests that they are RSVs that have been mistaken for reference taxa by the sklearn classifier.

Only two taxonomic categories were unique to the PacBio data set: *Microdorylaimus* DC sp1 (38 reads) and unclassified *Microdorylaimus* (3 reads). It is unclear why these taxonomic categories were not recovered in the Illumina or Sanger datasets, but they comprise less than 0.01% of PacBio reads. These taxonomic classifications could have resulted from PacBio-specific RSVs, or from PacBio-specific preprocessing steps.

Two taxonomic categories were shared by Illumina and Sanger but not by PacBio: *Acrobeloides sp.* CR-2010 and *Zeldia punctata*. These species could have been recovered from the Sanger data because of inconsistencies in manual sequence editing or failure to recognize chimeric

sequences upon visual inspection of electropherograms. Likewise, in the Illumina dataset, these species could result from RSVs of other *Acrobeloides* and *Zeldia* reads. Another possibility is that inherent differences between Illumina and Pacbio denoising processes resulted in the differential exclusion of particular sequence variants.

2.4.2 Artifactual sequence variants are not caused by Illumina 3' read errors or chimeras

As Illumina reads are incrementally trimmed, there is an observable increase in sequence diversity, as measured by the number of ASVs. It seems counterintuitive that more extensive trimming of reads would result in an increase in ASVs. Still, shorter reads have fewer 3' end-specific errors, ensuring that more reads pass quality filters. The result is an expanded sequence pool with increased random variation and a higher number of ASVs. However, it is important to note that the increased sequence variation resulting from additional reads passing filter does not increase the number of taxonomic categories being successfully classified. That is to say, the variation in the additional reads, likely produced by random error, does not alter the taxonomic classification of those reads. Like read trimming, successive reduction of chimera removal stringency (mfp) settings increases the number of reads passing filter, which increases the number of ASVs detected but does not significantly affect the number of taxonomic categories and species-level classifications.

The fact that increased taxonomic diversity among Illumina reads is unaffected by read trimming suggests that it is not caused by phenomena that cause deterioration of accuracy at

the 3’ ends of reads, such as phasing, pre-phasing, loss of context, or signal decay. It is more likely that this artifactual sequence diversity is caused by phenomena that cause base mismatches or miscalls, including polymerase mistakes and cross-talk between adjacent clusters (Wang, et al., 2017).

2.4.3 Agreement of diversity metrics between and within HTS technologies

Stacked bar plots show modest disagreement of relative abundance measures between corresponding pairs of BCDR samples in Illumina and PacBio datasets (**Figure 2.6**) and the corresponding sets of ten PCR replicates (**Figure 2.7**). The most abundant species among Illumina samples are also the most abundant species in the PacBio samples. Yet, the agreement of relative abundance measurements between technologies varies from species to species.

Bland-Altman plots for alpha diversity metrics among nine BDCR samples indicate that Illumina sequencing consistently overestimates the number of observed taxa relative to PacBio by ~15 taxa on average, which amounts to an overestimation of ~61%, given an average of ~25 taxa in PacBio BCDR samples. Despite the significant overestimation of observed taxa in Illumina samples, Bland-Altman plots demonstrate high agreement of Shannon and Simpson estimates between the Illumina and PacBio technologies.

Bland-Altman plots revealed varying degrees of systematic and proportional bias when assessing the agreement of relative abundance measurements for the six most common species

between corresponding sets of Illumina and PacBio PCR replicates, with the agreement being species-dependent. Some species had significantly divergent estimates between the datasets, while others showed variability more indicative of data dispersion than systematic differences between technologies.

CV values for all three alpha diversity metrics are slightly higher for Illumina than PacBio, indicating more variation among Illumina PCR replicates than among PacBio replicates, relative to their means. Estimated numbers of observed taxa are not very dissimilar within technologies, as shown by relatively low CV values (both below 9%) for Illumina and PacBio PCR replicate samples. There is even less variability for Shannon and Simpson indices within technologies, with CV values all under 3%. Shannon diversity is slightly higher overall among Illumina PCR replicates than PacBio PCR replicates, and Shannon diversity measures are more variable than Simpson diversity among Illumina PCR replicates.

The inflated taxonomic richness that we see in Illumina samples is due to an influx of RSVs (variants) during the sequencing process, and it has a more pronounced effect on Shannon diversity than on Simpson diversity. Simpson diversity gives more weight to the abundances of the most common taxa, making the additional low-abundance taxa in Illumina samples have a less pronounced effect on their Simpson indices. When analyzing Illumina sequence community data, alpha diversity metrics like Simpson, which are less sensitive to fluctuations in species richness, may provide more dependable diversity measures. Relative abundance estimates are less predictable than alpha diversity metrics but seem to have a high level of consistency

between Illumina and PacBio sequencing technologies. Despite the tendency of Illumina sequencing to artifactually inflate richness estimates, its higher sequencing capacity and lower cost per sample make it a potentially acceptable trade-off, especially when downstream analyses are not especially sensitive to variation in richness estimates.

2.4.4 Future use of relative abundance cutoffs

The putative artifactual taxa in our Illumina dataset all have extremely low abundances. In metabarcoding studies, it is a generally accepted best practice to filter out very low-abundance taxa to remove spurious or artifactual taxa (Littleford-Colquhoun et al., 2022). It is difficult to know where to set filtering cutoffs for our purposes because there are many biologically real taxa that are present in extremely low abundances in our BDCR samples. For example, *Carcharodiscus banaticus* is recovered by all three sequencing technologies and, therefore likely to be a biological reality (see **Table 2.3**), yet this species contributes just 27 reads (5.08×10^{-4} %) to the Illumina data set, four reads (8.56×10^{-4} %) to the PaBio data set, and one read (0.040%) to the Sanger data set. Imposing standard minimum read abundance cutoffs (10-20 reads, 0.01%-0.1% relative abundance) may risk eliminating real taxa along with artifactual taxa. Different cutoffs may be advisable when conducting biodiversity inventories as opposed to community ecology experiments.

2.4.5 Value of the previous reverse taxonomy study

Most species-level classifications in the study would not have been possible without the reverse taxonomy approach used in **Chapter 1** to characterize BDCR nematode communities. Of the 70 species-level classifications in the Illumina dataset, 22 were putative species originally found in the reverse taxonomy study, and they comprised ~79.8% of Illumina reads. Of the 40 species-level classifications in the PacBio dataset, 19 were taxa found in the reverse taxonomy study, and they comprised ~77.8% of PacBio reads. Of the 28 species-level classifications in the Sanger dataset, 16 were species found in the reverse taxonomy study, and they comprised ~90.4% of Sanger reads.

2.4.6 Effects of truncated sequence on taxonomic results

In this study, we used only forward Illumina reads, which were trimmed to 270bp, and we trimmed PacBio and Sanger reads to the same length. We did this to make more direct comparisons among Illumina, PacBio, and Sanger technologies. We opted not to use joined Illumina forward and reverse reads, with multiple ambiguous bases (N's) inserted between them, because we cannot differentiate the effects of missing data from the effects of technology-specific errors and biases.

An important consequence of using 270bp read fragments is that the community data in this chapter have significantly lower taxonomic resolution than those in Chapter 1, which were based on the full ~750bp amplicon. We know from the Chapter 1 results that the classification of full-length Sanger reads at 70% confidence yields eight more taxonomic categories and four

more species-level classifications than 270bp reads, with ~84.5% of Sanger reads receiving the same classification despite a reduction in amplicon length. Full-length PacBio reads at 70% confidence yield 176 more ASVs, ten more taxonomic categories, and ten more species-level classifications than 270bp PacBio reads. Nematode species inventories benefit most from using the full (~750bp) 5' 28S amplicon, which is only feasible using PacBio sequencing technology.

2.4.7 The need for mock community samples

Mock community samples will be needed to better understand how artifactual sequence variation might be mistaken for biological variation in our datasets. A variety of species would need to be cultured in order to ensure sufficient numbers of conspecific nematodes. Ideally, the cultured species would come from our BDCR study site. The composition of the mock community should at least be similar enough to the BDCR community to ensure comparable levels of inter- and intraspecific sequence variation. A well-characterized mock community would allow us to estimate recall and precision (as in Bokulich et al., 2018), as well as the sensitivity of our methods when detecting low-abundance taxa, or even single individuals.

Chapter 2 References:

Aird, D., et al. (2011). "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries." Genome biology **12**(2): 1-14.

Allhoff, M., et al. (2013). Discovering motifs that induce sequencing errors. BMC bioinformatics, Springer.

Blaxter, M. L. (2003). "Nematoda: genes, genomes and the evolution of parasitism." Adv Parasitol **54**: 101-195.

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

Bokulich, N. A., et al. (2018). "Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin." Microbiome **6**(1): 1-17.

Cacho, A., et al. (2016). "A comparison of base-calling algorithms for illumina sequencing technology." Briefings in bioinformatics **17**(5): 786-795.

Callahan, B. J., et al. (2016). "DADA2: High-resolution sample inference from Illumina amplicon data." Nature methods **13**(7): 581-583.

Edgar, R. C., et al. (2011). "UCHIME improves sensitivity and speed of chimera detection." Bioinformatics **27**(16): 2194-2200.

Eid, J., et al. (2009). "Real-time DNA sequencing from single polymerase molecules." Science **323**(5910): 133-138.

Fuks, G., et al. (2018). "Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling." Microbiome **6**(1): 1-13.

Haas, B. J., et al. (2011). "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons." Genome research **21**(3): 494-504.

Haider, D., et al. (2024). "Mock microbial community meta-analysis using different trimming of amplicon read lengths." Environmental Microbiology **26**(1): e16566.

Jeraldo, P., et al. (2014). "IM-TORNADO: a tool for comparison of 16S reads from paired-end libraries." PloS one **9**(12): e114804.

Kircher, M., et al. (2009). "Improved base calling for the Illumina Genome Analyzer using machine learning strategies." Genome biology **10**: 1-9.

Knight, R., et al. (2018). "Best practices for analysing microbiomes." Nature Reviews Microbiology **16**(7): 410-422.

Kozich, J. J., et al. (2013). "Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform." Applied and Environmental Microbiology **79**(17): 5112-5120.

Littleford-Colquhoun, B. L., et al. (2022). The precautionary principle and dietary DNA metabarcoding: commonly used abundance thresholds change ecological interpretation, Wiley Online Library.

Liu, Z., et al. (2007). "Short pyrosequencing reads suffice for accurate microbial community analysis." Nucleic acids research **35**(18): e120.

Liu, T., et al. (2020). "Joining Illumina paired-end reads for classifying phylogenetic marker sequences." BMC bioinformatics **21**: 1-13.

Maslen, N. (1980). "Additions to the nematode fauna of the Antarctic region with keys to taxa." British Antarctic Survey Bulletin **49**: 207-229.

Mayer, Pascal. (2011). “Isothermal amplification of nucleic acids on a solid support.” US patent 7972820 B2, Filed Jan 13th, 2011, and issued Jul. 5th, 2011.

O’Donnell, J. L., et al. (2016). "Indexed PCR primers induce template-specific bias in large-scale DNA sequencing studies." PloS one **11**(3): e0148698.

Öpik, M., et al. (2013). "Global sampling of plant roots expands the described molecular diversity of arbuscular mycorrhizal fungi." Mycorrhiza **23**: 411-430.

Prokopowich, C. D., et al. (2003). "The correlation between rDNA copy number and genome size in eukaryotes." Genome **46**(1): 48-50.

Porazinska, D. L., et al. (2009). "Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity." Molecular Ecology Resources **9**(6): 1439-1450.

Schirmer, M., et al. (2015). "Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform." Nucleic acids research **43**(6): e37-e37.

Taberlet, P., et al. (2012a). Environmental DNA, Wiley Online Library. **21:** 1789-1793.

Taberlet, P., et al. (2012b). "Towards next-generation biodiversity assessment using DNA metabarcoding." Molecular Ecology **21**(8): 2045-2050.

Taylor, D. L., et al. (2014). "A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning." Ecological monographs **84**(1): 3-20.

Wang, Q., et al. (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." Applied and Environmental Microbiology **73**(16): 5261-5267.

Wang, B., et al. (2017). "An adaptive decorrelation method removes Illumina DNA base-calling errors caused by crosstalk between adjacent clusters." Scientific reports **7**(1): 41348.

Werner, J. J., et al. (2012). "Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys." The ISME Journal **6**(7): 1273-1276.

Chapter 2 Tables and Figures:

Sample Number	Estimated Count
A122	4310
A123	5796
A124	5208
A125	4006
A126	4717
A127	5145
A128	4200
A129	5460
A130	6854

Table 2.1. Estimated counts of nematodes isolated from BDCR soil samples using Baermann funnel extraction.

Metric/Sequence Technology	Illumina	PacBio	Sanger
Total Number of Reads	5308406	472862	2480
ASVs/OTUs	1322 ASVs	233 ASVs	699 OTUs
Taxonomic Categories	93	55	36
Species-level Classifications	70	40	28

Table 2.2. Richness metrics for all BDCR samples as measured by Illumina, PacBio, and Sanger sequence technologies. Metrics include amplified sequence variants (ASVs) for Illumina and PacBio datasets, operational taxonomic units (OTUs) for the Sanger data set, total taxonomic categories, and taxonomic categories that were successfully classified to species at 70% confidence by the *sklearn* classifier in Qiime2.

Order	Family	Taxon	Illumina Count	Pacbio Count	Sanger Count
Dorylaimida	*Aporcelaimidae*	*Aporcelaimellus*_DC_sp6	33013	3658	26
Dorylaimida	*Aporcelaimidae*	*Aporcelaimellus_salicinus*	5	0	0
Dorylaimida	*Aporcelaimidae*	*Paraxonchium_laetificans*	85	13	0
Dorylaimida	*Leptonchidae*	*Utahnema*_DC_sp3	2733	399	2
Dorylaimida	*Leptonchidae*	*Utahnema*_DC_sp4	15427	1607	10
Dorylaimida	*Nordiidae*	*Kochinema_farodai*	92	0	0
Dorylaimida	*Nordiidae*	*Longidorella_penetrans*	20	0	0
Dorylaimida	*Nordiidae*	unknown_*Longidorella*_sp.	2133	279	10
Dorylaimida	*Nygolaimidae*	*Clavicaudoides_clavicaudatus*	90	7	0
Dorylaimida	*Qudsianematidae*	*Discolaimium*_sp._1_WJW-2018	11	0	0
Dorylaimida	*Qudsianematidae*	*Discolaimoides_symmetricus*	14	0	0
Dorylaimida	*Qudsianematidae*	*Ecumenicus*_DC_sp7	20761	2866	28
Dorylaimida	*Qudsianematidae*	unclassified_*Ecumenicus*	231	0	0
Dorylaimida	*Qudsianematidae*	*Microdorylaimus*_DC_sp1	0	381	0
Dorylaimida	*Qudsianematidae*	*Microdorylaimus_miser*	2	0	0
Dorylaimida	*Qudsianematidae*	unclassified_*Microdorylaimus*	0	3	0
Dorylaimida	*Telotylenchidae*	*Quinisulcius*_DC_sp1	2980	354	0
Dorylaimida	*Tylencholaimidae*	*Tylencholaimus_mirabilis*	3149	485	6
Dorylaimida	*Tylencholaimidae*	unclassified_*Tylencholaimus*	116	6	0
Dorylaimida	unknown *family*	*Carcharodiscus_banaticus*	27	4	1
Dorylaimida	unknown *family*	unknown_*Dorylaimida*_DC_sp2	3112	435	0
Dorylaimida	unknown *family*	unknown_*Dorylaimida*_sp.	2040	266	2
Dorylaimida	unknown *family*	unclassified_*Dorylaimida*	14921	1656	63
Rhabditida	*Anguinidae*	*Ditylenchus_anchilisposomus*	10	0	0
Rhabditida	*Anguinidae*	*Ditylenchus*_sp._85C1	55899	6907	2
Rhabditida	*Anguinidae*	unclassified_*Ditylenchus*	47	0	0
Rhabditida	*Aphelenchidae*	*Aphelenchus*_DC_sp1	23356	1843	6
Rhabditida	*Aphelenchidae*	*Aphelenchus*_DC_sp2	196517	33188	6
Rhabditida	*Aphelenchidae*	unclassified_*Aphelenchus*	1593	149	8
Rhabditida	*Aphelenchidae*	unknown_*Aphelenchus*_sp1	326	0	0
Rhabditida	*Aphelenchidae*	unknown_*Aphelenchus*_sp2	2	0	0
Rhabditida	*Aphelenchoididae*	*Aphelenchoides*_sp._NKZ223	5	2	0
Rhabditida	*Aphelenchoididae*	unknown_*Aphelenchoides*_sp.	46	0	0
Rhabditida	*Aphelenchoididae*	unclassified_*Aphelenchoididae*	2083	213	1

Rhabditida	Aphelenchoididae	Bursaphelenchus_sp._1_SAS-2019	9	0	0
Rhabditida	Aporcelaimidae	Axonchium_propinquum	113	21	1
Rhabditida	Cephalobidae	Acrobeles_cf._complexus_CR-2010	81	0	0
Rhabditida	Cephalobidae	Acrobeles_complexus	10	0	0
Rhabditida	Cephalobidae	Acrobeles_DC_sp1	575094	46370	684
Rhabditida	Cephalobidae	Acrobeles_DC_sp2	211220	11802	347
Rhabditida	Cephalobidae	Acrobeles_singulus	2878	160	3
Rhabditida	Cephalobidae	Acrobeles_sp._CR-2010	6	0	0
Rhabditida	Cephalobidae	Acrobeles_sp._JB-132	366	0	0
Rhabditida	Cephalobidae	unclassified_Acrobeles	3661	197	10
Rhabditida	Cephalobidae	Acrobeloides_DC_sp1	216556	12025	403
Rhabditida	Cephalobidae	Acrobeloides_DC_sp2	1382877	122891	188
Rhabditida	Cephalobidae	Acrobeloides_DC_sp3	6	0	0
Rhabditida	Cephalobidae	Acrobeloides_DC_sp4	202251	7490	125
Rhabditida	Cephalobidae	Acrobeloides_DC_sp5	20	0	0
Rhabditida	Cephalobidae	Acrobeloides_DC_sp6	54236	5170	5
Rhabditida	Cephalobidae	Acrobeloides_sp._CR-2010	647	0	8
Rhabditida	Cephalobidae	Acrobeloides_sp._DWF-1106	76869	4963	0
Rhabditida	Cephalobidae	Acrobeloides_sp._FHD001	2415	0	0
Rhabditida	Cephalobidae	Acrobeloides_sp._M31D	64	0	0
Rhabditida	Cephalobidae	unclassified_Acrobeloides	37722	13037	11
Rhabditida	Cephalobidae	Cephalobus_cubaensis	43	0	0
Rhabditida	Cephalobidae	Cervidellus_DC_sp1	11833	820	40
Rhabditida	Cephalobidae	Cervidellus_sp._SN-2010	78	0	0
Rhabditida	Cephalobidae	Cervidellus_sp.1_HMM2018	4641	313	0
Rhabditida	Cephalobidae	unknown_Cervidellus_sp.	260	12	0
Rhabditida	Cephalobidae	unclassified_Cervidellus	13	0	0
Rhabditida	Cephalobidae	Chiloplacus_DC_sp1	169148	17796	21
Rhabditida	Cephalobidae	Chiloplacus_demani	408	0	0
Rhabditida	Cephalobidae	Nothacrobeles_borregi	1935	95	0
Rhabditida	Cephalobidae	Nothacrobeles_spatulatus	6	0	0
Rhabditida	Cephalobidae	Stegelleta_DC_sp1	83	0	0
Rhabditida	Cephalobidae	Stegelleta_DC_sp2	20766	1374	11
Rhabditida	Cephalobidae	unclassified_Stegelleta	188	0	0
Rhabditida	Cephalobidae	Stegelletina_salinaria	18	0	0
Rhabditida	Cephalobidae	unclassified_Stegelletina	705	0	0
Rhabditida	Cephalobidae	unknown_Stegelletina_sp.	44786	3094	9
Rhabditida	Cephalobidae	unclassified_Zeldia	62	0	0
Rhabditida	Cephalobidae	Zeldia_punctata	121	0	2

Rhabditida	*Cephalobidae*	*Zeldia_spannata*	3071	161	4
Rhabditida	*Cephalobidae*	unclassified_*Cephalobidae*	57320	2823	25
Rhabditida	*Diplogasteridae*	unlassified_*Diplogasteridae*	2	0	0
Rhabditida	*Onchocercidae*	*Dirofilaria_immitis*	42	0	0
Rhabditida	*Panagrolaimidae*	*Panagrolaimus*_DC_sp1	1090302	97933	376
Rhabditida	*Panagrolaimidae*	unclassified_*Panagrolaimidae*	523337	56833	23
Rhabditida	*Panagrolaimidae*	unclassified_*Panagrolaimus*	3062	277	0
Rhabditida	*Panagrolaimidae*	unknown_*Panagrolaimidae*_sp.	3291	36	0
Rhabditida	*Rhabditidae*	*Diploscapter*_sp._JU359	55479	6492	0
Rhabditida	*Rhabditidae*	*Mesorhabditis_monhystera*	1292	56	11
Rhabditida	*Rhabditidae*	unclassified_*Mesorhabditis*	64	2	0
Rhabditida	*Rhabditidae*	*Protorhabditis*_sp._JB122	22	0	0
Rhabditida	*Rhabditidae*	unclassified_*Rhabditidae*	9	0	0
Rhabditida	*Sphaerulariidae*	*Bealius_pinus*	310	32	0
Rhabditida	*Sphaerulariidae*	unclassified_*Sphaerulariidae*	49	8	0
Rhabditida	*Sphaerulariidae*	*Veleshkinema_iranicum*	27	0	0
Rhabditida	*Strongyloididae*	unclassified_*Strongyloides*	156	5	0
Rhabditida	*Tylenchidae*	unknown_*Tylenchidae*_sp1_HMM2018	504	42	0
Rhabditida	unknown *family*	unknown_*Rhabditida*_sp.	24	0	0
Rhabditida	unknown *family*	unknown_*Rhabditida*_sp._3006ed	464	0	0
Rhabditida	*unknown family*	unclassified_*Rhabditida*	166368	5492	2
unknown order	unknown *family*	unclassified_*Chromadorea*	170	319	0

Table 2.3. Taxonomic categories of Illumina, PacBio, and Sanger sequence reads at 70% confidence using the *sklearn* classifier, as well as read counts for each. The "unclassified" designator was used when the *sklearn* naïve Bayesian classifier was not able to identify OTUs to a particular taxonomic rank at 70% confidence. Reference taxa that did not include species names are designated as "unnamed sp.". Two dorylaimidan reference taxa and one panagrolaimid reference taxon were designated as "unknown" with respect to family and/or genus level classification. Taxonomic categories that are classified to genus and higher levels may include sequence reads from more than one species. Illumina unique taxa are colored red. Pacbio-unique taxa are colored blue. Two taxa (green) are found in Illumina and Sanger datasets, but not PacBio.

Sample Name	Illumina			PacBio		
	Observed Taxa	Shannon	Simpson	Observed Taxa	Shannon	Simpson
A122	50	2.28618	0.84813	29	2.33074	0.84973
A123	31	1.89501	0.78946	21	1.88467	0.79867
A124	39	2.14432	0.83428	27	2.35737	0.87357
A125	42	2.13516	0.84716	25	2.08417	0.83313
A126	41	1.73249	0.72252	26	1.73752	0.72226
A127	40	1.94441	0.76305	26	1.84782	0.75954
A128	38	1.67612	0.67247	24	1.77200	0.70786
A129	37	2.09039	0.82028	24	2.11010	0.81723
A130	44	2.28487	0.86771	22	2.07017	0.83172
A123 replicate 1	34	1.86518	0.74380	21	1.91152	0.76808
A123 replicate 2	27	1.79061	0.75230	19	1.82493	0.77159
A123 replicate 3	34	1.86491	0.77620	23	1.88896	0.78993
A123 replicate 4	28	1.84043	0.76994	19	1.86842	0.78570
A123 replicate 5	29	1.85230	0.76370	21	1.89976	0.78455
A123 replicate 6	32	1.85431	0.76362	20	1.86943	0.77658
A123 replicate 7	31	1.85194	0.76207	20	1.87301	0.77835
A123 replicate 8	31	1.80703	0.73849	18	1.80911	0.75578
A123 replicate 9	28	1.76347	0.71352	18	1.81860	0.74530
A123 replicate 10	28	1.77651	0.72581	18	1.80821	0.73994

Table 2.4. Alpha diversity metrics calculated from nematode communities in the 9 BDCR soil samples and 10 PCR replicates by Illumina and PacBio sequencing technologies, including observed number of taxa, Shannon index, and Simpson index.

Diversity Measure	Comparison	Mean Difference	Lower Confidence Interval	Upper Confidence Interval	p-value	Significance
Observed Taxa	PacBio vs. Illumina	-15.33	-20.03	-10.63	<0.00001	***
Observed Taxa	Sanger vs. Illumina	-26.78	-31.48	-22.08	<0.00001	***
Observed Taxa	Sanger vs. PacBio	-11.44	-16.15	-6.74	<0.00001	***
Shannon Diversity	PacBio vs. Illumina	0.0006	-0.3244	0.3257	0.99999	ns
Shannon Diversity	Sanger vs. Illumina	-0.4153	-0.7403	-0.0902	0.01057	*
Shannon Diversity	Sanger vs. PacBio	-0.4159	-0.7409	-0.0909	0.01045	*
Simpson Diversity	PacBio vs. Illumina	0.0032	-0.0817	0.0881	0.99518	ns
Simpson Diversity	Sanger vs. Illumina	-0.0812	-0.1661	0.0037	0.06268	ns
Simpson Diversity	Sanger vs. PacBio	-0.0844	-0.1693	0.0005	0.05160	ns

Table 2.5. ANOVA and Tukey's HSD Test results for number of observed taxa, Shannon Diversity, and Simpson Diversity alpha diversity estimates from Illumina, PacBio, and Sanger sequence data sets.

Technology and Metric	Standard Deviation (SD)	Coefficient of Variation (CV)
Illumina Observed Taxa	2.57337	8.52109
Illumina Shannon Diversity	0.03883	2.12997
Illumina Simpson Diversity	0.02024	2.69504
PacBio Observed Taxa	1.63639	8.30656
PacBio Shannon Diversity	0.03879	2.08854
PacBio Simpson Diversity	0.01729	2.24601

Table 2.6. Variability of measurements for Illumina and PacBio sequencing technologies, across 10 PCR replicates, using 3 metrics: number of observed taxa, Shannon Diversity, and Simpson Diversity.

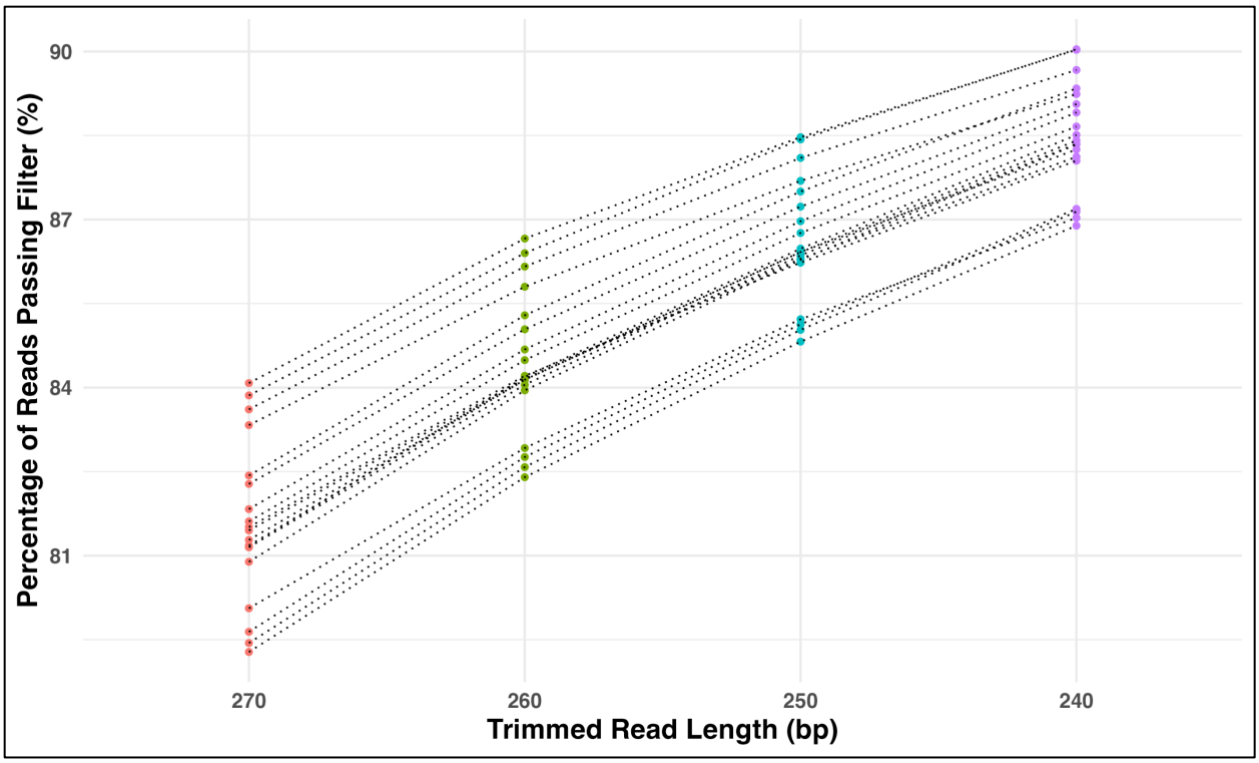

Figure 2.1. Effects of trimmed read length on percent of Illumina sequence reads passing filter. Each set of colored dots represents 9 BDCR soil samples and 10 PCR replicates, which were trimmed to a given length.

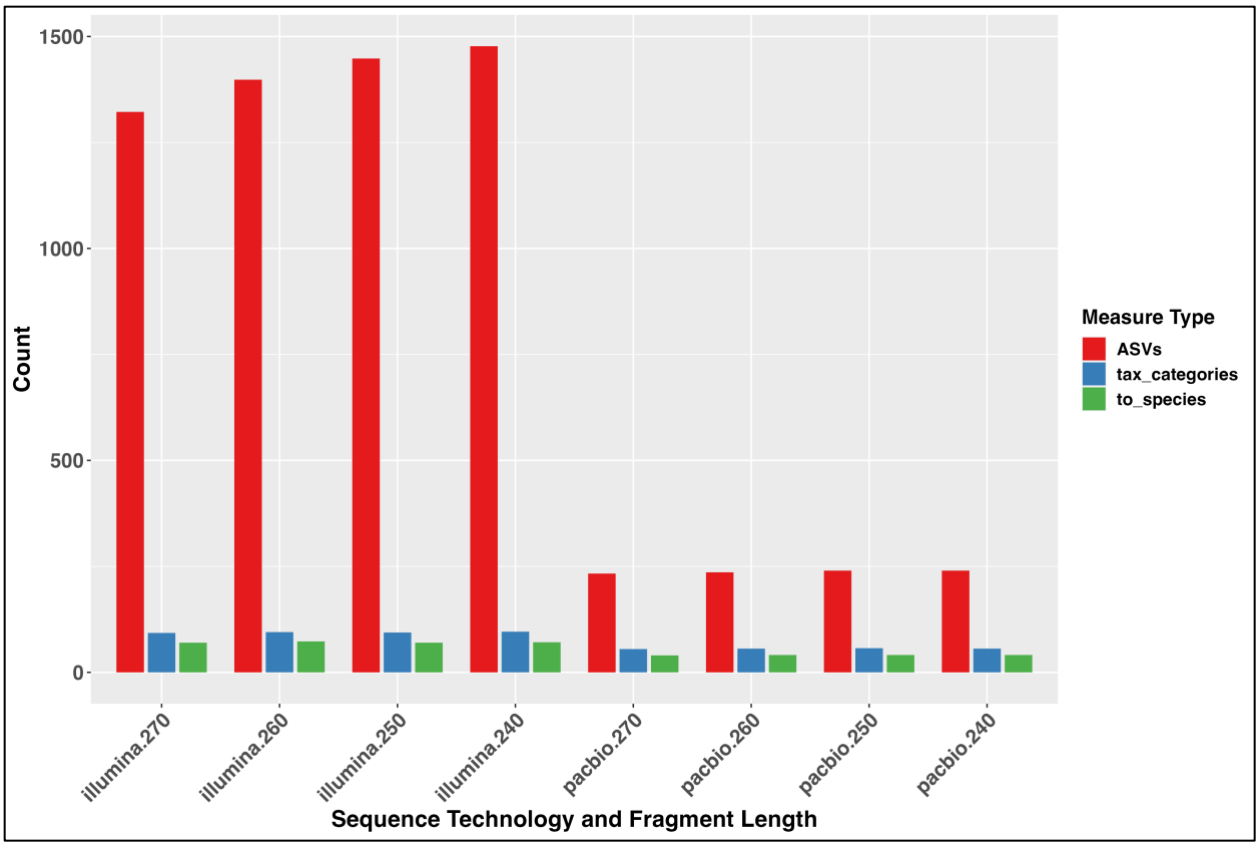

Figure 2.2. Effects of trimmed read lengths on Illumina sequence diversity. Diversity metrics include number of amplified sequence variants (ASVs), total number of taxonomic categories, and taxonomic categories successfully classified to species level at 70% confidence.

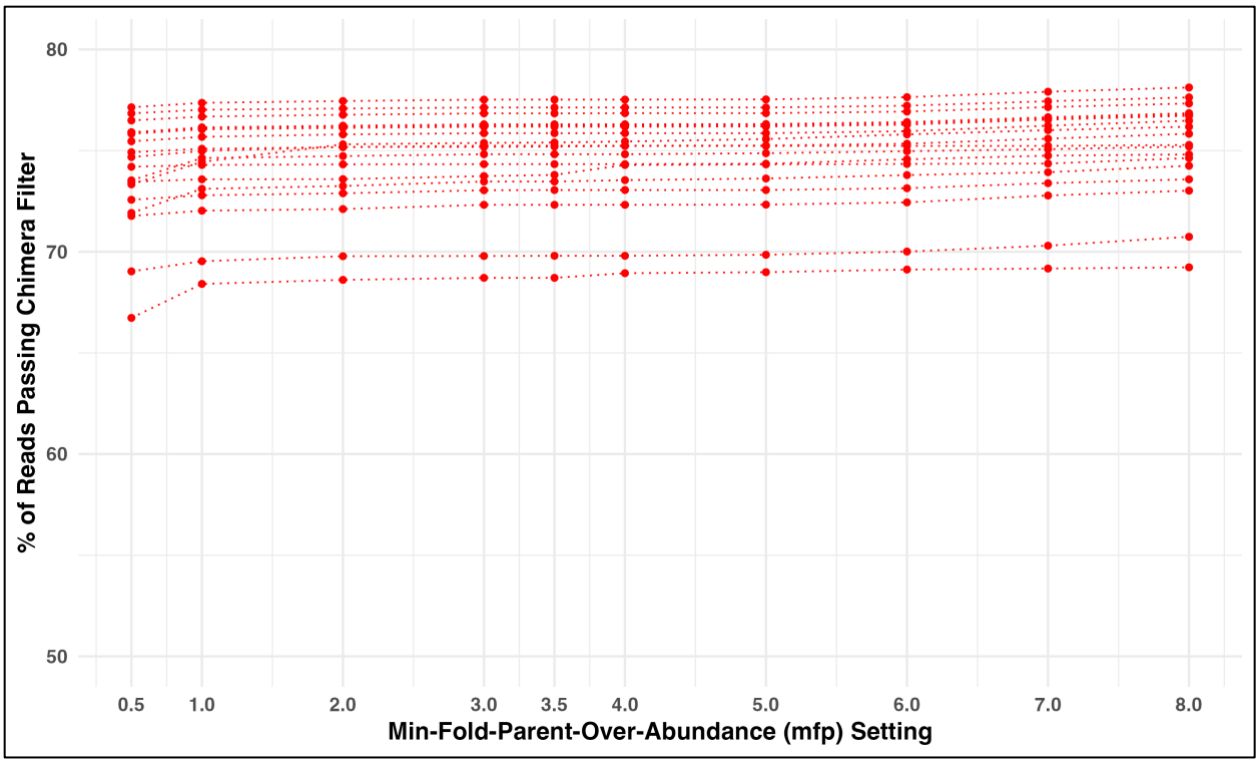

Figure 2.3. Total number of Illumina reads passing all filters as stringency of chimera removal increases from a setting of mfp=0.5 to mfp=8.0.

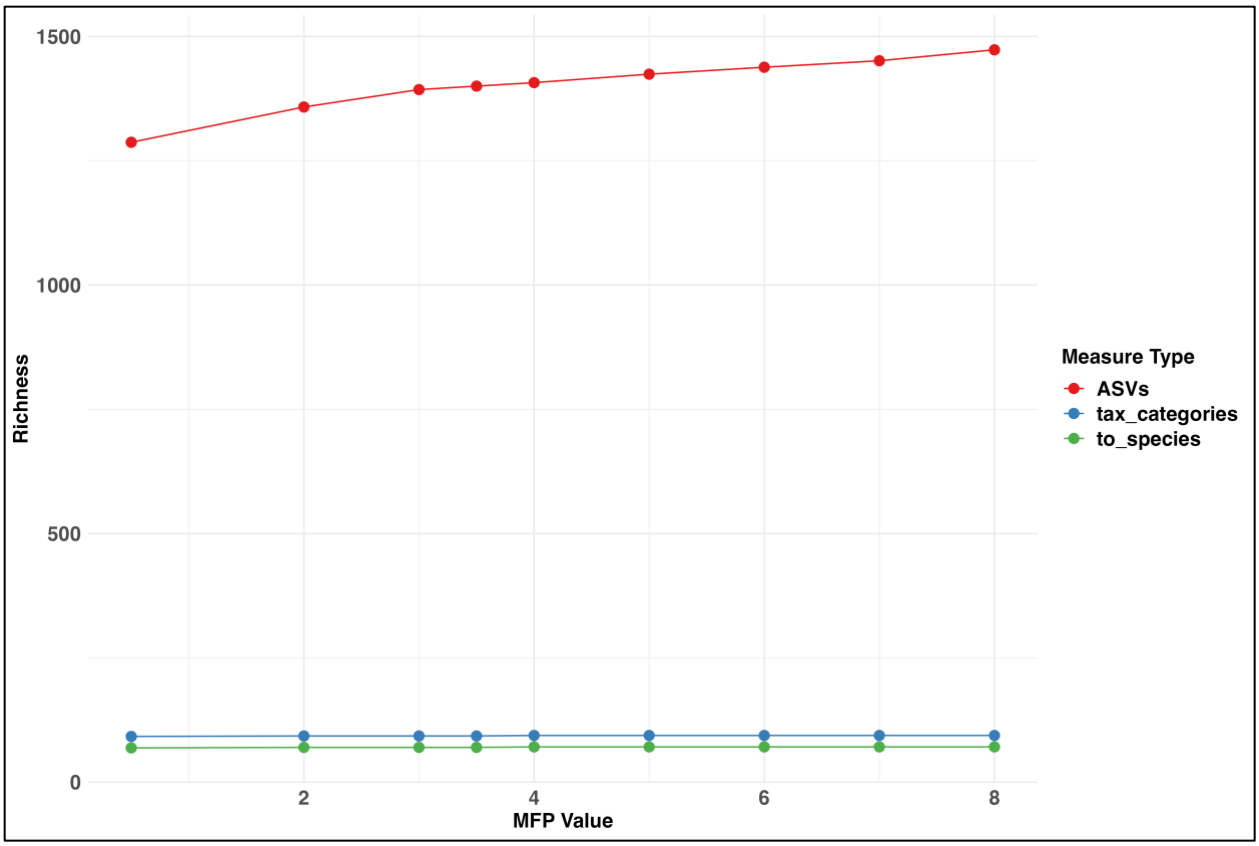

Figure 2.4. Total number of Illumina ASVs, total number of taxonomic categories, and number of species-level classifications (at 70% confidence) among all BDCR samples as stringency of chimera removal increases from an mfp=0.5 to mfp=8.0.

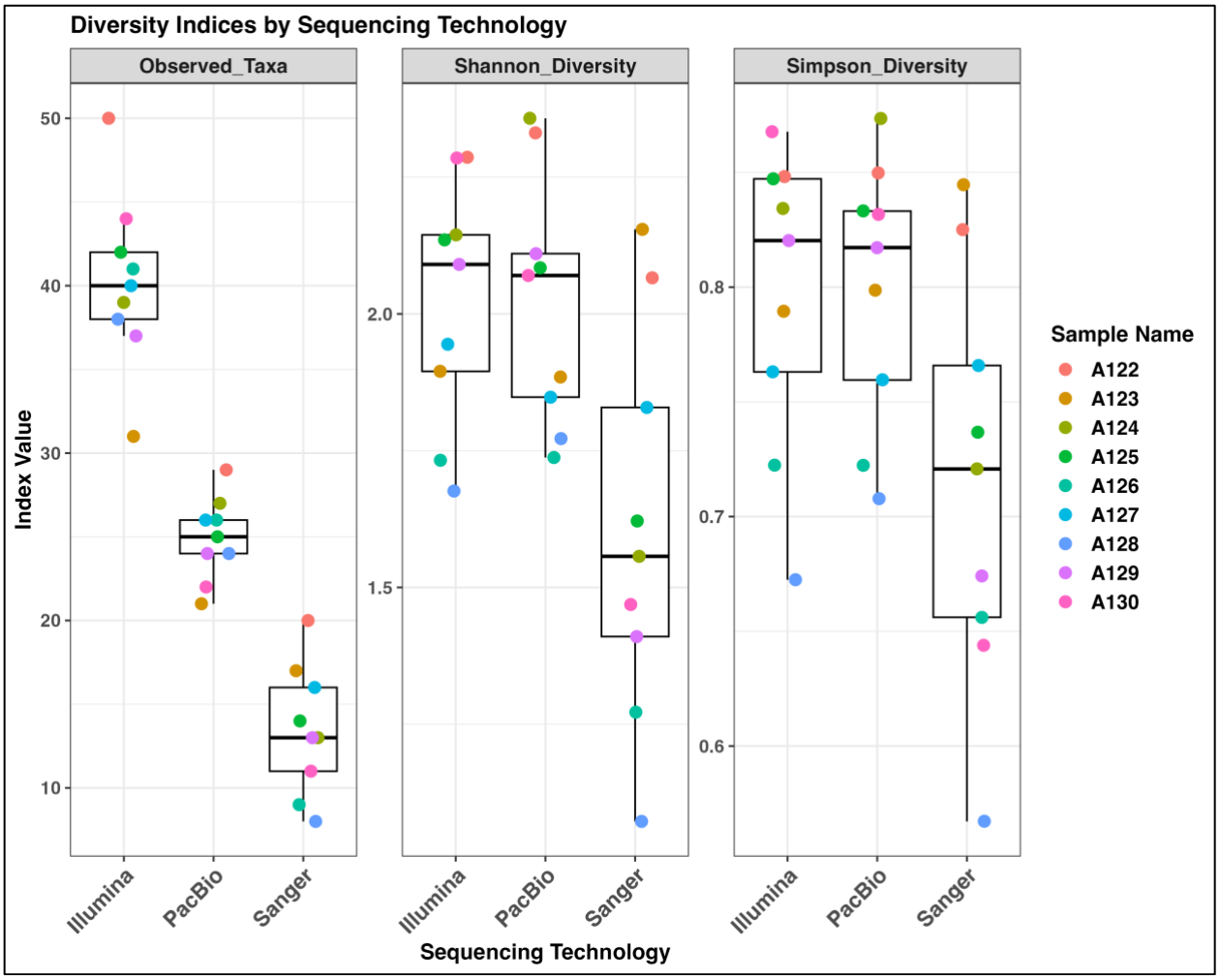

Figure 2.5. Boxplots showing estimated number of observed taxa (**A**), estimated Shannon diversity (**B**), estimated Simpson diversity (**C**) of nematode communities in BDCR soil samples.

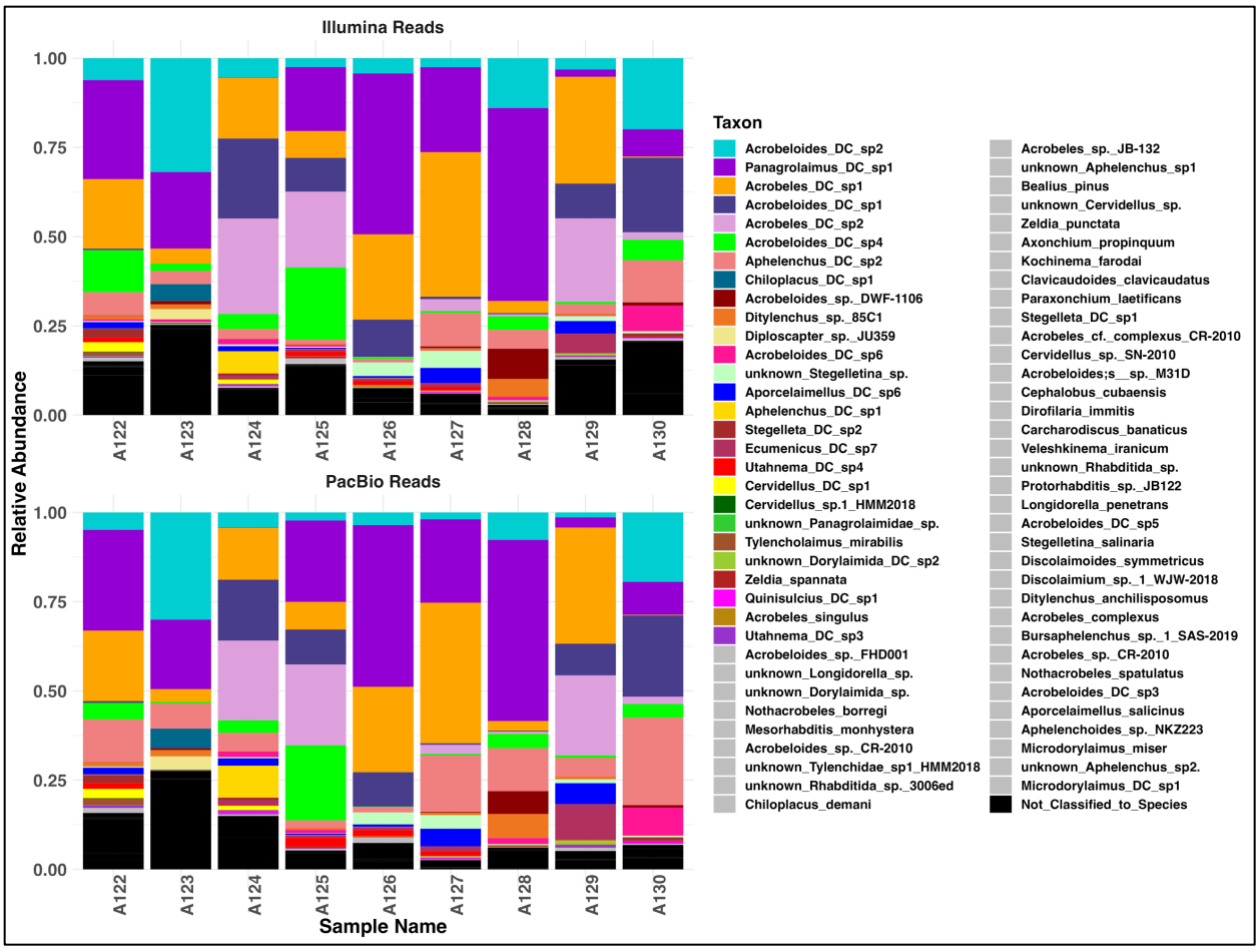

Figure 2.6. Stacked bar plot of Illumina and PacBio sequence classifications to species level at 70% confidence, by BDCR soil sample. Species with relative abundance less than 0.05% are colored grey. Taxonomic categories that could not be classified to species level are colored black.

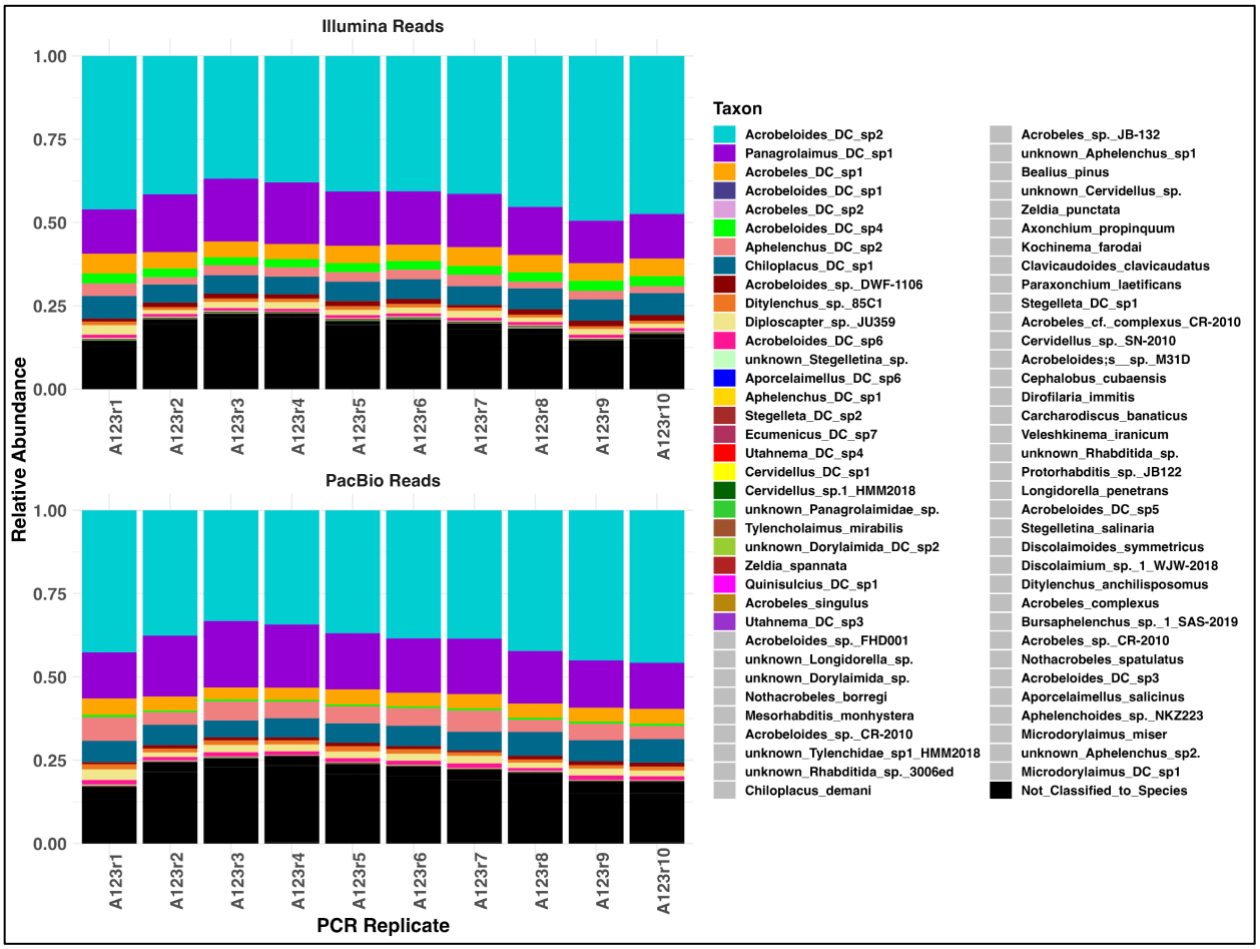

Figure 2.7. Stacked bar plot comparing 270bp Illumina and PacBio sequence classifications to species level at 70% confidence, for ten PCR replicates. Species with relative abundance less than 0.05% are colored grey. Taxonomic categories that could not be classified to species level are colored black.

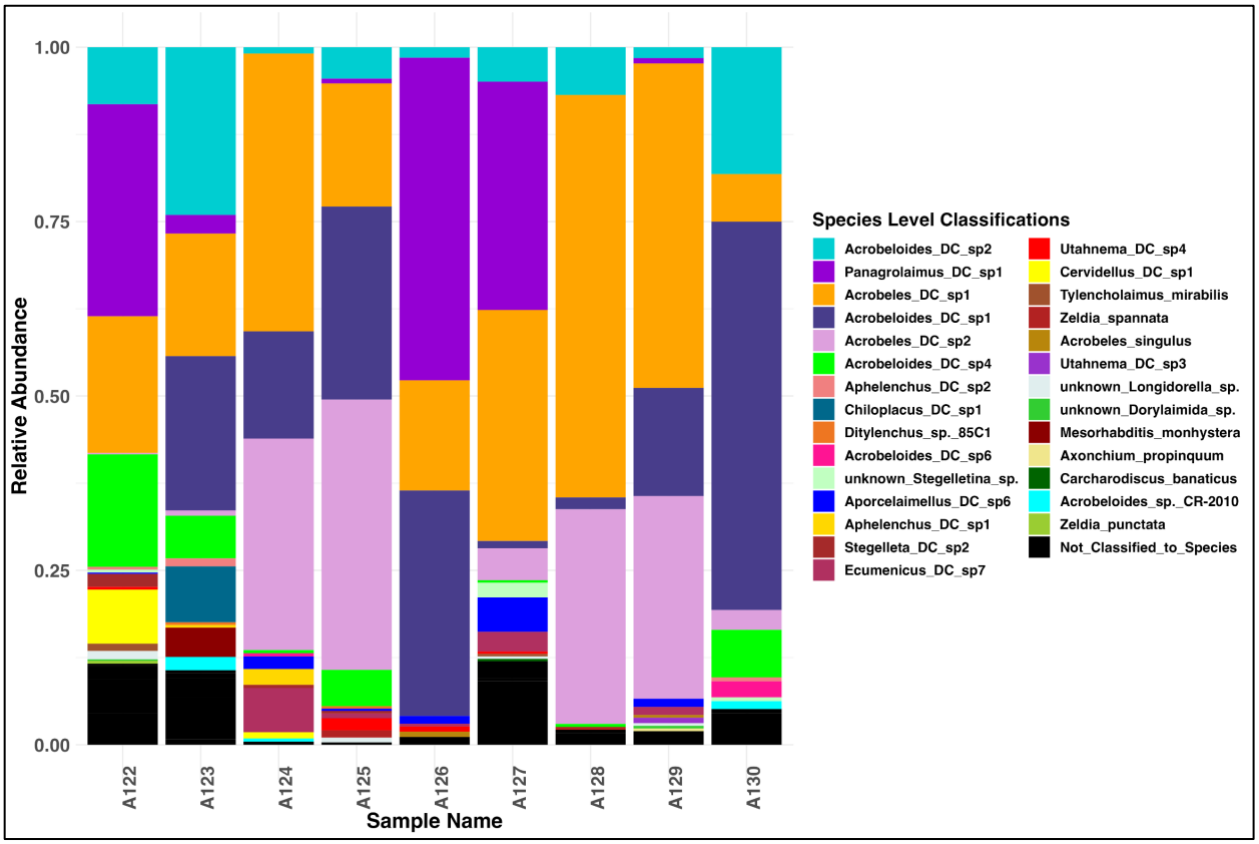

Figure 2.8. Stacked bar plot of 270bp Sanger sequence classifications to species level at 70% confidence, by soil sample. Taxonomic categories that could not be classified to species level are colored black.

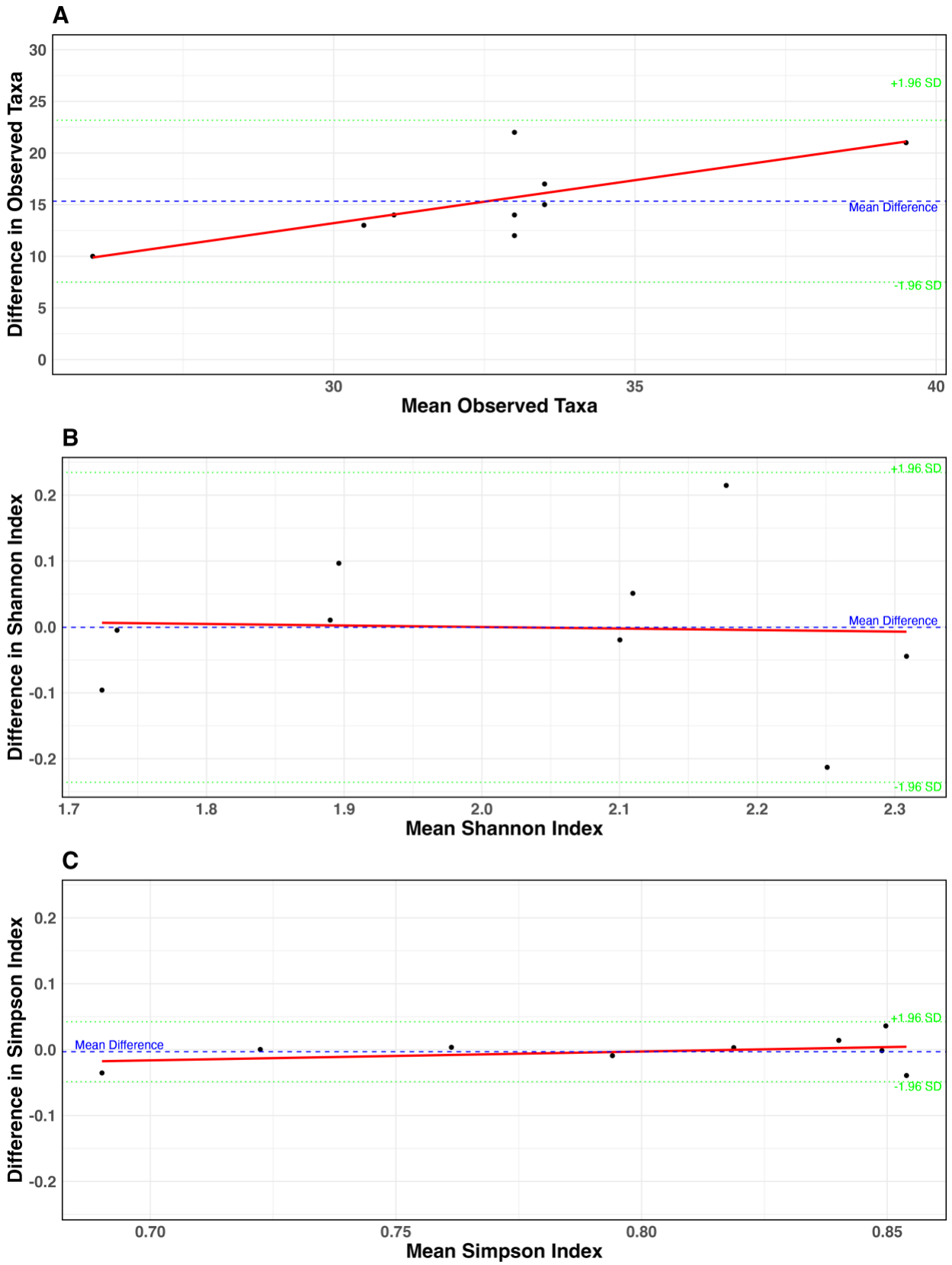

Figure 2.9. Bland-Altman plots depicting the agreement between Illumina and PacBio

sequencing technologies when estimating observed taxa (A), Shannon diversity (B), Simpson diversity (C) of nematode communities in BDCR soil samples. Blue dashed lines indicate the mean bias between Illumina and PacBio sequencing technologies across 10 PCR replicates. Means above zero indicate overestimates of species relative abundance by Illumina technology relative to PacBio, while means below zero indicate underestimation. Green dotted lines represent the upper (+1.96 SD) and lower (-1.96 SD) limits of agreement. The slopes of the lines of best fit (colored red) indicate the magnitude of proportional bias between technologies.

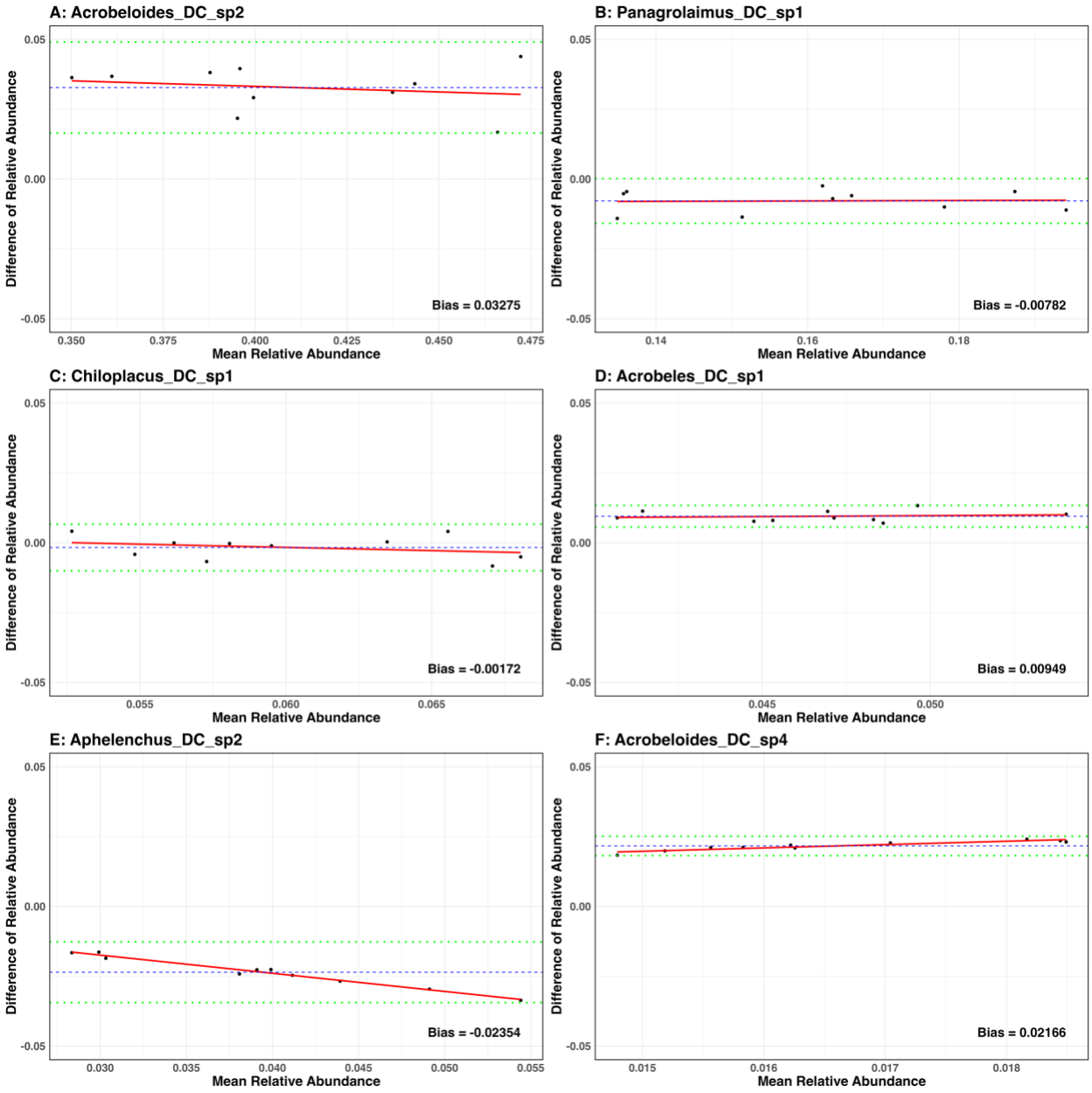

Figure 2.10. Bland-Altman plots depicting the agreement between Illumina and PacBio sequencing technologies when estimating the relative abundance of the 6 most common nematode species in BDCR soil samples. Blue dashed lines indicate the mean bias between Illumina and PacBio sequencing technologies across 10 PCR replicates. Means above zero indicate overestimates of species relative abundance by Illumina technology relative to PacBio, while means below zero indicate underestimation. Green dotted lines represent the upper (+1.96 SD) and lower (-1.96 SD) limits of agreement. The slopes of the lines of best fit (colored red) indicate the direction and magnitude of proportional bias between technologies.

Chapter 3: Nematode Community Structure in the Rhizospheres of Southern California Creosote (*Larrea tridentata*) and the Role of Soil Physicochemical Properties

Christopher A. Pagan[1] and Steven A. Nadler[1].

[1]Dept. of Entomology and Nematology, University of California, Davis, CA, U.S.A

3.1 Introduction

Nematodes are highly diverse and overwhelmingly abundant across marine, freshwater, and terrestrial biomes. With an estimated million species worldwide (Hugot et al., 2001), their small size and subtle morphological differences demand extensive time and specialized training for accurate identification and description. The increasing use of molecular tools in taxonomy (Ahmed et al., 2015) has revealed that some 'diagnostic' morphological traits are evolutionarily convergent or developmentally plastic (Smythe and Nadler, 2006), leading to a lack of monophyly in many taxa (Blaxter et al., 1998; De Ley and Blaxter, 2004). The immense diversity of nematodes, combined with these taxonomic challenges, makes cataloging and sampling species to determine their geographic distribution particularly difficult. As a result, very few biomes have been thoroughly sampled for nematode diversity (Wharton et al., 1989; Gattoni et al., 2022; Powers et al., 2021), hindering our understanding of their biogeography and distribution patterns.

High-throughput sequencing (HTS) technologies alleviate some of these challenges by enabling the parallel processing of hundreds of barcoded samples and, and thereby facilitating the analysis of complex biological communities. These technologies offer increased throughput, lower costs per sample, and heightened sensitivity. A significant drawback of HTS is the unpredictable relationship between species abundance and sequence read abundance, particularly for multicellular organisms, which makes relative abundance estimates unreliable (see **Chapter 2** intro). This current limitation affects the calculation of diversity indices and necessitates the use of indices that consider only the presence or absence of species. Despite these limitations, HTS remains a powerful tool for detecting species richness and is a valuable tool for biodiversity inventories.

PacBio sequencing has recently emerged as an ideal technology for amplicon sequencing and community analyses. It supports larger amplicons (>550bp), which were previously difficult to use in Illumina-based studies. The PacBio Sequel 8M platform can generate several million reads from these larger amplicons and the newer PacBio Revio platform produces tens of millions of reads. PacBio's larger read size enables the use of a wider selection of metabarcoding loci, with the main limiting factors being taxonomic coverage of primer pairs and ease of PCR amplification.

Additionally, the negative effects of non-target taxon amplification are mitigated by a higher sequence capacity. The limited capacity of older platforms meant that amplification of any non-target sequences would competitively reduce the number of target sequences in the final

dataset, thereby risking inadequate representation of target species. With current PacBio sequencing capacities (8M and Revio), it is possible to use metabarcoding primers with wider taxonomic coverage without concern for amplification of nontarget species. These additional sequences may even provide insights into ecological interactions, enhancing the overall utility of the dataset.

Formerly, most nematode species were thought to be cosmopolitan in their distributions. This assumption was based on their small size, high numerical abundance, and capacity for passive dispersal over wide geographic ranges (Ptatscheck et al., 2018; Ptatscheck and Traunsperger, 2020). More recent studies, however, suggest that endemism is more frequent among nematode taxa than previously thought (Zullini, 2018). Soil nematode abundance tends to be highest in temperate regions (30° to 55° latitude) and decreases toward the poles and the equator, with temperate coniferous forests exhibiting the highest mean nematode abundance (Song et al., 2017).

However, the relationship between nematode taxonomic richness and latitude is more complex and less consistent across studies. Some research indicates that nematode taxonomic richness does not follow the same clear latitudinal gradient observed in their abundance. While temperate coniferous forests may exhibit the highest mean nematode abundance, species richness does not necessarily peak in these regions. Contradictory findings exist regarding taxonomic richness, with some researchers estimating higher species-equivalent units near the

tropics (Porazinska et al., 2010) and others finding the highest genus-level richness in temperate deciduous forests (Song et al., 2017).

In desert ecosystems, nematode abundance and diversity are generally lower compared to other biomes. Desert soils, characterized by extreme temperatures, low moisture, and limited organic matter, present challenging conditions for nematode survival and reproduction. Despite these harsh conditions, certain nematode species have adapted to thrive in desert environments, often exhibiting specific physiological and behavioral adaptations to cope with arid conditions (Freckman, 1978). The presence of perennial desert plants, such as creosote bushes, can create microhabitats that support higher nematode densities and diversity (Freckman and Mankau, 1977) by depositing organic matter, concentrating moisture, releasing root exudates, and otherwise restructuring the surrounding soil (Pen-Mouratov, et al., 2003).

Soil physicochemical properties also significantly influence nematode abundance and community composition. Soil moisture, organic matter, pH, cation exchange capacity (CEC), and soil texture (percent sand, silt, and clay) are key factors affecting nematode populations. Soil moisture is critical, as nematodes are essentially aquatic organisms inhabiting wet substrates (Xiong et al., 2018). Organic matter serves as a primary food source, correlates with moisture retention, and influences soil's ability to hold and exchange ions (Solhenius, 1973). Soil pH affects the solubility of organic molecules and the bioavailability of nutrients and metals, which can be either beneficial or toxic to nematodes (Neina, 2019; Andersson et al., 2000). High CEC soils, which buffer pH more effectively and retain more moisture, tend to support higher

nematode abundances (Van Den Hoogen et al., 2019). Soil texture (percent sand, silt, and clay) affects the physical habitat of nematodes by influencing porosity, moisture retention, drainage, and aeration (Wallace, 1968; Kung et al., 1990).

Heavy metals such as copper (Cu), iron (Fe), manganese (Mn), and zinc (Zn) are essential micronutrients for plant and microbial metabolism, which can indirectly benefit nematode communities (Adamczyk-Szabela and Wolf, 2022; Nimnoi et al., 2024). However, soils with high concentrations of these metals can be considered polluted, leading to adverse effects on soil organisms. Nematodes may be affected by metal concentrations either directly through toxicity or indirectly through effects on their predators or food sources (Korthals et al., 1996b). Sensitivity to heavy metals varies among species of different functional guilds and life history strategies, with k-selected species generally being more sensitive than r-selected species (Georgieva et al., 2002; Korthals et al., 1996a; Ghanem et al., 2024; Garcia et al., 2022). The bioavailability of heavy metals is influenced by soil parameters such as pH, organic matter, and clay content (Dinic et al., 2019). Binding of metal ions by clay or dissolved organic carbon in soils is pH-dependent, with metal ions being more mobile and bioavailable in lower pH soils.

For this study, we collected 6 soil samples from 3 creosote bushes (2 samples per creosote bush) from 11 locations in Southern California, for a total of 66 soil samples. We employed a PacBio-based metabarcoding strategy with a ~750bp segment of the D1-D2 region of 28S ribosomal DNA, which has previously provided species-level resolution for approximately 85% of the BDCR nematode specimens in **Chapter 1**. This locus was amplified from pooled

nematode extracts and sequenced using the PacBio Sequel 8M platform. Additionally, we analyzed 14 soil physicochemical parameters. This study aims to accomplish the following:

1. Test the utility of our custom D1-D2 28S reference sequence database (**Chapter 1**) outside of Boyd Deep Canyon Reserve BDCR by conducting a biodiversity inventory of similar habitats in the deserts of Southern California.
2. Generate data on the geographic distribution of nematode species in the deserts of Southern California, enriching reference sequence databases and providing new prospective locations for future ecological studies.
3. Determine whether geographic location or soil physicochemical parameters are more predictive of nematode abundance and community composition.

3.2 Materials and Methods:

3.2.1 Soil sample collection

Samples were collected from beneath creosote bushes at 11 collection localities in the Sonoran and Mojave Deserts, ranging from -67m to 1170m elevation (**Table 3.1**). At each collection locality, three bushes were selected. At each selected bush, a ~9lb soil sample was taken from the north side and south side of the bush. Prior to sampling ~2cm of material was brushed away to remove cactus needles, thorns, and other dry plant matter. Samples were enclosed in plastic bags, transported back to UC Davis in coolers, and stored in a cold room at 10C. Samples were gently mixed to reduce spatial heterogeneity of nematodes in the soil.

3.2.2 Soil sample analyses

Soil samples were passed through a 60 mesh (0.25mm) sieve and submitted to the UC Davis Analytical Laboratory where the following analyses were done: Total C & N, pH, cation exchange capacity (CEC), total organic matter & organic carbon (Walkley-Black method), moisture retention (1 atm), and soil particle size analysis, as well as total Zn, Mn, Fe, and Cu.

3.2.3 Sample hydration

Approximately 850 cm^3 of soil was placed in 12 3oz paper cups that were punctured with a metal probe to make drainage holes. Eight rows of four holes were evenly spaced around the sides of the cups, and seven holes were punched in the bottom of each cup. Each of the 12 cups were filled to ~80% capacity with soil. The soil was then fully saturated with tap water, allowed to drain, and then incubated at 25C for 72 hours.

3.2.4 Nematode isolation

After hydration and incubation, soil samples were placed on four 6" Baermann funnels (~9oz per, 32x47cm Kimwipes, single-ply) at 28.5C in an incubator. Nematodes were collected from the funnels after 24 and 48 hours. The 24 hour and 48-hour nematode samples were counted separately before the water exchange protocol detailed below.

3.2.5 Water exchanges

Funnel catches (~40 ml) from each funnel were collected in a 250 ml beaker. The volume of liquid was brought to ~150 ml, and the beakers were placed on ice to encourage settling of the nematodes. After 30 minutes, the top 100 ml of liquid was aspirated off using an aspirator with a 5 µm spin filter (EMD Millipore) placed over the tip to avoid aspirating any nematodes. This process was repeated twice more, or until the liquid in the beaker was colorless.

3.2.6 Nematode counting

After the water exchange procedure, samples were decanted in a 100mm a square petri dish with a 6x6 grid. Under a stereoscope, all nematodes were counted in 6 randomly selected cells of the grid. A random number generator was used to select the X and Y coordinates (1-6) of cells. Counts were averaged over 6 cells and a total estimated worm count was extrapolated over the total area of the petri dish. After counting, nematodes from each sample were decanted into a 50ml polystyrene tube (Falcon) and spun at 1800rpm in a clinical centrifuge. Centrifugation and aspiration were used to pellet worms and reduce the liquid volume each sample until they could be transferred to a 0.5 ml microcentrifuge tube. The tubes were then centrifuged, and the supernatant was reduced again until the pellet of worms was barely submerged. The tubes were stored at -20C for future use.

3.2.7 DNA extraction and PCR

Pooled nematode samples were digested, using a modified PrepGEM protocol. The 0.5ml tubes containing frozen, pelleted nematodes were resuspended in 86ul of water. The following were added: 10ul of prepGEM orange buffer, 2ul of beta-mercaptoethanol, and 2ul of prepGEM enzyme. The samples were incubated at 75C for one hour, vortexed for ten seconds, briefly centrifuged, and incubated for another hour at 75C. The prepGem enzyme was then heat-killed at 95C for 8 minutes, and the samples were centrifuged at 5000 x g to pellet debris.

Supernatant from the prepGEM digests was transferred into 500ul of DNAzol reagent (Molecular Research Center, Cincinnati, Ohio) in a 1.5ml tube. DNA was precipitated with 250ul of 100% EtOH and 5ul of polyacryl carrier. Pellets were washed twice with 75% EtOH, and dried. Pellets were resuspended in 20-30ul of TE pH 8.0, depending on the number of nematodes that were digested.

Two-step PCR reactions were used to amplify a ~750bp piece of 5' 28S rDNA using the KOD XL polymerase kit. PCR reactions included 1.5ul of DNA extract, 3mM $MgCl_2$, 0.2mM of each dNTP, 0.5 units of KOD XL Polymerase, and 0.6 uM each of primers 391F and 799R (**Chapter 1**). Cycling conditions consisted of an initial melting step of 95C for 3 minutes, followed by 35 cycles of 95C for 30 seconds and 67 C for 1 minute, and then a final extension step at 72C for 6 minutes.

3.2.3 Library Preparation and Sequencing

Pacbio indexed (bc1001-bc1024) versions of primers 391F and 799R were ordered from Integrated DNA Technologies (San Diego, CA). The IDT Oligo Analyzer Tool (https://www.idtdna.com/calc/analyzer) was used to pair forward and reverse barcoded primers that had minimal complementarity (Δg > -6kcal/mol). Barcoded PCR products were generated in 50ul reactions using the KAPA Hifi Hotstart polymerase kit. Reactions included 3mM $MgCl_2$, 0.2mM of each dNTP, 1 unit of KAPA Hifi Hotstart polymerase, 1.2uM of forward and reverse barcoded primers, and the templates for these reactions was 1.5ul of 1/1000th dilution PCR products from each soil nematode sample. Cycling conditions consisted of an initial melting step at 95C followed by 24 cycles of 98C for 20 sec, 69C for 15 sec, 72C for 45 sec, and then a final extension step at 72C for 1 minute. SPRI beads (Byodynami) were used to clean 40ul of PCR product according to the manufacturer's instructions, and products were resuspended in 10ul of Tris-HCl pH 8.0. The DNA concentrations of all barcoded, bead-cleaned PCR products were measured using Qubit, made equimolar, and then pooled. The sequencing library was submitted to the QB3 sequencing facility at The University of California, Berkeley. There, Pacbio SmartBell adapters were added, and the library was sequenced on the Sequel 8M platform.

3.2.4 Bioinformatics

Circular consensus sequences were created from raw subread files using ccs, a PacBio proprietary program, with a min-passes setting of 20. Sequence data were demultiplexed using PacBio's proprietary software, lima, using a min-score setting of 80, min-end-score setting of 50, and min-ref-span setting of 0.75. Demultiplexed sequences were imported into the Qiime2

bioinformatic platform (Boleyn et al., 2019) and denoised using the DADA2 plugin (Callahan et al., 2016), with the default min-fold-over-parent-abundance setting of 3.5. The resulting ASVs were classified to a 70% level of confidence using the sklearn naïve Bayesian classifier. The filter table function was then used to select only features that were classified to Class *Enoplea* and Class *Chromadorea* with 70% confidence. Samples 244B and 252B were lost because of an error during library construction and their resulting sequence data are not included in the results below.

3.2.5 Data analyses

Statistical analyses were performed, and accompanying figures were produced in R (4.2.3). Classifier results were tallied in terms of read counts, ASV counts, and count of taxonomic categories. Taxonomic categories are defined as the lowest taxonomic level to which any ASV or group of ASVs are successfully classified at a confidence threshold of 70%. A stacked bar plot was generated to evaluate classifier performance by comparing the proportion of reads assigned to each taxonomic level at each collection location.

3.2.6 Sample richness, distribution of taxa, and trophic diversity

Richness was calculated for all samples, including total number of taxonomic categories classified, number of species-level classifications, and number of ASVs. Boxplots were generated to visualize multiple richness metrics for all samples and locations, and the iNext

package (Chao and Jost, 2102; Hsieh and Chao, 2016) was used to make ASV accumulation curves for all locations to assess completeness of sampling. Counts of taxa from all samples were aggregated by location, and a stacked bar plot was generated to visualize relative abundance of all taxa (by read count) among all locations.

The CompleUpset package was used to generate upset plots (Lex et al., 2014) depicting prevalence and distribution of ASVs and of species-level classifications among collection locations. To assess functional diversity of the nematode communities in our samples, trophic groups were assigned to OTUs based on their genus-level classifications (Bongers and Bongers, 1998). A pie chart was then generated to visualize proportions of trophic group assignments among all chromadorean and enoplean sequence reads. Stacked barplots were generated to visualize relative abundance of trophic groups among collection locations.

3.2.7 Evaluation of classifier success

Biogeographical data were combined with phylogenetic methods to explore unrecognized diversity among our ASVs and identify areas where our reference sequence database could be improved, particularly by increasing the number and diversity of sequences for underrepresented taxonomic groups. As a proof of concept, we focused on three putative species with the greatest number of ASVs: *Acrobeles DC sp2*, *Panagrolaimus DC sp1*, and *Robustodorus arachidis*.

Maximum likelihood trees were generated to investigate phylogenetic relationships among ASVs from these species, with alignments performed using MAFFT v7.525 (Algorithm: E-INS-i). Sequences from 181 ASVs classified as *Acrobeles DC sp2* and 190 ASVs classified as *Panagrolaimus DC sp1* were aligned together with all Family Cephalobidae and Panagrolaimidae reference sequences from our custom D1-D2 database, with each family functioning as an outgroup for the other. Sequences from 271 ASVs classified as *Robustodorus arachidis* were aligned separately with all Family Aphelenchoididae reference sequences from the custom database. After phylogenetic analysis, biogeographic data were added to determine which ASV groups were both sympatric and reciprocally monophyletic, indicating that they likely belong to distinct species. We also looked for evidence of location preference among monophyletic groups of ASVs. Phylogenetic trees were generated using IQ-Tree (Minh et al., 2020) and visualized with FigTree (Rambaut, 2010).

3.2.8 Effects of soil parameters on community composition

A boxplot was generated to visualize distributions of soil physicochemical measurements within and among sites. Some measurements were below the limits of tests conducted by the UC Davis Analytical Laboratory. CEC values reported as "<2.0" were set to 1.0. Silt % values reported as "<1.0" were set to 0.5. Total N measurements reported as "<0.020" were set to 0.01.

There are several inherent challenges in analyzing our high-throughput sequencing (HTS) data, particularly when assessing the effects of soil properties on nematode community composition and abundance. One major issue is that HTS technologies can be misleading with respect to read counts, as the mathematical relationship between the relative abundance of reads from species and their actual abundance in the original sample is unclear. This limitation necessitates the use of diversity indices that do not rely on relative abundance, such as those that handle binary (presence/absence) data, including Jaccard and Sørensen indices.

The multidimensional nature of our data poses additional challenges. Depending on the richness metric used, there are 2319 ASVs, 92 taxonomic categories, or 62 putative species comprising the biological community data. Additionally, the dataset includes 14 soil physicochemical variables with diverse distributions and modalities, some of which may require normalization. With only 64 useable samples, the ratio of variables to samples raises concerns about heteroscedasticity. Due to these concerns, our methods included dimension reduction steps to make subsequent analyses tractable.

The fossil package in R was used to calculate the pairwise Jaccard and Sørensen distances among nematode communities in all soil samples. These diversity indices are calculated similarly, but the Jaccard index gives equal weight to all taxa while Sørensen gives greater weight to shared taxa.

Non-metric multidimensional scaling (NMDS) analyses were done, using Jaccard and Sørensen distances, to visualize dissimilarity of nematode communities among soil samples. These analyses were performed with the vegan package in R. Environmental vectors (from soil physicochemical data) were fitted to the NMDS results to visualize correlations between community composition and soil parameters.

A principal component analysis (PCA) was performed using 14 soil physicochemical properties to reduce the dimensionality of the data and identify factors that explain the most variance in the data. Loadings were calculated for PCA 1, 2, and 3 for each soil parameter. Each parameter was assigned a weighted composite importance score, which was calculated as: $\sqrt{0.452(PC1)^2 + 0.190(PC2)^2 + 0.146(PC3)^2}$. Soil parameters were sorted into 3 categories of relative significance (high, medium, and low) by k-means clustering of their weighted composite importance scores.

Two beta regression generalized linear mixed models (GLMMs) were then created, using Jaccard and Sørensen distance between samples as response variables and the first three principal coordinates from soil physiochemical data as predictor variables, with a random effect grouping by location. Loadings were calculated and used to assess relative contributions of each soil physicochemical property to the overall variance among samples. The reason both Jaccard and Sørensen are used is that dissimilarity can be affected by low abundance, high diversity taxa, which might be present in some samples and not others because their odds of being detected are very low.

A Spearman's rank correlation test was done, using the cor.test function in R, to measure correlations between the 14 measured soil physicochemical properties and the estimated abundance of nematodes in each soil sample.

3.3. Results

3.3.1 Classification results

Out of 3,438,110 reads passing all filters, 59.16% were classified within Class Enoplea or Chromadorea (**Table 3.2**). Among these, 0.14% of reads could not be classified below class level, 2.77% were classified to order level but no lower, 4.30% were classified to family level but no lower, 9.77% were classified to genus level but no lower, and 83.00% were classified to species level.

There were 4092 ASVs assigned to all sequence reads by the DADA2 plugin in Qiime2, 2319 of which were assigned to Classes Enoplea and Chromadorea. Of these, 16 ASVs could not be classified below class level, 391 ASVs were classified to order but no lower, 306 ASVs were classified to family but no lower, 158 ASVs were classified to genus level but no lower, and 1448 ASVs were classified to species level.

ASVs were classified into 92 unique taxonomic categories, 62 of which were to species level at a confidence threshold of 70%. One taxonomic category, unclassified Chromadorea consisted of ASVs that could not be classified below class level. There were also 10 taxonomic categories that could not be classified below family level and 17 that could not be classified below genus level. Chromadorean outnumbered Enoplean taxonomic categories, 68 to 24.

The proportions of reads classified to species, genus, family, order, and class at each location are shown in **Figure 3.1**. A pie chart, depicting family-level diversity of reads among all samples (**Figure 3.2**) shows that Families Cephalobidae and Panagrolaimidae have the greatest relative abundance with 57.4% and 21.7% respectively. The next most abundant families are Leptonchidae (4.2%), Aphelenchidae (3.7%), and Anguinidae (2.9%). Families Qudsianematidae and Aphelenchoididae are both present at a relative abundance of 2.0% and Telotylenchidae at 1.1%. All remaining families are present at relative abundance under 0.5%.

3.3.2 Sample richness, distribution of taxa, and trophic diversity

Estimated nematode abundance and three different richness estimates are organized by location and shown in boxplots (**Figure 3.3**). These richness estimates include the number of ASVs, total taxonomic categories, and species-level classifications. MNP-KB, SB-N, and SB-W exhibited the highest average number of ASVs per sample, while MNP-HH had the lowest. DC-AH had the highest average number of taxonomic categories and species-level classifications per sample, whereas MNP-HH and DV-EC had the lowest in these respective categories. SB-W

had the highest average estimated number of nematodes per sample, with MNP-HH and MNP-KB having the lowest. Notably, one sample, 240B from DV-CC, was an extreme outlier with an estimated 18,634 nematodes, which is 4.43 standard deviations above the mean for all samples.

Rarefaction curves generated from ASV count data, aggregated by location, are shown in **Figure 3.4**. All samples reach a point of ASV saturation within the limits of their read counts. MNP-KB and SB-N both reach asymptotes at over 400 total ASVs across all their constituent samples. DV-CC has the lowest estimated ASV total at 141 ASVs.

The stacked barplot (**Figure 3.5**) depicts the relative abundance of species-level taxonomic classifications among all locations. The three most numerically abundant putative species (by read count) are *Acrobeles* DC sp2, *Panagrolaimus* DC sp1, and *Acrobeloides* DC sp1. All three of these putative species are classified to reference taxa that were discovered during the reverse taxonomy experiment in **Chapter 1**. Although *Acrobeles* DC sp2 is the dominant putative species at most locations, it is all but absent in the collection locations in Death Valley national Park (DV-CC and DV-EC), where *Panagrolaimus* DC sp1 is overwhelmingly dominant. In the Mojave National Preserve samples (MNP-HH and MNP-KB), *Acrobeles* DC sp2, *Panagrolaimus* DC sp1, and *Acrobeloides* DC sp1 are present, but they are all outnumbered by reads classified to the NCBI reference sequence: *Cervidellus* sp. 1 HMM2018. In the Twentynine Palms location (TNP), there was a much higher representation of *Utahnema* DC sp4 and *Ditylenchus* reads that could not be classified to species.

The upset plot of ASV distribution over collection locations (**Figure 3.6**) shows that the vast majority of ASVs (1218) occur at single locations, while just 3 ASVs occur at 9 or more locations. The upset plot depicting the distribution of species-level classifications across locations (**Figure 3.7**) shows that 11 putative species are found at just one location, while 13 putative species can be found at 9 or more locations.

In **Figure 3.8**, a pie chart depicting the diversity of trophic groups among all sequence reads, we see that microbivores are overwhelmingly dominant at a relative abundance of 81.5%. The next most abundant trophic groups are omnivore-predators and fungivores, at 6.7% and 5.9% respectively. Curiously, 0.20% of reads (9 ASVs), distributed over eight locations, are classified as the animal parasite *Dirofilaria immitis* at lower confidence scores (72%-87%). Reads that could not be classified to a taxonomic level where trophic group could be inferred comprised 1.5% of the total.

A stacked barplot of trophic diversity among collection locations (**Figure 3.9**) shows that microbivores were least dominant at the TNP location, as this was the location with the highest abundance of plant parasites, omnivore-predators, and fungivores. The MNP-KB location also had a relatively high abundance of omnivore-predators. Putative animal parasite reads were more abundant at the SB-N location compared to other locations. The DV samples had a higher representation of microbivores, with some fungivores, and a greatly reduced representation of all other trophic groups.

3.3.4 Soil measurements and nematode abundance estimates

The boxplot in **Figure 3.10** shows distributions of soil parameters over all 11 locations. Complete soil physiochemical data are reported in **supplemental materials**. Average soil parameter measurements for each location are reported in **Table 3.3** along with average nematode abundance per sample at each location. The abundance of organisms varies widely across locations, with SB-W and DV-CC showing the highest averages, while MNP-HH and MNP-KB have significantly lower abundances The soils in the DV-CC and ABW locations have the highest sand content, contributing to lower water retention (H2O cap). In contrast, MNP-KB and MNP-HH locations have higher clay and silt content, which likely enhances their water-holding capacity and cation exchange capacity (CEC). These locations also feature higher levels of organic carbon and organic matter. Locations like DV-EC and DC-AH have high concentrations of Fe, Mn, and Zn. In contrast, locations like ABW and TNP show lower concentrations of these micronutrients. The pH across locations varies slightly, with DC-LC and ABW showing more alkaline conditions, while MNP-HH and DC-AH have slightly more acidic soils.

3.3.5 Relationship of soil parameters to nematode community composition

The Jaccard NMDS plot (**Figure 3.11**) shows distinct clustering of nematode community samples by location, with some locations demonstrating very tight groupings. Notably, the Death Valley samples (DV-CC and DV-EC) are tightly clustered, indicating a high degree of similarity based on

species presence or absence. The Twenty-nine Palms (TNP) and Steele-Burnand West (SB-W) samples also form tight clusters, suggesting consistent community composition within these locations. However, there are a few outliers, such as one Mojave Hidden Hills Road sample (MNP-HH, brown), which is positioned in the bottom left, away from other samples from the same location. Samples from Deep Canyon (DC-AF, DC-LC, DC-AH) and Agave Hill (DC-AH) are more spread out, suggesting greater variability within these localities.

When comparing the Sørensen NMDS plot (**Figure 3.12**) to the Jaccard plot, a general trend of tighter clustering within locations is observed. The previous outlier Mojave Hidden Hills Road sample (MNP-HH, brown) now clusters closely with other Mojave National Preserve samples, indicating that this sample has a community composition more aligned with its peers when considering the shared presence of species. In contrast, the Death Valley samples, which were tightly clustered in the Jaccard plot, are more spread out along the NMDS1 axis in the Sørensen plot, though still tightly clustered along the second axis.

3.3.6 Correlation of soil parameters with nematode abundance

The Spearman rank correlation analysis revealed significant relationships between nematode abundance and several soil properties, with both positive and negative correlations observed (**Table 3.4**). Percent sand had a moderately positive correlation with nematode abundance (r_s = 0.521, $p < 0.001$), while Mn (r_s = 0.322, $p = 0.010$) and Zn (r_s = 0.294, $p = 0.018$) showed weak positive correlations. Percent clay (r_s = -0.521, $p < 0.001$), percent silt (r_s = -0.483, $p < 0.001$),

and water-holding capacity (r_s = -0.428, p = 0.000) had moderate negative correlations with nematode abundance, and CEC showed a weak negative correlation (r_s = -0.293, p = 0.019). In contrast, organic matter, organic carbon, total carbon, pH, N, Fe, and Cu did not show statistically significant correlations with nematode abundance (p > 0.05 for all).

3.3.7 Variation of soil properties among soil samples and collection locations

The first three principal components explained 45.2%, 19.0%, and 14.6% of the variation, respectively. Loadings are detailed in **Table 3.5**. PC1 is primarily influenced by organic carbon, organic matter, nitrogen, and water-holding capacity. Sand shows a relatively strong negative loading on PC1, suggesting an inverse relationship with these other properties. PC2 is more influenced by micronutrients, particularly zinc and iron, both of which show relatively high positive loadings. PC3 is characterized by positive loadings from cation exchange capacity, clay, pH and silt, with negative contributions from organic carbon and organic matter, indicating that PC3 is associated with soil texture and its related chemical properties. As shown by the k-means clustering of weighted composite importance scores in **Table 3.5**, 12 soil parameters had similar overall influence on the variation captured in the first three principal components, while two parameters, pH and Cu, had less explanatory power overall.

3.3.8 Influence of soil parameters on nematode community composition

The beta regression GLMM analysis results (**Table 3.6**) show that PC3 has the strongest effect on dissimilarity metrics in both models (Jaccard β= -0.202, Sørensen β= -0.206), with PC1 and PC2 contributing smaller but significant effects. The Jaccard model suggests a lower baseline dissimilarity (intercept = -0.422) compared to the Sørensen model (intercept = 0.223). Both models fit the data well (Jaccard AIC = -7617.3, Sørensen AIC = -5864.2), though the Jaccard model has greater variability (dispersion: 20.1 vs. 16.2). Overall, these findings show the importance of PC3-associated soil parameters in driving beta diversity, with the Sørensen index generally indicating greater dissimilarity between samples. The random effect for location shows similar variability in both models (SD= 0.1867 for Jaccard, 0.1862 for Sørensen), indicating a relatively small effect compared to the influence of the soil physicochemical properties as represented by their first three principal components.

3.3.9 Evaluation of sklearn classifier results and confidence scores

Counts of ASVs varied among species and locations. The three putative species with the greatest number of ASVs assigned to them are *Robustodorus arachidis (271 ASVs), Panagrolaimus* DC sp1 (190 ASVs), and *Acrobeles* DC sp2 (181 ASVs). These three species are found at 10, 11, and 11 locations, respectively. And, while *Panagrolaimus* DC sp1 and *Acrobeles* DC sp2 are the two most abundant putative species, accounting for 25.9% and 21.5% of all Nematoda reads, Robustodorus *arachidis* reads are only 1.6% of all Nematoda reads. The relative abundance of ASVs among all species level classifications is shown in **Figure 3.13**. More

detailed information is shown in **Table 3.7**, including AVS counts for all putative species and number of samples and locations in which each species-level classification occurs.

The distribution of confidence scores for species-level classifications among all ASVs, using boxplots across all samples, are shown in **Figure 3.14A**. The number of ASVs classified to each species varies, with some species having only a few ASVs and others having hundreds. These boxplots reveal variance in confidence values across all ASVs, irrespective of location. The bar graph in **Figure 3.14B** shows the variance of confidence value distributions for putative species across all locations (variance of variance). The distributions of confidence values for some species show substantial variance across locations, suggesting that the reliability of these classifications varies both within and among locations. Species with higher bars may warrant further investigation for potential over-classifications, where multiple species are grouped under a single classification. The three species showing the most variance in confidence values across locations are *Nothacrobeles triniglarus*, *Robustodorus arachidis*, and *Nothacrobeles borregi*. Of the 62 putative species, 36 show little to no variance in the confidence scores associated with the classification of their ASVs.

3.3.10 Phylogenetic and biogeographic analyses

As mentioned in **section 2.5**, we selected *Robustodorus arachidis*, *Panagrolaimus DC sp1*, and *Acrobeles DC sp2* for phylogenetic and biogeographical analyses of their constituent ASVs. These three putative species have the greatest numbers of ASVs assigned to them, and they

show high, medium, and low levels of variance in classification confidence scores across locations, respectively (**Figure 3.14A**). Furthermore, *R. arachidis* has the highest number of ASVs (271) assigned to it of all putative species, yet it accounts for just 1.6% of total reads. This is an unusually high level of sequence diversity over a small number of reads.

In the maximum likelihood tree including Family Aphelenchoididae and Family Aphelenchidae reference sequences from our custom database, the 271 *R. arachidis* ASV sequences form several distinct clades, some of which are not closely related to the *R. arachidis* reference sequence from our database (**Figure 3.15**). Among these, two ASV groups—one consisting of 27 ASVs and another of 151 ASVs—are positioned outside the clade containing all other ASVs and reference sequences from the genus *Robustodorus*, with high support values. In contrast, two other ASV groups, comprising 24 and 68 ASVs respectively, are more closely related to *Robustodorus* reference sequences in general. Additionally, there is a single ASV that is inferred to be sister to the *R. arachidis* reference sequence.

The group containing 27 ASVs is primarily associated with the MNP-KB location, where 18 of the ASVs are exclusively found. The remaining ASVs are distributed across other locations, including SB-N, SB-W, DC-AH, DC-LC, and TNP, either as single occurrences or in various combinations. While MNP-KB seems to be a key location for this group, they also exhibit a broader geographic range, albeit with less genetic diversity. The group containing 151 ASVs shows the widest geographic spread, with the majority of ASVs found in SB-N (87 ASVs), followed by SB-W (19 ASVs), DC-AF (23 ASVs), and ABW (16 ASVs). Some ASVs are present in combinations of these

locations, including DV-EC and TNP. The group containing 68 ASVs has the second broadest geographic distribution, with the majority of ASVs found in DC-AH (37 ASVs), followed by SB-N (14 ASVs) and SB-W (9 ASVs). Finally, the group containing 24 ASVs has 20 ASVs split evenly between DC-AH and MNP-KB, with the remaining 4 ASVs are found in MNP-HH, TNP, SB-N and SB-W.

In the tree generated from Family Cephalobidae reference sequences and the 181 ASVs classified as *Acrobeles DC sp2*, 176 of these ASVs form a clade with *Acrobeles complexus* NCBI reference sequences and *Acrobeles sp2* reference sequences from the Nadler Lab. (Nexus-formatted tree files in **supplemental materials**). These ASVs were nearly all classified with 100% confidence. The reference sequences they grouped with include HM055391.1, HM055389.1, HM055390.1, HM055394.1, RA124b.006, HM055388.1, DA132, RA124b.011, DA86, DA71, RA125.003, and RA131.006. Within this group, a subset of 44 ASVs was more closely related to the HM055388.1 *Acrobeles cf. complexus* CR-2010 reference sequence. These ASVs were distributed across eight locations, with some found exclusively in single locations and others across multiple combinations, including up to six different locations. Another subset of 102 ASVs was more closely associated with three *Acrobeles DC sp2* reference sequences that were generated during our reverse taxonomy study (**Chapter 1**). The latter ASVs were spread across nine unique locations.

Additionally, four ASVs form a clade with reference sequences from other *Acrobeles* species, including *A. maeneeneus*, *A. singulus*, and other unnamed isolates from NCBI, though these

ASVs were classified at varying confidence levels and exhibit lower support values, making their phylogenetic placement relative to the *A. complexus* and *A. sp2* clades uncertain. There was also a lone ASV of uncertain placement, which was classified at 75.7% confidence and found at the ABW location. Notably, there are many cephalobid reference sequences throughout the tree that were previously classified to the same genus or species but do not form monophyletic groups. This could mean that the classifications are correct but associated with incorrect taxonomy strings, that the group is genuinely non-monophyletic, or that there are some mistakes that were made by the sklearn classifier.

In the region of the cephalobid/panagrolaimid tree were Panagrolaimidae reference sequences group with 190 *Panagrolaimus DC sp1* ASVs, 189 of those ASVs form a clade with Genus *Panagrolaimus* reference sequences, including RA138.002 *Panagrolaimus sp1*, RA138.003 *Panagrolaimus sp1*, PalSPs15N55 *Panagrolaimus sp.*, sfPanSPj37N54 unclassified sp., and EU253569.1.3460 *Panagrolaimus sp. PS1159*. These ASVs resolve into two main groups. The first group consists of 103 ASVs, which are more closely associated with the reference sequences RA138.002 *Panagrolaimus sp1* and RA138.003 *Panagrolaimus sp1*, which were generated during our reverse taxonomy study (**Chapter 1**). The second group includes 84 ASVs that are more closely associated with the NCBI reference sequences EU253569.1.3460 *Panagrolaimus sp. PS1159*, sfPanSPj37N54 *unclassified sp.*, and PalSPs15N55 *Panagrolaimus sp..* Additionally, there are two ASVs that are found sister to the clade containing the two main panagrolaimid ASV groups. Both ASVs are found in the DC-AH

location. Lastly, there is a single ASV that groups with 15 NCBI reference sequences from the genus *Halicephalobus* and is found at the DV-CC location.

The 103 ASVs that group with the *Panagrolaimus DC sp1* reference sequences have the widest geographic distribution (9 locations), with 11 ASVs identified from ABW, 9 from DC-AF, 26 from DV-CC, and 30 from DV-EC. Additionally, 15 ASVs are found in both DV-CC and DV-EC. The remaining 13 ASVs are distributed across different combinations of locations, including ABW, DC-AF, DC-LC, DV-CC, DV-EC, MNP-KB, SB-N, and SB-W. The 84 ASVs that group with the *Panagrolaimus sp.* NCBI reference sequences are found in just three locations: 47 ASVs in MNP-HH, 1 ASV in both MNP-HH and MNP-KB, and 32 ASVs in DC-AH.

3.4. Discussion

3.4.1 Classification results

While nematode classification was the primary goal, the results also highlighted a significant number of non-nematode taxa, reflecting the wide taxonomic coverage of the primers used. Only 60% of all reads were classified to taxonomic ranks within Phylum Nematoda, indicating that the primers 391F and 799R also amplified a variety of non-nematode taxa. These include various types of algae (green, brown, golden), dinoflagellates, cercozoans, choanoflagellates, SAR clade organisms, oomycetes, apicomplexans, ciliates, tardigrades, and arthropods. The reverse primer, 799R, is highly conserved across Metazoa and has been used in various

phylogenetic and barcoding studies to resolve species-level diversity in groups such as pinworms (Okamoto, et al., 2009), crustaceans (Puillandre et al., 2011), insects (Pedro, et al., 2020), and annelids (Grosse et al., 2021). Despite this, some taxa were notably absent from our sequence data. For example, although rotifers were observed in our Baermann funnel catches and Rotifera sequences were included in the reference database, no reads were classified as Rotifera. This suggests a possible gap in primer coverage or sequence representation. Further investigation into the full coverage of these barcoding primers is needed, and exploring a more conserved forward primer could enhance broader taxonomic coverage.

Recent advances in PacBio sequencing technology, such as the Revio platform, offer a significant increase in sequencing capacity, generating a much higher volume of reads. As a result, the amplification of non-target taxa becomes less problematic, since the excess sequencing capacity ensures that target taxa are still adequately represented. This technology could allow for the successful sequencing of a wide variety of metazoan taxa from a single soil sample using just one set of primers.

In this study, approximately 1.5 million putative Nematoda reads were analyzed, with 83% successfully classified to the species level, despite the nascent and incomplete state of our reference sequence database and the lack of established benchmarks for classifier settings. Across all locations, we identified 2,319 ASVs, representing significant genetic diversity within these communities. Among these, 92 unique taxonomic groups were classified at a 70% confidence threshold, including 62 putative species: 19 described and 43 undescribed. The

ability to classify a high percentage of reads to species level demonstrates the utility of our current database for certain taxa, but there is also a clear need for continued refinement and expansion to improve accuracy and capture the full diversity of nematodes in these habitats.

Families Cephalobidae (57.4%) and Panagrolaimidae (21.7%) showed the greatest relative sequence read abundance among all samples. However, the proportion of Cephalobidae reads was smaller than we have observed in our previous studies of creosote nematode communities (**Chapter 1**). This can be explained by the overwhelming representation of Panagrolaimidae reads in samples from Death Valley National Park (DV-CC and DV-EC). These two locations alone accounted for over 76% of Panagrolaimidae reads.

Microbivores dominated all locations with 81.5% relative abundance. This is consistent with previous studies in similar habitats (Freckman and Mankau, 1977; Treonis et al., 2012; Pervez et al., 2023). Although Panagrolaimidae had a higher-than-usual relative abundance, the overall trophic group ratios align with prior observations. Following microbivores, omnivore-predators accounted for 6.7%, and fungivores for 5.9%. The lower classification success in these groups, particularly in DC-AH, may be attributed to the underrepresentation of rarer taxa in the reference sequence database. Our custom reference sequence database was supplemented with nematode sequences generated in a previous study of the DC-AF location, and that location serves as a standard of comparison for the proportion of reads successfully classified to species level. The DC-AH and TNP locations had lower proportions of species-level classifications than DC-AF, probably due to the higher abundance of omnivore predators and

plant parasites at these locations, such as *Utahnema* and *Ditylenchus* species. These taxa may be underrepresented in the reference database because they were either not present at DC-AF or were so rare that they were not sampled in our previous studies. An additional bias may arise from mixing soil samples before hydration and Baermann funnel extraction, potentially leading to the destruction of larger-bodied nematodes like dorylaimidans. Smaller or earlier life stages of these species are less likely to be affected, which could skew the observed community composition, particularly reducing the relative abundance of larger species.

While some rare taxa detected in this study likely represent true findings, others may be artifacts or misidentifications. An extremely low abundance taxon, *Caracharodiscus banaticus*, recovered using all three sequencing technologies in Chapters 1 and 2, was detected again in this study, occurring in one Anzo-Borego (ABW) sample, two Agave Hill (DC-AH) samples, and all six Kelbaker Road (MNP-KB) samples, with read counts ranging from 12 to 1012. In contrast, *Dirofilaria immitis* was found at a low relative abundance (0.20%) across eight locations, raising questions about its actual presence in these samples. Further investigation through BLAST analysis of nine associated ASVs suggested that these reads might be non-nematode, partially nematode, or potentially spurious or chimeric.

3.4.2 Geographic distribution of ASVs and putative species

The geographic distribution of ASVs and species-level classifications provides some insight into the diversity and structure of nematode communities across our collection locations. The

majority of ASVs showed a high degree of localization, with 1218 ASVs (52.5%) being found at just one location. This suggests that many ASVs are specific to particular environments or microhabitats, though this pattern could be partly from geographic structuring and partly from the presence of artifactual sequence variants from PCR and sequencing processes. Three ASVs were found at nine or more locations, indicating a few genetic variants with broader distributions.

At the species level, geographic distribution patterns also varied substantially. While some putative species were found at only one location, suggesting a degree of endemism, others were more widespread, occurring at nine or more locations. This broader distribution could indicate that these species are more cosmopolitan and capable of thriving in a variety of environments. However, the relationship between ASVs and species-level classifications can be challenging to interpret.

3.4.3 Phylogenetic and biogeographic analyses

Given the unexpectedly high number of ASVs assigned to *Robustodorus arachidis* relative to its low representation in sequence reads, we chose this species as a case study to investigate potential over-classification. To provide context and contrast, we also examined two other species, *Acrobeles* DC *sp2* and *Panagrolaimus* DC *sp1*, which not only have high ASV counts but are also the two most abundant putative species by read count. Additionally, *Acrobeles* is in the family Cephalobidae, known for its relatively low sequence diversity, while *Panagrolaimus*

belongs to Panagrolaimidae, a family with relatively high sequence diversity. These characteristics made these three species particularly valuable for gaining insights into the accuracy of our ASV classifications. Our analyses generated the following hypotheses for future testing:

Hypothesis 1: *Robustodorus arachidis* represents multiple species, some of which may not even belong to the same genus. The ASVs associated with *R. arachidis* form four reciprocally monophyletic clades, which are found in different combinations sympatrically across multiple collection locations.

Hypothesis 2: *Acrobeles DC sp2* ASVs represent either a single cosmopolitan species or two closely related species which are sympatric in many locations. This putative species is primarily composed of two ASV groups, which form distinct clades within a larger monophyletic clade that also includes *A. complexus* and *A. sp2* reference sequences. Of the 181 ASVs classified as *A.* DC *sp2*, 176 may represent either a single species or two closely related species.

Hypothesis 3: *Panagrolaimus* DC *sp1* ASVs represent either two allopatric species or two allopatric populations of the same species. The ASVs in this putative species are mainly divided into two large groups. The first group contains 103 ASVs distributed across 9 locations, while the second group has 84 ASVs found predominantly in two locations. There is minimal geographic overlap between the two ASV groups. The high sequence diversity observed in

Panagrolaimidae compared to other families may facilitate the detection of population-level differences within the same species.

3.4.4 Variation of soil properties among soil samples and collection locations

The PCA results and loadings show that organic carbon, nitrogen, water-holding capacity, and soil texture (percent sand, silt, and clay) account for most of the variation in soil properties across the study sites. PC1 reveals a clear inverse relationship between sand content and factors like organic matter and water-holding capacity, emphasizing the influence of soil texture on the observed differences. Sites such as MNP-HH, MNP-KB, and DV-CC exhibit the most significant contrasts in these properties. PC2 is influenced primarily by micronutrient levels, particularly zinc and iron, while PC3 reflects differences in soil texture variables, including clay, silt, and cation exchange capacity (CEC). These components describe how the soil samples differ in their physicochemical properties, with texture and organic matter being the major differentiating factors between sites.

3.4.5 Influence of soil parameters on nematode community composition

The NMDS analyses revealed consistent patterns in nematode community composition across different locations, with both the Jaccard and Sørensen NMDS plots showing distinct clustering by location. Despite some variations between these two distance metrics, all samples generally clustered with others from the same location, indicating the influence of location-specific

factors on community structure. There were a few outlier samples, but the overall consistency in clustering suggests a strong role of environmental and geographic factors in shaping these communities.

The Death Valley samples (DV-CC and DV-EC) exhibited the tightest clustering, suggesting a high degree of similarity in species presence and absence across these locations. This pattern was observed in both the Jaccard and Sørensen NMDS plots, though the clusters were slightly more dispersed along the NMDS1 axis in the Sørensen plot. This suggests that the environmental conditions in Death Valley are distinct enough from other locations that their nematode communities that are not only internally similar but also distinct from those of other samples.

Several environmental factors likely contribute to this distinctiveness. The DV-CC and DV-EC locations are the lowest in elevation among locations, at -67 meters and +155 meters, respectively. Additionally, DV-EC is geomorphically distinct, with creosote bushes growing out of desert pavement rather than the sandy hummocks found at all other collection locations. The soils at these sites also stand out, with DV-CC having the highest sand content and DV-EC having the highest copper concentrations and total carbon. Another possible factor, which was not measured in this study, is the variation in peak soil temperatures at collection depths. Extreme temperatures in parts of Death Valley could potentially exceed the thermal tolerance of certain nematode species, restricting their presence and contributing to the observed distinctiveness of these communities.

The MNP-HH and MNP-KB samples were generally more spread out in the Jaccard NMDS plot compared to other locations, with one of the 6 MNP-HH samples standing out as an extreme outlier, positioned far from its peers. However, in the Sørensen plot, this sample clustered more closely with other Mojave National Preserve samples. This is likely due to the presence of rare, low-abundance taxa in these samples. These taxa are so infrequent that they may be sampled inconsistently within and among locations, which can confound Jaccard distance calculations, as this metric gives equal weight to both shared and unshared taxa. In contrast, the Sørensen distance, which gives double weight to shared taxa, more accurately reflects the core community composition, leading to the clustering of MNP-HH with its peers. This shows how important it is to consider the idiosyncrasies of a study system before choosing diversity metrics.

The overlaid environmental vectors on the NMDS plots demonstrated internal logical consistency among the soil parameters. For instance, vectors for silt and clay were inversely correlated with sand, and organic matter, organic carbon, and nitrogen were positively correlated with one another, aligning with general expectations. The vectors also showed clear correlations with specific locations. For example, copper was most correlated with the Death Valley samples, reflecting the high copper content in these soils, while vectors for clay, silt, and water-holding capacity pointed towards the Mojave National Preserve samples, which had the highest measurements for these variables.

Overall, the NMDS plots reveal distinct clustering patterns of nematode communities that correspond with different soil properties and environmental factors across locations. The contrasting results between the Jaccard and Sørensen plots, particularly in their treatment of rare taxa, emphasize the importance of carefully selecting diversity metrics in ecological studies, particularly in environments where low-abundance species are prevalent.

The beta regression GLMM analysis indicates that PC3, which is associated with soil texture (clay, silt, CEC), has the most substantial impact on beta diversity as measured by dissimilarity metrics. The stronger effect of PC3 on both Jaccard and Sørensen indices implies that soil texture and related chemical properties are important factors influencing the differences in nematode communities between locations. While the exact mechanisms remain unclear, these findings suggest that variations in soil texture could be linked to differences in the composition and distribution of nematode communities across different sites. The negative loading of sand on PC1 and the strong influence of PC3 on community composition suggest that locations with higher sand content, like DV-CC, may foster different nematode communities compared to sites with more clay and silt, such as MNP-HH and MNP-KB.

Soil texture, particularly factors associated with PC3 (clay, silt, CEC), significantly influences nematode community composition across different locations. Sandy soils, which tend to have lower moisture retention, may favor species that are better adapted to rapid desiccation. These adaptations might include physiological traits, such as high constitutive levels of protective osmolytes like trehalose and glycerol (Crowe and Madin, 1974) or the ability to rapidly produce

them in response to decreasing moisture, as well as behavioral strategies that reduce water loss, such as clumping (Wharton, 2002). Additionally, the presence of organic matter in sandy soils may enhance survival by creating microhabitats with improved water retention, which could be particularly important for species that are less adapted to rapid desiccation. Further research into how anhydrobiotic phenotypes (Shannon, et al., 2005) interact with soil properties, along with measurements of soil relative humidity at sampling sites, would improve our understanding of the factors influencing nematode community composition across different soil types.

3.4.6 Correlation of soil parameters with nematode abundance

The Spearman rank correlation analysis provided some insight into how soil properties influence total nematode abundance. The significant correlations observed suggest that soil texture is a significant variable affecting nematode abundance, with a strong positive correlation between sand content and nematode density. In contrast, the negative correlations with clay and silt content indicate that finer-textured soils may prevent the accumulation large numbers of nematodes.

Interestingly, water-holding capacity and CEC, which are typically associated with fertile soils, also showed a negative correlation with nematode abundance. This counterintuitive result suggests that in arid environments, drier, sandier soils might offer better support for some species assemblages. One possible explanation is that soils with higher moisture retention

could lead to a more diverse (nematode and non-nematode) community, possibly increasing competition or predation, and thereby reducing overall nematode abundance.

Micronutrients like manganese (Mn) and zinc (Zn) showed weak but significant positive correlations with nematode abundance, indicating that while these elements may support nematode communities, their influence is secondary to that of soil texture. Other chemical properties, such as organic matter, organic carbon, pH, nitrogen, and copper, did not show significant correlations, suggesting that these factors are less critical in determining nematode densities in these habitats.

Overall, the results indicate that soil texture, particularly sand content, has a significant influence on nematode abundance. The higher nematode densities observed in sandy soils could be another indication that these environments favor species more adapted to rapid desiccation. In contrast, the lower nematode densities observed in finer-textured soils could result from slower drying rates, which might benefit species less adapted to rapid desiccation but also expose them to increased competition or predation. Given that we collect soil samples in a fully desiccated state and rehydrate them to allow nematodes to reanimate, the species we detect are those that have successfully endured an anhydrobiotic state. If certain taxa struggle to reanimate under laboratory conditions, they might be underrepresented in our data. Additionally, our abundance estimates could be influenced by some degree of development and reproduction during the rehydration period, though the extent to which this occurs is not well known and likely varies by species.

3.4.7 Future strategies for improving the reference sequence database

To address the shortcomings of our reference sequence database, particularly the issue of non-monophyly among certain taxa, we need to refine our reference taxonomy files. Non-monophyly in reference taxa can result from genuine para- or polyphyly within the taxa or from inaccuracies in existing databases. To improve the reliability of our classifications, it may be necessary to acknowledge these discrepancies and add specific labels that indicate potential non-monophyly among the reference sequences. This would allow us to continue using some sequences while clearly flagging their possible limitations.

A possible strategy for enhancing the taxonomic coverage of our reference database involves the integration of phylogenetic methods with geographic data. By analyzing phylogenetic patterns and geographic distributions of ASVs, we can identify potentially novel taxa that are not represented in the current database. These ASVs can guide targeted recollection efforts at specific locations, where we can resample these potentially new taxa. Employing reverse taxonomic methods, we would then document both the molecular and morphological characteristics of these organisms, with the assistance of expert taxonomists. This iterative process of expanding the reference database will gradually improve classifier accuracy.

3.4.8 Summary

This study shows that soil physicochemical variables, particularly soil texture, organic matter, and water retention, significantly influence nematode community composition, affecting both richness and beta diversity. Soil texture, especially sand content, emerged as a significant factor that was positively corelated with nematode abundance, while finer-textured soils with more clay and silt were negatively corelated. The use of the 28S D1-D2 metabarcoding locus provided species-level classifications in many nematode groups, although its utility varies among taxa depending on how well those taxa are represented in our custom reference database. By combining biogeographical data with phylogenetic methods, we identified areas where the reference sequence database could be improved, revealing gaps in our taxonomic coverage, and pointing to specific ASVs that may represent unrecognized diversity. Future work should focus on refining the database by incorporating more diverse reference sequences from under-sampled taxa and addressing non-monophyly among reference taxa.

Chapter 3 References:

Adamczyk-Szabela, D. and W. M. Wolf (2022). "The impact of soil pH on heavy metals uptake and photosynthesis efficiency in Melissa officinalis, Taraxacum officinalis, Ocimum basilicum." Molecules **27**(15): 4671.

Ahmed, M., et al. (2015). "Nematode taxonomy: From morphology to metabarcoding." Soil Discussions **2**(2): 1175-1220.

Andersson, S., et al. (2000). "Leaching of dissolved organic carbon (DOC) and dissolved organic nitrogen (DON) in mor humus as affected by temperature and pH." Soil Biology and Biochemistry **32**(1): 1-10.

Blaxter, M. L., et al. (1998). "A molecular evolutionary framework for the phylum Nematoda." Nature **392**(6671): 71-75.

Bolyen, E., et al. (2019). "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2." Nature biotechnology **37**(8): 852-857.

Bongers, T. and M. Bongers (1998). "Functional diversity of nematodes." Applied soil ecology **10**(3): 239-251.

Callahan, B. J., et al. (2016). "DADA2: High-resolution sample inference from Illumina amplicon data." Nature methods **13**(7): 581-583.

Chao, A. and L. Jost (2012). "Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size." Ecology **93**(12): 2533-2547.

Crowe, J. H. and K. Madin (1974). "Anhydrobiosis in tardigrades and nematodes." Transactions of the American Microscopical Society: 513-524.

De Ley, P. and M. L. Blaxter (2004). A new system for Nematoda: combining morphological characters with molecular trees, and translating clades into ranks and taxa. Proceedings of the Fourth International Congress of Nematology, 8-13 June 2002, Tenerife, Spain, Brill.

Dinić, Z., et al. (2019). "Prediction models for bioavailability of Mn, Cu, Zn, Ni and Pb in soils of Republic of Serbia." Agronomy **9**(12): 856.

Freckman, D. W. (1978). "Ecology of anhydrobiotic soil nematodes." Dry Biological Systems: 345-357.

Freckman, D. and R. Mankau (1977). "Distribution and trophic structure of nematodes in desert soils." Ecological Bulletins: 511-514.

Ghanem, M. F., et al. (2024). "Impact of metal polluted sewage water on soil nematode assemblages in agricultural settings of Aligarh, India." Soil Ecology Letters **6**(1): 230193.

Garcia, N., et al. (2022). "Diversity of plant parasitic nematodes characterized from fields of the French national monitoring programme for the Columbia root-knot nematode." PloS one **17**(3): e0265070.

Gattoni, K., et al. (2022). "Context dependent role of abiotic and biotic factors structuring nematode communities along two environmental gradients." Molecular Ecology.

Georgieva, S. S., et al. (2002). "Nematode communities under stress: the long-term effects of heavy metals in soil treated with sewage sludge." Applied soil ecology **20**(1): 27-42.

Grosse, M., et al. (2021). "Describing the hidden species diversity of Chaetozone (Annelida, Cirratulidae) in the Norwegian Sea using morphological and molecular diagnostics." ZooKeys **1039**: 139.

Hsieh, T., et al. (2016). "iNEXT: an R package for rarefaction and extrapolation of species diversity (H ill numbers)." Methods in Ecology and Evolution **7**(12): 1451-1456.

Hugot, J.-P., et al. (2001). "Biodiversity in helminths and nematodes as a field of study: an overview." Nematology **3**(3): 199-208.

Korthals, G. W., et al. (1996a). "Long-term effects of copper and pH on the nematode community in an agroecosystem." Environmental Toxicology and Chemistry: An International Journal **15**(6): 979-985.

Korthals, G. W., et al. (1996b). "Short-term effects of cadmium, copper, nickel and zinc on soil nematodes from different feeding and life-history strategy groups." Applied soil ecology **4**(2): 107-117.

Kung, S.-P., et al. (1990). "Soil type and entomopathogenic nematode persistence." Journal of Invertebrate Pathology **55**(3): 401-406.

Lex, A., et al. (2014). "UpSet: visualization of intersecting sets." IEEE transactions on visualization and computer graphics **20**(12): 1983-1992.

Minh, B. Q., et al. (2020). "IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era." Molecular Biology and Evolution **37**(5): 1530-1534.

Neina, D. (2019). "The role of soil pH in plant nutrition and soil remediation." Applied and environmental soil science **2019**: 1-9.

Nimnoi, P., et al. (2024). "Insights into soil nematode diversity and bacterial community of Thai jasmine rice rhizosphere from different paddy fields in Thailand." PeerJ **12**: e17289.

Okamoto, M., et al. (2009). "Phylogenetic relationships of rodent pinworms (genus Syphacia) in Japan inferred from 28S rDNA sequences." Parasitology International **58**(4): 330-333.

Pedro, P. M., et al. (2020). "Culicidae-centric metabarcoding through targeted use of D2 ribosomal DNA primers." PeerJ **8**: e9057.

Pen-Mouratov, S., et al. (2003). "Seasonal and spatial variation in nematode communities in a Negev desert ecosystem." Journal of Nematology **35**(2): 157.

Pervez, R., et al. (2023). "Nematode diversity and community structure from high-altitude district Leh (Ladakh), India." Indian Phytopathology **76**(2): 559-567.

Porazinska, D. L., et al. (2010). "Ecometagenetics confirm high tropical rainforest nematode diversity." Molecular Ecology **19**(24): 5521-5530.

Ptatscheck, C., et al. (2018). "The extent of wind-mediated dispersal of small metazoans, focusing nematodes." Scientific reports **8**(1): 6814.

Ptatscheck, C. and W. Traunspurger (2020). "The ability to get everywhere: dispersal modes of free-living, aquatic nematodes." Hydrobiologia **847**(17): 3519-3547.

Puillandre, N., et al. (2011). "Barcoding type specimens helps to identify synonyms and an unnamed new species in Eumunida Smith, 1883 (Decapoda: Eumunididae)." Invertebrate Systematics **25**(4): 322-333.

Rambaut, A. (2010). "FigTree v1. 3.1." (No Title).

Shannon, A. J., et al. (2005). "The anhydrobiotic potential and molecular phylogenetics of species and strains of Panagrolaimus (Nematoda, Panagrolaimidae)." Journal of Experimental Biology **208**(12): 2433-2445.

Smythe, A. B. and S. A. Nadler (2006). "Molecular phylogeny of Acrobeloides and Cephalobus (Nematoda: Cephalobidae) reveals paraphyletic taxa and recurrent evolution of simple labial morphology." Nematology **8**(6): 819-836.

Sohlenius, B. (1980). "Abundance, biomass and contribution to energy flow by soil nematodes in terrestrial ecosystems." Oikos: 186-194.

Song, D., et al. (2017). "Large-scale patterns of distribution and diversity of terrestrial nematodes." Applied soil ecology **114**: 161-169.

Treonis, A., et al. (2012). "Soil nematodes and their prokaryotic prey along an elevation gradient in the Mojave Desert (Death Valley National Park, California, USA)." Diversity **4**(4): 363-374.

Van Den Hoogen, J., et al. (2019). "Soil nematode abundance and functional group composition at a global scale." Nature **572**(7768): 194-198.

Wallace, H. (1968). "The dynamics of nematode movement." Annual Review of Phytopathology **6**(1): 91-114.

Wharton, D. and I. Brown (1989). "A survey of terrestrial nematodes from the McMurdo Sound region, Antarctica." New Zealand Journal of Zoology **16**(3): 467-470.

Wharton, D. A. (2002). "16. Nematode survival strategies." Biol. Nematodes **389**: 20.

Xiong, D., et al. (2020). "Nonlinear responses of soil nematode community composition to increasing aridity." Global Ecology and Biogeography **29**(1): 117-126.

Zullini, A. (2018). "Cosmopolitism and endemism in free-living nematodes." Biogeographia–The Journal of Integrative Biogeography **33**.

Chapter 3 Tables and figures:

Collection Locality	Soil Sample	GPS Coordinates	Elevation	Sampling Date
BDCR alluvial fan (DC-AF)	A205	N33°39.528', W116°22.249'	251m	12/3/18
	A205B	N33°39.529', W116°22.247'	251m	12/3/18
	A206	N33°39.519', W116°22.238'	252m	12/3/18
	A206B	N33°39.520', W116°22.238'	252m	12/3/18
	A207	N33°39.500', W116°22.246'	253m	12/3/18
	A207B	N33°39.501', W116°22.245'	253m	12/3/18
BDCR lower Deep Canyon (DC-LC)	A212	N33°38.585', W116°22.758'	342m	12/4/18
	A212B	N33°38.581', W116°22.757'	343m	12/4/18
	A213	N33°38.593', W116°22.744'	343m	12/4/18
	A213B	N33°38.593', W116°22.744'	343m	12/4/18
	A214	N33°38.594', W116°22.787'	339m	12/4/18
	A214B	N33°38.595', W116°22.788'	339m	12/4/18
BDCR Agave Hill (DC-AH)	A218	N33°38.335', W116°23.920'	836m	12/5/18
	A218B	N33°38.337', W116°23.920'	837m	12/5/18
	A219	N33°38.339', W116°23.909'	837m	12/5/18
	A219B	N33°38.340', W116°23.908'	838m	12/5/18
	A220	N33°38.338', W116°23.917'	837m	12/5/18
	A220B	N33°38.338', W116°23.916'	837m	12/5/18
New Ironage Road	A231	N34°10.074', W115°44.016'	379m	12/9/18

Twentynine Palms (TNP)	A231B	N34°10.075', W115°44.015'	379m	12/9/18
	A232	N34°10.072', W115°44.013'	379m	12/9/18
	A232B	N34°10.073', W115°44.012'	379m	12/9/18
	A233	N34°10.078', W115°44.009'	379m	12/9/18
	A233B	N34°10.079', W115°44.009'	379m	12/9/18
Mojave National Preserve near Hidden Hills Road (MNP-HH)	A234	N34°47.597', W115°36.557'	1170m	12/10/18
	A234B	N34°47.598', W115°36.557'	1170m	12/10/18
	A235	N34°47.593', W115°36.548'	1169m	12/10/18
	A235B	N34°47.594', W115°36.547'	1169m	12/10/18
	A236	N34°47.579', W115°36.553'	1169m	12/10/18
	A236B	N34°47.579', W115°36.554'	1169m	12/10/18
Mojave National Preserve near Kelbaker Road (MNP-KB)	A237	N35°08.803', W115°44.113'	1127m	12/10/18
	A237B	N35°08.804', W115°44.112'	1127m	12/10/18
	A238	N35°08.784', W115°44.115'	1128m	12/10/18
	A238B	N35°08.784', W115°44.114'	1128m	12/10/18
	A239	N35°08.783', W115°44.108'	1129m	12/10/18
	A239B	N35°08.783', W115°44.109'	1129m	12/10/18

(Table continued)

Collection Locality	Soil Sample	GPS Coordinates	Elevation	Sampling Date
Death Valley National Park near Copper Canyon (DV-CC)	A240	N36°06.914', W116°44.922'	-67m	12/12/18
	A240B	N36°06.914', W116°44.921'	-67m	12/12/18
	A241	N36°06.920', W116°44.910'	-67m	12/12/18
	A241B	N36°06.920', W116°44.911'	-67m	12/12/18
	A242	N36°06.908', W116°44.886'	-67m	12/12/18
	A242B	N36°06.907', W116°44.885'	-67m	12/12/18
Death Valley National Park near Echo Canyon (DV-EC)	A243	N36°26.406', W116°49.136'	155m	12/12/18
	A243B	N36°26.406', W116°49.137'	155m	12/12/18
	A244	N36°26.404', W116°49.141'	155m	12/12/18
	A244B	N36°26.404', W116°49.141'	155m	12/12/18
	A245	N36°26.307', W116°49.317'	136m	12/12/18
	A245B	N36°26.306', W116°49.317'	136m	12/12/18
BLM Land Anza-Borrego Wilderness (ABW)	A246	N32°45.176', W116°02.633'	195m	8/30/19
	A246B	N32°45.176', W116°02.633'	195m	8/30/19
	A247	N32°45.181', W116°02.635'	195m	8/30/19
	A247B	N32°45.181', W116°02.635'	195m	8/30/19
	A248	N32°45.174', W116°02.625'	195m	8/30/19
	A248B	N32°45.174', W116°02.625'	195m	8/30/19
North of	A249	N33°14.515', W116°23.325'	215m	8/31/19

Steele-Burnand Desert Research Center (SB-N)	A249B	N33°14.515', W116°23.325'	215m	8/31/19
	A250	N33°14.523', W116°23.326'	215m	8/31/19
	A250B	N33°14.523', W116°23.325'	215m	8/31/19
	A251	N33°14.520', W116°23.332'	215m	8/31/19
	A251B	N33°14.520', W116°23.331'	215m	8/31/19
West of Steele-Burnand Desert Research Center (SB-W)	A252	N33°14.399', W116°23.437'	212m	8/31/19
	A252B	N33°14.398', W116°23.437'	212m	8/31/19
	A253	N33°14.394', W116°23.437'	212m	8/31/19
	A253B	N33°14.393', W116°23.436'	212m	8/31/19
	A254	N33°14.391', W116°23.445'	212m	8/31/19
	A254B	N33°14.392', W116°23.445'	212m	8/31/19

Table 3.1. Creosote soil sample collection locality abbreviations, sample numbers associated with each location, GPS coordinates, and elevation of each sample.

Class	Order	Family	Genus	Species	Trophic Group	Sequence Count
Chromadorea	Rhabditida	Cephalobidae	*Acrobeles*	DC sp2	microbivore	391653
Chromadorea	Rhabditida	Panagrolaimidae	*Panagrolaimus*	DC sp1	microbivore	323852
Chromadorea	Rhabditida	Cephalobidae	*Acrobeloides*	DC sp1	microbivore	111524
Chromadorea	Rhabditida	Cephalobidae	*Cervidellus*	unnamed sp. 1 HMM2018	microbivore	101059
Chromadorea	Rhabditida	Cephalobidae	*Acrobeloides*	unclassified	microbivore	67993
Chromadorea	Rhabditida	Cephalobidae	*Acrobeles*	DC sp1	microbivore	60727
Chromadorea	Rhabditida	Cephalobidae	unclassified	unclassified	microbivore	55285
Enoplea	Dorylaimida	Leptonchidae	*Utahnema*	DC sp4	omnivore-predator	44726
Chromadorea	Rhabditida	Anguinidae	*Ditylenchus*	unclassified	plant parasite	42919
Chromadorea	Rhabditida	unclassified	unclassified	unclassified	undefined	32040
Chromadorea	Rhabditida	Aphelenchidae	*Aphelenchus*	DC sp2	fungivore	26039
Chromadorea	Rhabditida	Aphelenchoididae	*Robustodorus*	*arachidis*	fungivore	24662
Chromadorea	Rhabditida	Aphelenchidae	*Aphelenchus*	DC sp1	fungivore	18332
Enoplea	Dorylaimida	Leptonchidae	*Utahnema*	DC sp3	omnivore-predator	18281
Chromadorea	Rhabditida	Cephalobidae	*Acrobeloides*	DC sp4	microbivore	15799
Enoplea	Dorylaimida	Qudsianematidae	*Ecumenicus*	unnamed sp.	omnivore-predator	13536
Chromadorea	Rhabditida	Cephalobidae	*Stegelletina*	unclassified	microbivore	13133
Chromadorea	Rhabditida	Cephalobidae	*Acrobeloides*	DC sp6	microbivore	12288
Enoplea	Dorylaimida	unknown	unknown	unnamed sp.	omnivore-predator	10625
Enoplea	Dorylaimida	unclassified	unclassified	unclassified	omnivore-predator	9789
Chromadorea	Rhabditida	Telotylenchidae	Quinisulcius	unclassified	plant parasite	8353
Enoplea	Dorylaimida	Qudsianematidae	*Ecumenicus*	DC sp7	omnivore-predator	8210
Enoplea	Dorylaimida	Qudsianematidae	*Microdorylaimus*	DC sp1	omnivore-predator	7133
Chromadorea	Rhabditida	Telotylenchidae	Quinisulcius	DC sp1	plant parasite	6517
Chromadorea	Rhabditida	Aphelenchidae	*Aphelenchus*	unnamed sp.	fungivore	6221
Chromadorea	Rhabditida	Cephalobidae	*Cervidellus*	unclassified	microbivore	5856
Chromadorea	Rhabditida	Aphelenchidae	*Aphelenchus*	unclassified	fungivore	5758
Chromadorea	Rhabditida	Cephalobidae	*Placodira*	*lobata*	microbivore	5501
Chromadorea	Rhabditida	Cephalobidae	*Acrobeles*	unnamed sp. JB-132	microbivore	4947

Chromadorea	Rhabditida	Cephalobidae	*Zeldia*	unnamed sp. JB-140	microbivore	4338
Chromadorea	Rhabditida	Cephalobidae	*Acrobeloides*	DC sp5	microbivore	4250
Chromadorea	Rhabditida	Aphelenchoididae	unclassified	unclassified	fungivore	4230
Chromadorea	Rhabditida	Cephalobidae	*Nothacrobeles*	*triniglarus*	microbivore	3547
Enoplea	Dorylaimida	Aporcelaimidae	*Aporcelaimellus*	DC sp6	omnivore-predator	3508
Chromadorea	Rhabditida	Rhabditidae	*Mesorhabditis*	*monhystera*	microbivore	3050
Enoplea	Dorylaimida	Dorylaimida	*Carcharodiscus*	*banaticus*	omnivore-predator	2912
Enoplea	Dorylaimida	Tylencholaimidae	*Tylencholaimus*	*mirabilis*	omnivore-predator	2603
Chromadorea	Rhabditida	Panagrolaimidae	unclassified	unclassified	microbivore	2519
Chromadorea	Rhabditida	Cephalobidae	*Stegelletina*	unnamed sp.	microbivore	2435
Chromadorea	unclassified	unclassified	unclassified	unclassified	undefined	2068
Chromadorea	Rhabditida	Cephalobidae	*Nothacrobeles*	*borregi*	microbivore	1759
Chromadorea	Rhabditida	Onchocercidae	unclassified	unclassified	animal parasite	1706
Chromadorea	Rhabditida	Onchocercidae	*Dirofilaria*	*immitis*	animal parasite	1335
Chromadorea	Plectida	Plectidae	*Plectus*	unnamed sp.	microbivore	1256
Chromadorea	Rhabditida	Cephalobidae	*Nothacrobeles*	*spatulatus*	microbivore	1164
Chromadorea	Rhabditida	Anguinidae	*Ditylenchus*	unnamed sp. 85C1	plant parasite	1155
Chromadorea	Rhabditida	Cephalobidae	*Acrobeles*	unclassified	microbivore	1107
Enoplea	Dorylaimida	Qudsianematidae	*Discolaimus*	*major*	omnivore-predator	1071
Chromadorea	Rhabditida	Cephalobidae	*Acrobeles*	unnamed sp. CR-2010	microbivore	1030
Chromadorea	Rhabditida	Telotylenchidae	unclassified	unclassified	plant parasite	902
Chromadorea	Rhabditida	Aphelenchoididae	*Robustodorus*	*helicus*	fungivore	833
Chromadorea	Rhabditida	Panagrolaimidae	*Panagrolaimus*	unclassified	microbivore	661
Chromadorea	Rhabditida	Rhabditidae	*Mesorhabditis*	*paucipapillata*	microbivore	640
Chromadorea	Rhabditida	Telotylenchidae	Quinisulcius	DC sp2	plant parasite	586
Chromadorea	Rhabditida	Telotylenchidae	Telotylenchus	sp. CCN-2014	plant parasite	553
Enoplea	Dorylaimida	Qudsianematidae	*Discolaimus*	unclassified	omnivore-predator	518
Chromadorea	Rhabditida	Cephalobidae	*Stegelleta*	DC sp2	microbivore	499
Chromadorea	Rhabditida	Pratylenchidae	*Pratylenchus*	unclassified	plant parasite	455

Chromadorea	Rhabditida	Nothotylenchidae	*Nothotylenchus*	*similis*	plant parasite	448
Chromadorea	Rhabditida	Panagrolaimidae	unknown	unnamed sp.	microbivore	437
Enoplea	Dorylaimida	Aporcelaimidae	*Aporcelaimellus*	unclassified	omnivore-predator	420
Chromadorea	Rhabditida	unclassified	unknown	unnamed sp.	undefined	392
Enoplea	Dorylaimida	Nordiidae	*Longidorella*	unnamed sp.	omnivore-predator	337
Chromadorea	Rhabditida	Tylenchidae	unknown	unnamed sp. 1 HMM2018	plant parasite	331
Enoplea	Dorylaimida	Nygolaimidae	*Solididens*	unnamed sp.	omnivore-predator	279
Enoplea	Dorylaimida	Qudsianematidae	unclassified	unclassified	omnivore-predator	254
Chromadorea	Plectida	Plectidae	*Plectus*	unclassified	microbivore	188
Chromadorea	Rhabditida	Cephalobidae	*Acrobeloides*	unnamed sp. DWF-1106	microbivore	158
Chromadorea	Rhabditida	Cephalobidae	*Cervidellus*	*cervus*	microbivore	128
Chromadorea	Rhabditida	Cephalobidae	*Stegelleta*	DC sp1	microbivore	120
Enoplea	Dorylaimida	Tylencholaimidae	*Tylencholaimus*	unnamed sp. 1 HMM2018	omnivore-predator	78
Chromadorea	Plectida	Plectidae	*Anaplectus*	*porosus*	microbivore	76
Chromadorea	Rhabditida	Rhabditidae	*Caenorhabditis*	*elegans*	microbivore	73
Chromadorea	Rhabditida	Cephalobidae	*Acrobeles*	*singulus*	microbivore	69
Chromadorea	Rhabditida	Cephalobidae	*Acrobeloides*	unnamed sp. FHD001	microbivore	67
Chromadorea	Rhabditida	Aphelenchoididae	*Robustodorus*	unclassified	fungivore	57
Chromadorea	Rhabditida	Cephalobidae	*Cervidellus*	unnamed sp. JB-138	microbivore	51
Chromadorea	Rhabditida	Anguinidae	*Ditylenchus*	unnamed sp. 20181130DIS3	plant parasite	32
Chromadorea	Rhabditida	Merliniidae	*Nagelus*	*obscurus*	plant parasite	27
Enoplea	Dorylaimida	Belondiridae	*Axonchium*	*propinquum*	omnivore-predator	26
Chromadorea	Rhabditida	Aphelenchoididae	*Aphelenchoides*	unclassified	fungivore	25
Enoplea	Dorylaimida	unknown	unknown	DC sp2	omnivore-predator	24
Enoplea	Dorylaimida	Nygolaimidae	unclassified	unclassified	omnivore-predator	17
Chromadorea	Rhabditida	Merliniidae	unclassified	unclassified	plant parasite	16
Chromadorea	Rhabditida	Diplogasteridae	unclassified	unclassified	omnivore-predator	15

Enoplea	Dorylaimida	Dorylaimidae	unclassified	unclassified	omnivore-predator	12
Enoplea	Dorylaimida	Tylencholaimidae	*Tylencholaimus*	unclassified	omnivore-predator	10
Chromadorea	Rhabditida	Cephalobidae	*Cervidellus*	*neftasiensis*	microbivore	9
Enoplea	Dorylaimida	Aporcelaimidae	*Aporcella*	*vitrinus*	omnivore-predator	9
Enoplea	Dorylaimida	Qudsianematidae	*Ecumenicus*	unclassified	omnivore-predator	8
Chromadorea	Rhabditida	Cephalobidae	*Acrobeloides*	unnamed sp. CR-2010	microbivore	2
Chromadorea	Rhabditida	Onchocercidae	*Dirofilaria*	unclassified	animal parasite	2
Total reads						1509645

Table 3.2. List of recovered taxonomic categories among all samples in decreasing order of read count, along with their putative trophic groups. We are defining "taxonomic category" as the lowest taxonomic rank to which the sklearn classifier has successfully classified an ASV or group of ASVs to a confidence level of 70%. Higher rank taxonomic categories are likely to include ASVs from multiple unclassified species.

Location	Avg abun.	Avg org C (%)	Avg org matter (%)	Avg clay (%)	Avg silt (%)	Avg sand (%)	Avg H20 cap (%)	Avg CEC (meq/100g)	Avg Cu (ppm)	Avg Fe (ppm)	Avg Mn (ppm)	Avg Zn (ppm)	Avg C (%)	Avg N (%)	Avg pH
DC-AF	3062	0.56	0.97	4.33	9.33	86.33	3.47	5.45	5	24418	408	69	0.62	0.06	7.91
DC-LC	2056	0.39	0.68	4.33	9.33	85.50	3.93	5.87	12	19768	283	52	0.51	0.04	8.25
DC-AH	1378	0.67	1.17	6.17	10.50	83.33	5.60	9.58	8	25350	445	64	0.70	0.06	7.53
TNP	2875	0.32	0.55	4.33	6.17	89.50	2.87	4.97	9	7093	187	27	0.35	0.03	8.00
MNP-HH	452	0.83	1.43	7.33	18.00	74.67	6.77	10.02	9	10155	308	43	0.80	0.08	7.32
MNP-KB	589	0.82	1.41	8.50	15.17	76.33	7.67	10.82	8	10753	294	34	1.36	0.09	8.19
DV-CC	7566	0.78	1.34	4.17	1.08	94.83	3.52	2.05	11	10228	465	43	1.10	0.06	7.71
DV-EC	791	0.70	1.20	5.80	14.00	80.20	5.44	2.38	18	9396	288	29	2.18	0.07	8.03
ABW	1873	0.28	0.48	4.17	3.00	92.83	3.22	2.30	5	11092	193	42	0.31	0.03	8.16
SB-N	2356	0.60	1.04	5.83	7.83	86.33	5.25	7.92	13	21528	325	63	0.84	0.05	8.05
SB-W	9576	0.63	1.09	5.80	10.80	83.40	5.84	8.76	10	21526	361	64	0.78	0.06	8.00

Table 3.3. Average nematode abundance and soil physicochemical property measurements of all samples in each location. Measurements for individual soil samples with each location can be found in supplemental materials.

Property	Spearman Correlation Coefficient	P-Value
Sand	0.5213	0.00001
Mn	0.32174	0.00953
Zn	0.29449	0.01817
Fe	0.20555	0.10323
Organic matter	0.00816	0.94896
Organic carbon	0.00768	0.95195
pH	-0.04837	0.70425
Cu	-0.05047	0.69209
N	-0.0808	0.52561
C	-0.11074	0.38367
CEC	-0.29263	0.01895
H_20 holding capacity	-0.42822	0.00042
Silt	-0.48335	0.00005
Clay	-0.52143	0.00001

Table 3.4. Spearman rank correlation coefficients and associated p-values for correlation of soil physicochemical properties relative to nematode abundance in soil samples.

Soil Property	PC1 (45.2%)	PC2 (19.0%)	PC3 (14.6%)	Weighted Composite Importance	K-means Cluster
Org.C	0.33835	0.0604	-0.33163	0.26172	High
Org.matter	0.33821	0.06185	-0.33151	0.26168	High
Sand	-0.33118	0.12074	-0.31367	0.25828	High
N	0.35601	-0.01058	-0.23649	0.25588	High
Zn	0.04564	0.56902	0.13227	0.25498	High
H20.capacity	0.34649	-0.08607	0.2283	0.25156	High
Clay	0.3228	-0.13156	0.27911	0.24852	High
Silt	0.31626	-0.11127	0.30536	0.24733	High
Fe	0.04065	0.5314	0.18898	0.24416	High
CEC	0.29276	0.13869	0.32669	0.24078	High
C	0.25762	-0.21834	-0.33219	0.23488	High
Mn	0.14053	0.42443	-0.22798	0.22526	High
pH	-0.17929	-0.1722	0.27462	0.17656	Medium
Cu	0.05225	-0.24056	-0.06975	0.11375	Low

Table 3.5. Soil property loadings and weighted importance. This table lists soil physiochemical parameters with their corresponding loadings on the first three principal components (PC1, PC2, PC3), which represent 45.2%, 19.0%, and 14.6% of the variance, respectively. Weighted composite importance scores were calculated for each soil parameter, and then sorted into three significance categories (high, medium, and low) using K-means clustering.

	Model	Estimate	Std. Error	z-value	P-value	AIC	BIC
Fixed Effects							
Intercept	Jaccard	-0.422	0.06	-7.048	<0.0001	-7617.3	-7579.4
PC1 Difference	Jaccard	-0.063	0.004	-15.939	<0.0001		
PC2 Difference	Jaccard	-0.039	0.006	-6.351	<0.0001		
PC3 Difference	Jaccard	-0.202	0.007	-29.742	<0.0001		
Intercept	Sørensen	0.223	0.06	3.713	0.0002	-5864.2	-5826.4
PC1 Difference	Sørensen	-0.061	0.004	-15.238	<0.0001		
PC2 Difference	Sørensen	-0.035	0.006	-5.499	<0.0001		
PC3 Difference	Sørensen	-0.206	0.007	-30.437	<0.0001		
Random Effects (SD)							
Location Intercept	Jaccard	0.1867					
Location Intercept	Sørensen	0.1862					
Model Fit Statistics							
Dispersion	Jaccard	20.1					
Dispersion	Sørensen	16.2					

Table 3.6. Summary of beta regression generalized mixed models that were used to estimate the effects of differences in the first three principal components (PCA1, PCA2, PCA3), calculated from soil physiochemical properties, on Jaccard and Sorenson community distance indices among the 64 soil samples.

Putative Species	Location Occurrence	Sample Occurrence	#ASVs	% of Nematoda Reads
Acrobeloides_DC_sp1	11	52	79	7.3874
Panagrolaimus_DC_sp1	11	33	190	21.4522
Robustodorus_arachidis	10	38	271	1.6336
Utahnema_DC_sp4	10	38	46	2.9627
Acrobeles_DC_sp2	9	46	181	25.9434
unknown_Dorylaimida_sp.	9	41	81	0.7038
Acrobeles_DC_sp1	9	39	53	4.0226
Ecumenicus_DC_sp7	9	26	29	0.5438
Quinisulcius_DC_sp1	9	17	23	0.4317
Aphelenchus_DC_sp2	9	19	20	1.7248
Acrobeloides_DC_sp6	9	36	16	0.8140
Nothacrobeles_borregi	9	21	9	0.1165
Nothacrobeles_triniglarus	8	13	19	0.2350
Zeldia_sp._JB_140	8	22	11	0.2874
Dirofilaria_immitis	8	14	9	0.0884
Acrobeloides_sp4	7	29	24	1.0465
Utahnema_DC_sp3	7	25	14	1.2109
Acrobeloides_sp._FHD001	7	9	8	0.0044
Acrobeloides_DC_sp5	6	14	20	0.2815
Aporcelaimellus_DC_sp6	6	17	8	0.2324
Cervidellus_sp._1_HMM2018	5	28	80	6.6942
Acrobeles_sp._JB_132	5	15	21	0.3277
unknown_Aphelenchus_sp.	5	22	16	0.4121
unknown_Ecumenicus_sp.	5	15	16	0.8966
Placodira_lobata	5	15	12	0.3644
Nothacrobeles_spatulatus	5	7	8	0.0771
Solididens_sp._DoryN.n	5	9	8	0.0185
unknown_Longidorella_sp.	5	8	8	0.0223
Microdorylaimus_DC_sp1	5	11	7	0.4725
Quinisulcius_DC_sp2	5	8	7	0.0388
Nothotylenchus_similis	5	7	6	0.0297
unknown_Stegelletina_sp.	5	9	1	0.1613
Ditylenchus_sp._85C1	4	10	22	0.0765
unknown_Panagrolaimidae_sp.	4	7	6	0.0289
Stegelleta_DC_sp2	4	6	5	0.0331
unknown_Plectus_sp.	3	3	11	0.0832
Tylencholaimus_mirabilis	3	9	8	0.1724

Caenorhabditis_elegans	3	3	1	0.0048
Discolaimus_major	3	4	1	0.0709
Aphelenchus_DC_sp1	2	7	27	1.2143
Robustodorus_helicus	2	6	18	0.0552
Carcharodiscus_banaticus	2	6	7	0.1929
Stegelleta_DC_sp1	2	6	7	0.0079
Ditylenchus_sp._20181130DIS3	2	3	4	0.0021
unknown_Tylenchidae_sp._1_HMM2018	2	2	4	0.0219
Acrobeles_singulus	2	3	3	0.0046
Acrobeloides_sp._DWF_1106	2	4	3	0.0105
Cervidellus_cervus	2	6	3	0.0085
unclassified_Tylencholaimus	2	2	3	0.0007
Acrobeles_sp._CR_2010	2	8	2	0.0682
Anaplectus_porosus	1	1	1	0.0050
Aporcella_vitrinus	1	1	1	5.96E-04
Axonchium_propinquum	1	1	1	0.0017
Cervidellus_neftasiensis	1	1	1	5.96E-04
Mesorhabditis_paucipapillata	1	1	1	0.0424
Nagelus_obscurus	1	1	1	0.0018
Tylencholaimus_sp._1_HMM2018	1	1	1	0.0052
unknown_Dorylaimida_DC_sp2	1	1	1	0.0016
unknown_Rhabditida_sp.	1	2	1	0.0260

Table 3.7. Species-level classifications in descending order of number of occurrences among all locations. Sample occurrence, number of ASVs, and relative abundance of reads representing each putative species are also included in the table.

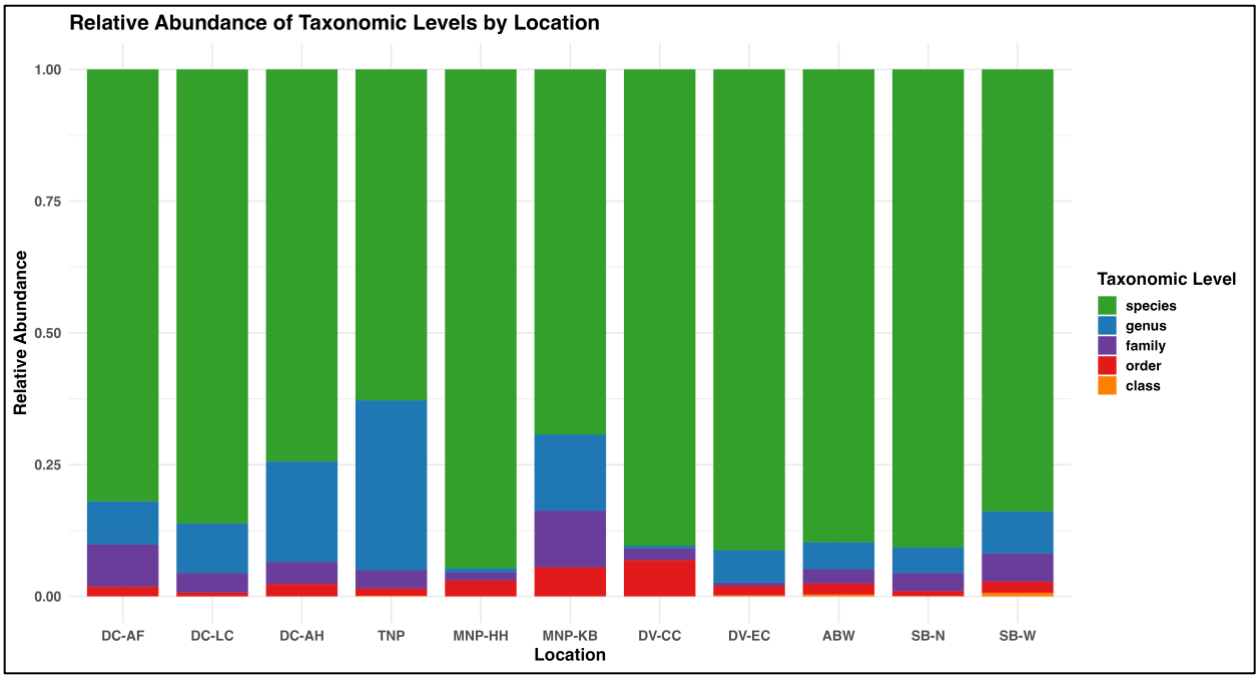

Figure 3.1. Proportion of reads classified to species, genus, family, order, and class level at each collection location. See Table 1 for location codes.

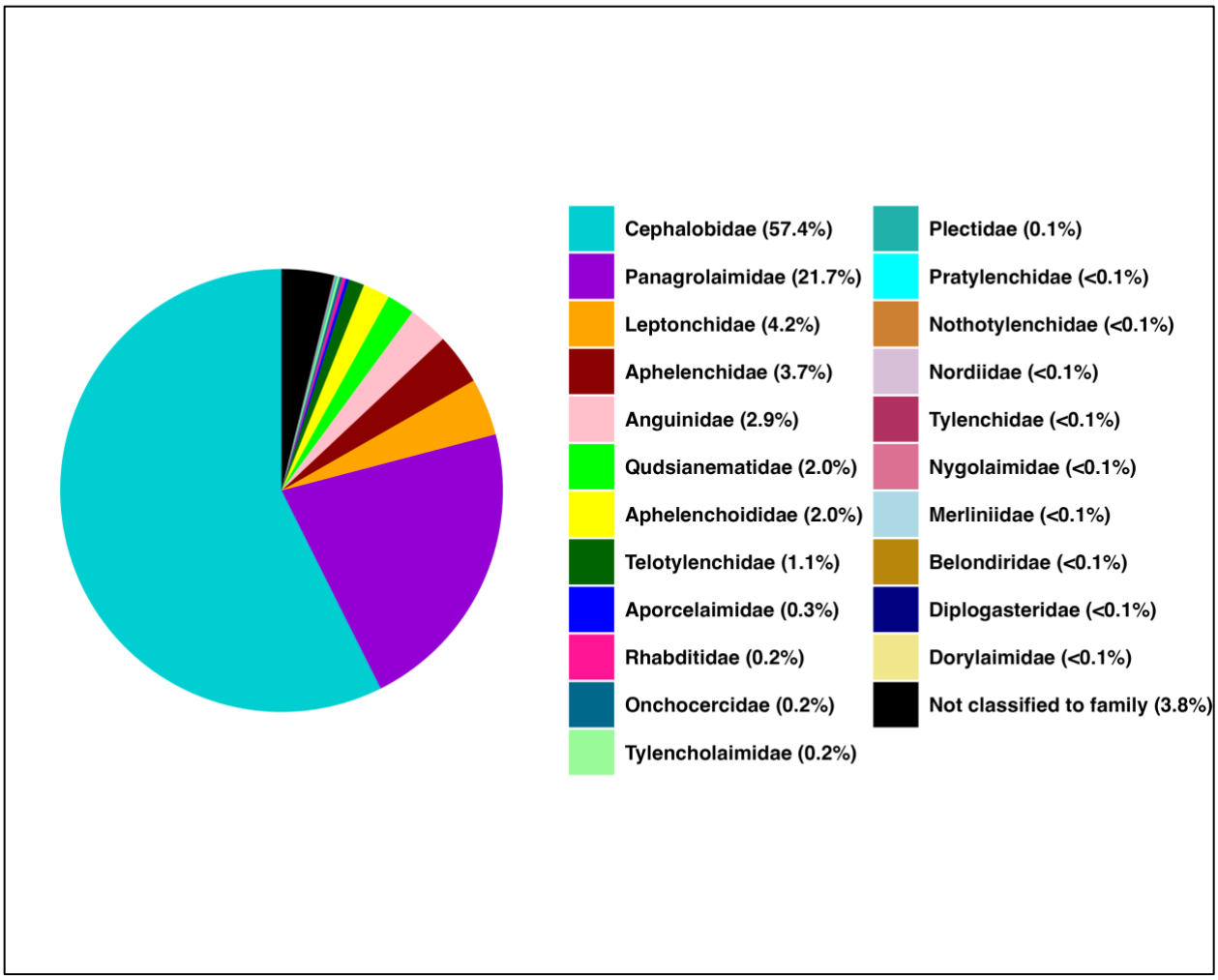

Figure 3.2. Pie chart showing family-level relative abundance among all sequence reads that were classified within Classes Chromadorea and Enoplea. Sequence reads that were not successfully classified to family level at 70% confidence are colored black.

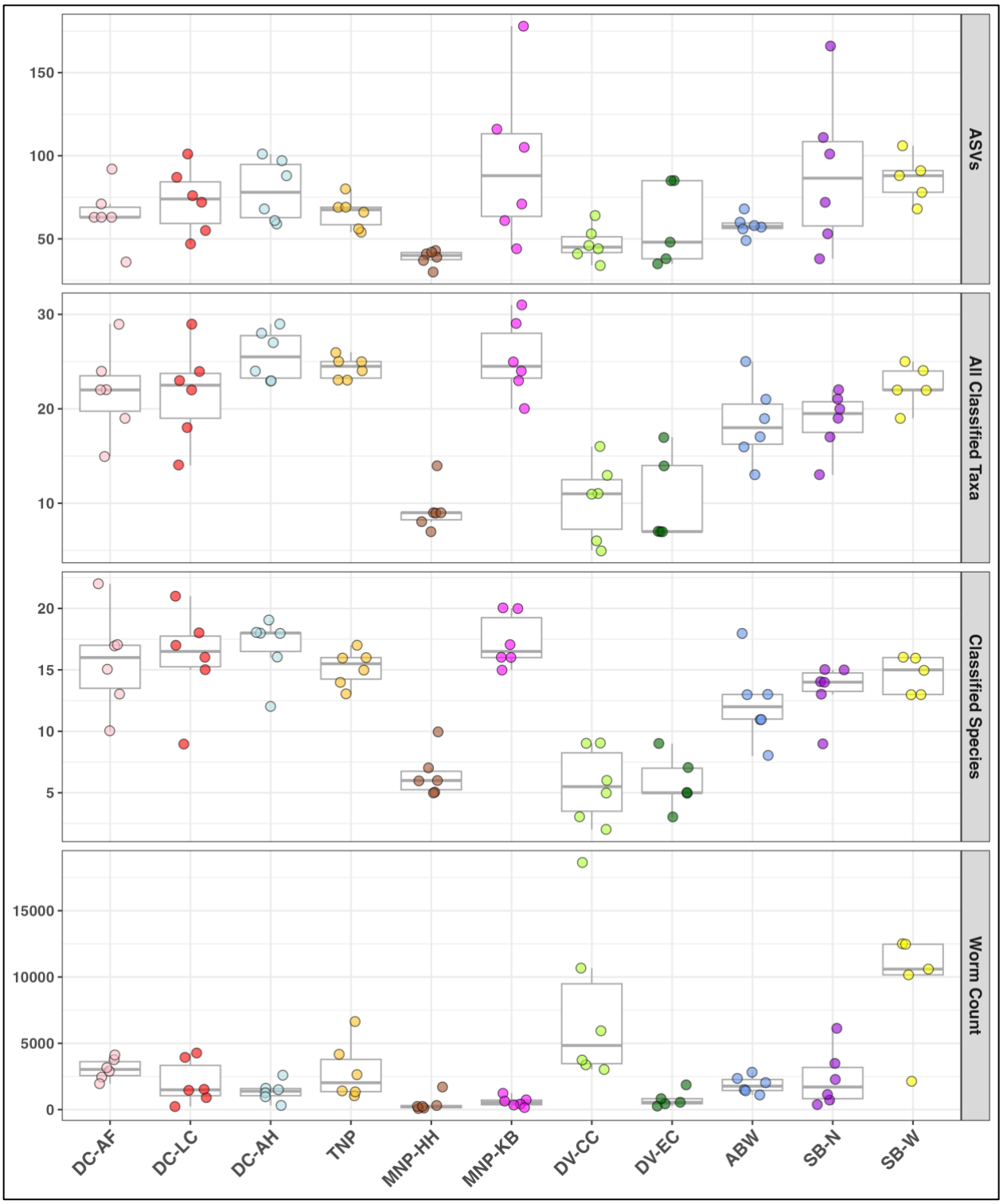

Figure 3.3. Boxplot showing number of ASVs, total number of taxonomic categories, and number of species-level classifications, as well as estimated number of worms in soil samples at all collection locations.

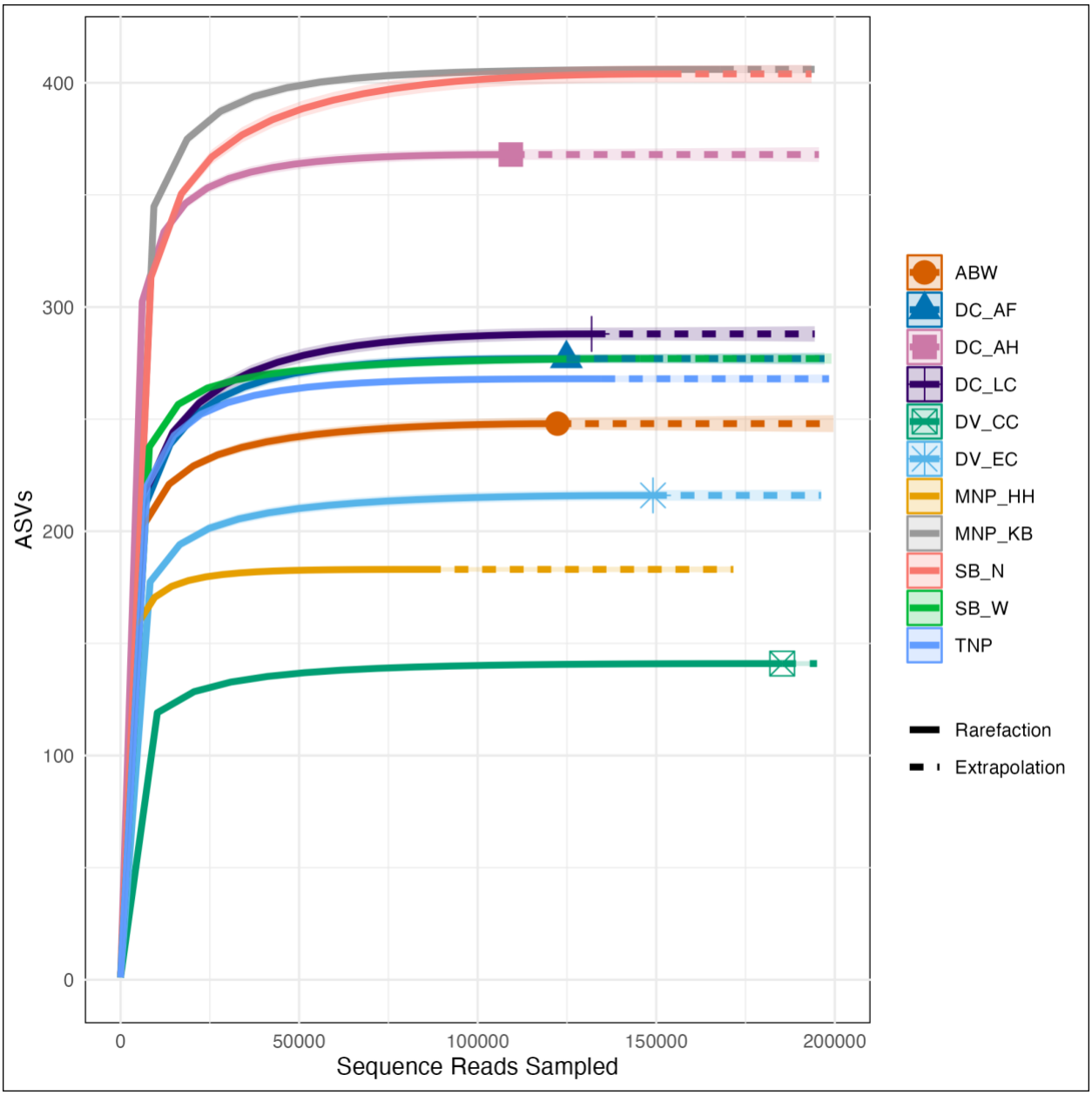

Figure 3.4. ASV accumulation curves by location. Solid lines indicate reads sampled from the actual data, while dotted lines predict (extrapolated) diversity if sampled beyond the limits of available reads.

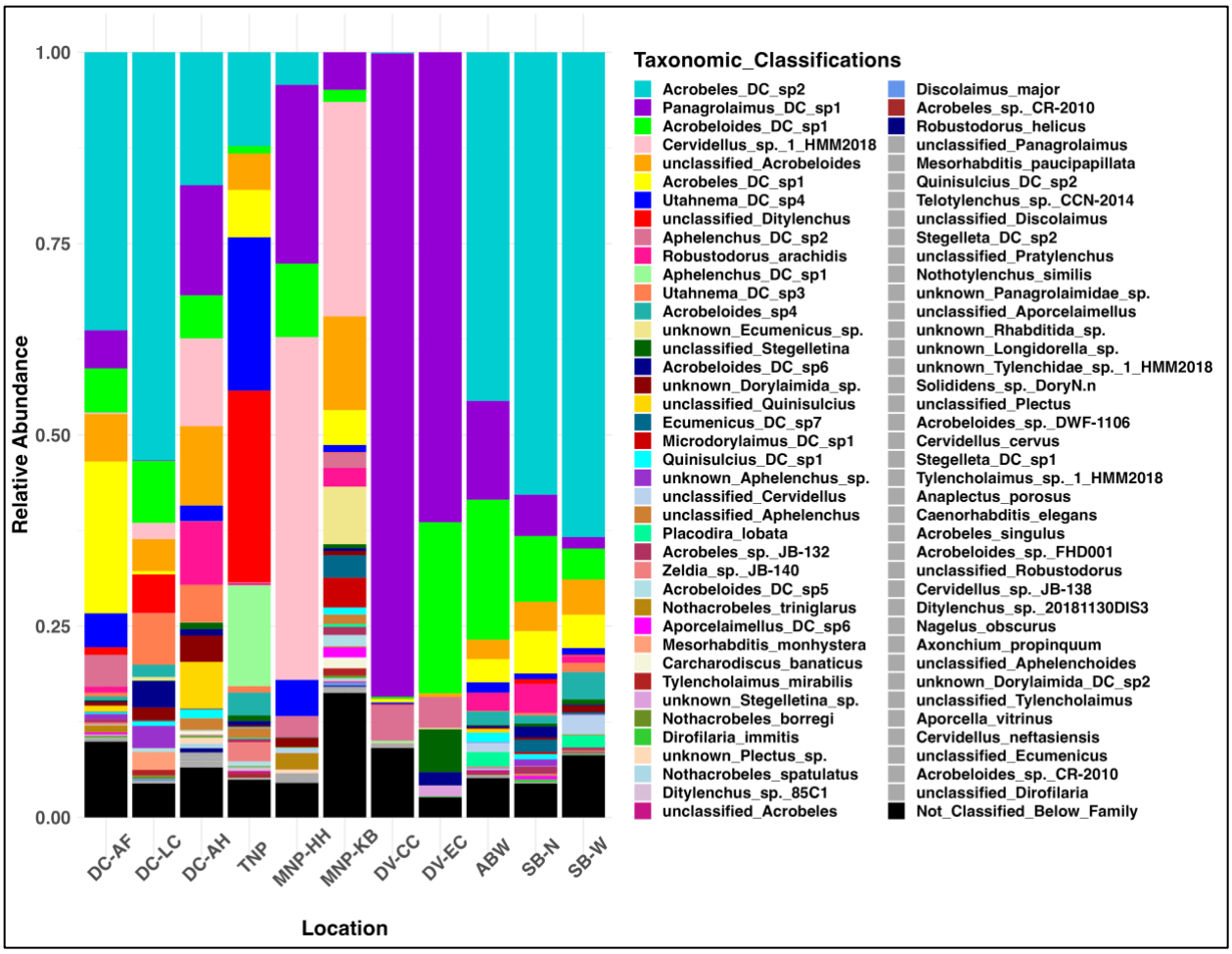

Figure 3.5. Stacked barplot of sequence classifications (70% confidence), by soil sample. Taxonomic categories are in order of decreasing relative abundance. Species with relative abundance less than 0.05% are colored grey. Taxonomic categories that could not be classified below family level are colored black.

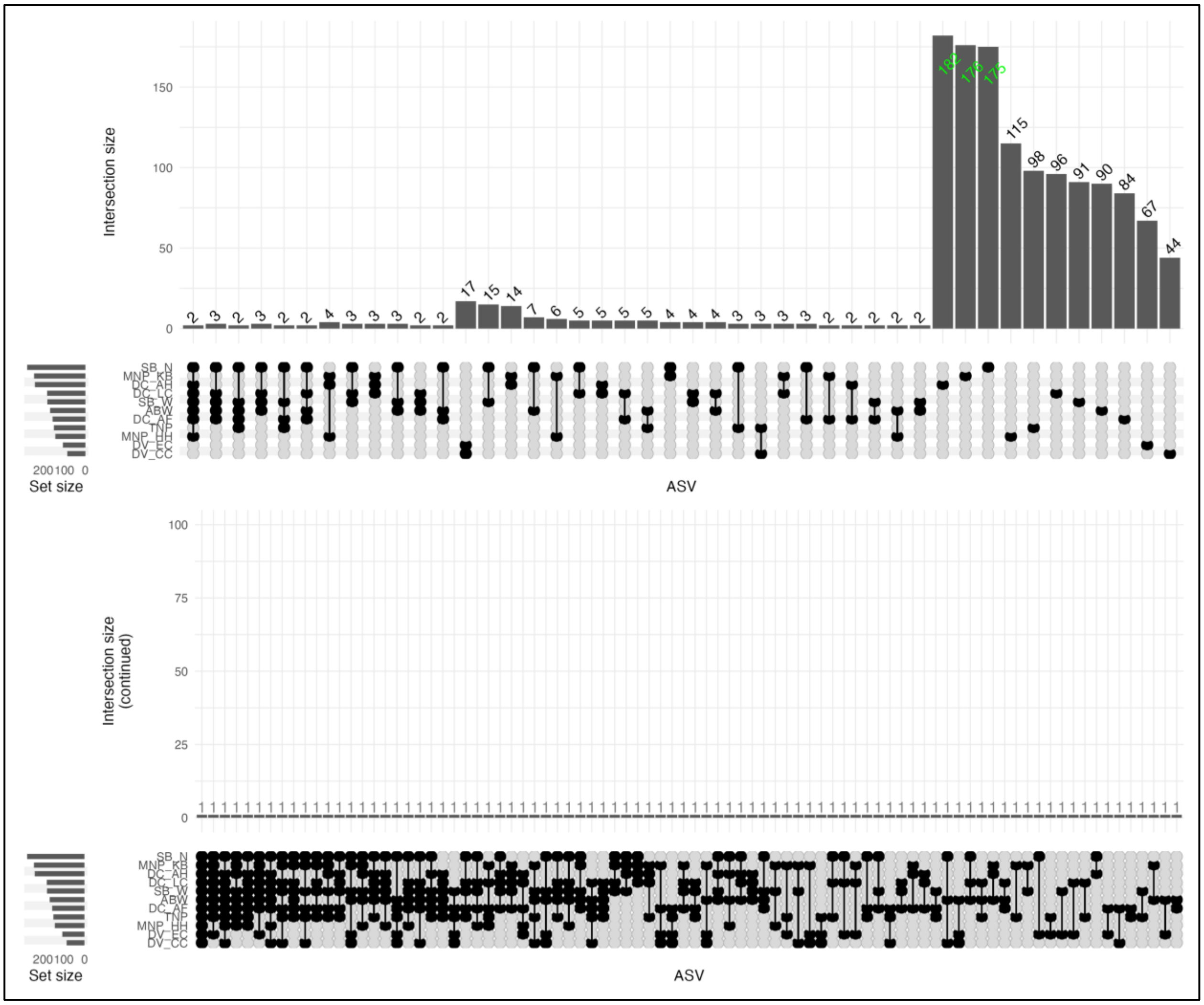

Figure 3.6. UpSet plot illustrating the overlap of ASVs across different sampling locations. The horizontal bar plot (left panel) represents the total number of ASVs detected at each location. The matrix (bottom panel) shows the locations included in each intersection, indicated by connected circles. The vertical bar graph (top panel) displays the number of ASVs shared among the intersecting locations (intersection size).

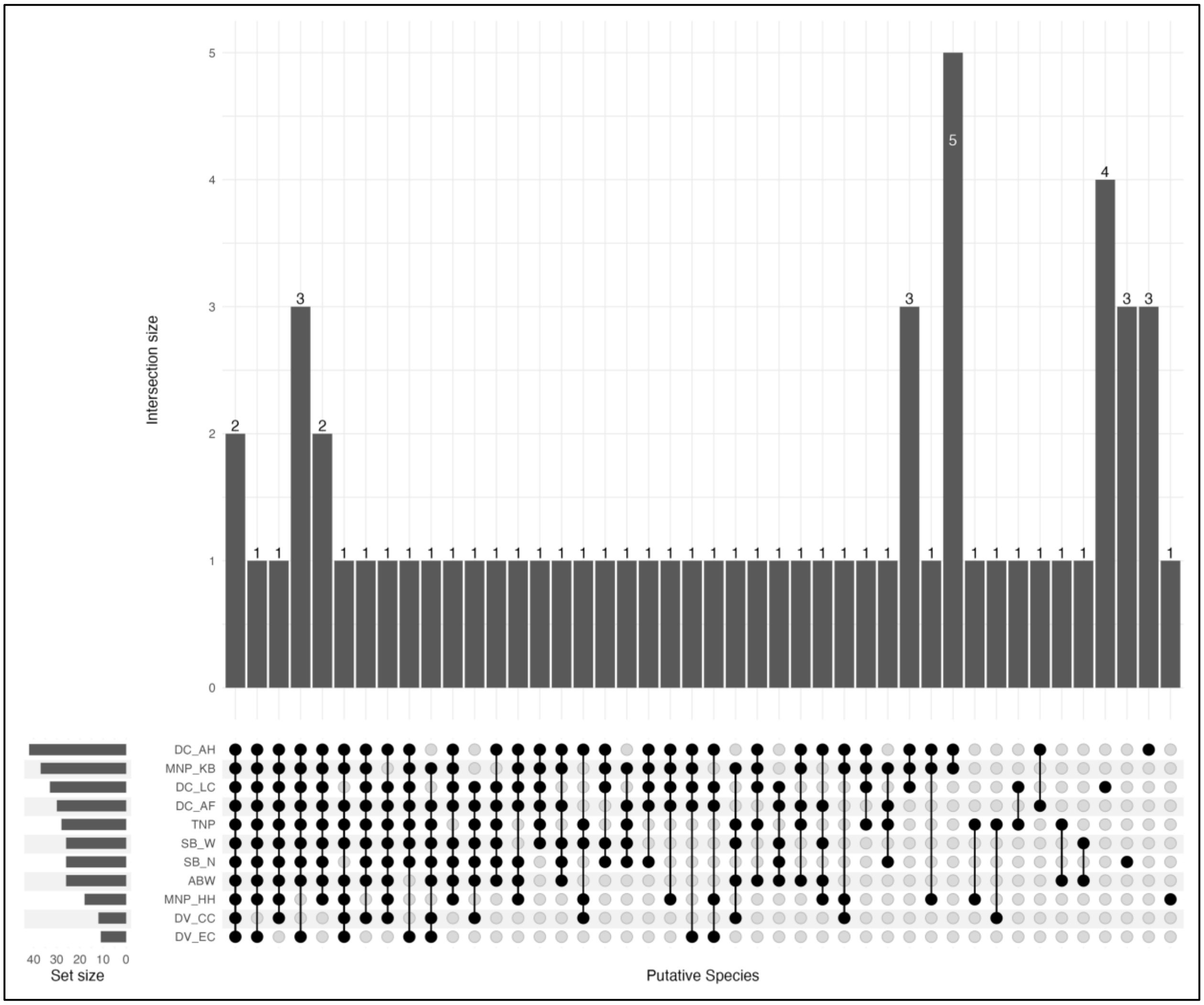

Figure 3.7. UpSet plot illustrating the overlap of species-level classifications across different sampling locations. The horizontal bar plot (left panel) represents the total number of putative species detected at each location. The matrix (bottom panel) shows the locations included in each intersection, indicated by connected circles. The vertical bar graph (top panel) displays the number of putative species shared among the intersecting locations (intersection size).

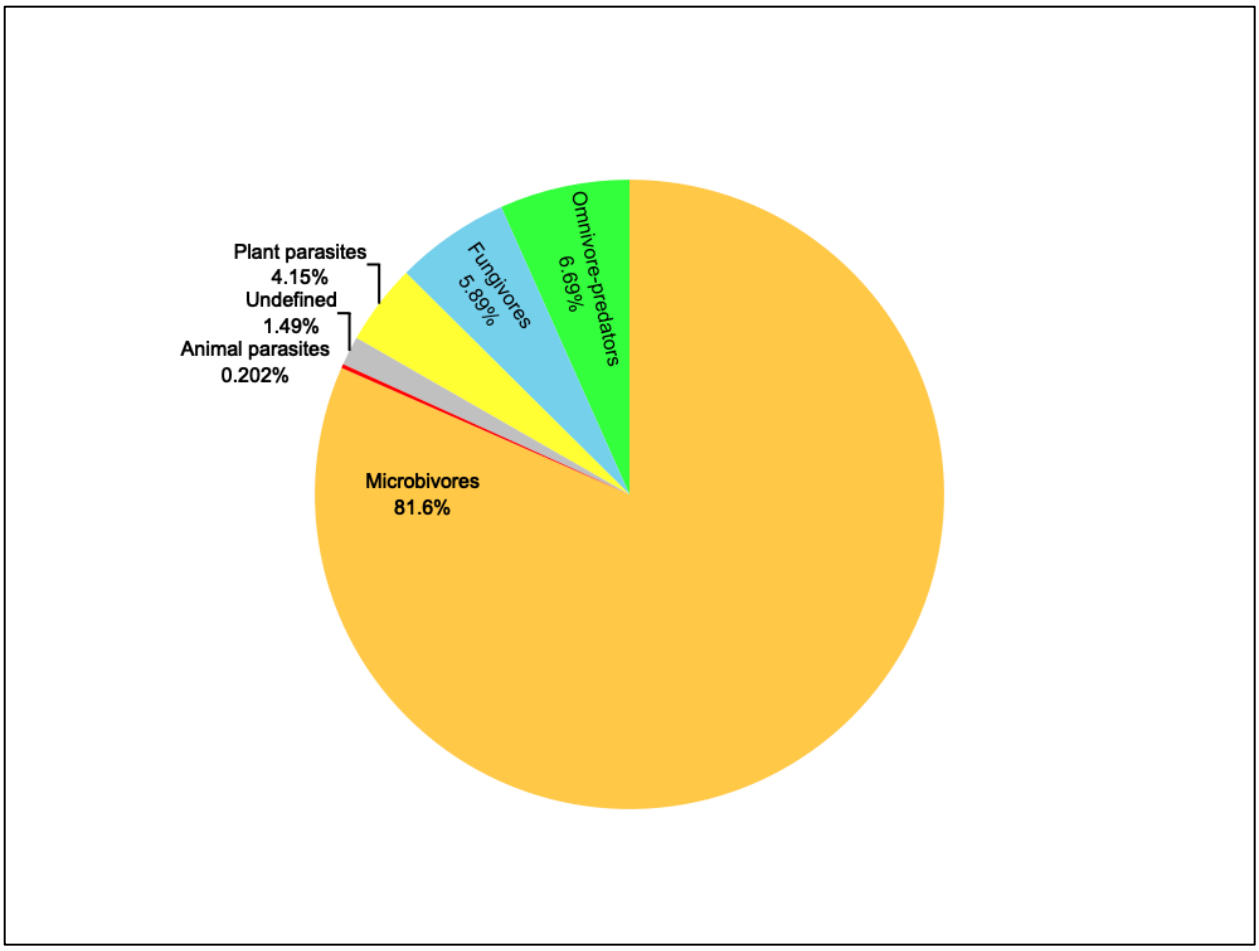

Figure 3.8. Piece chart depicting proportions of trophic group categories among all sequence reads in all samples.

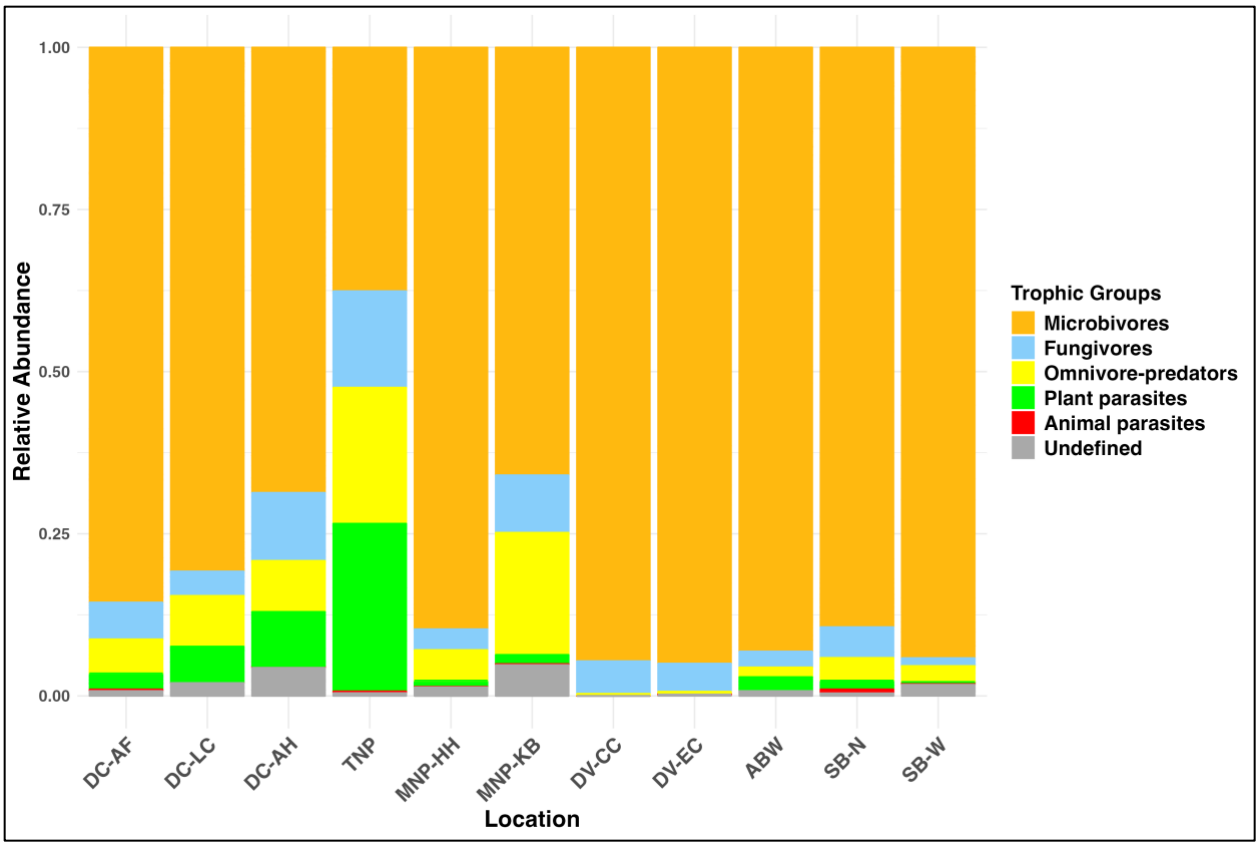

Figure 3.9. Stacked barplot of relative abundances of different trophic groups among collection locations. Reads that were not classified to a taxonomic rank that was clearly indicative of a trophic group were labeled as “Undefined” and colored grey.

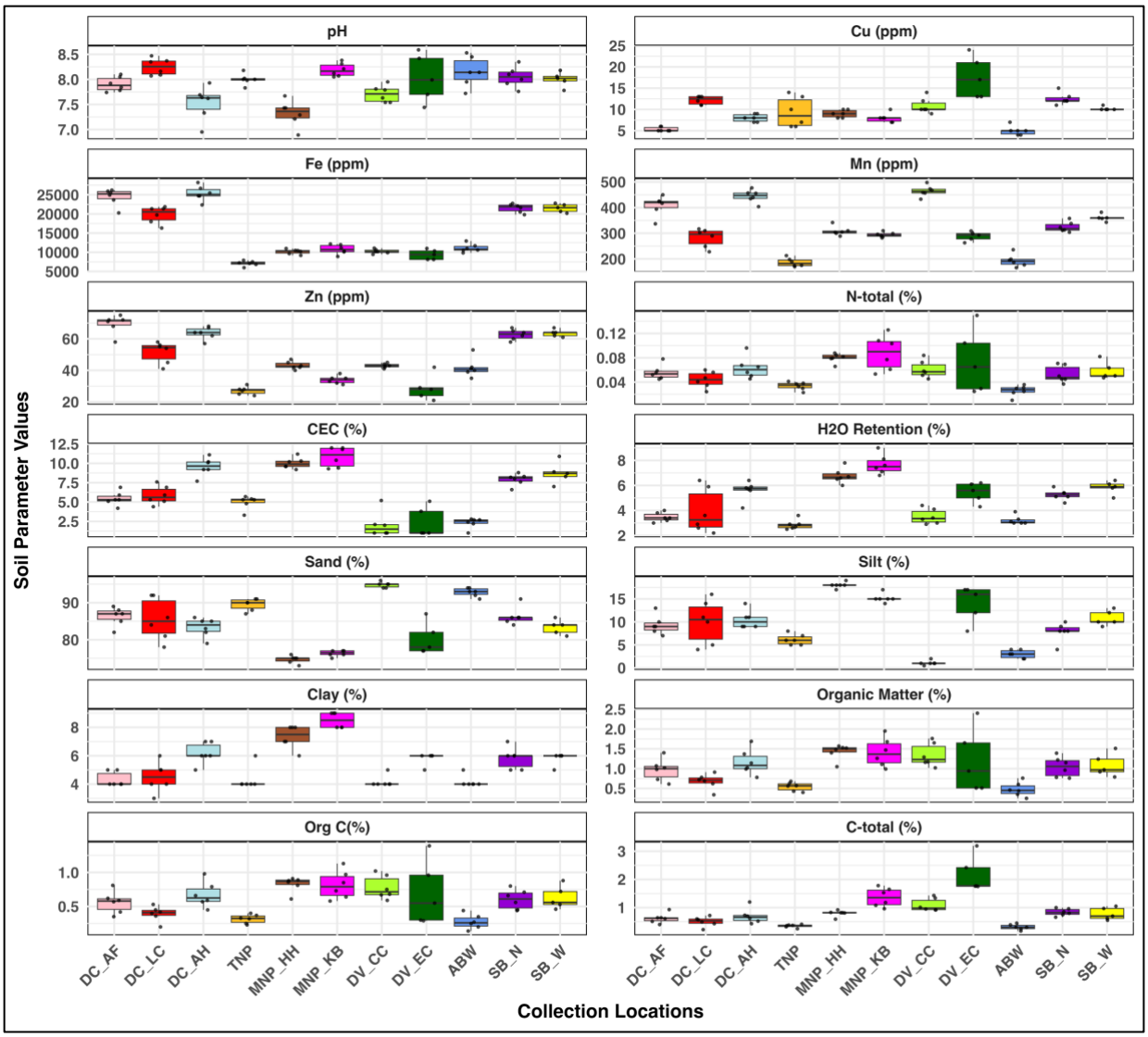

Figure 3.10. Boxplot depicting soil physicochemical properties across all locations. Each facet represents a different soil parameter, with units specified where applicable. Data points represent individual soil sample measurements.

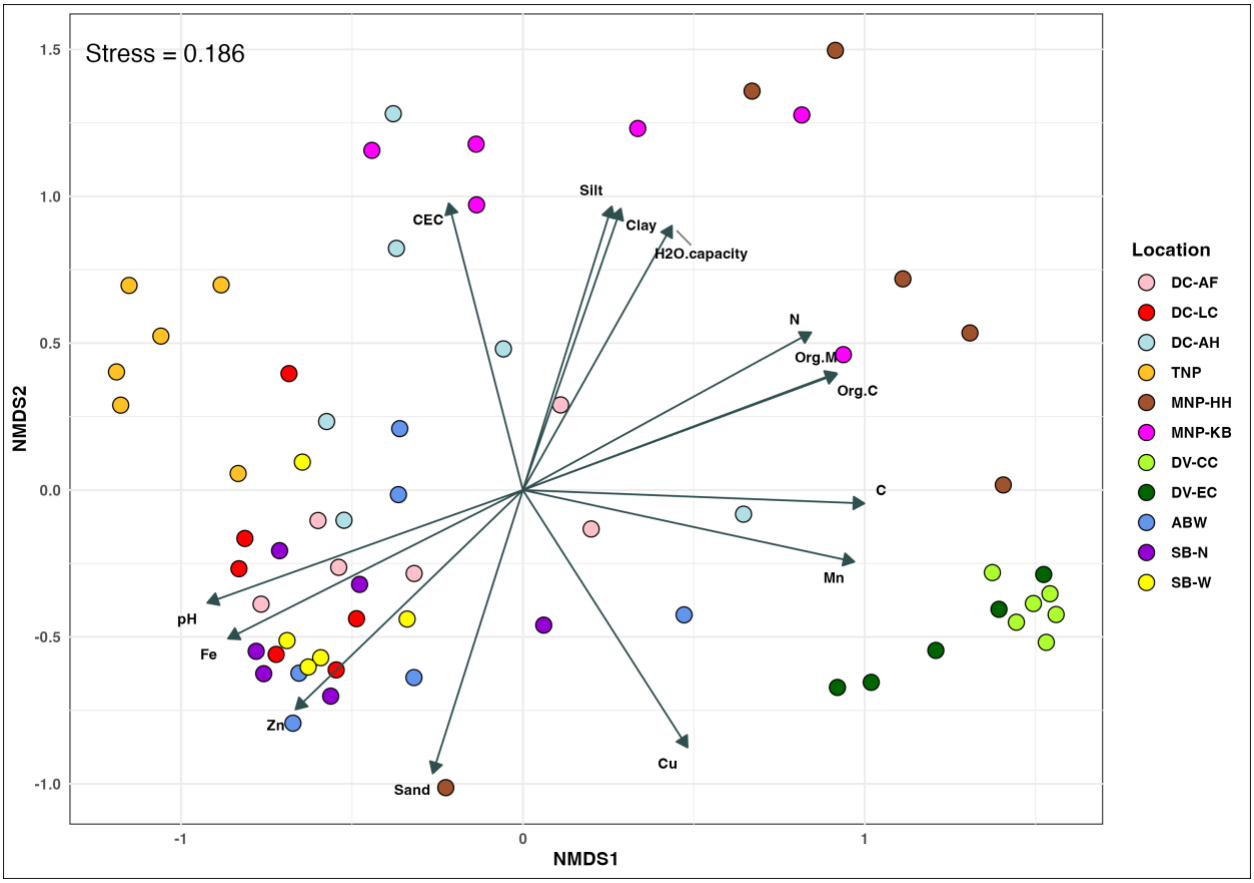

Figure 3.11. Non-metric Multidimensional Scaling (NMDS) ordination plot of nematode community composition based on Jaccard distance. Samples are represented by points, colored according to their geographic locations. Environmental vectors, derived from soil physicochemical data, are overlaid to indicate the direction and magnitude of correlations between soil parameters and community composition.

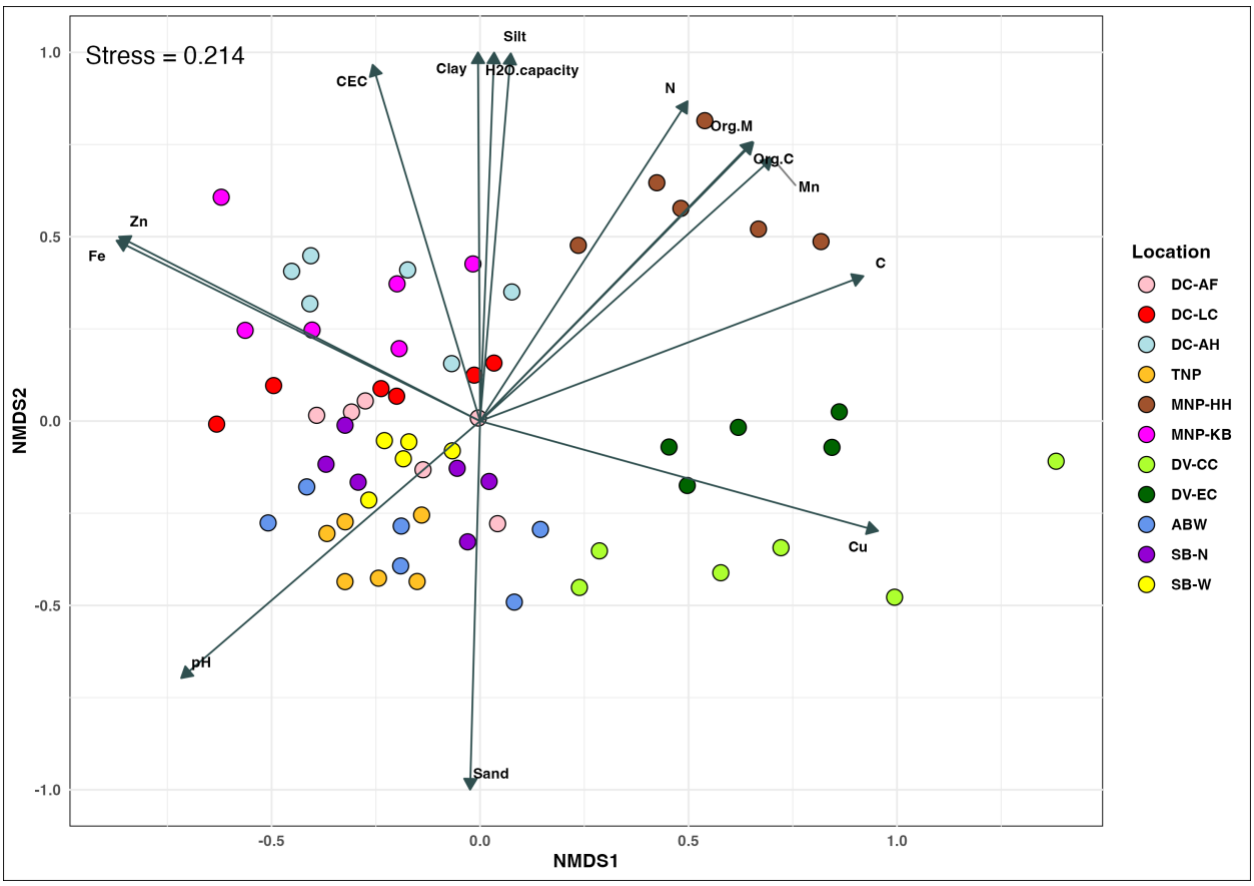

Figure 3.12. Non-metric Multidimensional Scaling (NMDS) ordination plot of nematode community composition based on Sørensen distance. Samples are represented by points, colored according to their geographic locations. Environmental vectors, derived from soil physicochemical data, are overlaid to indicate the direction and magnitude of correlations between soil parameters and community composition.

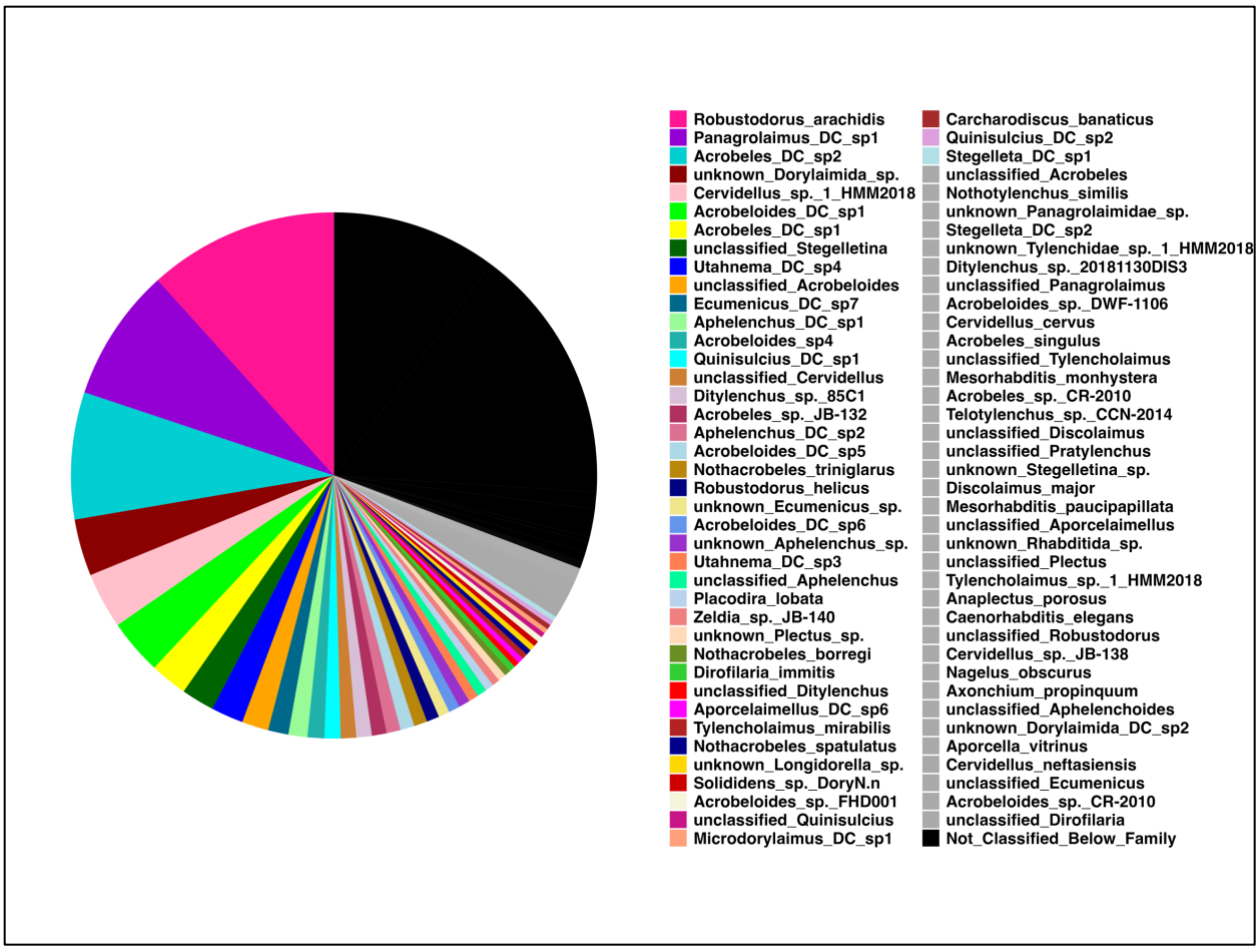

Figure 3.13. Pie chart depicting proportion of the 2319 ASVs that were assigned to each taxonomic category. Taxa with less than 0.25% of ASVs are colored grey. Taxa that could not be classified below Family level are colored black.

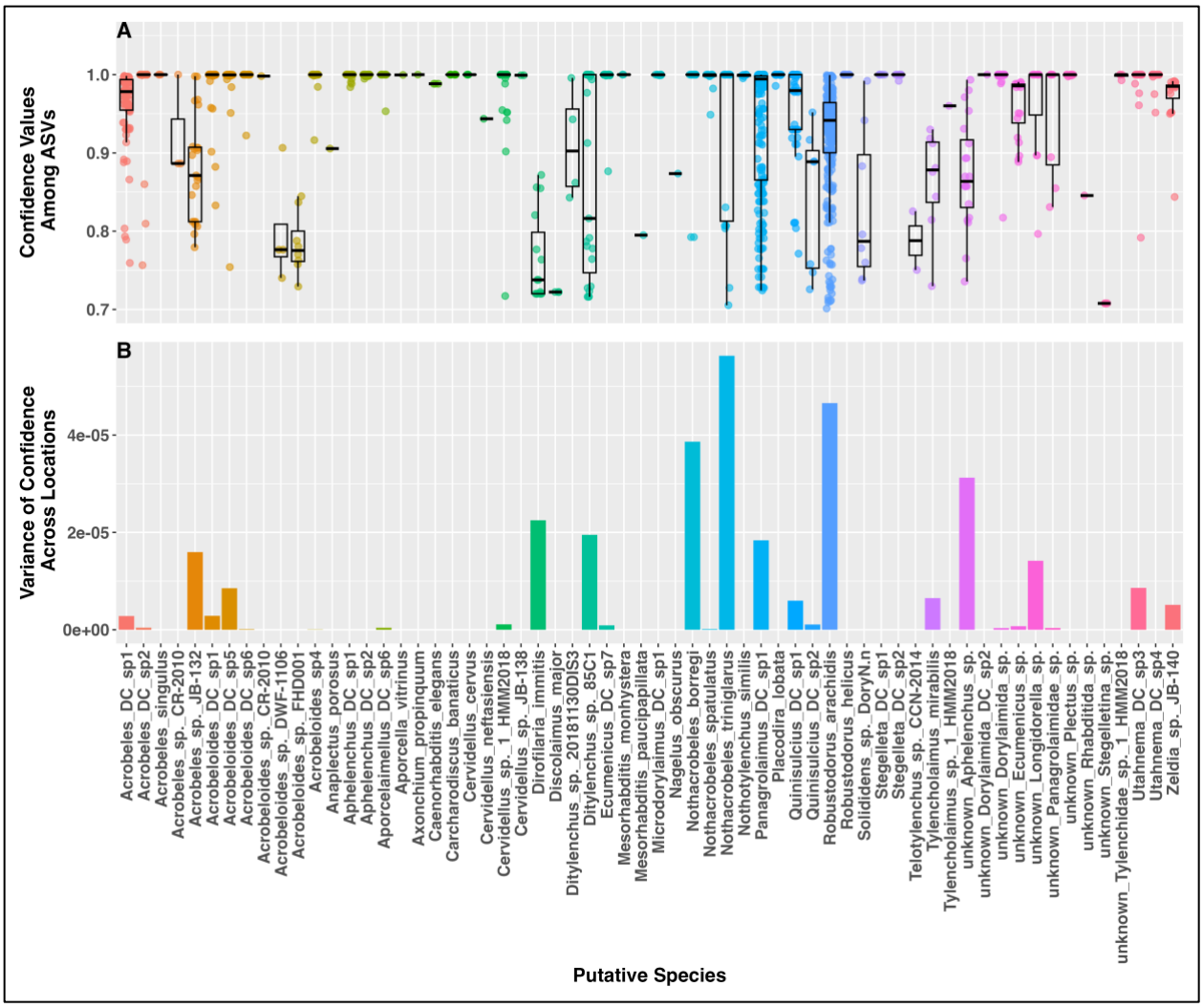

Figure 3.14. (**A**) Boxplot of confidence scores for each species-level classifications across all soil samples. Dots represent confidence scores of individual ASV that have been classified to the corresponding species on the x-axis. (**B**) Bar plot showing variance of confidence scores among collection locations for each putative species. Higher bars indicate inconsistency of confidence values among collection locations where a putative species is found.

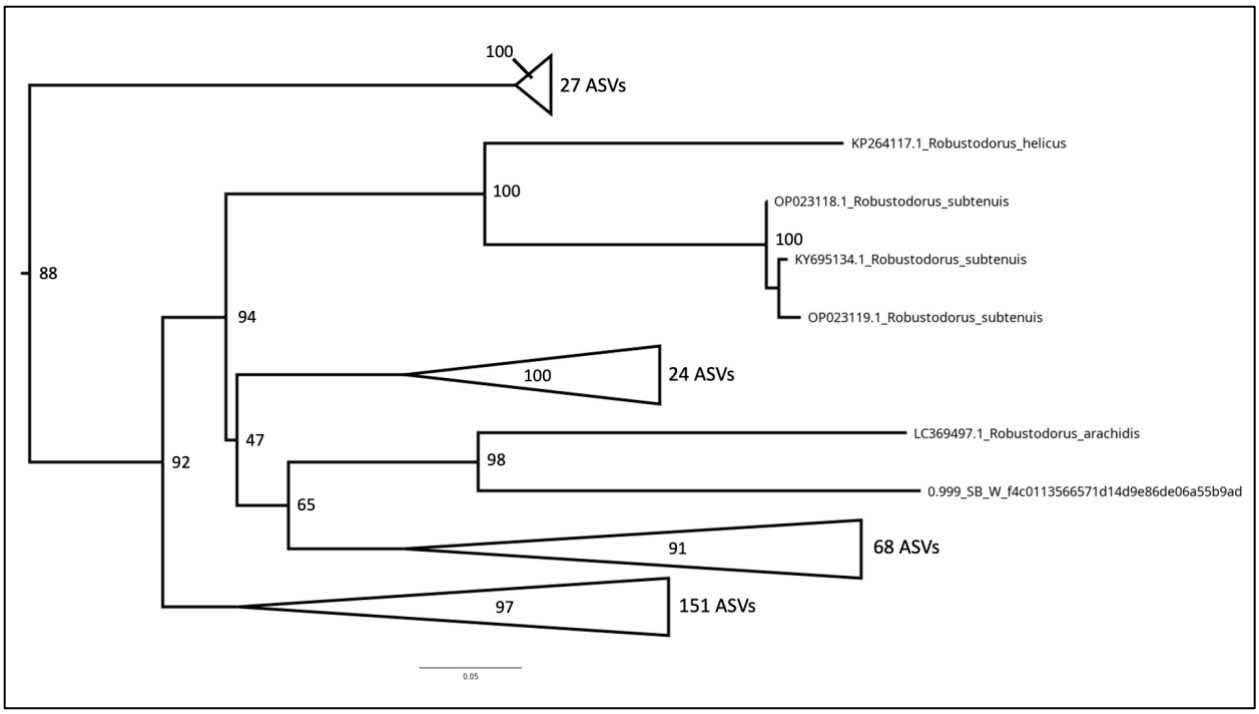

Figure 3.15. Maximum likelihood tree made from Aphelenchoididae and Aphelenchidae reference sequences and putative *Robustodorus arachidis* ASV sequences. Only the clade containing 5 *Genus Robustodorus* reference sequences and the 271 associated ASVs is shown. One ASV, designated by a unique alphanumeric identifier, is found siter to the *Robustodorus arachidis* refeence sequence and the remaining 270 ASVs resolve into 4 monophyletic ASV groups. Triangles represent collapsed ASV groups. The distance metric is average number of substitutions per site.