High-throughput amplicon sequencing of rRNA genes requires a copy number correction to accurately reflect the effects of management practices on soil nematode community structure

B. J. Darby, T. C. Todd, M. A. Herman

Digital Object Identifier (DOI): doi:10.1111/mec.12480

Publisher's Link: http://onlinelibrary.wiley.com/doi/10.1111/mec.12480/abstract

High-throughput amplicon sequencing of rRNA genes requires a copy number correction to accurately reflect the effects of management practices on soil nematode community structure

B. J. Darby[1,2], T. C. Todd[3], M. A. Herman[2]

[1]Department of Biology, 10 Cornell St. Stop 9019, University of North Dakota, Grand Forks, ND 58202

[2]Division of Biology, Kansas State University, Manhattan, KS 66506

[3]Department of Plant Pathology, Kansas State University, Manhattan, KS 66506

Correspondence: Brian Darby, 10 Cornell St. Stop 9019, Grand Forks, ND 58202,

brian.darby@und.edu

Keywords (4-6): tallgrass prairie, nematodes, Konza prairie, bar-coding, nitrogen enrichment, burning

Running title: Amplicon sequencing of grassland nematodes

Abstract

Nematodes are abundant consumers in grassland soils, but more sensitive and specific methods of enumeration are needed to improve our understanding of how different nematode species affect, and are affected by, ecosystem processes. High-throughput amplicon sequencing is used to enumerate microbial and invertebrate communities at a high level of taxonomic resolution, but the method requires validation against traditional specimen-based morphological identifications. To investigate the consistency between these approaches, we enumerated nematodes from a 25-year field experiment using both morphological and molecular identification techniques in order to determine the long-term effects of annual burning and nitrogen enrichment on soil nematode communities. Family-level frequencies based on amplicon-sequencing were not initially consistent with specimen-based counts, but correction for differences in rRNA gene copy number using a genetic algorithm improved quantitative accuracy. Multivariate analysis of corrected sequence-based abundances of nematode families was consistent with, but not identical to, analysis of specimen-based counts. In both cases, herbivores, fungivores, and predator/omnivores generally were more abundant in burned than non-burned plots, while bacterivores generally were more abundant in non-burned or nitrogen enriched plots. Discriminate analysis of sequence-based abundances identified putative indicator species representing each trophic group. We conclude that high-throughput amplicon sequencing can be a valuable method for characterizing nematode communities at high taxonomic resolution as long as rRNA gene copy number variation is accounted for and accurate sequence databases are available.

Introduction

The tallgrass prairie was the dominant vegetation of much of the North American Great Plains, but the four percent of the original area that remains (Samson and Knopf 1994) is fragmented and mostly limited to a few reserves. Many of these remnants experience environmental conditions that are not representative of a native prairie habitat, such as nitrogen enrichment, restricted burning, and absence of native grazing. Frequent burning removes surface litter, promotes warm season C_4 grasses, reduces soil moisture, and deters woody encroachment and invasion by exotics (Collins et al. 1998). In nitrogen-limited systems such as the tallgrass prairie, elevated nitrogen levels tend to reduce plant species richness (Gibson et al. 1993).

Nematodes are a diverse and functionally important component of grassland ecosystems (Bardgett et al. 1999; Todd et al. 2006, Yeates et al. 2009). They occupy multiple levels of the soil food web, with individual species specializing on bacteria, fungi, plant roots, or other soil invertebrates as a food source (Yeates et al. 1993). Nematodes moderate their prey populations, mobilize organic-bound nutrients, and influence plant communities (Ingham et al. 1985; Wardle et al. 2004). Nematode communities are responsive to various disturbances and land management practices, and have been utilized widely as ecological indicators (Neher 2001). In tallgrass prairies, burning regime and nitrogen status are important determinants of nematode community structure, with total numbers of herbivores and bacterivores responding positively to both types of disturbance (Todd et al 1996). It is typical to characterize nematode community responses at a coarse level of resolution (e.g. trophic group or family), although

evidence is accumulating that responses within trophic or family groupings are not uniform (Fiscus and Neher 2002; Jones et al. 2006).

Nematodes from the soil are difficult, and in some cases impossible, to identify to species. Many species cannot be identified if only juveniles are present or if one sex is absent. This is part of the reason that we have an incomplete understanding of the dynamics of grassland nematode species. A sensitive, high-resolution method for assessing nematode community composition would promote a more comprehensive understanding of how the soil environment affects the distribution of nematode species, as well as how different nematode species affect the soil ecosystem. High-throughput amplicon sequencing (elsewhere called "metagenomic" or "metagenetic" pyrosequencing) has been proposed as a promising solution for rapid identification of nematode communities (Porazinska et al. 2009, Creer et al. 2010). Briefly, bulk DNA is purified from a community of nematodes that have been extracted from soil. A phylogenetically informative locus (such as 18S SSU rRNA [Donn et al. 2011] or the Internally Transcribed Spacer Unit ITS [Powers et al. 1997]) is amplified from the community DNA and amplicons are sequenced on a next-generation, massively parallel sequencing platform. Amplicon pyrosequencing usually produces several thousand reads per sample that are screened for quality metrics, sorted into taxonomic groups, and identified taxonomically based on existing databases. While this approach has produced suitable qualitative results (i.e. identification of species assemblages), significant challenges remain for obtaining quantitative data (i.e. relative abundances of species present) for use in community-level analyses and soil food web diagnostics (Porazinska et al. 2009). Variation in rRNA gene copy number among nematode species likely accounts for much of the quantification problem because the PCR

amplification and emulsion PCR for pyrosequencing appears to be fairly reproducible (Porazinska et al. 2010).

The primary objective of this study was to determine the long-term effects of annual burning and elevated nitrogen inputs on soil nematode community composition. Early responses of nematodes to short-term (eight years) of annual burning and nitrogen enrichment showed that bacterivores were more abundant in nitrogen enrichment plots, but total herbivores were more abundant in the burned plots than non-burned plots only in the absence of nitrogen enrichment (Todd 1996). We hypothesized that the effect of annual burning and nitrogen enrichment on nematode trophic groups will be more distinct after 25 years in that herbivores and fungivores will be more abundant in burned than non-burned plots across both seasons. We used high-throughput amplicon sequencing to identify the families and species that are potential indicators of burning or nitrogen enrichment. A second objective of this study was to develop the analytical methods that are necessary to determine whether high-throughput amplicon sequencing could accurately replace traditional specimen-based enumerations. We hypothesized that, given adequate correction factors for variation in rRNA operon copy number, amplicon sequencing data could be converted to “virtual specimen” counts that would provide comparable results to actual specimen counts.

Methods

Sampling regime, amplification and sequencing

This experiment utilized an established set of experimental plots at the Konza Prairie Biological Station LTER site in the Flint Hills region of northeastern Kansas, USA (Todd 1996; Jones et al. 2006). The field site consisted of sixty-four 12.5 m × 12.5 m plots arranged in four

replicate blocks that had been assigned a factorial treatment structure (presence/absence) of burning (annually in spring), mowing (annually in summer), nitrogen (10g N/m^2 annually in spring), and phosphorous (1g P/m^2 annually in spring) amendments. For this experiment, we sampled only from the 16 plots (four plots per block) that were not mowed and had not received additional phosphorous. Thus, we had two treatments (with and without spring burning, with and without spring nitrogen enrichment) and four replicate plots per treatment combination, or 16 plots total. The design structure was a split plot, with burning as the whole plot treatment and nitrogen as the subplot treatment.

In spring (June) and fall (October) of 2010, 20 soil cores (2.4 cm dia. from 0 to 10 cm) from each plot were collected and pooled to produce one composite sample per plot. Nematodes were extracted by sucrose-flotation from three subsamples of each composite sample (each approximately 100 cm^3 in volume). In two of the subsamples ("A" and "B"), nematodes were counted and 100 individuals were identified to family by microscopy using temporary fresh-mounts of live specimens ("specimen-based counts"). These individuals were returned to their respective samples, and the family identifications were attributed to one of four main feeding groups: herbivores, fungivores, bacterivores, and predator-omnivores (Todd 1996). The third subsample was preserved in DESS solution (Yoder et al. 2006) for archiving. Thus, we use for this experiment two subsamples from each of 16 plots collected in two seasons, or 64 subsamples total. In addition to the 64 subsamples from field-extracted nematodes, two additional "test" samples were created by picking 10 individuals each from 11 different species in culture that were isolated from Konza prairie: *Oscheius tipulae (*strain KS599*), Oscheius sp.* FVV-2 (KS600), two *Mesorhabditis sp. MR1 (KS601), Mesorhabditis sp. MR2*

(KS602), *Rhabditis sp. RA5 (*KS594*)*, *Protorhabditis sp. RA9*, *Pristionchus pseudaerivorous (KS596)*, *Rhabditophanes sp. RA8 (KS597)*, *Panagrolaimus sp. (KS598)*, *Cephalobus sp. (CE1)*, and *Acrobeloides sp. (KS586)*. All ten individuals were females from a bleach-synchronized cohort that were in their first day as adults.

The individuals in the temporary fresh-mounts were recovered live, returned to all of the other individuals that had been extracted (plus the co-extracted plant, fungal, and non-nematode invertebrates), and genomic DNA was extracted from both subsamples A and B separately (and the two test samples) using the MO-BIO Tissue and Cells DNA kit (Carlsbad, CA). This kit utilizes bead-beating with jagged-edged garnet beads to disrupt tissue and was used according to the manufacturer's specification except that glass Pasteur pipettes were used to transfer nematodes in the bead-beating solution to prevent adherence to the pipette surface. A 350 bp region of the 18S small-subunit (SSU) rRNA gene (containing two highly variable regions) was amplified from each sample's genomic DNA. The forward primers used for amplification included a 30-bp sequencing region specific for 454-titanium sequencing (5'- CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG -3'), a 10-bp multiplex identifier (MID), and a 24-bp forward primer that anneals to eukaryotic 18S rRNA (5'- GGT GGT GCA TGG CCG TTC TTA GTT -3'). The 10-bp MID barcode sequences were designed with the following attributes: 1) no barcode began with "G" (the last nucleotide of the sequencing key), 2) no nucleotides were repeated in tandem, and 3) all MID sequences differed from all other MID sequences by at least 3 nucleotides to facilitate error detection and correction. The reverse primer included a 26-bp sequencing primer (5'- CCT ATC CCC TGT GTG CCT TGG CAG TCT CAG -3') and a 22-bp reverse primer (5'- AGC GAC GGG CGG TGT GTA CAA A -3'). The forward primer used here is identical to NF1 used

in Porazinska et al. (2009). The reverse primer anneals 15 base pairs away from the primer 18Sr2b used in Porazinska et al. (2009) because 18Sr2b is known to not amplify a number of nematodes (mostly Rhabditidae) that have been previously isolated from Konza Prairie. Samples were amplified in 30 µl reactions with a final concentration of 1X high-fidelity buffer (containing 1.5 mM MgCl), 200 µM dNTPs, 0.2 µM each primer, 0.02 U Phusion high-fidelity DNA Polymerase (New England Biolabs Inc.) in the following cycle: 5 min at 98 °C, followed by 35 cycles of 20 seconds at 98 °C, 10 sec at 65 °C, 30 sec at 72 °C, and final extension for 7 min at 72 °C. Seventeen unique MID barcodes were used during amplification; one MID was used for both subsamples of the test sample, and the remaining sixteen MIDs were randomly assigned to one "A" subsample and one "B" subsample (but always of two different field samples). Amplicons from each subsample were cleaned and normalized using the SequalPrep Normalization Kit (Invitrogen) using two binding steps. The normalized amplicons from all "A" subsamples, plus one of the control samples, were pooled, cleaned once more with Purelink PCR clean-up kit (Invitrogen), and sequenced unidirectionallly on one half (one region) of a 454 pico-titre plate. Similarly, the normalized amplicons from all "B" subsamples, plus the other control sample, were pooled, cleaned, and sequenced on the other half (region) of the pico-titre plate. Both halves of the pico-tire plate were sequenced on the same run using Titanium chemistry according to manufacturer's specifications.

Read processing

We used custom designed scripts to pre-process reads prior to analysis. This pipeline was performed in a BASH script that was run in Ubuntu 11.04 Natty Narwhal. First, reads were screened utilizing several of the tools in BioPython library (Cock et al. 2009) for the following

quality criteria: 1) valid 10-bp barcode, 2) no ambiguous base calls ("N"), and 3) final read length between 250 and 450 bp. Filter-passed reads were trimmed of amplification primers using Cutadapt 1.0 (Martin 2011). Trimmed reads were de-replicated with USEARCH (i.e. clustered to 100% similarity, Edgar 2010) and screened for potential chimera using the reference-guided UCHIME feature of USEARCH. The reference database we used included 26,894 non-redundant eukaryotic 18S rRNA sequences (of which 1,331 were nematode sequences) from the ARB-SILVA database (Pruesse et al. 2007), obtained on-line 21 April 2012. The non-redundant set of chimera-free read clusters was searched against the reference database using SSAHA2 (Ning et al. 2001) to obtain the nearest match and assign this taxonomy to all identical reads. The reference database was also de-replicated prior to use, meaning that if two different species (accessions) were identical within the approximately 350 bp target region then only one was kept. This also means that a "species" in this manuscript is defined as the unique reference sequence to which other reads from the sequencing data are the best match. One alternative approach is to cluster non-chimeric sequences by a certain similarity threshold and match this "operational cluster" to a reference sequence. Preliminary attempts with this approach yielded inaccurate identifications from what appeared to be clusters comprised of reads that shared similar homopolymer extension sequencing errors even though they originated from different species.

Copy number estimation

A traditional closed form analytical solution to copy number correction was not possible because an unequal number of 18S amplicons were amplified from the samples and the amplicons were comprised of an unequal proportion of non-nematode reads. A number of

alternative optimization approaches are available, and we used a Genetic Algorithm (Morrall 2006, Figure 1) to estimate relative rRNA gene copy number using the specimen-based and sequence-based dataset. A genetic algorithm can be described as an iterative optimization algorithm, or a highly parallelized "guess-and-test", in which multiple plausible solutions are allowed to mutate until the system converges on one sufficiently optimal solution that maximizes a fitness function. Thus, the results of the iterative Genetic Algorithm do not necessarily produce a “correct answer” in the same sense of a closed-form algebraic solution. Instead, it produces a result, which after several thousand generations of exploring the entire potential parameter space, offers the best available explanation between the sequence-based data and the specimen-based data. In the present application, the "guess" is a prediction of rRNA gene copy number for each taxon, and the "test" is the difference between actual specimen-based counts and the predicted specimen-counts based on a given copy number solution. A population of copy number estimates is allowed to “mutate” in small increments until the population converges onto an optimized solution as defined by a minimal sums of squared errors (SSE). This process was performed first on the data from the test samples, grouped by species, and secondly on the field samples, grouped by family. The GA for the field samples was performed on the spring and fall samples separately. Specific parameters of the genetic algorithm, and the original code implemented in MATLAB (R2012b, The MathWorks, Natick, MA) are provided in Supplementary Methods. The copy number estimates from the genetic algorithm are relative copy numbers per individual (rCNPI) and not absolute copy number per individual (aCNPI). For example, if rCNPI for family A equals 200 and equals 100 for family B, then aCNPI may not be known for either species, but we at least know that family A

has twice as many copies per individual as family B. We should also distinguish copy number per individual from copy number per [haploid] genome (CNPG), which is the number of tandem rRNA operons on the chromosome that contains the rRNA operon. Absolute copy number per genome (aCNPG) would be the measurement reported from qPCR estimations of copy number that are scaled to a known single-copy gene.

Comparison of specimen-based vs. sequence-based counts

To compare the two methods of enumeration (morphological specimen identifications vs. molecular sequence identifications) we created two datasets: specimen-based and sequence-based counts. The specimen counts (from morphological identifications at the family level of taxonomic resolution) were computed as individuals per 100 g of dry soil after correction for soil moisture content that was measured on each individual sample. We then created "virtual specimen" counts by computing the total abundance of nematodes for each sample from the specimen counts and multiplying by the relative proportion of sequencing read counts after correction for rCNPI based on the Genetic Algorithm. (The original sequencing data was identified bioinformatically to species, but was then grouped by family for direct comparison with the specimen counts.) The virtual specimen counts represent the dataset that would be obtained if a researcher only had 1) total abundance counts, 2) read counts from high-throughput amplicon sequencing, and 3) a reasonably accurate copy number estimate for each taxon (which does not currently exist for nematodes). Thus, both datasets are in the same units (individuals per 100 g of dry soil), at the same level of taxonomic resolution (family), of which the families can also be binned into groups (Supplementary Information Table S1). After removal of chimera and non-nematode sequences, two subsamples (one each from two

different field samples) were left without any nematode sequences and were removed from both datasets. Both datasets were log-transformed ($y = \log(x+1)$) to meet assumptions of normality and homoscedasticity prior to all analyses. To describe the sampling efficiency of the high-throughput amplicon sequencing, we also constructed sample-based species accumulation curves based on 100 randomizations in EstimateS (Version 8.2, R.K. Colewell, http://purl.oclc.org/estimates; see also Colewell et al. 2012).

We conducted three analyses on both datasets in parallel to determine if the virtual specimen dataset would be an adequate alternative to the actual specimen count dataset. First, we tested the effect of season, burning, and nitrogen effects, plus all two-way interactions, on the abundance of trophic groups by linear mixed model analysis using the restricted maximum likelihood (REML) method of PROC MIXED (Statistical Analysis Software, Release 9.3, SAS Institute, Cary, NC, USA) using a split-plot variance structure (Todd et al. 1999). Secondly, we tested the effect of season, burning, and nitrogen effects, plus all two-way interactions, on the abundance of families by linear mixed model analysis using the same split-plot variance structure and restricted maximum likelihood (REML) method of PROC MIXED. Finally, we performed principal components analysis (PCA) using PROC PRINCOMP to discriminate the communities at the level of family. The first two principal components were analyzed by linear mixed model analysis (as above) to determine if the principal components varied by season, burning, nitrogen, or any two-way interactions.

Identification of putative indicator species

We also wanted to test whether the relative abundance of species in the sequencing counts were sufficient to discriminate treatment effects from the field experiment. To test this,

we converted the read counts of species in the sequencing data to relative abundance data (Using their respective family-wise copy number correction factors), and performed linear discriminate analysis (LDA, using PROC DISCRIM). Discriminate analysis is a classification technique that finds a linear combination of species abundances that most reliably discriminate the treatments of interest. To select the species to include, we first performed stepwise selection (forward with optional backward) using PROC STEPDISC (Statistical Analysis Software, Release 9.3, SAS Institute, Cary, NC, USA).

Results

Amplicon sequencing results

We obtained 955,608 quality-filtered reads (in the field samples) of which 86.1% were determined to be non-chimeric, 50.9% were metazoans, and 42.7% were of nematode origin (Table 1). Of the 407,908 nematode reads, 49,955 were unique, non-redundant sequences that matched a total of 129 different 18S rRNA accessions in the reference database; 117 of the accessions matched two or more reads (Supplemental Materials Table S1). The families recovered from amplicon sequencing were largely consistent with specimen data and the genera recovered were qualitatively consistent with what is expected from tallgrass prairie communities (Todd et al. 2006). However, the proportional representation of these families was poorly correlated between the specimen counts and the sequence counts (r = 0.01, p = 0.94). For example, less than 4% of the specimen-based counts were Rhabditidae, but this family represented more than 90% of the sequencing reads for the spring sampling and nearly 30% of the sequencing reads for the fall sampling (Table 2). Tylenchidae was the most common family represented in the specimen-based counts, but represented 0.2% and 1.8% of the

sequencing reads for spring and fall sampling, respectively. Spring samples were dominated by Rhabditidae sequences, mostly from the genera *Rhabditis*, *Oscheius*, and *Mesorhabditis*. The spring samples were less diverse overall, with lower richness as indicated by species accumulation curves (Fig. 2).

Copy number correction

The genetic algorithm was used as an optimization tool to estimate a correction factor for differences in relative rRNA gene copy number so that we could convert sequence-based abundances to their specimen-based "virtual" equivalent. First, the genetic algorithm was tested on the manually constructed test communities of 10 individuals that were picked from laboratory cultures. This run converged on a strongly supported solution within 1,000 generations and improved little beyond that (Fig. 3). The two independent amplifications of this test community were consistently biased from the actual equivalent numbers of input specimens per species (Table 3). The genetic algorithm generated copy number estimates with a range of nearly 100-fold difference between the largest (*Rhabditis* sp. RA5) and smallest (*Protorhabditis* sp. RA9) estimates. Virtual specimen counts were computed for the eleven species in the test samples and the average deviation of the virtual counts from their known counts (of 10 specimens each) was 1.23, meaning that, on average, the virtual counts were within 12.3% of their actual known specimen counts. Next, the genetic algorithm was performed separately for spring and fall samples, and for replicate subsamples from each field plot. Copy number estimates for families generally were similar between subsamples, and were strongly correlated ($r = 0.98$, $p < 0.0001$; data not shown) for families representing ≥ 1% of specimen and sequence counts, suggesting that consistent copy number estimates are

obtainable with the genetic algorithm procedure. In contrast, family-level copy number estimates for spring vs. fall samples were uncorrelated (Table 2; $r = 0.06$, $p = 0.79$). The relative copy number estimates that were obtained from the genetic algorithm at the level of family confirms the observation that Rhabditidae were the most disproportionately overrepresented taxa in the spring sequence-based counts relative to the specimen-based counts. The raw sequence-based counts were then corrected for rRNA gene copy number bias to produce a virtual specimen count as was done for the test samples. Copy number corrections resulted in proportional representation of virtual counts (Table 2, "%Virt") that were more similar ($r = 0.36$, $p = 0.15$ for spring; $r = 0.83$, $p < 0.0001$ for fall) to actual specimen counts (Table 2, "%Spec"), than were the original sequence counts (Table 2, "%Seq").

Trophic- and family-level treatment effects

To address the primary objective of this study (to determine the long-term effects of annual burning and nitrogen enrichment on the soil nematode community), we analyzed the specimen counts at the level of trophic group and family using univariate linear mixed model analysis followed by multivariate analysis of families using Principal Components Analysis. In the linear mixed model analysis of trophic group abundance, herbivores, fungivores, and predator/omnivores were more abundant in burned plots than non-burned plots ($p < 0.05$, Table 4). In the case of herbivores, which were affected by a significant burn × season interaction, the burn effect was more pronounced in the spring than in the fall. In contrast, bacterivores were more abundant in nitrogen-supplemented plots than non-amended plots while predator/omnivores were more abundant in non-amended plots than in nitrogen supplemented plots ($p < 0.05$, Table 4). The families Aporcelaimidae, Belondiridae,

Criconematidae, Hoplolaimidae, Mononchidae, Paratylenchidae, Pratylenchidae, Prismatolaimidae, and Tylenchidae were more abundant in burned plots than non-burned plots (Supplementary Table S2), while Rhabditidae were less abundant in burned plots than non-burned plots. The families Pratylenchidae and Rhabditidae were more abundant in nitrogen-supplemented plots than in non-amended plots, while Aporcelaimidae, Belondiridae, Criconematidae, and Tylencholaimidae were less abundant in nitrogen-supplemented plots than in non-amended plots. Following Principal Components Analysis of the specimen-based counts, the first principal component axis explained 21% of the variance (Fig. 4A) and divided burned plots from non-burned plots with a strong burn effect ($F_{1,44} = 31.36$, $p < 0.0001$) as well as a significant season effect ($F_{1,44} = 13.16$, $p = 0.0007$). All herbivore and predator/omnivore families had a positive value on the first principal component (Supplementary Table S3). The second principal component axis explained and additional 10% of variance and divided nitrogen supplemented plots from non-amended plots, with a strong nitrogen effect ($F_{1,44} = 31.77$, $p < 0.0001$). Most bacterivore and fungivore families had a positive value on the second principal component (Supplementary Table S3).

To address the second objective of this study (to determine whether high-throughput amplicon sequencing could accurately replace traditional specimen-based enumerations), we repeated the univariate and multivariate tests using virtual specimen counts, which were the copy-number corrected sequence-based read counts. In general, the non-adjusted sequencing read counts were not consistent with the actual specimen counts at the level of family abundance (e.g., compare dissimilarity between "Virtual Counts" and "Specimen Counts" in Table 2). The disproportionate over-abundance of Rhabditidae introduced an artifact that

resulted in a significant "Season" effect for most taxa in linear mixed model analysis (Supplementary Table S2) and also caused the first principal component of PCA to be more influenced by season than it was for the specimen counts (Fig. 4B). Analysis of virtual counts was more similar to the actual specimen counts than were the non-adjusted sequence counts. Linear mixed model analysis of trophic groups on virtual counts picked up the significant Burn effect for herbivores, fungivors, and predator/omnivores that existed for actual specimen counts, but not the Nitrogen effect for bacterivores or predator/omnivores (Supplementary Table S4). Season was less commonly a significant effect on linear mixed model analysis of individual families using virtual counts than it was using sequence counts (Supplementary Table S2). Principal Components Analysis of the virtual specimen counts at the level of family was also more similar to the specimen counts than to the non-adjusted sequence counts. Season ($F_{1,44}$ = 5.15, $p < 0.0284$) and burning ($F_{1,44}$ = 13.61, p = 0.0006) were distinguished along the first principal component (Fig. 4C), which explained 25% of the variance. Season ($F_{1,44}$ = 24.12, $p <$ 0.0001), season*burning ($F_{1,44}$ = 4.63, p = 0.0372) and nitrogen ($F_{1,44}$ = 6.44, p = 0.0150) were distinguished along the second principal component (Fig. 4C), which explained an additional 10% of the variance. Eigenvector weights for absolute and virtual family counts were positively correlated for the first principal component (r = 0.52, p = 0.01), and negatively correlated for the second principal component (r = -0.68, p = 0.0003).

Identification of putative indicator species

We performed linear discriminate analysis of species' relative abundance (for fall samples only) to determine which species were indicators of annual burning and/or nitrogen enrichment. Stepwise selection identified 14 species whose relative abundance was determined

by linear discriminate analysis (LDA) to be most characteristic of annual burning and/or nitrogen enrichment (Table 5). Responses generally were consistent with previously reported trophic- and family-level responses to burning and nitrogen enrichment. For example, the bacterivores *Chiloplacus* sp. KJC, *Pristionchus sp.*, *Wilsonema schuurmansstekhoveni*, and *Plectus aquaticus* were most abundant in plots with added nitrogen but without annual burning, while *Dorylaimellus virginianus*, and *Helicotylenchus varicaudata*, *Rhabdolaimus aquaticus*, and *Aglenchus agricola* were most abundant in plots with annual burning but without nitrogen enrichment.

Discussion

Ecological effects of annual burning and nitrogen enrichment on nematode species

One objective of this study was to determine the long-term effects of annual burning and elevated nitrogen inputs on soil nematode community composition. We found that 25 years of annual burning increased the abundance of herbivores, fungivores, and predator/omnivores, relative to non-burned plots. Conversely, 25 years of annual nitrogen application increased the abundance of bacterivores, and decreased the abundance of predator/omnivores, relative to non-amended plots. The abundance of certain key families was consistent with these patterns: Aporcelaimidae, Belondiridae, Criconematidae, Hoplolaimidae, Mononchidae, Paratylenchidae, Pratylenchidae, and Tylenchidae were more abundant in burned plots than non-burned plots (Supplementary Table S2), while Rhabditidae (bacterivores) were more abundant in nitrogen-supplemented plots than in non-amended plots. These results are consistent with our understanding of the effect of burning and fertilizer amendments on grassland soil food webs. Annual burning in a tallgrass prairie like Konza tends to increase the

productivity of warm-season C4 grasses (Collins et al. 1998) and increases plant root growth. This increases belowground root biomass and promotes greater invertebrate herbivore populations. Nitrogen-based fertilizer treatments allow microbes in N-limited soils to decompose substrates that have high carbon content. This increases the productivity and biomass of soil bacteria and, consequently, their bacterial-feeding nematode predators. Enrichment-type bacterivores, such as the families Rhabditidae and Panagrolaimidae, benefit from nitrogen enrichment more than basal-type bacterivores (such as the family Cephalobidae). Our findings are also an extension of the short-term effects of burning and nitrogen additions that have been published previously for the same experiment (Todd 1996). Through the first nine years of the experiment, total bacterivores were more abundant in nitrogen enrichment plots, but total herbivores were more abundant in burned plots than non-burned plots only in the presence of nitrogen enrichment.

We further identified fourteen species by linear discriminate analysis whose relative abundance was indicative of annual burning or nitrogen enrichment. The responses of putative indicator species based on sequence-based counts generally reflected reported family-level responses. For example, dominant herbivorous nematode taxa of the tallgrass prairie (Hoplolaimidae, Criconematidae) typically respond positively to burning, while bacterivorous taxa typically respond positively to nitrogen enrichment (Todd 1996; Todd et al. 2006). These trends were present in the specimen-based counts from the present study and were reproduced for selected species in the discriminate analysis of sequence-based counts. Expected positive responses to burning also were observed for selected omnivorous species in the Dorylaimida.

Species-level responses to disturbance are unlikely to be reliably predictable based on those of broader taxonomic categories. In fact, variable disturbance responses within nematode trophic groups and families are well documented (Fiscus and Neher 2002; Jones et al. 2006). There are at least three mechanisms to explain these observations. First, disturbance-specific changes in the plant community are likely to induce subsequent changes in the nematode assemblage that are indirectly related to the original disturbance, particularly for herbivorous species. Thus, nematode responses will be mediated by plant responses and ultimately determined by the host specificity of the taxa involved and by competitive interactions among species. Second, environmental optima may differ among members of the same functional group, such that environmental changes produce divergent responses in the presence of a unified overall food-web response. Finally, it should be noted that the trophic habits of many nematode taxa are poorly documented. Erroneous characterization of trophic behavior will result in unexpected responses relative to other members of the functional group. Regardless of the mechanism, it is clear that greater taxonomic resolution is desirable for identifying putative indicator taxa, and that high-throughput amplicon sequencing offers a promising approach to achieve this goal.

Comparison of amplicon sequencing method to specimen counts

A second objective of this study was to determine whether high-throughput amplicon sequencing produced accurate community composition data that was comparable to the traditional morphological identifications. We applied both specimen-based morphological identifications and sequence-based molecular identifications on the same set of samples. We found amplicon sequencing to be qualitatively valid as the identifications of the sequence-

based counts were taxonomically representative of the families and genera known to exist in tallgrass prairie soils. The high-throughput sequencing data did allow us to enumerate nematodes to species, which is not necessarily possible with morphological identifications due to the constraints of time, labor, expertise, or because some species are cryptic or can only be identified accurately with specimens of a particular sex or developmental stage. Orr and Dickerson (1966) identified 238 species in 61 samples collected from a nearby Flint Hills pasture using meticulous morphological identifications. Our amplicon sequencing procedure identified 129 different species (i.e. GenBank accessions), of which 117 were represented by more than a single read. We believe that 117 species is a reasonable, if not conservative, estimate of the total number of species sampled in this area at these time points (spring and fall of one year). The total number of species found across the prairie landscape, however, is expected to be greater than our estimate because, 1) the 350-bp amplicon product only includes two variable regions, so not all species can be resolved with this locus, and 2) more species would accumulate from sampling of more sites, more diverse habitats, and multiple years.

Although amplicon sequencing resulted in a reasonable but not necessarily complete survey of the taxa present, the relative proportions of the families present in the sequence-based counts failed to match the specimen-based counts. Rhabditidae were the most over-represented sequences relative to their specimen counts. We believe that Porazinska et al. (2010) experienced a similar phenomenon in which two species of *Oscheius* were the most abundant read clusters in tropical rain forest samples from Costa Rica (obtained through amplicon sequencing), even though this genus was not even among the top ten taxa of specimen based counts in comparable samples (Powers et al. 2009). We suspected that rRNA

gene copy number variation was the primary cause for this bias. To compensate for this artifact, we used an iterative genetic algorithm to estimate rRNA gene copy number. Estimated sequence-based abundances of families were improved, but still not perfect, relative to specimen-based abundances following correction for copy number variation. Furthermore, multivariate analysis of the effects of annual burning and nitrogen enrichment on family abundances produced similar results for virtual specimen counts and actual specimen counts. We conclude that sequence-based counts from high-throughput amplicon sequencing cannot reflect accurately the proportional abundance of specimens in a sample if it does not adequately correct for copy number variation. rRNA copy number correction was already demonstrated to be effective for amplicon sequencing of microbial 16S (Kembel et al. 2012), but is not currently a part of the typical analysis pipeline for amplicon sequencing of eukaryotes (Bik et al. 2012). Copy number variation is a problem for any molecular method that relies on quantitative amplification of rRNA genes, such as clone libraries, AFLP, TRFLP, and qPCR quantification. Vervoort (2012) solved this problem by calibrating family- and genus-specific qPCR assays to standards of known numbers of hand-picked individuals, but our preliminary analysis suggests that copy numbers can vary by greater than 100-fold among species within a family. Our approach is one example of an approach that can be used, but future efforts will benefit from an even finer resolution morphological identification of at least a subset of the community. One approach might be to pool a small proportion (e.g. 10%) of specimens into one or a few pooled samples that can be both sequenced and morphologically identified, thus providing a copy number correction estimate that can be applied to the remaining samples that are sequenced.

Implications of rRNA copy number variation for understanding nematode ecology

Copy number differences between species are inconvenient for certain amplification-based molecular tools, but it is still an interesting biological phenomenon that could prove to be ecologically informative. A large number of tandem operons in a genome are thought to facilitate the transcription of more rRNA subunits (Weider et al. 2005). More ribosomes in the cytoplasm increases the capacity of a cell to translate mRNA into proteins. Ribosomal RNA operon copy numbers vary from 1 to 15 between species of bacteria (Klappenbach et al. 2001), and may vary by as much as 100 to 1000-fold between species of nematodes and other microinvertebrates (Long and Dawid 1980). Ribosomal RNA copy number is positively correlated with genome size (Prokopowich et al. 2003) and colonizer-type life history traits (Klappenbach et al. 2000) and may be linked to the stoichiometry of phosphorous limitations in rapidly growing species (Elser et al. 2000). In the present study of grassland nematodes, the Rhabditidae had the greatest rRNA gene copy numbers among the grassland nematodes examined. Rhabditidae are enrichment-type microbivores that are capable of very rapid growth and reproduction. Their "boom-and-bust" lifestyle means that they are in relatively low abundance in prairie soils most of the time, followed by brief periods of rapid population growth and dispersal after enrichment events. As such, Rhabditidae are indicators of acute nitrogen enrichment (Ferris and Bongers 2006) and perform important ecological functions as they disperse microbes throughout the soil, regulate the microbial communities on which they feed, and influence the rate and conditions in which soil organic matter is transformed into inorganic and dissolved organic nitrogen (Freckman 1988). Unfortunately, they are difficult and sometimes impossible to identify from field specimens if a certain age or sex is not available.

Perhaps amplicon sequencing can be used to leverage the bias of rRNA gene copy number and characterize these enrichment-type bacterivores that are scarce in most soils but abundantly represented in amplicon sequencing due to their high copy number. Characterizing rRNA gene copy numbers could increase our understanding of individual species and serve as a genomic signature of certain ecological traits such as fecundity, generation time, or developmental rates.

It is also possible for amplicon sequencing to reduce the apparent richness of a particular treatment or habitat if a high-copy-number species is present. If one (or a few) high-copy-number species are present in a sample, then most of the sequencing reads will be from the high-copy-number species and few reads are left for the low-copy-number species. A possible example of this potential artifact is the difference in species richness that we found between spring and fall samples (Fig. 2). On average, 93% of the sequences from the spring samples were Rhabditidae, while only 29% of the fall samples were Rhabditidae. We cannot necessarily determine whether spring samples were indeed less taxonomically diverse than fall samples (as would be implied by Fig. 2), or if the abundance of Rhabditidae sequences in the spring samples simply limited the detection of species with less abundant rRNA gene fragments.

There is a clear need to characterize rRNA operon copy numbers for a range of species, as nematode ecologists have done for colonizer-persistor values (Bongers 1990) and for feeding habits (Yeates et al. 1993). This will benefit a variety of molecular identification tools that rely on quantitative amplification of a rRNA gene, but it may be inappropriate to extrapolate any given estimate of copy number per individual (including ours) to other communities because: 1) there are differences in species identity, and our test samples demonstrate that there could be significant variability in copy number between species of the same family or even genus, and 2)

the predominant developmental stage of each species could vary seasonally and affect the overall cell count due to the relative proportion of juveniles, sperm, and developing eggs. Indeed, it may not even be possible to obtain a single set of copy number correction factors that can be applied universally to all communities at all times. Here we outline the different sources of rRNA gene copy number variation and propose the following nomenclature to be used when reporting copy number estimates. The 18S rRNA gene is located in the rRNA gene operon along with 5.8S and 28S in tandem copies. We define the number of tandem rRNA operons in the chromosome that contains the rRNA operon as the "copy number per [haploid] genome" (CNPG), and the total number of tandem copies in a whole organism as the "copy number per individual" (CNPI). Thus, the CNPI is equal to CNPG times the number of cells, times their ploidy level (1 for haploid germ cells, 2 for diploid somatic cells, including developing eggs and embryos, and 3 for triploid cells). If a species' reproductive cycle is tied to plant phenology, or if peak population growth is seasonal, the number of rRNA gene copies in an individual coming from germ cells could also be seasonal. In this case, the perceived copy number per individual, when averaged across many individuals, could appear to vary seasonally within a species due to developmental and reproductive maturity. We are not aware of known instances where CNPG varies between cells of an individual or even between individuals of a population (but see Zhang 1990, McTaggert et al. 2007). Note that within this understanding copy number variation, we do not believe that CNPI should necessarily depend on biomass *per se*. If, for example, biomass is statistically correlated with somatic cell count, then it would be possible for CNPI to appear statistically confounded with biomass. The copy number estimates that we use (Table 2, 3) are specific to this study and can only be interpreted relative to another species

or family in the dataset. This is a limitation of the GA approach that cannot be overcome without an absolute estimate. In contrast, absolute copy number estimates would need to be scaled to a known single-copy gene or to a concentration standard (such as in Lee et al. 2008). Some species (especially, for example, Rhabditis sp. RA5) can have relative copy number values up to several orders of magnitude greater than other representative prairie taxa, even within the same family. Previous estimates of rRNA gene copy number (CNPG) in nematodes range from 55 in *Caenorhabditis elegans*, to 280 in *Panagrellus silusiae*, to 300 in *Ascaris lumbricoides* (Long and Dawid 1980), and our estimates of rRNA gene copy number largely support this magnitude of variation. Regardless of how copy numbers are expressed, it will be helpful to include one or more species that are a culturable, frequently occurring, well-defined biological species with a cosmopolitan distribution so as to facilitate the calibration of estimates between different methods, research groups, or geographical areas. In the present case, we used *Oscheius tipulae* because it fits all of these characteristics (Baille et al. 2008, Felix et al. 2001).

Recommendations and prospects

Molecular techniques such as high-throughput amplicon sequencing will not replace the need for traditional specimen-based morphological identification of nematodes. Instead, molecular techniques emphasize the need for intimate collaboration and sustained cross-training between ecologists and taxonomists. Amplicon sequencing does have the capacity to enumerate communities to finer taxonomic resolution than what is typically feasible for traditional morphological methods, but the approach is currently limited by the taxonomic resolution available in the 18S rRNA loci, the variability in rRNA copy numbers, and the

completeness of public sequence databases. We outline three specific recommendations to the research community:

1) High-throughput amplicon sequencing may be better suited as a complement to, rather than a replacement of, traditional specimen-based enumeration methods. Applying both traditional and molecular methods to the same samples may offer both the taxonomic resolution of amplicon sequencing and the accurate relative proportions of specimen counts. Molecular sequences also provide a putative species designation for specimens that cannot otherwise be identified due to their sex or developmental stage. Therefore, morphological identification of a given sample might proceed faster if a list of species in that sample (or season, or field site) is available from molecular data. Rather than perform amplicon sequencing on all samples of an experiment, it may be more prudent to pool portions of all samples from a block, season, or site, perform amplicon sequencing to obtain a species list, and then use this species list to inform the morphological identifications.
2) rRNA gene copy numbers will need to be quantified for diverse nematode species if amplicon sequencing data is to be used quantitatively. In the process these three practical questions need to be addressed: 1) is it possible to estimate copy number in non-culturable individuals isolated from field samples, 2) is copy number phylogenetically autocorrelated (so that we can attribute known copy numbers to related species), and 3) is copy number correlated to ecological or physiologically relevant traits? It may not be possible to generate one single copy number correction

for all taxa. Instead, it may be necessary to empirically generate site-specific copy number corrections using a subset of each community at each sampling point.

3) High-throughput amplicon sequencing will continue to benefit from adding well-curated, full-length sequences to publically available databases from specimens that have been morphologically identified and documented by digital vouchering approaches such as video capture and editing (VCE, De Ley and Bert 2002, De Ley et al 2005). Automated analysis pipelines of high throughput sequencing are only as accurate as the databases on which they rely, and our analysis benefited from sequences that came from morphologically-identified specimens collected directly from Konza prairie (http://nematode.unl.edu/konzinfo.htm).

In summary, we found that annual burning increased the overall abundance of herbivorous, fungivorous, and predatory/omnivorous nematodes, and that nitrogen enrichment increased the overall abundance of bacterivores. We have also identified significant issues related to the use of high-throughput amplicon sequencing in nematode ecological studies, and have demonstrated potential approaches to compensate for these limitations. While sequencing facilitates the identification of indicator taxa at a high level of taxonomic resolution, quantification of taxa abundances for use in standard ecological indices (e.g. Maturity, Enrichment, and Structure Indices) remains problematic. Efforts to account for overrepresentation of Rhabditid sequences relative to specimen counts were only moderately successful. In contrast, the most abundant family in tallgrass prairie, the Tylenchidae, remained severely underrepresented in sequence counts relative to specimen counts. Both of these

families are important indicators of nitrogen enrichment and other environmental disturbances (Todd et al. 2006). It is our opinion that high-throughput amplicon sequencing can still be a valuable method for characterizing nematode communities at high taxonomic resolution if we can estimate rRNA gene copy number in nematodes and maintain accurate and complete sequence databases. In light of these issues, we suggest that high-throughput amplicon sequencing be used as a complementary approach to, rather than as a replacement for, traditional nematode community and soil food web diagnostics. Inferences based on sequencing results must account for both the strengths and weaknesses of the methodology.

Acknowledgements

This work was funded by a grant from the National Science Foundation (NSF EF 0723862) and by the Kansas State University Targeted Excellence program.

References

Baille D, Barriere A, Felix MA (2008) *Oscheius tipulae*, a widespread hermaphroditic soil nematode, displays a higher genetic diversity and geographical structure than *Caenorhabditis elegans*. *Molecular Ecology* **17**, 1523-1534.

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

Bik HM, Porazinska DL, Creer S*, et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution* **27**, 233-243.

Bongers T (1990) The maturity index: An ecological measure of environmental disturbance

based on nematode species composition. *Oecologia* 83, 14-19.

Cock PJA, Antao T, Chang JT*, et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423.

Collins SL, Knapp AK, Briggs JM, Blair JM, Steinauer EM (1998) Modulation of Diversity by Grazing and Mowing in Native Tallgrass Prairie. *Science* 280, 745-747.

Colwell RK, Chao A, Gotelli NJ, et al. (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology 5, 3-21.

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

De Ley P, Bert W (2002) Video Capture and Editing as a Tool for the Storage, Distribution, and Illustration of Morphological Characters of Nematodes. *Journal of Nematology* 34, 296-302.

De Ley P, De Ley IT, Morris K*, et al.* (2005) An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. *Philosophical Transactions of the Royal Society B-Biological Sciences* 360, 1945-1958.

Donn S, Neilson R, Griffiths B, Daniell T (2011) Greater coverage of the phylum Nematoda in SSU rDNA studies. *Biology and Fertility of Soils* **47**, 333-339

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*.

Elser JJ, Sterner RW, Gorokhova E*, et al.* (2000) Biological stoichiometry from genes to ecosystems. *Ecology Letters* 3, 540-550.

Felix MA, Vierstraete A, Vanfleteren J (2001) Three biological species closely related to

Rhabditis (*Oscheius*) *pseudodolichura* Korner in Osche, 1952. Journal of Nematology 33, 104-109.

Ferris H, Bongers T (2006) Nematode Indicators of Organic Enrichment. *Journal of Nematology* 38, 3-12.

Fiscus DA, Neher DA (2002) Distinguishing sensitivity of freeliving soil nematode genera to physical and chemical disturbances. Ecological Applications, 12, 565–575.

Freckman DW (1988) Bacterivorous Nematodes and Organic-Matter Decomposition. *Agriculture Ecosystems & Environment* 24, 195-217.

Gibson, DJ, Seastedt, TR, Briggs, JM (1993) Management practices in tallgrass prairie: large- and small-scale experimental effects on species composition. *Journal of Applied Ecology* 30, 247-255.

Ingham RE, Trofymow JA, Ingham ER, Coleman DC (1985) Interactions of bacteria, fungi, and their nematode grazers: effects on nutrient cycling and plant growth. *Ecological Monographs* 55, 119-140.

Jones KL, Todd TC, Wall-Beam JL*, et al.* (2006) Molecular Approach for Assessing Responses of Microbial-Feeding Nematodes to Burning and Chronic Nitrogen Enrichment in a Native Grassland. *Molecular Ecology* 15, 2601-2609.

Kembel SW, Wu M, Eisen JA, Green JL (2012) Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS Comput Biol* **8**, e1002743.

Klappenbach JA, Dunbar JM, Schmidt TM (2000) RRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology* 66, 1328-1333.

Lee C, Lee S, Shin S, Hwang S (2008) Real-time PCR determination of rRNA gene copy number: absolute and relative quantification assays with *Escherichia coli*. *Applied Microbiology and Biotechnology* 78, 371-376.

Long EO, Dawid IB (1980) Repeated genes in eukaryotes. *Annual Review of Biochemistry* 49, 727-764.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17.

McTaggart SJ, Dudycha JL, Omilian A, Crease TJ (2007) Rates of Recombination in the Ribosomal DNA of Apomictically Propagated *Daphnia obtusa* Lines. *Genetics* 175, 311-320.

Morrall D (2006) Ecological Applications of Genetic Algorithms. In: *Ecological Informatics*, pp. 69-83. Springer Berlin Heidelberg.

Neher DA (2001) Role of nematodes in soil health and their use as indicators. Journal of *Nematology* 33, 161-168.

Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: A Fast Search Method for Large DNA Databases. *Genome Research* 11, 1725-1729.

Orr CC, Dickerson OJ (1966) Nematodes in True Prairie Soils of Kansas. *Transactions of the Kansas Academy of Science* 69, 317-334.

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

Porazinska DL, Giblin-Davis RM, Faller L*, et al.* (2009) Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Molecular Ecology Resources* 9, 1439-1450.

Powers TO, Todd TC, Burnell AM*, et al.* (1997) The rDNA internal transcribed spacer region as a taxonomic marker for nematodes. *Journal of Nematology* **29**, 441-450.

Powers TO, Neher DA, Mullin P*, et al.* (2009) Tropical nematode diversity: vertical stratification of nematode communities in a Costa Rican humid lowland rainforest. *Molecular Ecology* 18, 985-996.

Prokopowich CD, Gregory TR, Crease TJ (2003) The correlation between rDNA copy number and genome size in eukaryotes. *Genome* 46, 48-50.

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

Samson F, Knopf, F (1994) Prairie conservation in North America. *Bioscience* 44, 418-421.

Todd TC (1996) Effects of management practices on nematode community structure in tallgrass prairie. *Applied Soil Ecology* 3, 235-246.

Todd TC, Powers TO, Mullin PG (2006) Sentinel nematodes of land-use change and restoration in tallgrass prairie. *Journal of Nematology* 38, 20-27.

Vervoort MTW, Vonk JA, Mooijman PJW, Van den Elsen SJJ, Van Megen HHB, et al. (2012) SSU Ribosomal DNA-Based Monitoring of Nematode Assemblages Reveals Distinct Seasonal Fluctuations within Evolutionary Heterogeneous Feeding Guilds. *PLoS ONE* 7, e47555. doi:10.1371/journal.-pone.0047555

Wardle DA, Bardgett RD, Klironomos JN, Setala H, van der Putten WH, Wall DH (2004) Ecological linkages between aboveground and belowground biota. *Science* 304, 1629-1633.

Weider LJ, Elser JJ, Crease TJ*, et al.* (2005) The functional significance of ribosomal (r)DNA

variation: impacts on the evolutionary ecology of organisms. *Annual Review of Ecology, Evolution, and Systematics* 36, 219-242.

Yeates GW, Bongers T, de Goede RGM, Freckman DW, Georgieva SS (1993) Feeding habits in soil nematode families and genera - An outline for soil ecologists. *J. Nematology* 25, 315-331.

Yeates GW, Ferris H, Moens T, Van Der Putten WH (2009) The role of nematodes in ecosystems. In: Nematodes as Environmental Indicators (eds. Wilson MJ, Khakouli-Duarte T), pp. 1-44. CABI, Cambridge, MA.

Yoder M, De Ley IT, King IW*, et al.* (2006) DESS: a Versatile Solution for Preserving Morphology and Extractable DNA of Nematodes. *Nematology* 8, 367-376.

Zhang QF, Saghai Maroof MA, Allard RW (1990) Effects on adaptedness of variations in ribosomal DNA copy number in populations of wild barley (*Hordeum vulgare* ssp. *spontaneum*). *Proceedings of the National Academy of Sciences* 87, 8741-8745.

Data Accessibility

Raw sequencing reads, read processing scripts, summarized specimen-based and sequence-based datasets, and genetic algorithm (MATLAB code): Dryad entry doi:xxxxx.

Author Contributions

All authors designed and performed the research, BJD analyzed the data (with input from TCT and MAH), and all authors wrote and edited the manuscript.

Figures

Figure 1. Schematic flowchart of Genetic Algorithm.

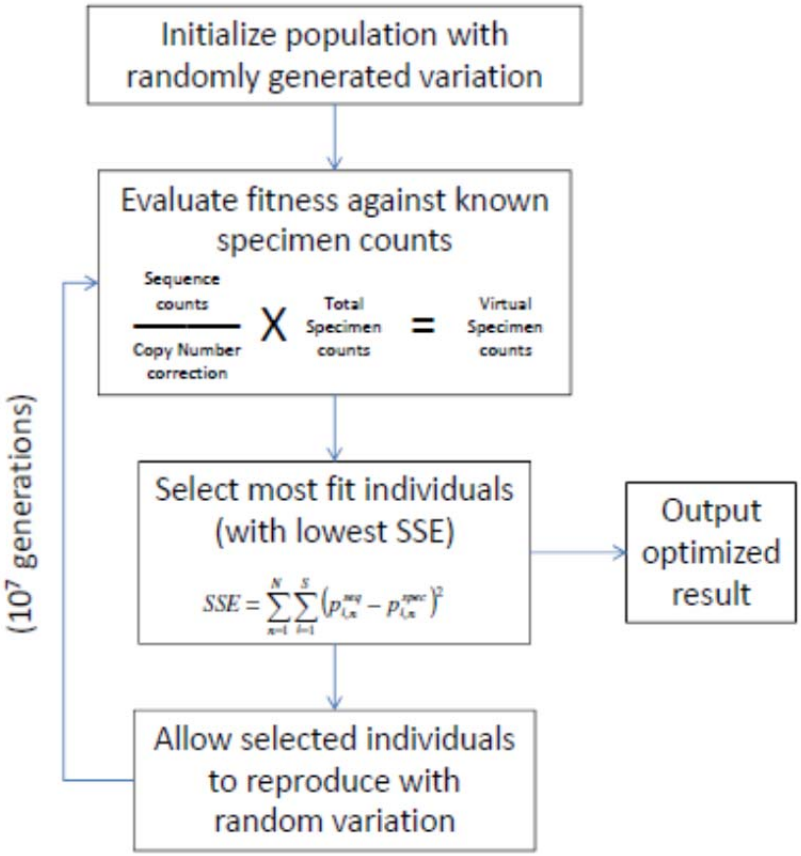

Figure 2. Sample-based species accumulation curves following 100 simulated re-samplings of sequence-based data from spring (solid line) and fall (broken line) samples (n = 32 subsamples for each season).

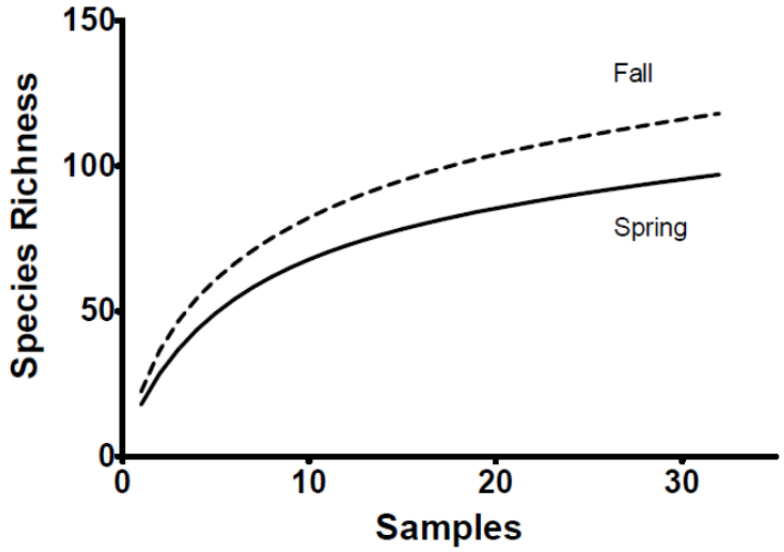

Figure 3. Progress of genetic algorithm for test samples. Copy number (rCNPI) estimations for 11 species (pooled into the same test sample) are shown through 10,000,000 generations of the iterative genetic algorithm. The final (lower right-hand) panel illustrates the time course of the fitness function (SSE); note the double-$\log_{10}$ scale. Letters indicate the species: (A) Rhabditis sp. RA5, (B) Pristionchus pseudaerivorous, (C) Rhabditophanes sp., (D) Oscheius tipulae, (E) Mesorhabditis sp. MR1, (F) Oscheius sp. FVV-2, (G) Panagrolaimus sp., (H) Mesorhabditis sp. MR2, (I) Cephalobus sp., (J) Acrobeloides sp., (K) Protorhabditis sp. RA9.

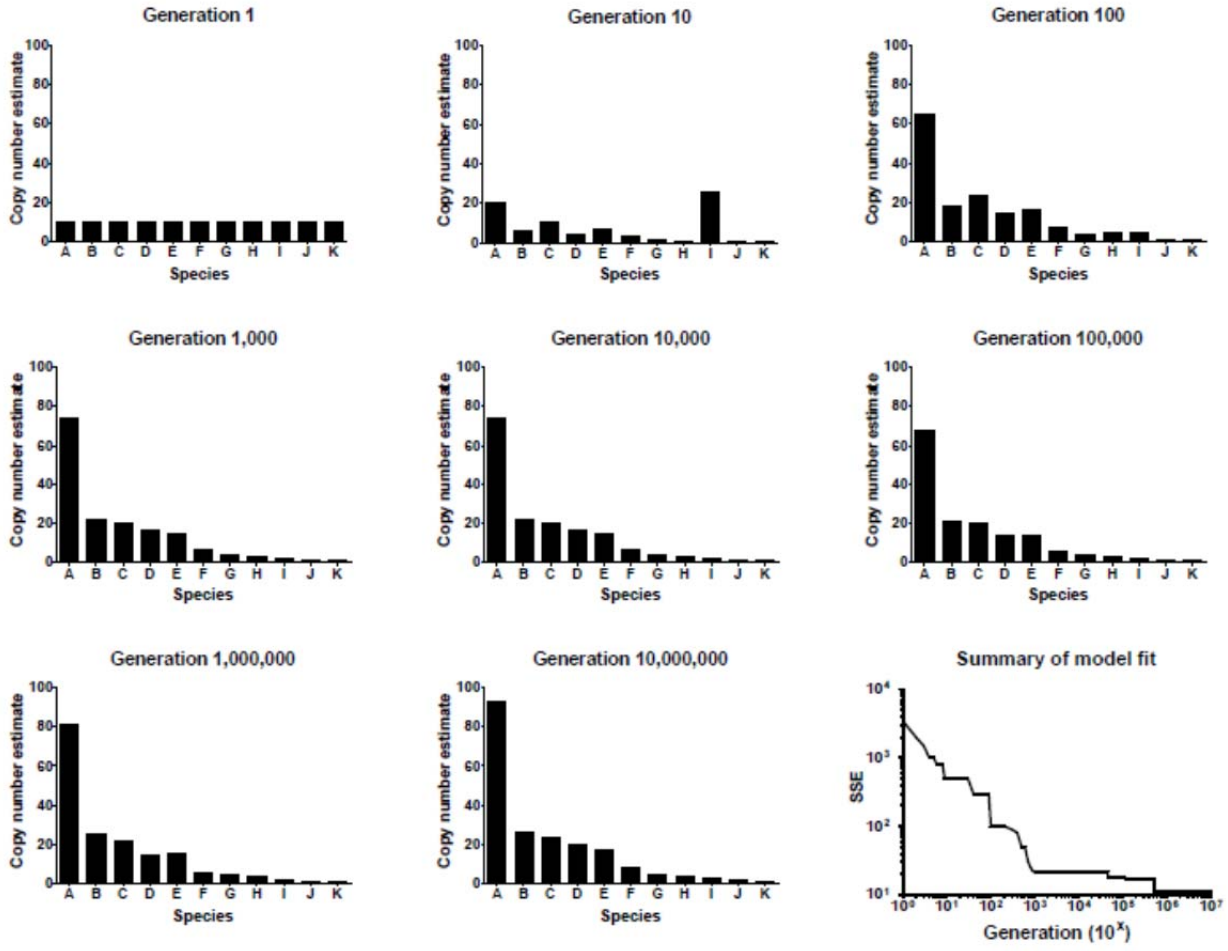

Figure 4. Principal Components Analysis of specimen-based counts, sequence-based read counts, and copy number-corrected virtual nematode counts. A) PCA of specimen-based data with symbols representing least-squares means (n = 8, 4 plots each with two subsamples) of samples from Spring (S) or Fall (F), with burning (+B, red symbols) or without burning (-B, black symbols), and with nitrogen enrichment (+N, circles), and without nitrogen enrichment (-N, squares). B) PCA of sequencing read counts, with symbols representing least-squares means (n = 8) of samples as coded in (A). C) PCA of virtual specimens (copy number adjusted read counts), with symbols representing least-squares means (n = 8) of samples as coded in (A).

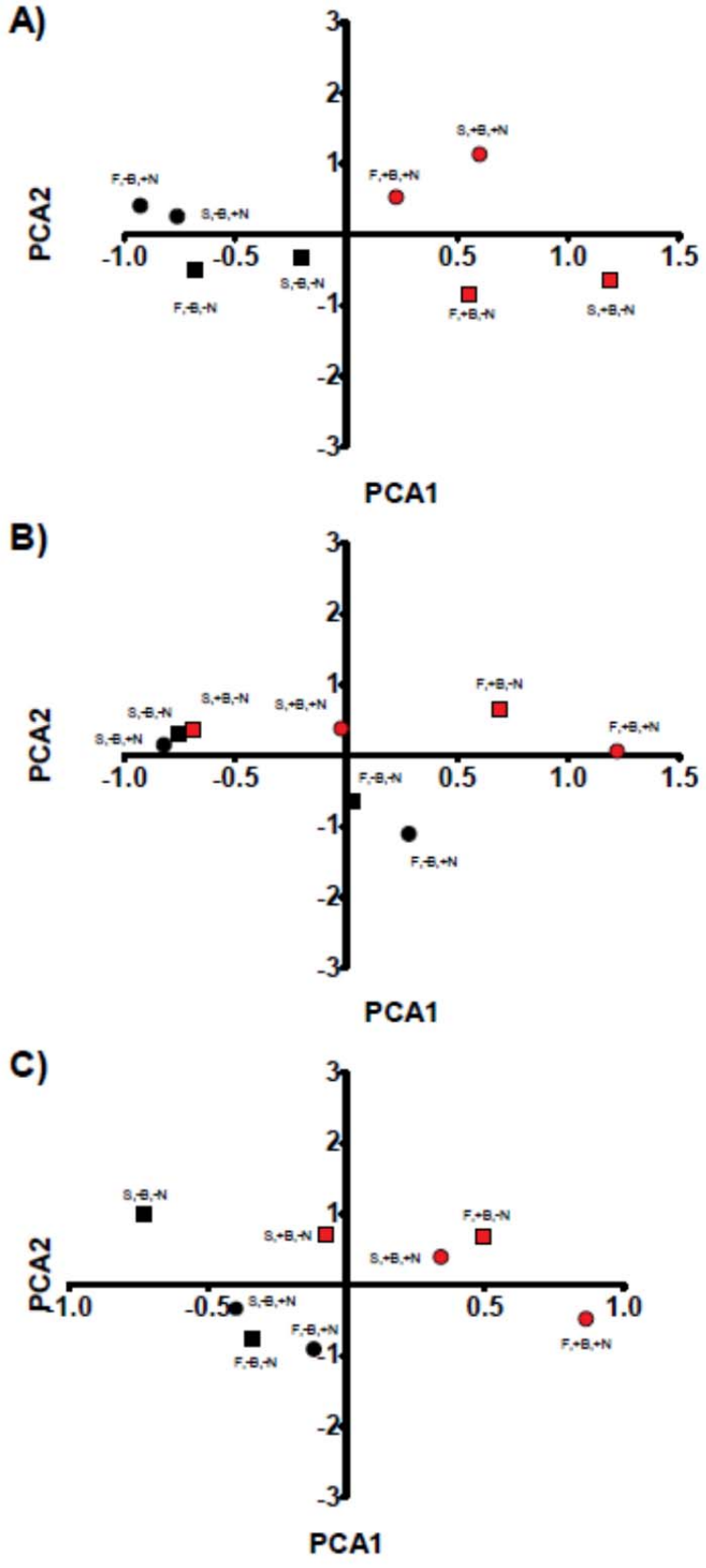
A)
PCA2
PCA1
F,-B,+N
S,-B,+N
S,-B,-N
F,-B,-N
S,+B,+N
F,+B,+N
F,+B,-N
S,+B,-N
B)
PCA2
PCA1
S,-B,-N
S,-B,+N
S,+B,-N
S,+B,+N
F,+B,-N
F,+B,+N
F,-B,-N
F,-B,+N
C)
PCA2
PCA1
S,-B,-N
S,+B,-N
S,+B,+N
F,+B,-N
S,-B,+N
F,-B,+N
F,-B,-N
F,+B,+N

Table 1. Summary of high-throughput amplicon sequencing reads and clusters.

	Reads from Field Samples	Reads from Test Samples	Total Non-Redundant Clusters
Quality-passed	955,608	51,025	146,623
Non-chimeric	822,464	38,400	111,119
Metazoa	486,441	38,398	63,250
Nematodes	407,908	38,397	49,955
Rhabditidae	335,457	26,492	31,612

Table 2. Improvement of sequence-based counts relative to specimen-based counts by copy number correction factors. "%Spec" and "%Seq" represent the overall relative abundance of specimen-based and sequence-based counts, respectively, for each family represented (averaged across all samples of either "Spring" or "Fall"). "rCNPI" indicates the relative Copy Number Per Individual estimates determined from the genetic algorithm, and "%Virt" represents the overall relative abundance of virtual specimen counts (sequence-based counts corrected for relative copy number, averaged across all samples) for each family.

	Spring				Fall			
Family	%Spec	%Seq	rCNPI	%Virt	%Spec	%Seq	rCNPI	%Virt
Anguinidae	0.8	0.0	1.0	1.2	1.0	1.3	3.5	4.0
Aphelenchidae	1.8	0.1	1.0	0.9	2.1	0.8	3.1	2.3
Aphelenchoididae	2.7	0.0	1.0	1.1	4.2	0.1	3.4	0.4
Aporcelaimellidae	0.7	0.5	1.6	2.7	0.7	2.6	11.7	2.8
Belondiridae	2.3	0.1	2.0	1.3	1.6	1.1	6.8	1.8
Cephalobidae	5.5	0.2	1.0	4.3	6.7	4.7	6.5	6.0
Criconematidae	11.3	0.2	1.0	5.2	7.6	1.9	3.2	6.4
Diphtherophoridae	0.6	0.0	34.4	0.0	0.2	0.8	32.0	0.2
Hoplolaimidae	17.3	1.6	1.0	13.6	20.4	8.5	3.5	23.1
Leptonchidae	0.3	0.0	1.0	0.1	0.1	0.0	40.3	0.0
Longidoridae	0.5	1.0	1.8	4.8	0.7	5.4	9.9	5.1
Meloidogynidae	1.1	0.0	1.0	0.8	0.6	0.3	50.8	0.0
Mononchidae	0.3	0.6	13.1	0.7	0.6	4.5	29.3	1.8
Panagrolaimidae	0.2	0.0	5.3	0.0	0.8	2.2	16.0	1.0
Paratylenchidae	7.7	0.0	1.0	2.2	11.0	3.7	5.7	4.3
Plectidae	4.4	0.5	1.9	8.2	2.4	6.3	8.1	6.4
Pratylenchudae	3.0	0.2	1.0	1.5	1.2	0.6	8.3	0.5
Prismatolaimidae	1.0	0.2	1.0	1.5	1.3	0.5	1.7	3.2
Qudsianematidae	1.3	0.8	53.2	0.3	1.3	16.4	219.0	3.0
Rhabditidae	3.7	93.3	1044.7	24.2	3.7	28.5	33.3	7.7
Tylenchidae	28.1	0.2	1.0	5.6	27.0	1.8	1.0	13.9
Tylencholaimidae	0.5	0.0	1.0	0.8	0.1	0.1	9.4	0.2
Correlation with %Spec:		-0.02		0.36		0.07		0.83

Table 3. rRNA copy number estimates from test sample specimens. The test sample contained exactly 10 adults of each species. Two replicate amplifications were sequenced ("RepA" and "RepB") and their relative copy number per individual (rCNPI) were estimated with the genetic algorithm (Fig. 2).

Species	Strain	RepA	RepB	rCNPI	#Ind	VirtA	VirtB
Rhabditis sp. RA5	KS594	8,950	8,726	92.7	10	10.0	10.3
Pristionchus pseudaerivorous	KS596	2,517	2,611	26.6	10	9.8	10.7
Rhabditophanes sp.	KS597	2,704	2,450	24.0	10	11.7	11.2
Oscheius tipulae	KS599	1,581	1,570	19.6	10	8.3	8.8
Mesorhabditis sp. MR1	KS601	1,702	1,654	16.8	10	10.5	10.8
Oscheius sp. FVV-2	KS600	679	669	8.5	10	8.3	8.6
Panagrolaimus sp. UK1	KS598	480	376	4.4	10	11.3	9.3
Mesorhabditis sp. MR2	KS602	402	345	3.6	10	11.6	10.5
Cephalobus sp.	CE1	248	235	2.7	10	9.5	9.5
Acrobeloides apiculata	KS586	153	131	1.8	10	8.8	8.0
Protorhabditis sp. RA9	RA9	100	114	1.0	10	10.3	12.5
	Total:	**19,516**	**18,881**		**110**	**110**	**110**

Table 4. Summary of linear mixed model analysis on trophic group abundance of specimen-based data. A) F-values on log(x+1)-transformed specimen counts, with abundance of herbivores (HERB), fungivores (FUNG), bacterivores (BACT) or predator/omnivores (PRED/OMNI) as a dependant variable. B) Least-Squares Means for Burn vs. Non-burned treatment. C) LS Means for Ambient vs. Nitrogen-enriched treatments.

A)	Effect	df	HERB	FUNG	BACT	PRED/OMNI
	Season	1,8	0.21	0.59	1.33	0.31
	Burn	2,8	19.49***	29.23***	0.07	14.47***
	B × S	2,8	9.14**	0.54	0.01	0.01
	Nitrogen	1,8	0.91	1.27	5.67*	3.84*
	N * S	2,8	0.02	2.08	1.57	0.48
	B × N	2,8	0.02	0.03	2.79	0.09

B)	Burn means	HERB	FUNG	BACT	PRED/OMNI
	Non-Burned	133.2^{A}	148.9^{A}	96.2	10.0^{A}
	Burned	372.1^{B}	285.7^{B}	91.6	50.6^{B}

C)	Nitrogen means	HERB	FUNG	BACT	PRED/OMNI
	Ambient	199.2	220.8	75.3^{A}	34.7^{A}
	N-enriched	248.4	192.8	116.9^{B}	15.3^{B}

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 5. Species recovered in high-throughput amplicon sequencing whose relative abundance was determined to be characteristic of annual burning and/or nitrogen enrichment. Columns represent the relative abundance of the species (from virtual counts of the fall samples only) in plots with (+B) or without -B) burning, and with (+N) or without (-N) nitrogen enrichment. Species included in the model were selected by stepwise Discriminate Analysis selection.

Species	Accession #	-B,-N	-B,+N	+B,-N	+B,+N
Aglenchus agricola	FJ969113	0.13	0.08	0.69	0.55
Aporcelaimellus sp. F5	AJ875155	0.34	3.14	4.21	1.84
Chiloplacus sp. KJC	HQ130507	0.00	0.95	0.00	0.00
Dorylaimellus virginianus	AY552969	0.19	0.00	3.87	0.27
Helicotylenchus varicaudatus	EU306354	0.41	0.10	0.88	0.04
Odontolaimus sp. OdLaS	FJ969131	0.00	0.62	0.66	0.20
Panagrolaimus detritophagus	GU014546	0.00	1.82	0.71	6.20
Paractinolaimus sp. PM	AY552975	0.37	0.41	0.08	0.21
Plectus aquatilis	AF036602	0.32	1.89	0.91	0.10
Pristionchus sp.	AY146554	1.76	12.46	0.64	1.70
Psilenchus sp. CA12	EU130840	0.00	1.08	0.00	0.09
Pungentus silvestris	AY284788	0.09	0.35	0.50	0.16
Rhabdolaimus aquaticus	FJ969139	0.00	0.00	0.74	0.01
Wilsonema schuurmansstekhova	AJ966513	1.54	2.65	0.13	1.78