Discrimination of Plant-Parasitic Nematodes from Complex Soil Communities Using Ecometagenetics

Dorota L. Porazinska, Matthew J. Morgan, John M. Gaspar, Leon N. Court, Christopher M. Hardy, and Mike Hodda

ABSTRACT

Many plant pathogens are microscopic, cryptic, and difficult to diagnose. The new approach of ecometagenetics, involving ultrasequencing, bioinformatics, and biostatistics, has the potential to improve diagnoses of plant pathogens such as nematodes from the complex mixtures found in many agricultural and biosecurity situations. We tested this approach on a gradient of complexity ranging from a few individuals from a few species of known nematode pathogens in a relatively defined substrate to a complex and poorly known suite of nematode pathogens in a complex forest soil, including its associated biota of unknown protists, fungi, and other microscopic eukaryotes. We added three known but contrasting species (Pratylenchus neglectus, the closely related P. thornei, and Heterodera avenae) to half the set of substrates, leaving the other half without them. We then tested whether all nematode pathogens—known and unknown, indigenous, and experimentally added—were detected consistently present or absent. We always detected the Pratylenchus spp. correctly and with the number of sequence reads proportional to the numbers added. However, a single cyst of H. avenae was only identified approximately half the time it was present. Other plant-parasitic nematodes and nematodes from other trophic groups were detected well but other eukaryotes were detected less consistently. DNA sampling errors or informative errors or both were involved in misidentification of H. avenae; however, the proportions of each varied in the different bioinformatic pipelines and with different parameters used. To a large extent, false-positive and false-negative errors were complementary: pipelines and parameters with the highest false-positive rates had the lowest false-negative rates and vice versa. Sources of error identified included assumptions in the bioinformatic pipelines, slight differences in primer regions, the number of sequence reads regarded as the minimum threshold for inclusion in analysis, and inaccessible DNA in resistant life stages. Identification of the sources of error allows us to suggest ways to improve identification using ecometagenetics.

Discrimination of Plant-Parasitic Nematodes
from Complex Soil Communities Using Ecometagenetics

Corresponding author: M. Hodda; E-mail address: mike.hodda@csiro.au

* The e-Xtra logo stands for “electronic extra” and indicates that the online version contains five supplemental figures.

http://dx.doi.org/10.1094/PHYTO-08-13-0236-R
© 2014 The American Phytopathological Society

Vol. 104, No. 7, 2014 749
Table 1. Summary of biotic structure of substrates

<table>
<thead>
<tr>
<th>Substrate</th>
<th>Pratylenchus spp.</th>
<th>Heterodera avenae</th>
<th>Total</th>
<th>PPN</th>
<th>FLN</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potting mix</td>
<td>1,000/1,500</td>
<td>1 cyst</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Banana</td>
<td>1,000/1,500</td>
<td>1 cyst</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sugar</td>
<td>500/750</td>
<td>1 cyst</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forest</td>
<td>150/250</td>
<td>1 cyst</td>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5 PPN = plant-parasitic nematodes, FLN = free-living nematodes, and Other = other eukaryotes.
6 Pratylenchus neglectus/P. thornei.
7 In amended substrates only.
8 Total = all nematodes.
match to the query on the 3' end. Generally, length-dependent OCTUs were BLAST-matched (2) against the NCBI nonredundant database and assigned the GenBank accession number of the best match. The distribution patterns of GenBank-assigned unique sequences across samples and the abundances of reads of each unique sequence were used to filter out unique sequences with errors and singletons (a unique sequence represented by only one read and present in only one sample). All possible three-way alignments were used to identify potential chimeras. Only valid unique sequences were then used for further analysis. Each valid unique sequence (OTU) was taken to represent one species, species complex, or group of cryptic or closely related species if their sequences were identical.

OCTUs generated by both bioinformatic pipelines from SSU rDNA by pyrosequencing are not necessarily equivalent to species (39); therefore, OCTUs were linked to putative species using Head-Tail patterns (41). In brief, this technique relies on the observation that, in metagenetic datasets generated from SSU rDNA by pyrosequencing, a single species is usually represented by a series of OCTUs and each OCTU by multiple sequencing reads (39). The most frequently observed OCTU of a species—the Head—should be characterized by the best BLAST matches to a database reference sequence, and less frequently observed OCTUs—the Tail—have slightly variant sequences to the same matching reference sequence. Although the number of Tail OCTUs differs among species, predictable Head-Tail patterns emerge when the scores are sorted and putative species can be identified. For analysis of numbers of reads, this means that all reads of each OCTU (Head and Tail) linked to a putative species were summed to generate abundance per species per sample. For both pipelines, heads with less than three reads were not linked back to any species and instead discarded.

Each putative species was then checked by comparing it across the replicate samples, across the two gaskets, and finally against other OTUs generated by the pipeline: species with nearly matching sequences but only detected in one replicate, gasket, or pipeline were collapsed to a single species.

Nematode species were sorted into plant parasites and free living using the classification of Yeates et al. (59).

**Sources of error for H. avenae.** The sources of error were assessed by comparing whether H. avenae was added to the sample, whether any DNA originating from H. avenae was present, and whether H. avenae was identified by a pipeline.

To determine whether H. avenae sequences were detectable in the raw DNA sequence data for each sample, we conducted a detailed examination of every sequence independently before processing by the pipelines. First, raw reads were assigned to samples of origin based on the presence of MID tag sequences at both ends. Reads that could not be assigned to samples using this procedure were discarded. All identical reads remaining were then collapsed into a total of 177,869 unique sequences. These unique sequences were initially screened by performing pairwise alignments against the known H. avenae SSU DNA sequence. Sequences that were <95% identical to this reference sequence were regarded as highly unlikely to have originated from H. avenae and were not investigated further. All 12,986 remaining sequences were then compared against the NCBI nonredundant nucleotide database using MEGABLAST. The 6,598 sequences that returned H. avenae as the best match were retained, and the total number of H. avenae reads in each sample was calculated. For each pipeline, samples were then classified as (i) True Positive if H. avenae was added to the sample, reads were detected in the raw data, and identified by the pipeline; (ii) True Negative if H. avenae was NOT added to the sample, reads were NOT detected in the raw data, and NOT identified by the pipeline; (iii) Pipeline False Negative if H. avenae was added to the sample, reads were NOT detected in the raw data, but they were NOT identified by the pipeline; (iv) Pipeline False Positive if H. avenae was NOT added to the sample, reads were NOT detected in the raw data, but they were identified by the pipeline; (v) DNA Sampling False Positive if H. avenae was added to the sample, but reads were NOT detected in the raw data using the procedure described immediately above, and NOT identified by the pipeline; (vi) DNA Sampling False Negative if H. avenae was NOT added to the sample, but reads were detected in the raw data using the procedure described immediately above, and identified by the pipeline; (vii) Combined False Positive if H. avenae was added to the sample, reads were NOT detected in the raw data, however the pipeline gave the opposite result to the raw data and identified H. avenae; and (viii) Combined False Negative if H. avenae was NOT added to the sample but reads were detected in the raw data yet the pipeline gave the opposite result and did not identify H. avenae. A schematic summary is presented in Table 2.

Intraspecific variability is known in the primer sequence for many nematodes, including H. avenae (55), which could cause errors in metagenetic identifications. Potential errors caused by single nucleotide substitutions in the primer sequence were therefore investigated using APDP with slightly relaxed criteria for retention of reads, termed APDP(A+T). In this version of the pipeline, one difference was allowed in the primer sequence before a sequence was discarded. (A proofreading polymerase was used to ensure that these minor changes in the primer sequences did not prevent annealing and replication of a variant sequence.) In this version of the pipeline, one difference was allowed in the primer sequence before a sequence was discarded.
sequence during the PCR. Without easing criteria, the variation in primer sequence would have been detected by APDP and discarded as a sequencing error. Therefore, the APDP(A+T) pipeline demonstrates the effect of small differences in primer sequence on identification. Error checking in OCTUPUS is not based on strict congruence of the primer sequence; therefore, this procedure was not replicated for OCTUPUS.

Pipeline errors were further investigated for their sensitivity to relatively rare organisms such as H. avenae (here represented by a single cyst among many nematodes and other organisms), and whether this could be altered. Pipelines have a threshold for the number of reads of a particular sequence that are regarded as real. Singletons—sequences represented by a single read only—are generally regarded as likely to be chimeras or errors, whereas many reads of the same sequence are unlikely to arise through the stochastic processes generating chimeras or sequencing errors. Of course, a single read may be a real if it comes from a relatively rare species with few copies of the SSU rRNA gene among many abundant species with many copies contributing most of the DNA for sequencing. Furthermore, the probability of a single read being real will increase as the number of different organisms—and, hence, different sequences—increases in a sample relative to the number of reads. Because a single cyst of H. avenae was added to samples known to contain many other nematodes, the possibility that real reads were being discarded was tested by varying the threshold at which pipelines discard sequences as probable chimeras or sequencing errors. Pipelines may discard in an early step to simplify further processing (APDP) or a late step to ensure that only unusual singletons are eliminated and not slight variants (OCTUPUS). The effect on pipeline errors (see definition above) of varying this threshold was tested for each of the pipelines by rerunning the pipeline with the threshold set at 1, 2, and 3 reads.

Statistical analysis. All analyses were conducted using Genstat 14 (54). Numbers of nematodes and species were analyzed by regression, with Fisher’s protected least significant difference at \( P = 0.05 \) used for multiple comparisons. Numbers of nematodes added and sequence reads were analyzed by linear regression of logarithmically transformed data (log \( x + 1 \)). Presence or absence in samples was analyzed by logistic regression. In addition to the standard alerts incorporated into Genstat, plots of residuals versus fitted values were visually inspected to confirm absence of heteroscedasticity.

RESULTS

Nematode counts by microscope. The total number of nematodes extracted from the substrates differed significantly between the different substrates (\( F = 88.36, \text{df} = 3,120, P < 0.001 \)) (Fig. 1). As expected, the total number was the lowest within the potting mix and slightly but not significantly higher in banana soil, significantly higher in the sugarcane soil, and significantly higher again in the forest soil. The amended substrates showed a similar pattern of increase, except with significantly more individuals as expected (\( F = 195.14, \text{df} = 1,120, P < 0.001 \)) (Fig. 1). The increases were =500 nematodes per sample, slightly less for the forest soil (Fig. 1).

Molecular data. In the molecular analysis, the gaskets differed in their total numbers of reads (373,144 from gasket 1 and 439,267 from gasket 2). The pipelines differed in their interpretation of the total reads. In total, =85.5% passed the quality assessment within OCTUPUS and =74.5% passed the first-step filtering within APDP. Among these high-quality reads, OCTUPUS identified =54% as nematodes, =17% as other eukaryotes, and 30% as chimeras, errors, or unidentifiable. APDP identified 37% as nematodes, 20% as other eukaryotes, and 42% as chimeras, errors, or unidentifiable. The total number of high-quality reads differed between pipelines (\( F = 28.02, \text{df} = 1,118, P < 0.001 \)), between gaskets (\( F = 9.22, \text{df} = 1,118, P < 0.001 \)), and among the different substrates (\( F = 29.76, \text{df} = 3,118, P < 0.001 \)) but not between the amended and unamended substrates (\( F = 3.41, \text{df} = 1,118, P = 0.067 \)) or between amended and unamended substrates among the different substrates (amendment–substrate interaction: \( F = 2.17, \text{df} = 3,118, P = 0.095 \)).

In the unamended substrates, few plant-parasitic nematode species were found per sample in the potting mix, significantly more in the banana and forest, and significantly more again in the sugarcane soil (\( F = 926.6, \text{df} = 3,110, P < 0.001 \)) (Fig. 2A). The number of other nematode species increased from potting mix to banana to sugar to forest (\( F = 277.5, \text{df} = 3,110, P < 0.001 \)) (Fig. 2B). The total number of species of other eukaryotes (especially fungi) was greatest in the potting mix, lowest in sugarcane, and intermediate in banana and forest (\( F = 156.2, \text{df} = 3,110, P < 0.001 \)) (Fig. 2C). Thus, the unamended substrates varied in their abundances and diversities of plant-parasitic nematodes, other nematodes, and other eukaryotes, as summarized in Table 1.

Adding three species to the sample increased the number of plant-parasitic nematode species recovered by very close to three species for all substrates (\( F = 205.1, \text{df} = 3,110, P < 0.001 \)) (Fig. 3).
2A); there was no difference between substrates (\(F = 0.06, df = 3,110, P = 0.980\)). Two of the added species were always identified correctly as *P. neglectus* and *P. thornei* (Fig. 3A and B). However, the third species was correctly identified as *H. avenae* in only 50 to 60% of the positive samples, depending on the informatic pipeline (Fig. 3C). The species was correctly identified as absent in 78 to 100% of the negative samples, again depending on the pipeline. Incorrect identifications of *H. avenae* varied among the different substrates but were not related to the abundance or diversity of any of the other groups in any simple manner.

Adding the extra species changed the numbers of other nematode species and other eukaryotic species detected (\(F = 25.69, df = 1,110, P < 0.001\) and \(F = 70.21, df = 1,110, P < 0.001\), respectively) (Fig. 2B and C). However, the effects were not uniform across substrates. Adding nematodes generally increased diversity of other nematodes and increased it more where initial diversity was higher (\(F = 10.69, df = 3,110, P < 0.001\)) (Fig. 2B). Adding nematodes decreased diversity of other eukaryotes and decreased it more where initial diversity was higher (\(F = 13.49, df = 3,110, P < 0.001\)) (Fig. 2C).

The recovered sequences of all three species added were identical to the known sequences of the nematodes input. Most of the high-quality sequences of plant-parasitic nematodes other than those added were very similar to named sequences in GenBank, with approximately half identical and most >98% similar (Table 3). However, some of the species names associated with the sequences differed between pipelines. Differences in the species names were least for nematodes (e.g., for plant parasites, 3 species different of 36 total) and greater for other eukaryotes (e.g., for fungi, almost half different).

Altogether, 55% of the plant-parasitic nematode species not added experimentally were found only in a particular substrate but were in all replicates within that substrate; the remainder were found in only some replicates within a substrate (Table 4). Most differed only between substrates, not among gaskets, pipelines, or between amended and unamended samples (Table 4). One species differed between gaskets, four between pipelines, six according to whether nematodes were added experimentally or not, and the identifications of three species were affected by combinations of gasket, pipeline, and amendment. Other types of nematodes and other eukaryotes varied considerably with substrate but most did not differ between amended and unamended samples; differences in total numbers of species (noted above) were the result of differences in a few species only (Supplemental Figures 1 to 5).

For each of the species added, the total number of sequence reads varied between pipelines and among substrates. For each of the species of *Pratylenchus* added, there was a strong correlation between number of reads and numbers of nematodes added; however, the relationship was logarithmic and varied significantly between species, pipelines, and substrates (Fig. 4).

**Sources of error for *H. avenae*.** For *H. avenae*, all pipelines identified >65% of samples correctly, with OCTUPUS worst and APDP(A+T) best (Fig. 5). All pipelines identified correctly more unamended samples (75 to 90%) than amended samples (50 to 59%). OCTUPUS had the highest pipeline error, and most of these were false positives; APDP had fewer incorrect results, with APDP(A+T) having even fewer incorrects, and all were false negatives.

In searching the raw data, sequences from *H. avenae* were not detected in 12 amended samples (where they should have been), and were detected in 3 unamended samples (where they should not have been) (Tables 2 and 5). OCTUPUS returned the same result as the detailed search eight times (DNA search errors) but reversed seven (combined errors); most of the reversals were amended samples in which *H. avenae* was not identified in the detailed search of raw data (Fig. 5). APDP returned different results to the three unamended positives from the raw data search (combined negatives). APDP(A+T) recorded all detections the same as the detailed searches of raw data; there were no combined errors (Fig. 5). Thus, APDP(A+T) unequivocally identified all 15 errors as DNA sampling errors, whereas APDP identified 12 as DNA sampling errors and 3 as combined errors, and OCTUPUS identified 8 as DNA sampling errors and 7 as combined errors.

Increasing the number of reads at which sequences were eliminated as likely chimeras or sequencing errors resulted in slightly fewer false-positive pipeline errors for OCTUPUS, and slightly more false-negative pipeline errors for APDP (Table 6).

**DISCUSSION**

The results show that the ecometagenetic technique first identified the complexity of the soil eukaryotic communities within the substrates, then correctly detected and identified the three...
Fig. 3. Number of reads of A, *Pratylenchus neglectus*; B, *P. thornei*; and C, *Heterodera avenae* in each sample of each substrate recovered by each pipeline. Eight samples are arranged in order from left to right within each substrate: sample 1 gasket 1; sample 2 gasket 1; …sample 1 gasket 2; sample 2 gasket 2…and so on.
species added under all four scenarios of substrate complexity. For both the *Pratylenchus* spp., the identifications—positive and negative—were unequivocally correct. Furthermore, the number of reads gave a good indication of the number of nematodes present. For *H. avenae*, the results were not as good. However, considering that the species was present as a single, resistant, resting, immobile cyst among many nematodes, fungi, and other organisms, correct identification of >50% of positive samples and >75% of negative samples is impressive. By comparison, standard, nonmolecular methods for extraction and identification of cyst nematodes are seldom 100% efficient and may be only 60 to 70% (6,45,50).

For diagnosis of nematodes in real situations, there are many potential methodological considerations (23). In the current experiment, the effects of many of the factors likely to influence the results obtained by ultrasequencing have been investigated, including their interactions. The results indicate that further

<table>
<thead>
<tr>
<th>Species</th>
<th>N</th>
<th>100%</th>
<th>99%</th>
<th>98%</th>
<th>&lt;98%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Added species</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Plant-parasitic nematodes</td>
<td>33</td>
<td>12</td>
<td>8</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Free-living nematodes</td>
<td>126</td>
<td>31</td>
<td>21</td>
<td>14</td>
<td>60</td>
</tr>
<tr>
<td>Other eukaryotes</td>
<td>186</td>
<td>73</td>
<td>37</td>
<td>20</td>
<td>54</td>
</tr>
</tbody>
</table>

**Fig. 4.** Relationship of number of A, *Pratylenchus neglectus* and B, *P. thornei* with number of sequence reads identified by OCTUPUS and APDP pipelines. For *P. neglectus*, ln(number of reads) = number of nematodes × 0.002820 + 3.229 for OCTUPUS, and ln(number of reads) = number of nematodes × 0.002820 + 3.005 for APDP (%$R^2 = 78.7$, df = 2.61, $P < 0.001$). For *P. thornei*, ln(number of reads) = number of nematodes × 0.003422 + 3.511 for OCTUPUS, and ln(number of reads) = number of nematodes × 0.003422 + 3.005 for APDP (%$R^2 = 82.7$, df = 2.61, $P < 0.001$). To clarify the graph, position of the points and line for APDP are offset by 10 units to the right.
investigation of some of these factors in realistic situations is desirable. Not all factors were investigated completely but comprehensive experimental analyses of all potential influences in any systematic way would require a prohibitively large study; investigation of the interactions of these factors as well would involve an even larger study. The main factors potentially affecting diagnostics using ultrasequencing are, we suggest, included in the current investigations.

One of the main potential factors causing bias in results from ultrasequencing is that other species of nematodes and other organisms may or may not be present, and in small or large numbers. Using real substrates with a range of natural complexity, extracted using a common method, showed that this can be an issue under these realistic field circumstances and not just under controlled experimental conditions. Non-plant-feeding nematodes and other eukaryotes rather than plant-parasitic nematodes seemed to be most affected.

The number of species in a sample influences the diversity of DNA present. Having a greater diversity of DNA can cause errors in several ways, including interference or competition between DNA in PCR reactions, misidentifications because of identical sequences, or bioinformatic confusion between sequences that represent different species as opposed to variants of species already identified. To a certain extent, this can be controlled by using a selective extraction method but, if the target species are not known beforehand, this is impossible. Even so, other organisms are commonly present as contaminants, even in quite selective extraction methods. One of the main attractions of ultrasequencing is that it allows simultaneous detection of many species; therefore, other methods will be more suitable for highly directed detection of particular species where there are few or no other species present.

The present investigations used the extremely common Whitehead tray technique for extraction of all organisms except the experimentally added *H. avenae* cysts. This is among the simplest nematode extraction methods and is in wide use in many laboratories. It was preferred over potentially more efficient but specialized methods such as elutriators or flotation techniques because the more specialized methods are used less and generally in laboratories dedicated particularly to nematodes. Using Whitehead trays gave a realistic range of other organisms in the samples (such as protists—particularly ciliates—and rotifers), as well as any nematode parasites such as *Pasteuria* spp. and organisms included in the gut contents in of the nematodes. Rates of nematode extractions can differ considerably from different substrates using Whitehead trays (18,23,24,29,30).

Another factor potentially biasing results from ultrasequencing is the number of individuals of a particular species. The amount

![Fig. 5. Percentage of all sample identifications of *Heterodera avenae* (*N* = 32 positive and 32 negative) with experimental treatment manual searches of raw data and identification by pipeline all in agreement (all true), pipeline making an identification at variance with experimental treatment and manual search of raw data (pipeline error), manual search of raw data at variance with experimental treatment and this error propogulated by the pipeline (DNA sampling error), and manual search of raw data at variance with experimental treatment and pipeline (combined error).](image)

<table>
<thead>
<tr>
<th>Influence</th>
<th>No</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gasket</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Pipeline</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Amendment</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Pipeline–amendment interaction</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Gasket–pipeline interaction</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Substrate only</td>
<td>5</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>18</td>
<td>33</td>
</tr>
</tbody>
</table>

* Substrate refers to the number of species that differed only between substrates and not according to any of the other influences within the substrate.
of DNA present is related to the numbers of individuals in both absolute and relative terms. Similar to the diversity of DNA, the amount of DNA can influence results, as outlined above. In the present experiment, the influence of the amount of DNA was investigated using the different numbers of the *Pratylenchus* spp. added to the substrates, and the single cyst of *H. avenae* added afterward.

The numbers of *Pratylenchus* spp. added to samples correspond approximately to abundances above, at, and below the threshold for plant damage in cereals, their main local hosts (52). Different extraction efficiencies among substrates, however, meant that the numbers added to the different substrates were not related uniformly to the numbers extracted: this is a feature of nematode extraction in practice (6). Nevertheless, to imitate nonexperimental conditions, the *Pratylenchus* spp. were added before extraction to ensure that they were used with all associates, and that any selection for particular individuals during extraction was included.

For reasonably large numbers of vermiciform nematodes of well-characterized species, it appears that, under most of the circumstances tested, ultrasequencing detected the nematodes consistently. Dose-response curves could be constructed by adding precise numbers of nematodes to samples after extraction but improved detection of presence at low numbers when DNA is inaccessible may add more value for quarantine situations.

A single cyst of *H. avenae* was used because it represents a lower limit for detection. It was added after the extraction process to ensure that just the single cyst was present in the sample: there was no opportunity for hatching or loss in the extraction process. No nematode extraction method is 100% efficient (6,23–25,45) and adding the nematode prior to extraction may have meant that it was not in the extract. It would also have necessitated using a different extraction technique because cyst nematodes are not extracted by Whitehead trays (unless they hatch to vermiform juveniles). A single vermiciform nematode may have been damaged during introduction, meaning that a viable nematode was not present in the final sample from which DNA was extracted but the extremely tough cyst could be introduced with little prospect of mechanical damage.

Another reason for using *H. avenae* was to test the effect of different types of nematodes. The detailed analysis of results for *H. avenae* suggest that extraction of DNA from cyst nematodes may require different, more disruptive, techniques than the bead-beating used. Such techniques may, of course, affect the absolute and relative amounts of DNA from vermiciform nematodes and other organisms; therefore, further trials may be required.

The presence of closely related species may bias results as well via mechanisms similar to those of the other biasing factors. In the present investigations, this was tested using the two species of *Pratylenchus*. The results for these species may not apply for all closely related species but testing all possible sets of closely related species of pest nematodes would be an impossible job because there are so many (46). Furthermore, the results indicate that the presence of sibling species may not be the biggest issue for ultrasequencing.

**Detection of Pratylenchus spp.** For large numbers of nematodes, even for the closely related species *P. neglectus* and *P. thornei*, detection was consistent in amended samples and not affected by the gasket or bioinformatic pipeline; that is, there were no false positives or false negatives. For each *Pratylenchus* sp., the number of reads was quite closely correlated with the numbers input (%$R^2$ values >80%) (Fig. 4) but the relationship was nearly linear only on a logarithmic scale. This form of relationship means that the errors and uncertainty of predicted values become larger as the number of nematodes and, consequently, the number of reads increases. Hence, estimates of abundance based on numbers of reads should be regarded as qualitative estimates only, particularly for high abundances. At the other end of the scale, the intercept of the regression line with the Y-axis is theoretically an estimate of the number of reads expected when no nematodes are present. This threshold was not an issue in the case of the two *Pratylenchus* spp. because they were always detected correctly but the number of reads regarded as possible when no nematodes are present did change the errors in detection of *H. avenae* (Table 6). Examination of re-

### TABLE 5. Comparison of identification of *Heterodera avenae* from individual replicate samples using whether a cyst was experimentally placed in the tube containing nematodes (Correct), manual searches of raw data (RD), the OCTUPUS pipeline (OCTUPUS), the unaltered APDP pipeline (APDP), and the APDP pipeline with criteria eased to allow one nucleotide different in the primer sequence (“APDP(A+T)”)*

<table>
<thead>
<tr>
<th>Amendment, substrate, gasket (1 or 2), and rep (1–4)</th>
<th>Potting</th>
<th>Banana</th>
<th>Sugar</th>
<th>Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Unamended</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OCTUPUS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APDP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APDP(A+T)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amended</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OCTUPUS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APDP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APDP(A+T)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Symbols and abbreviations: + represents positive identifications and – represents negative identifications. Shaded cells represent pipeline false positives (Pipe F-pos), cells with borders represent pipeline false negatives (Pipe F-neg), cells with symbols in bold represent DNA sampling errors (SE) (positive or negative according to the symbol), cells with either shading or borders and bold symbols represent combined errors (CE), and unmodified cells with unmodified contents represent true positives or negatives according to the symbol within.

### TABLE 6. Effect of increasing the stringency of *Heterodera avenae* detection in the raw data on pipeline errors by different analysis pipelines

<table>
<thead>
<tr>
<th>Reads*</th>
<th>OCTUPUS</th>
<th>APDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>False positive</td>
<td>False negative</td>
<td>False positive</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Minimum reads for *H. avenae* presence.
gressions similar to those for the Pratylenchus spp. may yield better estimates of the optimum threshold for regarding reads as real.

The relationship between abundance and number of reads was very different for the two added species: the numbers of reads of P. neglectus were an order of magnitude smaller than P. thornei (Fig. 4A and B) even though the numbers of individuals in the samples was only approximately two-thirds (Table 1). Differences in extraction efficiency (related to different mobility) may be involved but more likely it is a different number of rDNA copies in the two species. The number of rDNA copies can vary significantly even between closely related species and, thus, different nematode species can generate very disparate numbers of reads per individual (17,39,41). Unfortunately, the number of rDNA copies for these two and most other nematode species are yet to be characterized and, until this is done, inferences about abundance may be limited to relative comparisons across samples at best. Variability in numbers of copies of the rRNA gene can also occur within species (26,31) but is also unknown for most nematodes. The nonlinearity of the relationship between number of reads and abundance may also relate to the number of other organisms in the samples. This could have two effects. One is to dilute the relative proportion of each species rDNA in a sample; in pyrosequencing, the number of reads may be more related to proportions in a sample than to absolute abundance. The other effect is to create a larger pool of different sequences. Pipelines may have more difficulty discriminating errors and real species as the real pool of different sequences increases and the differences between sequences in the underlying sample become smaller. There is also a greater chance of incorrectly assigning a sequence.

Differences in amplification efficiency during the PCR step may also be involved, because these can vary considerably (28). The relationships between nematode abundance and number of reads were statistically different but the magnitude of the differences between the different pipelines and substrates was small. This indicates that, once a relationship is determined for a species, it may be used as a rough estimator of abundance. A test of whether the relationship is similar for different populations of P. neglectus and P. thornei will confirm this.

**Detection of H. avenae.** Detection and identification of H. avenae was less reliable than that of Pratylenchus spp.: it was not detected in ≈50 to 80% of the samples receiving the amendment, depending on the pipeline or search of raw data (Table 5). In addition, H. avenae was detected in several samples processed by OCTUPUS where it should not have been (Fig. 3C). Because OCTUPUS accepted more reads than APDP, it may fail to recognize errors sufficiently; however, the price may be inferring the presence of species where there are none (false positive). By contrast, APDP employs more conservative criteria to better detect errors and discards more reads than OCTUPUS but the price may be failing to identify real low-abundance species, leading to some false negatives.

**Types of errors.** The results also showed that such promising results can be skewed by many different factors. We have distinguished between errors relating to DNA sampling procedures and those derived from the informatic analyses of the sequence data using eight categories (see Materials and Methods). For most of the categories, the interpretations are clear: i and ii represent correct interpretations, iii and iv represent pipeline errors, and v and vi represent DNA sampling errors which are then promulgated by the pipeline. The interpretation of categories vii and viii is less clear and there are several indistinguishable possibilities. One is a DNA sampling error resulting in no H. avenae sequences where there should be some, followed by a pipeline error incorrectly identifying H. avenae even though sequences were really absent. Another possibility is an extremely variant but real sequence not detected by the search (maybe not even on GenBank) but identified as most likely associated with H. avenae by the pipeline, or a sequence that was not really a variant of H. avenae being incorrectly identified as a variant by the search, then being correctly identified as another species by the pipeline. There is also the possibility of an error in the alignment or search resulting in an incorrect call on the presence of H. avenae sequences in the individual search (34), while the pipeline correctly identified the sequences. Further investigations will be required to distinguish these possibilities.

Identifying the sources of potential errors is important to direct development of protocols that will mitigate the errors because different strategies will be needed to address the different types of errors. Strategies to avoid DNA sampling errors include better lab protocols to avoid contamination, alternative DNA extraction techniques, and increased sequencing depth, whereas strategies for avoiding informatics errors include higher or lower sensitivity to sequence differences and improvements to error detection.

**Sampling errors.** Some of the sampling error false negatives for H. avenae may be related to the small number of nematodes added: a single cyst as opposed to a thousand each of the veriform Pratylenchus spp. The result may be very low abundance in the DNA pool and, consequently, no sequences in the raw data. However, the difference in the number of nematodes added may be less than first appears because a single cyst contains many eggs and juveniles (53). In real situations, the quality of cysts can vary, as can their contents of very few to up to 600 eggs (49). This would also need to be considered in deploying ulrasequencing but is probably not a factor in the present experimental results because all the cysts came from a uniform culture, where conditions were good for the nematodes and the cysts contained many eggs (J. Lewis, personal communication). For practical use, though, one of the supposed advantages of molecular methods is the ability to detect small numbers of a target pest (14); therefore, improving detection at low abundances is an area for future refinement of ecometagenetic methods.

Perhaps more likely is that the poorer detection of H. avenae is related to its biology making DNA extraction hard. In contrast to the active, vermiform Pratylenchus spp., the globose cyst of H. avenae is a dead female body with a hardened cuticle filled with eggs. It is a survival stage protecting the live eggs inside against harsh or unsuitable conditions such as drought or absence of hosts lasting from months to years (32). As such, the cyst is very resistant to mechanical damage, such as that used in bead beating to extract DNA. In addition, the eggs encased inside the cyst are themselves designed for survival. Their chitinous eggshells are also resistant to mechanical breakage and release of DNA. Considering this resistance to damage, plus the small number added, plus the presence of a slightly different priming sequence, the ability of our methods to detect H. avenae in half or more of the samples where it was present is actually rather surprising. Although sample processing with bead-beating appears reliable for veriform life forms of nematodes (38,41), the present results indicate that further careful evaluation may be required for resistant life stages.

In many nonexperimental situations, vermiform juveniles are present together with cysts. In these cases, detection of H. avenae may match the consistency of that for the veriform Pratylenchus spp. However, in many other applications such as quarantine testing or testing before planting this may not be the case. In these cases, improved methods of DNA extraction may need to be developed.

The differences in results for the Pratylenchus spp. and H. avenae are important. In many situations, it will be unknown which nematode pests are present; therefore, methods must be robust to accommodate different biological characteristics. Any successful deployment to real situations will need to be able to reliably detect a range of nematodes; however, our results indicate that further work will be required to achieve this.
Sampling error false positives could be caused by presence of target DNA through laboratory contamination or the presence of unsequenced and probably undescribed species with identical or near identical sequences.

**Informatic errors.** For informatic errors, false negatives are probably caused by *H. avenae* sequences being incorrectly identified as errors and removed, while false positives are probably caused by the error-correction approaches in some pipelines, which have been shown to create novel sequences not observed in the raw data (19). Comparison of results from the different pipelines allows more precise identification of these errors because the processing of sequence reads varies so much between the pipelines.

One of the main differences may be in the algorithms used to detect the errors and chimeras common in metagenetic data (40). Our results show that these differences can generate conflicting results and, ultimately, misrepresent the underlying reality. For instance, in OCTUPUS, reads are initially checked for quality by a Lucy-trim phred-based algorithm assessing how well nucleotides are called. By contrast, APDP uses 100% primer and MID tag matching. This resulted in different pools of reads accepted for the next-step analyses.

Differences also occurred between pipelines in the grouping of sequences into clusters because different criteria are used. The reads are clustered at 99% identity in OCTUPUS and 100% identity in APDP. This affected the number and identity (nucleotide order) of final “consensus” sequences. How and when these sequences were then matched against reference databases also produced subtle differences in the names associated with them.

OCTUPUS showed a propensity for false positives and APDP for false negatives (Fig. 5; Table 2). There were few false negatives in OCTUPUS but those present probably resulted from the minimum number of reads required to call a series of similar sequences an OCTU. As discussed above, sequences represented by only 1 or 2 reads often represent errors and are treated as such (39). However, they can be real but rare; therefore, the only way to ensure that such sequences are not discarded may be to ensure that they are more frequent than the false sequences. This may involve more PCR or higher read coverage.

For APDP, there were more false negatives and they had a different origin: the PCR primer used in this experiment had a single-nucleotide difference from the *H. avenae* reference sequence, making all reads with this mismatch undetectable. When this biological factor was accounted for by allowing for one mismatch in the expected primer sequence, APDP(A+T) recorded only a single false negative (Tables 2 and 5). This result suggests that the false negatives are not a result of the sensitivity of the analysis as in OCTUPUS. Careful investigations of the effects of this change in APDP on detection of other species are necessary before it could be recommended for samples where the target species was unknown or, more specifically, the primer sequence variation in nematode species is unknown. Allowing one base mismatch may reduce the false-negative rate but also reduce the discrimination between other species. Only a comprehensive survey of all nematodes is likely to decide this, and such a survey was beyond the resources available in the present study.

In contrast to false negatives, OCTUPUS had a high false-positive rate. This may be explained by the pipeline combining sequences originating from different species into single OCTUs, creating sequences in some samples that were not observed in the raw DNA sequence data. This behavior has been observed in other analysis tools that similarly correct or change the raw data during processing, in order to account for errors that occur through the data generation procedure (19). In fact, these corrections of the data may be justified in some cases; some of the combined errors for OCTUPUS reversed an apparent absence of the sequence for *H. avenae* in samples where it should have been present to give an ultimately correct result (Table 2, combined errors). However, the price for this was identification of *H. avenae* in samples where it was really absent.

Because APDP filters the raw reads rather than changes the data, it is less susceptible to this form of error (33). Hence, there were few false positives.

Because neither informatics pipeline was unerring for detection of *H. avenae*, and their errors were largely complementary, data analysis with more than one bioinformatic pipeline may be desirable. Indeed it may improve interpretation. For example, where a target such as *H. avenae* is detected by both pipelines, it could be regarded as a true positive. Absence in both pipelines could be regarded as a true negative. Where only one pipeline detects a target, a sample could be considered conditionally positive and recommended for further testing using methods optimized for the species in question. With the methods used in this experiment, 20% of samples would require further testing; however, this procedure should increase confidence in interpretation of results and make identifications more robust.

Despite these significant differences in how each pipeline handled the data and some samples with opposite results, most results and read patterns produced by the pipelines were largely congruent.

Other bioinformatic pipelines have been suggested for ultra-sequencing data, particularly from prokaryotes (e.g., QIME and MOTHUR) (8,44). These pipelines, like OCTUPUS, are based on clustering methods. All three pipelines have been benchmarked against each other, with interpretation of data largely congruent (4,12,58). The goal of the present article was not pipeline benchmarking but, rather, the ability of ecometagenetics to detect and identify pathogens; therefore, the benchmarking process was not repeated because it is an extensive project in itself. The APDP pipeline takes a fundamentally different approach to error screening, which may produce rather different results (33), and so was chosen to test the potential effect of different pipelines within the limited resources available. An expectation of using ecometagenetics for detection of pathogens is that the results should be consistent regardless of the pipeline. Our results demonstrate that this is not the case but it needs to be addressed if using ecometagenetics for screening for pathogens. We have presented comparisons of the pipelines not to suggest one or the other pipeline is better but to show that, in practical, real-world situations, pipelines can make a difference.

**Detection of plant parasites other than those experimentally added.** An advantage of the ecometagenetic approach to pest detection is that it allows detection of many species simultaneously. In addition to the known species added experimentally, other plant-parasitic nematodes were detected quite consistently. The total numbers of plant-parasitic nematode species detected followed the pattern expected, with potting mix having very few species other than those added and the other substrates having high diversity.

The identities of the species were also close to those known from the various substrates. In the banana soil, most of 11 plant-parasitic nematode species identified in our study occur commonly in the soil, together with the dominant *Radopholus similis* (35,36).

In the sugarcane soil, seven species of plant-parasitic nematodes are known (51) and our study identified five of them. We also identified six new species, two of which were close relatives to species known to be present in the sugarcane soil (*P. zeae* and *T. annulatus*); the rRNA gene sequences for *P. zeae* and *T. annulatus* overlapping our amplicons are not on the NCBI database. Other new species in the sugarcane soil included additional species of the genus *Meloidogyne*, which is both notoriously difficult to identify and has many new species being described (9,13,43).

As expected, the forest soil contained very different plant-parasitic nematode fauna, with little overlap with the agricultural
soils, and uniquely contained some species from the family Criconematidae, many of which are typical of undisturbed soils and rarely do well under conditions of intensive agricultural management (42). Species unknown to science are common among the nematodes and many other eukaryotic plant pathogens such as fungi (21,37). In all, <5% of the estimated total number of species in either group has been formally described. Of the ≥24,000 described nematode species, only ≥1,800 have sequences deposited on GenBank (21). Plant-parasitic nematodes causing economic damage in Europe or North America are much more likely to have sequences available than other, free-living species. (Damaging animal or human parasites also have many sequences on GenBank.) The vast majority of nematode species have never been sequenced. Given the proportion of all species with no sequences available and the Palaeartic geographic biases in the sequences that are available, it is, in fact, surprising that the lists of plant-parasitic nematodes in our study are so close to what is known from the soils. This apparent success should be tempered by the observation that many substrates and species are uncharacterized, presenting very real possibility that species—particularly closely related ones—could be misidentified because they may be indistinguishable based on any one part of the genome (D. Porazinska unpublished data). There is also the same propensity to morphologically misidentify what are really unknown indigenous species as damaging cosmopolitan, European, or North American species, because the latter are well known and described in the literature and morphological keys, whereas the former are not.

Identification of free-living nematodes and other eukaryotes. Generally, the results for free-living nematodes and other eukaryotes show that errors can occur in metagenomic methods. The apparent—but not real—changes in their diversity in samples when the three species were added experimentally illustrate this (Fig. 2B and C). In the case of free-living nematodes, the apparent changes in diversity may come from several sources. One source is that sequencing errors may become more common in the higher presence of closely related rRNA gene sequences (16,40). Another source may be variation among copies of the sequence: the bioinformatic pipelines may identify a variant read as a different species. These sources of inflation of diversity estimates may be compounded by other artifacts generated by the algorithms used to identify clusters of reads as separate and distinct species (60). The observation that the changes in response to adding the three species of nematodes were opposite in free-living nematodes and other eukaryotes shows the potential intricacies of the interactions of these and other potentially confounding or conflicting errors in metagenomic methodologies. Clearly, more investigations are warranted before deployment of the techniques to wider soil communities.

In considering organisms other than nematodes, the selection of the gene region is also a factor. In the case of fungi, there was often little congruence between the results from the different pipelines because multiple high-identity matches were available for each sequence. This lack of convergence most likely indicates that the gene region used in this study, while appropriate for nematodes, was less suitable for fungi. Refinement of the gene region, or possible use of multiple gene regions, may be desirable for expanding the technique to a wider range of taxa.

The techniques used in our experiments were tailored to nematodes (and other organisms) removed from the substrates before DNA was extracted. This is a potential source of artifacts in results for other eukaryotes. A potential solution to this issue is DNA extraction directly from the substrate (3). However, this option can introduce other biases and errors; therefore, further investigation is still required (12).

Conclusion. Metagenomic techniques have considerable potential to recover a wide range of organisms simultaneously (10,12). This characteristic makes them useful for studies of soil sustainability and health. Our results show that there are many challenges in verifying the methods for many organisms in complex mixtures in complex substrates. However, we have shown that, for well-known target species of nematodes such as may be found in biosecurity situations, species can be detected in both simple and complex substrates. We have further shown that sources of errors can be identified and, therefore, investigated to optimize the technique for use with wider ranges of organisms under different scenarios. Many suggestions for such optimization have been suggested by our experiment and analysis.

ACKNOWLEDGMENTS

We thank the Office of the Chief Executive CSIRO for funding the Distinguished Visiting Fellowship that made this work possible; the CSIRO Biosecurity Flagship for operating funds; R. Eichner (Australian Department of Agriculture Fisheries & Forestry, Melbourne), S. Buyan (Bureau of Sugar Experiment Stations, Bundaberg), T. Pattison (Department of Agriculture Forestry & Fisheries Queensland, South Johnstone), and G. Stirling (BioCrops, Brisbane) for supplying substrates; and J. Lewis (SARDI) for supplying nematode cultures.

LITERATURE CITED

as drivers of chimerism in nSSU metagenetic analyses. Nucleic Acids Res. 40:466.