Journal of Nematology 44(1):18–25. 2012.

The Nature and Frequency of Chimeras in Eukaryotic Metagenetic Samples.

Dorota L. Porazinska,[1] Robin M. Giblin-Davis,[1] Way Sung,[2] W. Kelley Thomas[2]

Abstract: Pyrosequencing of an artificially assembled nematode community of known nematode species at known densities allowed us to characterize the potential extent of chimera problems in multi-template eukaryotic samples. Chimeras were confirmed to be very common, making up to 17% of all high quality pyrosequencing reads and exceeding 40% of all OCTUs (operationally clustered taxonomic units). Typically, chimeric OCTUs were made up of single or double reads, but very well covered OCTUs were also present. As expected, the majority of chimeras were formed between two DNA molecules of nematode origin, but a small proportion involved a nematode and a fragment of another eukaryote origin. In addition, examples of a combination of three or even four different template origins were observed. All chimeras were associated with the presence of conserved regions with 80% of all recombinants following a conserved region of about 25bp. While there was a positive influence of species abundance on the overall number of chimeras, the influence of specific-species identity was less apparent. We also suggest that the problem is not nematode exclusive, but instead applies to other eukaryotes typically accompanying nematodes (e.g. fungi, rotifers, tardigrades). An analysis of real environmental samples revealed the presence of chimeras for all eukaryotic taxa in patterns similar to that observed in artificial nematode communities. This information warrants caution for biodiversity studies utilizing a step of PCR amplification of complex DNA samples. When unrecognized, generated abundant chimeric sequences falsely overestimate eukaryotic biodiversity.

Key words: chimera, biodiversity, metagenetics, multi-template PCR, nematode, ultrasequencing.

Recent advances in high-throughput sequencing have been transforming the field of microbial ecology by uncovering vast numbers of completely new and uncharacterized species (Amaral-Zetler et al., 2010; Fonseca et al., 2010; Porazinska et al., 2010). Because the microbial world is still largely unknown, each novel sequence has the potential for representing a true novel species. It is well recognized, however, that sequence novelty can arise artificially through PCR artifacts (e.g. chimeras) and sequencing errors and thus when unrecognized result in false perceptions of microbial diversity (Wang and Wang, 1997; Reeder and Knight, 2009). While attempts to understand the nature of these problems and consequently to correct them have been made in the prokaryotic community (Quince et al., 2009; Edgar et al., 2011; Haas et al., 2011), parallel attempts in the eukaryotic world have been lagged. In principle, PCR/sequencing artifacts leading to formation of artificial novel sequences in eukaryotic microbiota might be similar to those in prokaryotes. However, the nature of eukaryotes expressed by e.g. their multicellularity, the presence of reproductive organs and cells, or large but yet unknown copy numbers of rDNA genes used in metagenetics could amplify these artifacts. Because the extent of knowledge about microscopic eukaryotic diversity is minimal, differentiating between truly novel sequences from those that arise artificially is problematic. Typically, molecular databases that are used for identity matching contain only a small fraction of all potential species signatures. For instance, out of the predicted >1 million nematode species (Lambshead, 2004), <4% is formally described (Hallan, 2007) and only ~6,000 are molecularly characterized (NCBI) and present in DNA sequence databases. In the absence of comprehensive databases, a better understanding of how chimeras form and their nature, frequency, and identity is of utmost importance. Here, we elucidate these issues by analyzing three 454-generated pyrosequencing datasets that include two replicate datasets from artificially assembled but realistic nematode communities of known identities and densities (Porazinska et al., 2009; 2010a) and one dataset consisting of a total of 12 true environmental samples. We define a chimera (hybrid, recombinant) as a DNA molecule that arises from at least two different DNA parental molecules during a PCR amplification of mixed (e.g. different species) DNA templates. The most common mechanism leading to a formation of a chimeric sequence is incomplete extension of one template that anneals to another in a next PCR cycle (Edgar et al., 2011). Precisely because of the fact of going through at least one less round of PCR amplification than the parental sequences, chimeric sequences are intuitively assumed to be less frequent. Theoretically, they can be formed at any point along the sequence.

Materials and Methods

Description of artificial nematode community samples: Two out of seven available datasets (1rA and 3rA, Porazinska et al., 2010a) were used for detailed chimera analysis. Both datasets came from artificially assembled nematode DNA samples of 41 species of known identity and abundance described elsewhere (Porazinska et al., 2009). In terms of nematode abundance, two nematode species were represented by ten and eleven individuals/species

Received for publication December 15, 2011.
[1]Fort Lauderdale Research and Education Center, University of Florida, IFAS, 3205 College Avenue, Fort Lauderdale, FL 33314, USA.
[2]Hubbard Center for Genome Studies, University of New Hampshire, 35 Colovos Rd., Durham, NH 03824, USA.
The authors thank the anonymous reviewers for suggestions that improved the manuscript. This project was supported by USDA/CSREES – TSTAR 2006-04347 (FLA-FTL-04544). Data files (SFF) of the data output generated by GS FLX are available at Short Read Archive, NCBI with the following accession number SRX012333.
E-mail: dorotalp@ufl.edu
This paper was edited by Nancy Kokalis-Burelle.

(*Acrostichus pura* and *Bursaphelenchus mucronatus* 167), one nematode species by two individuals/species (*Diplogastrellus metamasius* 198), and the remaining species by one individual/species. Specifics for DNA extraction and PCR conditions of the ~ 400bp SSU (rapidly evolving sequence of ribosomal DNA coding for the small subunit of rRNA) region are provided by Porazinska et al. (2009; 2010a). The 1rA and 3rA datasets differ in that 1rA was derived from a single in-house mPCR (multi-template) replicate (1r) and 3rA from three separate in-house mPCR replicates that were pooled together (3r). Both in-house amplified samples were independently emulsion-amplified (emPCR) and sequenced on GS FLX generating reads at ~250bp at the Interdisciplinary Center for Biotechnology Research (ICBR) at the University of Florida. Metagenetic data analysis established that 3 out of 41 nematode species failed to amplify (Porazinska et al., 2009, 2010a).

Earlier analysis of the seven datasets established that the results (quantitative and qualitative recovery of species) were consistently repeatable and reproducible and independent of the number of in-house pooled PCR replicates (1 vs. 3) or pyrosequencing runs (Porazinska et al., 2010a). Therefore, for in detail chimera analysis we selected two datasets as representative replicates (1rA and 3rA) that maximized potential differences in chimera outcomes to ensure the robustness of our conclusions. While these two datasets are not *per se* novel, the analyses regarding chimeras (the nature, frequencies, identities, position of breakpoints, distribution of palindromes) is entirely unique.

Sequence processing: Sequences generated from the 454 GS FLX (as above) were processed using a bioinformatics pipeline OCTUPUS (**O**perational **C**lustered **T**axonomic **U**nits for **P**arallel-tagged **U**ltra **S**equencing) (Fonseca et al., 2010). OCTUPUS scanned sequences for quality using Lucy-trim (based on phred-base-calling algorithm that checks sequence quality) with default parameters (Chou et al., 2001) and screened them for a minimum length of 200 bp. Sequences were clustered to OCTUs (Operational Clustered Taxonomic Units) at within OCTU identity cut off levels ranging from 95 to 99% using MEGABLAST (pairwise sequence comparison algorithm) (Zhang et al., 2000) with the minimum overlapping match length set at default 28 nucleotides. MUSCLE (reiterative multiple sequence alignment algorithm) was used on grouped sequences to generate a list of "fixed" (consensus) OCTUs that were blast-matched (pairwised compared and matched to a sequence in a database) (Altschul et al., 1997) against the NCBI database (National Center for Biotechnology Information, repository for molecular biology information), expanded by the nematode reference sequences from our control experiments (Porazinska et al., 2009, 2010a) and nematode reference sequences from Costa Rica (Powers et al., 2009). The reference sequences (~400 bp) for all constituent nematode species within the artificial and Costa Rican communities were constructed by directly PCR-amplifying each nematode species individually and following with traditional Sanger sequencing in both directions (Porazinska et al., 2009). All OCTUs were analyzed for the presence of putative chimeras by comparing the total length of each OCTU sequence against the total length of the Sanger generated sequences present in the database. Because MEGABLAST uses a greedy algorithm that matches as much length as possible, OCTU sequences that match at high identity but incomplete length are likely to be chimeric. Because 99% identity cut off was determined to be the minimum level at which all expected nematode species become recovered (Porazinska et al., 2010c), OCTUs generated at this level were used for in depth analysis of the chimeric process.

All OCTUs flagged as chimeras were individually manually analyzed. All sequences within each chimeric OCTU were aligned to a Sanger generated sequence to which the OCTU sequence showed a significant (>93%) BLAST-match along the 5'-end (Fig. 1) using MEGA 4 with default parameters. Therefore, if a theoretical chimeric OCTU consisted of 5 sequences, each of the 5 sequences was visually compared against the sequence identified in the BLAST-match. These 5 paired alignments were then analyzed for: 1. the position (bp location) of the chimeric breakpoint on the BLAST-match Sanger sequence (e.g. 181, 181, 182, 182, 183bp); 2. the length (bp) of the 3'-end deviating from the predicted 5'-end of the sequence (e.g. 55, 57, 60, 60, 58bp); and 3. the identity of the 3'-end part of the sequence using blast as described above against the NCBI database (e.g. 50 bp 3'end matching *B. mucronatus*). These three types of data collected for each chimeric sequence within each chimeric OCTU were then used to describe the

FIG. 1. An example alignment of three sequences: *Rhabditidoides* n. sp 243 (top), a chimeric OCTU sequence (middle) matching *Rhabditidoides* n. sp 243 on a 5'-end, but deviating from that sequence on the 3'-end, and *Caenorhabditis elegans* (bottom), matching the 3'-end of the chimeric OCTU sequence. Encircled nucleotide illustrates the location of the chimeric breakpoint where the chimeric OCTU changes its identity from *Rhabditidoides* to *C. elegans*. The breakpoint follows a conserved region of 25 nucleotides located between 153 -177bp on *C. elegans*.

nature and frequency of chimeras. To determine whether chimeras form randomly, chimeric OCTUs of six 5'-end representative nematode species (*Bursaphelenchus mucronatus*, *B. abruptus*, *B. platzeri*, *Longidorus*, *Teratocephalus* sp. and *Zeldia punctata*) were visually compared between 1rA and 3rA datasets for the exactness of the identity of 3'-ends and the position of breakpoints. To establish potential explanations for chimera formation, Sanger generated sequences of all nematode species used in the experiment were analyzed for the presence of palindromes using Find Palindromic Sequences (www.BioPHP.org) set to the minimum and maximum length at 6 and 12 nucleotides, respectively.

Environmental samples: A survey of nematode diversity within three habitats (soil, litter, and canopy) in a tropical rainforest at La Selva Biological Station, Costa Rica was conducted in March 2007. Sampling design, sample processing, DNA extraction and sequencing, and sequence processing have been described in detail in Porazinska et al. (2010b). The dataset consisted of 4 soil, 4 litter, and 4 canopy samples (12 total metagenetic samples). Flagging of chimeras has been performed as described above. To understand the relationship between the numbers of reads and OCTUs and the numbers of chimeric reads and chimeric OCTUs of specific taxonomic identities, Pearson correlation was performed using an add-ins StatistiXL program.

Results

Artificial nematode samples: The contribution of chimeric OCTUs in either artificial dataset was consistently high regardless of the within-OCTU-identity cut off and ranged between 32 and 56% (Fig. 2). Identified chimeras accounted for ~2% of all high quality sequencing reads at 95% identity cut off to ~17% at 99% (Table 1). From all chimeric OCTUs (~250 bp on average), ~30% did not involve any nematode species. Among the 70% that did, ~65% represented a nematode-nematode combination of species used in the experiment and 35% a nematode-else combination (e.g. fungus, bacteria, collembolan, human DNA, or an unknown sequence). Typically, chimeras consisted of two parental DNA origins (5'-end different from 3'-end) in direct orientation, but a combination of three DNA origins (<3%) was also observed.

An average length of the 3'-end deviating from the predicted nucleotide sequence of the 5'-end was ~60bp (Table 1, Fig. 3) and 3'-ends of <10 bp were generally not chimeric, but rather indicated a presence of potential end sequencing errors. Approximately 49% of chimeric OCTUs were singletons, 23% doubletons, and 28% contained 3-85 sequencing reads (Fig. 4). In comparison, 96% of the non-chimeric OCTUs were made up of the exact same categories (57% singletons, 15% doubletons, and 23% 3-85 sequencing reads, data not shown).

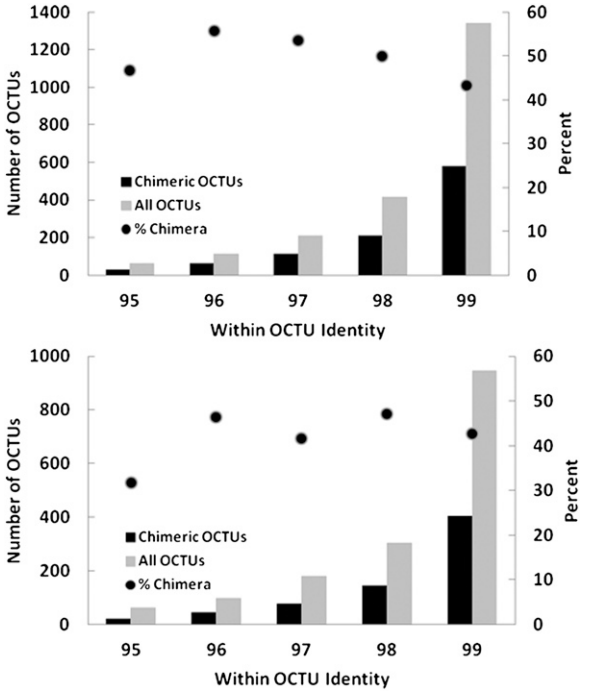

Fig. 2. Numbers and percentages of chimeric OCTUs as a function of different cut off levels for Within OCTU Identity. A) 1rA dataset, B) 3rA dataset.

The apparent chimeric breakpoints were linked to two regions (Table 1) within the 250 bp barcoding Sanger-generated consensus sequence. These two regions (~40 bp apart) were preceded by the presence of a highly conserved (across all nematode species used in the study) site (Fig. 1). More chimeric breakpoints (87%) were associated with the longer (25 bp) conserved site than with the shorter conserved site (Table 1, Fig. 5). No palindromes were detected at either of these two conserved regions, in contrast to the remaining parts of Sanger-generated nematode sequences (Fig. 6). While almost all nematode species were involved in a chimeric combination, at least 5 species (e.g. *Plectus acuminatus*) were consistently absent on the 5'-end, and at least 3 species were not detected on the 3'-end (Table 1). The chimeras were not predetermined because species specific combinations at specific breaking points were not observed across the two artificial nematode community datasets.

Although an average number of chimeric OCTUs per species was 10 and 7 for 1rA and 3rA datasets, respectively, as many as 78 OCTUs (*Acrostichus sp.*) and as few as 1 OCTU (e.g. *Bursaphelenchus tusciae*) per species were observed, similarly to non-chimeric OCTUs (Porazinska et al., 2010c). The number of chimeric OCTUs per species (and indirectly, the total number of chimeric reads) was significantly correlated ($P < 0.005$) with the total number of reads each species generated (r = 0.81 and 0.84 for 1rA and 3rA, respectively). In general, the most abundant species (10 individuals/species) formed more chimeras than less abundant species

TABLE 1. Characteristics of chimeric sequences in high quality >199 reads in two independent datasets: 1rA and 3rA. 1rA involved just single PCR replicate of a DNA template consisting of 38 known nematode species in known frequencies, and 3rA was made up by pooling of three PCR replicates from the same DNA template. Chimeric breakpoints followed two conserved regions within NF1-18Sr2b: CTTCTTA (6bp) and ATAACAGGTCTGTGATGCCCTTAGA (25bp).

	1rA	3rA
Total reads	11482 (100%)	11624 (100%)
Total chimeric reads	2019 (18%)	1708 (15%)
Nematode chimeric reads	1611 (14%)	1323 (11%)
Total OCTUs	1338 (100%)	946 (100%)
Total chimeric OCTUs	579 (43%)	404 (43%)
Nematode chimeric OCTUs	411 (31%)	284 (30%)
Average length of the 3'-end deviating from the predicted 5'-end	55 bp	64 bp
Average # of reads/chimeric OCTU	4	5
Chimeric OCTUs following CTTCTTA (109-115bp)	52 (13%)	53 (19%)
Chimeric OCTUs following ATAACAGGTCTGTGATGCCCTTAGA (153-177bp)	359 (87%)	231 (81%)
Nematode species not detected in chimeras on the 5'-end		*Bursaphelenchus anatolius*
	Bursaphelenchus hylobianum	*Bursaphelenchus hylobianum*
	Koerneria n. sp. 227	*Koerneria* n. sp. 227
		Aduncospiculum halicti 94
	Acrostichus puri	*Acrostichus puri*
		Halicephalobus n. sp. 696
	Plectus acuminatus	*Plectus acuminatus*
	Hoplolaimus galeatus	*Hoplolaimus galeatus*
	Paractinolaimus sp.	*Paractinolaimus* sp.
	Prismatolaimus sp.	*Prismatolaimus* sp.
	Belonolaimus longicaudatus	*Belonolaimus longicaudatus*
Nematode species not detected in chimeras on the 3'-end	*Aphelenchoides besseyi*	*Aphelenchoides besseyi*
	Belonolaimus longicaudatus	
	Bursaphelenchus platzeri	*Bursaphelenchus platzeri*
		Halicephalobus n. sp. 696
	Hoplolaimus galeatus	*Hoplolaimus galeatus*

(1 individual/species) in both datasets. However, among the less abundant species, the number of chimeric OCTUs varied (1 - 24) and there was no consistent pattern across the two datasets.

Environmental samples: Out of 171,861 of all high quality reads, using an approach of incomplete length match of ≥10 bp on the 3'-end between a queried OCTU sequence and a database sequence, at least 14% were categorized as chimeras. A great majority on the 5'-end were identified as nematodes (58%) followed by mites (22%), and tardigrades (5%) (Fig. 7). A wide array of other eukaryotic taxa such as copepods, springtails, rotifers, flatworms, fungi, annelids, or insects made up anywhere from 3 to 1% of all chimeric reads. Groups of organisms representing <1% of chimeric reads included spiders, millipedes, centipedes, snails, and plants, but also chordates, branchiopods, and stramenopiles. The identities of the 3'-end fragments were not determined.

In terms of OCTUs, out of the total 15,658, more than 40% were identified as chimeric. The proportionate distribution of the identities of chimeric OCTUs was similar to that of the chimeric reads. The abundance of chimeric reads and chimeric OCTUs for all different taxonomic groups were significantly ($P < 0.0005$) correlated with non-chimeric reads (r = 0.99) and non-chimeric OCTUs (r = 0.99).

DISCUSSION

One of the central goals of ecology is to understand the estimates and patterns of biodiversity. While features of macroorganismal diversity are relatively known, microbiotic diversity remains greatly uncharacterized and unexplored. Advances in molecular high-throughput platforms have been fundamentally transforming the perception of prokaryotic and eukaryotic microbiotic diversity in aquatic and terrestrial environments (Huber et al., 2007; Roesch et al., 2007; Porazinska et al., 2010b) with estimates exceeding previous expectations. However, simultaneous advances in bioinformatics tools suggest problems that could result in false perceptions of biodiversity. For instance, PCR and pyrosequencing errors or alignment strategies and clustering methods have been shown to erroneously inflate microbial diversity by several orders of magnitude (Quince et al., 2009; Huse et al., 2010). Microbiotic diversity can also be misrepresented through the presence of chimeric sequences formed during PCR amplification of multi-template samples. The average frequency of chimeric molecules can range from 4% - 70% (Robison-Cox et al., 1995; Lahr and Katz, 2009) but most recent reports utilizing novel methods for evaluating high-throughput quality of sequences suggest levels exceeding 90% (Edgar et al., 2011; Quince et al., 2011).

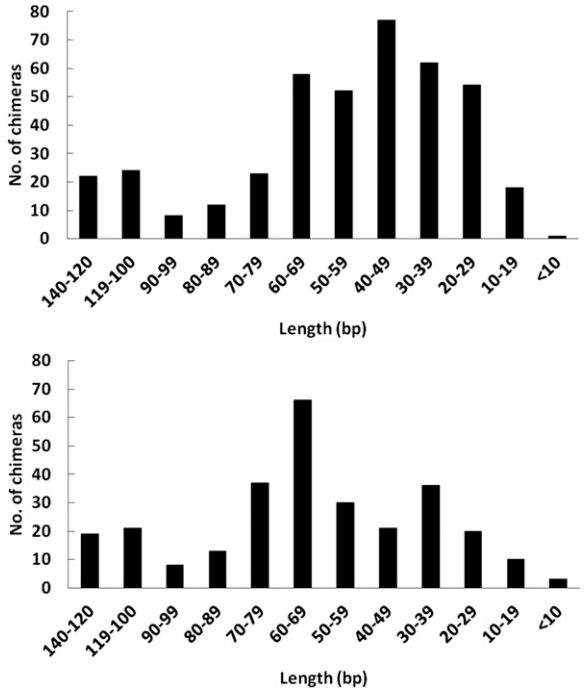

FIG. 3. Frequency of 3'chimeric ends of specific lengths in two independent datasets A) 1rA dataset, B) 3rA dataset.

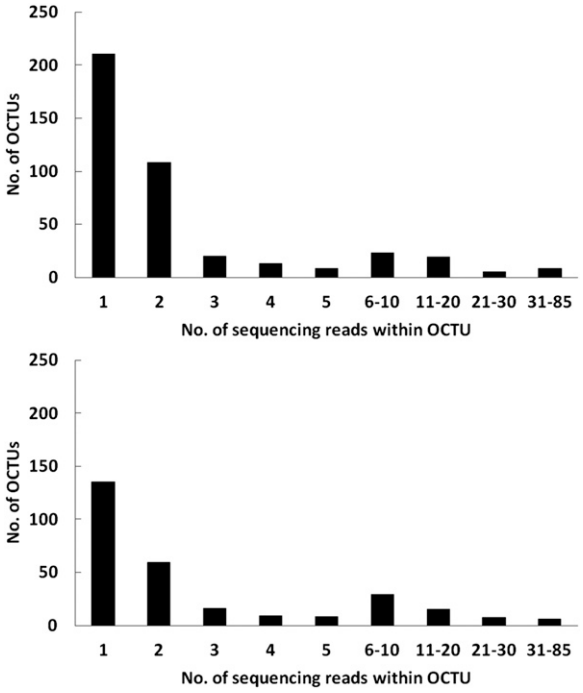

FIG. 4. Frequency of chimeric OCTUs with specific number of sequencing reads in two independent datasets A) 1rA dataset, B) 3rA dataset.

Because each variant sequence in an environmental sample potentially represents a unique species, unrecognized chimeric molecules could be mistaken as new divergent lineages, artificially inflating diversity estimates. Initial approaches for tagging chimeras were selective to rather small datasets supported by a reference sequence database. Typically, reads that did not match the reference database were split into half-fragments and compared back against the database. All half-fragments from the same read but matching different reference sequences were categorized as chimeras, inspected, and then manually confirmed. This approach, however, when applied to our high-throughput sequencing control datasets underestimated the number of chimeric molecules by almost two orders of magnitude (Porazinska et al., 2010a) because chimeras involving unequal fragments (e.g. 70% 5'-end and 30% 3'-end) failed to be tagged. Most recent advances in microbial high-throughput data analysis already allow for significantly more precise chimera tagging (Huber et al., 2004; Ashelford et al., 2006; Nilsson et al., 2010; Edgar et al., 2011; Haas et al., 2011; Quince et al., 2011), but protocols are 16S/ITS specific, require currated/multiple aligned reference databases (unavailable for microscopic eukaryotes), and depend on advanced bioinformatics skills or high computational power. As opposed to the above protocols generally relaying on scores of three-way alignments (query vs. two parental sequences), we targeted these short 3'-ends by using MEGABLAST and generating a list of incomplete length matches. The analysis of all these incomplete length matches (starting with a single bp difference) in our control datasets allowed us to characterize in detail their nature and frequency.

Our results indicate that the frequency of chimeric sequencing reads is consistently high (~17%) and confirms the results from prokaryotic studies. Typically, they were formed between 2 parent sequences, but ~3% of all chimeras involved >2 different DNA origins. The great majority of chimeras predictably consisted of two nematode origins, but combinations between a nematode and another eukaryote were also present. Given that our primers are not nematode exclusive but readily amplify a variety of other eukaryotic phyla (Porazinska et al., 2009) and nematode species selected for the control experiments were maintained under xenic culture conditions, this result was not unexpected.

High-throughput sequencing reads are often processed by clustering them into OTUs (operational taxonomic units) at different cut off levels that potentially correspond to different levels of taxonomic organization (e.g. 97% is equivalent to "species" in prokaryotic studies). Because the majority of chimeric reads are very unique and "divergent", each chimeric sequence can form its own OTU and consequently exceed 40% of all OCTUs clearly overestimating the true diversity at all cut-off levels. While the majority of chimeric OCTUs consisted of single and double reads, many were covered by >80 reads indicating that approaches targeting singleton OCTUs (OTUs) as potential chimeras are insufficient in recognizing the extent of chimera problems. Moreover, because chimeric and non-chimeric OCTUs can be characterized by similar depth (number

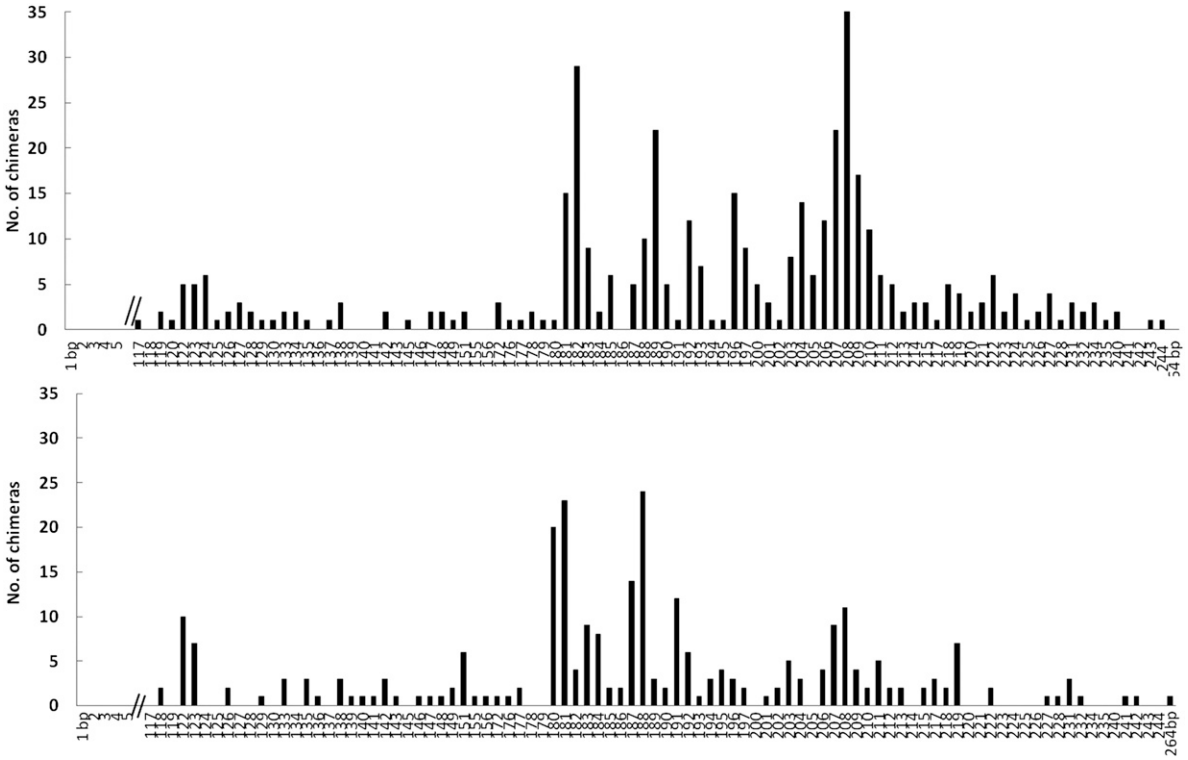

FIG. 5. Frequency of chimeras at specific chimeric breakpoints along the SSU barcoding region in two independent datasets A) 1rA; and B) 3rA. The entire region spans ~400bp and first bp of the 5' NF1 forward primer is designated as 1. First conserved region associated with chimera formation is located between 109-115 bp and the second region between 153-177bp. A) 1rA dataset, B) 3rA dataset.

of sequences), defining a specific read number threshold for chimeras seems unrealistic.

It has been shown that the frequency of chimeras depends on several factors including PCR conditions (e.g. number of cycles, extension time, type of DNA polymerase) and DNA template concentrations (Wang and Wang, 1997; Lahr and Katz, 2009). Our results suggest that the number of chimeras may be associated with the characteristics of the sequenced region. The apparent chimeric breakpoint locations were clearly nonrandom, but rather clustered around conserved regions. There were two regions like that within 250 bp reads, and over 80% of chimeras showed a bias towards the longer conserved region (25 bp) at about 153-177 bp on *C. elegans*, and thus producing mostly chimeras of unequal parental fragments of rather long 5'- and short 3'-ends making it difficult to correctly tag chimeras. Although this general pattern was observed in both datasets, the exact location of chimeric breakpoints and specific species combinations were different in these

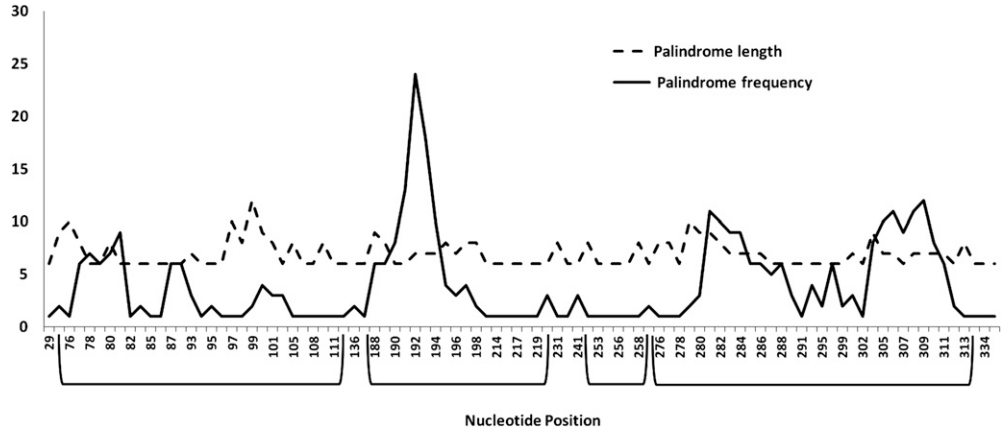

FIG. 6. An average length and frequency of palindromic sequences at specific nucleotide positions on Sanger generated sequences of all the nematode species used in the control experiments illustrating the absence of palindromes within the conserved regions (located at 109-115 bp and 153-177 bp) and their wide-spread presences elsewhere. Brackets illustrate continuous palindromic bp positions within the DNA diagnostic region.

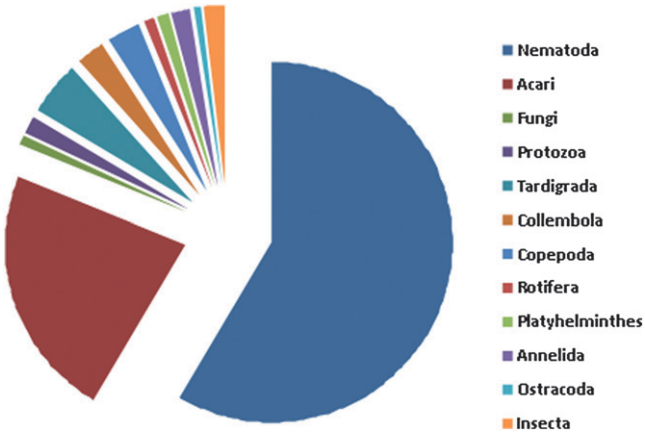

FIG. 7. Distribution of chimeric reads of different taxonomic identities in environmental samples of the Costa Rican tropical rainforest.

two datasets suggesting that the process of chimera formation is, as expected, a random rather than a specifically predictable and repeatable set of events. Given a rather wide spread of palindromes along Sanger generated sequences and their absence within the conserved regions, the length of conserved regions appears to be the most important aspect of where and why chimeras form. Consequently, a selection of barcoding regions with minimal presence of conserved sites would be strongly advised.

The complexity of microbial communities has been recently recognized as another factor driving the process of chimera formation (Hass et al., 2011; Quince et al., 2011). Although frequency may depend on the richness and evenness of communities, specific characteristics of constituent species and their abundances may also play a role. Similar to Haas et al. (2011), the number of chimeras generated by each species was associated with that species abundance in our artificial communities. The identity of species was generally less important because species representing all phylogenetic clades were prone to chimera formation. Because species represented by 1 individual formed different numbers of chimeras across the two datasets, an idea of specific species propensity for generating chimeras is rather unlikely. In addition, a bias towards chimeras between closely-related species was not apparent. It is possible, however, that this result stems from a lower divergence of the 3'region where most chimeras occurred. In other words, a portion of reads recognized as non-chimeric could be indeed chimeric, but are unrecognized because the 3'ends of some of the closely related species are identical. Studies utilizing specific tagging of individuals are needed to resolve the estimates of chimeras between closely related species.

An analysis of sequencing reads in environmental samples confirmed the results observed in the artificially assembled communities with similarly high counts of chimeras among reads and OCTUs. Because nematodes typically dominate in abundance and richness over other components of eukaryotic microbiota, predominance of chimeras of nematode origin was not surprising. A more revealing result relates to the overall quality and quantity of chimeras, including those of non-nematode origin. We identified chimeras representing a wide array of eukaryotic taxa ranging from fungi to invertebrates to plants with quantities of chimeric reads and chimeric OCTUs of specific taxonomic identities linked to non-chimeric reads and non-chimeric OCTUs. Since abundances of sequencing reads and OCTUs generally reflect quantities of specific organisms in samples (Porazinska et al., 2010a), abundance of chimeras are most likely linked to abundances of organisms as well. Given that priming sites of the diagnostic regions used in our study are fairly conservative across different phyla, an amplification of even a single nematode should not be expected to produce no chimeras. Guts and cuticles of nematodes are routinely loaded with DNA of other taxonomic groups like fungi or plants. While low sensitivity of standard PCR and Sanger sequencing can keep chimeras undetected (they are outnumbered by the dominating nematode sequence), their presence can be easily revealed with high-throughput sequencing platforms. In the original test of the 454 technology for nematode diversity assessments, Porazinska et al. (2009) included one multitemplate sample formed by pooling amplicons of separately amplified individuals of different nematode species. Although the frequency of chimeras in this sample was minimal (1.5%, unpublished data), their presence was undeniable.

Overall, we estimate that the abundance of chimeras is high, but it should be emphasized that estimates presented here are very conservative because chimeras of closely related species might not be recognizable. Chimeras did not appear to form at random locations of the amplified DNA region, but rather clustered around conserved sites. Although chimeras of nematode origin were most abundant, they were not nematode exclusive, and instead spanned a wide array of eukaryotic taxa. This information warrants caution because abundant chimeric sequences can mistakenly overestimate eukaryotic biodiversity. It should be noted, however, that when proper chimera recognition protocols are implemented, the high chimeric abundance in environmental samples does not prevent the use of pyrosequencing in biodiversity studies.

LITERATURE CITED

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J. 2006. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. Applied and Environmental Microbiology 72:5734–5741.

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

Nilsson, R., Abarenkov, K., Veldre, V., Nylinder, S., De Wit, P., Borsche, S., Alfredsson, J., and Kristiansson, E. 2010. An open source chimera checker for the fungal ITS region. Molecular Ecology Resources 10:1076–1081.

Chou, H., and Holmes, M. 2001. DNA sequence quality trimming and vector removal. Bioinformatics 12:1092–1104.

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

Fonseca, V. G., Carvalho, G. R., Sung, W., Johnson, H. F., Power, D. M., Neill, S. P., Packer, M., Blaxter, M. L., Lambshead, P. J. D., Thomas, W. K., and Creer, S. 2010. Second-generation environmental sequencing unmasks marine metazoan biodiversity. Nature Communications DOI: 10.1038/ncomms1095.

Haas, J. B., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergen, E., Methe, B., and Desantis, T. Z., The Human Microbiome Consortium, Petrosino, J. F., Knight, R., and Birren, B. W. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Research doi:10.1101/gr.112730.110.

Hallan, J. 2007. Synopsis of the Described Nematoda of the World. http://insects.tamu.edu/research/collection/hallan/Nematoda/Family/0NematodaIndex0

Huber, J. A., Welch Mark, D. B., Morrison, H. G., Huse, S. M., Neal, P. R., Butterfield, D. A., and Sogin, M. L. 2007. Microbial population structures in the deep marine biosphere. Science 318:97–100.

Lahr, D. J. G., and Katz, L. A. 2009. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. Biotechniques 47:857–863.

Lambshead, P. J. D. 2004. Marine nematode biodiversity. Pp. 436–467 *in* Z.X. Chen, S.Y. Chen, and D.W. Dickson, eds. Nematology: Advances and perspectives Vol. 1: Nematode morphology, physiology and ecology. CABI Publishing.

Porazinska, D. L., Giblin-Davis, R. M., Farmerie, W., Powers, T. O., Kanzaki, N., Faller, L., Morris, K., Sung, W., and Thomas, W. K. 2009. Evaluating high throughput sequencing as a method for metagenomic analysis of nematode diversity. Molecular Ecology Resources 9:1439–1450.

Porazinska, D. L., Sung, W., Giblin-Davis, R. M., and Thomas, W. K. 2010a. Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. Molecular Ecology Resources 10:666–676.

Porazinska, D. L., Giblin-Davis, R. M., Esquivel, A., Powers, T. O., Sung, W., and Thomas, W. K. 2010b. Ecometagenetics confirms high tropical rainforest nematode diversity. Molecular Ecology 19:5621–5630.

Porazinska, D. L., Giblin-Davis, R. M., Powers, T. O., Sung, W., and Thomas, W. K. 2010c. Linking operational clustered taxonomical units (OCTU) from parallel ultra sequencing (PUS) to nematode species. Zootaxa 2427:55–63.

Powers, T. O., Neher, D. A., Mullin, P., Esquivel, A., Giblin-Davis, R. M., Kanzaki, N., Stock, S. P., Mora, M. M., and Uribe-Lorio, L. 2009. Tropical nematode diversity: vertical stratification of nematode communities in a Costa Rican humid lowland rainforest. Molecular Ecology 18:985–996.

Quince, C., Lanzen, A., Curtis, T. P., Davenport, R. J., and Turnbaugh, P. J. 2011. Removing noise from pyrosequenced amplicons. BioInformatics 12:38 doi:10.1186/1471-2105-12-38.

Quince, C., Lanzen, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., Read, L. F., and Sloan, W. T. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. Nature Methods 6:639–641.

Reeder, J., and Knight, R. 2009. The 'rare biosphere': reality check. Nature Methods 6:636–637.

Robison-Cox, J. F., Bateson, M. M., and Ward, D. M. 1995. Evaluation of nearest-neighbor methods for detection of chimeric small-subunit rRNA sequences. Applied and Environmental Microbiology 61:1240–1245.

Roesch, L. F., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K., Kent, A. D., Daroub, S. H., Camargo, F. A. O., Farmerie, W. G., and Triplett, E. W. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. International Society for Microbial Ecology Journal 1:283–290.

Wang, G. C. Y., and Wang, Y. 1997. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. Applied and Environmental Microbiology 63:4645–4650.

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