Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses

V. G. Fonseca, B. Nichols, Delphine Lallias, C. Quince, G. R. Carvalho, D. M. Power, S. Creer

To cite this version:

V. G. Fonseca, B. Nichols, Delphine Lallias, C. Quince, G. R. Carvalho, et al.. Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. Nucleic Acids Research, 2012, 40 (9), pp.1-9. 10.1093/nar/gks002 . hal-02647717

HAL Id: hal-02647717
https://hal.inrae.fr/hal-02647717

Submitted on 29 May 2020

Nucleic Acids Research Advance Access published January 25, 2012

Nucleic Acids Research, 2012, ***1–9***
doi:10.1093/nar/gks002

Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses

V. G. Fonseca[1,2], B. Nichols[3], D. Lallias[1], C. Quince[3], G. R. Carvalho[1], D. M. Power[2] and S. Creer[1,*]

[1]Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Environment Centre Wales, Bangor University, Deiniol Road, Gwynedd LL57 2UW, UK, [2]Centre of Marine Sciences, CCMAR-CIMAR Associate Laboratory, University of Algarve, Gambelas, Faro 8005-139, Portugal and [3]School of Engineering, University of Glasgow, Rankine Building, Oakfield Avenue, Glasgow G12 8LT, UK

Received August 4, 2011; Revised November 23, 2011; Accepted December 23, 2011

ABSTRACT

Eukaryotic diversity in environmental samples is often assessed via PCR-based amplification of nSSU genes. However, estimates of diversity derived from pyrosequencing environmental data sets are often inflated, mainly because of the formation of chimeric sequences during PCR amplification. Chimeras are hybrid products composed of distinct parental sequences that can lead to the misinterpretation of diversity estimates. We have analyzed the effect of sample richness, evenness and phylogenetic diversity on the formation of chimeras using a nSSU data set derived from 454 Roche pyrosequencing of replicated, large control pools of closely and distantly related nematode mock communities, of known intragenomic identity and richness. To further investigate how chimeric molecules are formed, the nSSU gene secondary structure was analyzed in several individuals. For the first time in eukaryotes, chimera formation proved to be higher in both richer and more genetically diverse samples, thus providing a novel perspective of chimera formation in pyrosequenced environmental data sets. Findings contribute to a better understanding of the nature and mechanisms involved in chimera formation during PCR amplification of environmentally derived DNA. Moreover, given the similarities between biodiversity analyses using amplicon sequencing and those used to assess genomic variation, our findings have potential broad application for identifying genetic variation in homologous loci or multigene families in general.

INTRODUCTION

Second-generation pyrosequencing of environmental DNA has provided unique insights into prokaryotic (1,2) and eukaryotic (3,4) molecular diversity and ecology. Massive parallel pyrosequencing has the potential to produce a large volume of data relatively cheaply and with an unprecedented read depth, generating millions of DNA sequences within a matter of hours (5). Despite advantages of high throughput sequencing, a major challenge is to determine the extent to which sequences produced from pyrosequencing-amplified regions of marker genes correspond to biological diversity. Recently, studies have recognized that biodiversity levels have become inflated due to artifacts associated with sample processing including both the PCR amplification and the pyrosequencing itself (6–8). PCR amplification with universal primers applied to genes conserved across phyla, such as the ribosomal nuclear small subunit (nSSU), is commonly used to identify microbial eukaryotes in natural environments. The extreme conservation of primer binding sites (9) and the availability of extensive database resources (10) has resulted in the nSSU being the most widely used marker for studying the molecular taxonomy of a diverse range of eukaryotes. Target taxa range from all protist kingdoms (11) to metazoan microorganisms (4), that are dominated by the Nematoda (12). In such analyses, one of the most commonly reported sources of sequence artifacts associated with highly homologous nSSU genes from environmental DNA samples is the formation of chimeric sequences during PCR amplification (8,13–16).

Chimeric sequences, or chimeras, are generated when incomplete extension occurs during PCR amplification and the resulting amplicon re-anneals to a foreign DNA strand and is copied to completion in the following PCR cycles. Chimeras are composed of two or more

[*] To whom correspondence should be addressed. Tel: +44(0)1248 382302; Fax: +44(0)1248 370731; Email: s.creer@bangor.ac.uk

phylogenetically distinct parental sequences and have been shown to occur in PCR-amplified nSSU data sets with frequencies of 30–70% (6,17,18) thus leading to false diversity estimates and false novel taxa. The critical factors that seem to affect PCR-generated recombination are the number of PCR cycles, PCR extension time, template concentration, *Taq* DNA polymerases and amplicon size (18–21). Chimera formation can be minimized experimentally by PCR optimization, nonetheless, no method has yet proved to be entirely effective. The importance of detecting chimeras is such that a plethora of bioinformatic software has also been developed, such as Chimera_Check (22), Bellerophon (13), CCode (23), Pintail (24), Mallard (17), Chimera Slayer (6) and Perseus (8). With the exception of Perseus, such approaches will only detect evident induced chimeras (25) and their accuracy for chimera detection has not been rigorously tested (6) or is still at an early stage, especially given recent advances in environmental DNA sequencing approaches. Although metagenetic (4,9) analyses are clearly based on complex and phylogenetically diverse assemblages, the roles of sample richness and phylogenetic diversity in driving chimera formation are largely unknown.

Wang and Wang (18,26) tested how sequence similarity between cloned 16S rRNA genes or mixed bacteria genomic DNA can influence PCR-based chimera formation. Nonetheless, these investigations were performed on a very small scale, did not consider sample richness and pre-dated the current second-generation sequencing perspective of amplicon pool diversity. The overarching aim here is to (i) analyze the effect of richness, evenness and genetic diversity on chimera formation and link this to diversity estimates and (ii) understand how chimeras are formed with respect to variable genetic diversity and secondary structure of the parent nSSU molecule. To this end, a nSSU metagenetic data set was generated by 454 Roche pyrosequencing of control pools of closely and distantly related nematode mock communities of known identity and richness.

MATERIAL AND METHODS

Sample preparation

To test if chimera formation during PCR reactions was associated with taxon richness or with phylogenetic distance, 74 Sanger-sequenced single nematode species were blast aligned to a contemporary Nematoda phylogenetic framework (27). Subsequently, the sequences were aligned using ClustalX and pairwise distances (p-distance) were calculated using MEGA 4.1 (28). Based on the phylogenetic affinities of the nematode sequences, subsets of closely related [mean pairwise divergence (MPD) of 25%, referred to as 'phylogenetically close'] and distantly related (MPD of 40%, referred to as 'phylogenetically distant') nSSU controls were generated by pooling the DNA extracts of 12, 24 or 48 individuals.

DNA extraction and preparation

DNA extraction from DESS-preserved (29) single worms was performed using a DNeasy Blood & Tissue Kit (Qiagen Inc), following the manufacturer's instructions. After extraction, all DNA was eluted in 40 µl of AE buffer and samples were stored at −20°C until use. The DNA extracts from all single individuals were quantified using a Nanodrop spectrophotometer and diluted to 0.5 ng/µl, and five replicates of the 12, 24 and 48 individuals were selected for the closely and distantly related treatments.

PCR amplification and sequencing analysis

The primers SSUFO4 forward (5′-GCTTGTCTCAAAG ATTAAGCC-3′) and SSUR22 reverse (5′-GCCTGCTGC CTTCCTTGGA-3′) were used to amplify ~450 bp of the nSSU rDNA (18S rDNA) region (30). Fusion primers were then developed according to Fonseca *et al.* (4). PCR amplification reactions and the thermocycle for the targeted nSSU region were optimized. Optimized reactions were performed using 0.25 ng/µl of genomic DNA template in 3 × 40 µl reactions using *Pfu* DNA polymerase (Promega) for each of the closely and distantly related nematode pools (12, 24 and 48 individuals) and all individual DNA extracts. PCR thermocycle conditions consisted of a 2-min denaturation step at 95°C followed by 35 cycles (thus facilitating the generation of chimeras) (6,18,26) of 1 min at 95°C, 45 s at 55°C, 3 min at 72°C and a final extension of 10 min at 72°C. Negative controls (ultrapure water only) were included for all amplification reactions. Electrophoresis of triplicate PCR products was undertaken on a 2% gel with Top Vision™ LM GQ Agarose (Fermentas), and the expected 450-bp fragment was purified using the QIAquick Gel Extraction Kit (Qiagen), following the manufacturer's instruction. All purified PCR products were quantified with an Agilent Bioanalyser 2100 and diluted to the same concentration (10 ng/µl). PCR amplifications from single nematodes and pooled nematodes were sequenced in a single direction (A-Amplicon) on a quarter and three-quarters of a plate, respectively, using a 454 Roche GSFLX (454 Life Sciences, Roche Applied Science) sequencing platform at Liverpool University's Centre for Genomic Research, UK.

Denoised reads and detection of chimeric PCR molecules

Pyrosequencing reads derived from 454 Roche data contain a substantial number of errors (referred to as *noise*), which includes sequencing errors mainly derived from the inclusion or deletion of single bases in homopolymer runs of 3 bp or longer, PCR single base substitutions and PCR chimeras (14,31). AmpliconNoise was used to remove noise from the pyrosequencing data; this comprises filtering, flowgram and sequence clustering steps. It has been shown to reduce noise by ~50% in environmental data sets (8). Subsequently, chimeras were identified using Perseus (8); this algorithm generates a Chimera Index (CI) for each read that is ≥0 with higher values corresponding to reads that are most likely to be chimeric. Perseus by pairwise alignments to all sequences of greater than or equal abundance identifies the most likely parent sequences of the candidate read and the most likely break point. Logistic regression is then used to classify chimeras so that the pyrosequencing data output lists chimeric and non-chimeric sequences. The lower the probability of the sequence evolving naturally, the higher the CI (8).

Perseus finds break point positions in the two parent sequences. To compare across the whole data set, it is necessary to fix these positions relative to a reference sequence. To do this, a four-way alignment between each chimeric sequence, its two parents and the *Caenorhabditis elegans* reference sequence was formed (GenBank/EMBL accession number EU196001). The most likely break point was identified by minimizing the number of differences between the sections of the parents contributing to the chimera and the chimera itself. The position of each break point on the reference sequence was then recorded and from this, the frequency breaks occurring at each position could be calculated. MFold RNA-folding software was used to predict the potential role on chimera formation of the secondary structure of the 18S rDNA amplicon region (32).

Generation of operational taxonomic units

Denoised mock nematode community data from which chimeras had been removed was used to identify Operational Taxonomic Units (OTUs). OTUs were calculated using a complete linkage-clustering algorithm, measuring the distance between the most distant members in each cluster, at a 99% identity cut-off. The number of OTUs generated was then used to determine the effect of taxa richness on chimera formation within a sample. Although numbers of reads within treatments varied this did not have a significant (ANOVA, $P > 0.05$) effect on chimera frequencies, number of OTUs and/or Shannon Index. However, all the analyses were also performed on a normalized data set by subsampling equal read numbers from each treatment and observations were found to be congruent with the non-normalized data (data not shown).

Statistical analysis

Species richness, or in this case OTU richness, takes no account of the evenness of the distribution. An index with better properties is the Shannon index (33). This increases with more taxa but also as the distribution of abundances across taxa becomes more even. The Shannon Index of biodiversity was established for each sample using Vegan R (34). To analyze the relationship between overall chimera percentage and the explanatory variables (e.g. phylogenetic relatedness, richness, diversity, number of reads), a linear model was fitted to the data, giving a multiplicative coefficient for each explanatory variable. An ANOVA was then performed to statistically determine which of the variables had an effect on the chimera percentage. Variables without a significant effect on chimera percentage were removed and the model was refitted to give accurate ANOVA results. A probability (P-value) <0.05 was considered significant.

RESULTS AND DISCUSSION

Here, it was possible to assess the effect of OTU richness, evenness and phylogenetic relatedness on chimera formation in a nSSU second-generation sequencing environmental data set. Moreover, a rigorous experimental design and bioinformatic analysis facilitated the identification of chimera breakpoint frequencies within the parent nSSU molecule. The 'mock community' contained equivalent concentration of 18S rDNA genes of individual nematodes of known identity and richness (GenBank Accession Numbers JN968213–JN968286). Nematodes are the most abundant phylum of meiofaunal environmental samples (4) representing a major part of biodiversity and perform numerous essential roles in ecosystems processes (35,36). The nematodes that were chosen included both phylogenetically distant and closely related species to emulate a likely environmental assemblage. Amplicons were sequenced on a Roche 454 GSFLX platform and generated a total of 339 515 pyrosequence reads (Genbank/EMBL/DDBJ short read archive accession number SRA043810.1). AmpliconNoise (14) generated 236 406 reads after removing errors arising from PCR and pyrosequencing errors and truncating sequences to a uniform 200 bp. Chimeras were detected in denoised 'mock community' data using the Perseus algorithm (8) and ~42% of the sequences were classified as chimeric and were removed before taxon richness was assessed by clustering sequences into OTUs at a 99% identity threshold for each data set. A summary of the mean number of reads for each data set, denoised sequences, chimera percentages and OTUs is given in Table 1.

Denoised sequences contained between ~15% and 60% of chimeras in some pools, confirming that 35 cycle PCRs do indeed generate numerous chimeras as already confirmed by previous studies (18,19,25) even within a small mock environmental data set. The results stress the importance of a chimera removal step to allow an accurate estimation of OTU numbers and robust estimates of biodiversity levels in environmental samples (8,15,37).

Overall, the mean OTU numbers were approximately double the number of unique nematode species in each pool. This could be associated not only with sequencing artifacts but also because organisms frequently contain multiple copies of heterogeneous nSSU genes (38). To assess the impact on the data set of multi-copy nSSUs, all single nematode PCR products were amplified

Table 1. Mean numbers of OTUs, denoised sequences, chimera percentages and reads for the pools of close and distantly related nematodes with 48, 24 and 12 individuals, respectively

Species phylogeny	Species number	OTUs at 99%	Denoised sequences	Chimera (%)	Read number
Close	48	87.6	138.40	35.60	13 882.20
Close	24	40.4	63.20	34.55	3809.00
Close	12	35.8	42.80	14.57	6159.80
Distant	48	63.2	161.00	58.98	5657.80
Distant	24	53.6	119.00	53.57	10 134.20
Distant	12	34.4	58.20	39.93	7638.20

with unique MID-tag sequences. Of the 74 MID-tagged single nematode amplifications, 61 were single copy 18S rDNA and 10 were double copy but all taxa were represented by a similar total number of sequences in PCR reactions (data not shown).

A striking observation was the difference in chimera formation between close and distantly related nematode assemblages. In the latter, the mean percentage of sequences classified as chimeric was 55% for the 24 and 48 species pools and was significantly higher (ANOVA, $P < 0.001$) than the equivalent pools of closely related nematode assemblages that had 35% chimeras. In line with the previous observation, the mean percentage of chimeras was significantly lower (ANOVA, $P < 0.01$) in the 12 species pools irrespective of the similarity or distance of the individuals in the nematode assemblages (Table 1). These results suggest that chimera formation in the 5′ region of nSSU amplicon pools is significantly higher in more phylogenetically diverse and richer data sets.

Although the studies were on a smaller scale, Qiu *et al.* (19) and Wang and Wang (26) analyzed chimera formation with bacterial rRNA clones and also found that PCR artifacts and chimera frequency increased as species diversity increased. To further confirm this, OTU diversity within the two nematode assemblages (close and distant pools) was expressed using the Shannon Index (33). OTU diversity (Shannon Index) and OTU richness (OTUs numbers) showed a significant effect (ANOVA, $P < 0.01$, $P < 0.001$) on chimera frequency further supporting the hypothesis that more diverse and richer samples generate a higher frequency of chimeric molecules. Additionally, diversity (Shannon index) had a significant effect both on the closely ($P < 0.05$, $P = 0.0324$) and distantly ($P < 0.01$, $P = 0.0023$) related nematode assemblages, and had a positive relationship with chimera frequency (Figure 1).

Analysis of chimera breakpoint occurrence in nSSU amplicon sequences revealed that regions with higher nucleotide sequence similarity had significantly higher breakpoint frequencies ($P < 0.001$, $P = 0.00039$) (Figure 2a and b). Indeed, in studies with bacteria using the 16S rRNA gene a large number of competing templates with fairly high sequence similarity generated more chimeras (6,16,26). Presumably, one explanation for this phenomenon may be the priming of strand synthesis by prematurely terminated templates in the next PCR round.

Different copies of the nSSU genes from the same organism may differ by up to 6.5% (18,38), and in the present study alignment of close and distantly related nematodes indicated an overall inter-specific sequence divergence of 10%, evidently enough to generate chimeras. To better reflect an environmental data set, in the present study, an alignment of 10 representatives of every phylum that is represented by meiofaunal taxa was performed and a 23% overall mean inter-specific sequence distance was observed for the same nSSU region. In fact, Wang and Wang (18) suggested that, despite some degree of nucleotide mismatching, partly terminated heterologous 16S rDNA templates can often be completed in the subsequent polymerization step resulting in chimeras. Thus, the possibility that formation of chimeric sequences between different copies of the nSSU genes was also likely to occur (6,16,26) is now confirmed with this experiment. It is probably the degree of sequence similarity within each individual that may determine chimera breakpoint formation. This is an issue inherently associated with the fact that the nSSU gene is a multicopy gene, and intra-specific variability might have a determinant effect on chimera formation, especially when sample richness is quite high. Haas *et al.* (6) and Wang and Wang (26) verified that more similar 16S rDNA genes more readily form chimeras but they did acknowledge the possibility of chimera formation among more divergent species. In fact, the latter phenomenon is evident for the first time in the present study, where chimeras are more often generated among richer, phylogenetically diverse samples, although the region where the chimera forms has to have sufficient conservation to favor hybridization and chimera formation. Our alternative perspective of chimera formation in mock communities may be locus specific, or may be related to our larger sample sizes (n = 12–48) that are predicted to emulate more closely, true environmental samples.

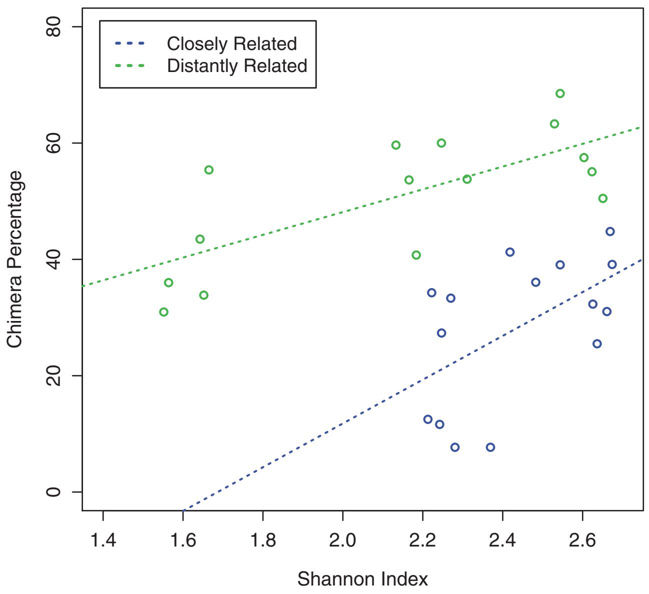

Figure 1. Chimera percentage and Shannon Index of closely and distantly related pools of nematodes.

Chimeras are generally composed of two true sequences, occasionally more (8), with a discrete break point where the transition from one sequence to another occurs. In the present data set, the distribution of chimera breakpoints showed a similar pattern across closely and distantly related nematode assemblages, with a mean peak of frequency at the first 140 bp of the selected nSSU region (Supplementary Figure S1). Although GC content is thought to correlate with chimera formation due to inefficient strand separation and susceptibility to secondary structure formation, a detailed analysis of the parent chimeric sequences at the breakpoints did not reveal a significant correlation between GC rich regions and

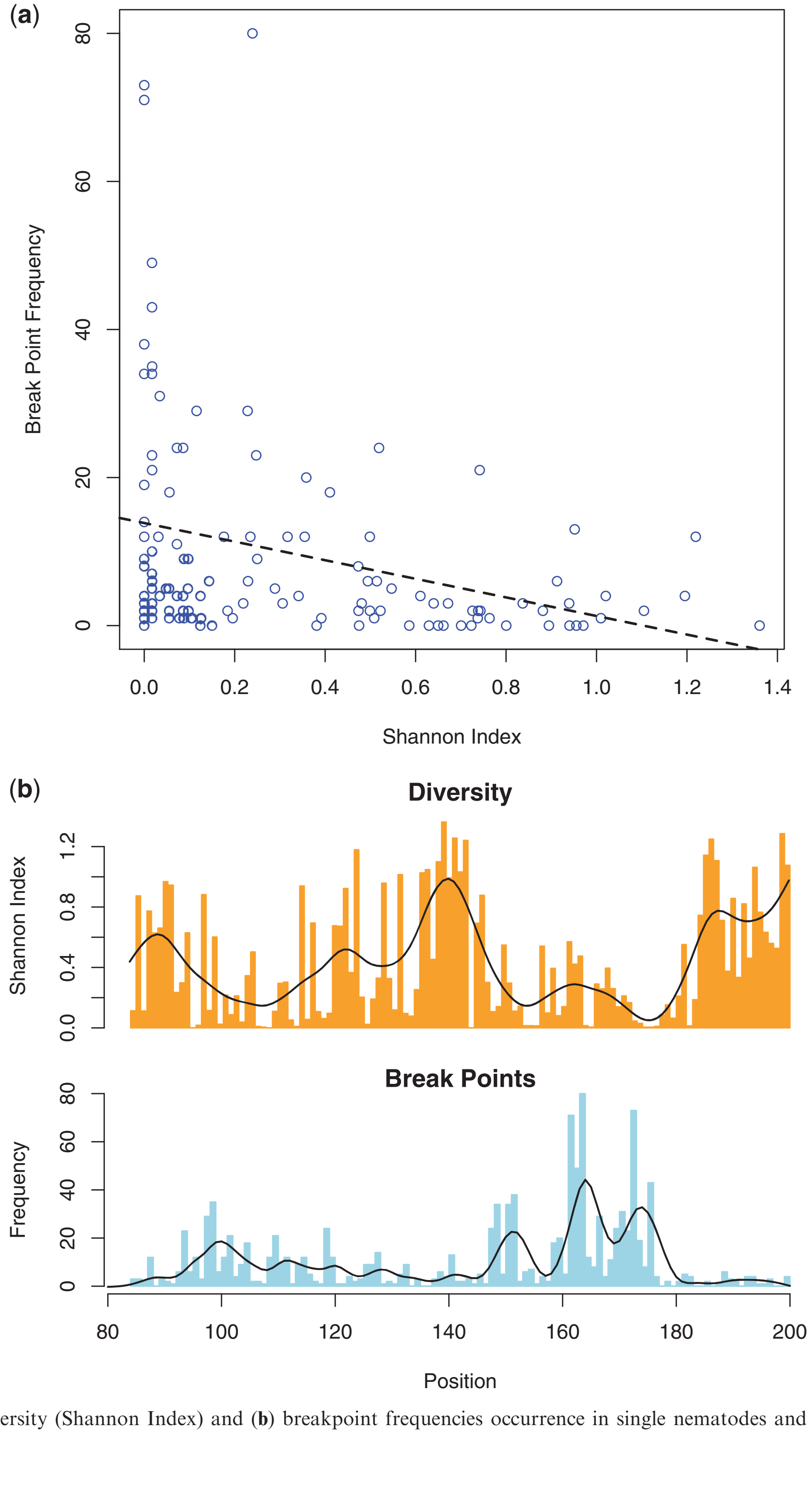

Figure 2. (**a**) Nucleotide diversity (Shannon Index) and (**b**) breakpoint frequencies occurrence in single nematodes and parental chimeric sequences, respectively.

chimera frequencies. To further investigate the breakpoint region, the secondary structure of the amplified nSSU fragment was modeled in 12 single nematode sequences at 55°C and 65°C folding temperatures (Supplementary Table S1). Analysis of the nSSU secondary structure showed that the regions where the breakpoints occurred coincided with hairpin loop structures at both temperatures, although at 65°C regions of secondary structure were less abundant (Supplementary Table S1 and Figure 3).

Hairpin-loops are common motifs in nSSU gene secondary structure due to their importance in ribosome folding and function (39) and their presence requires greater energy for melting to occur during PCR and

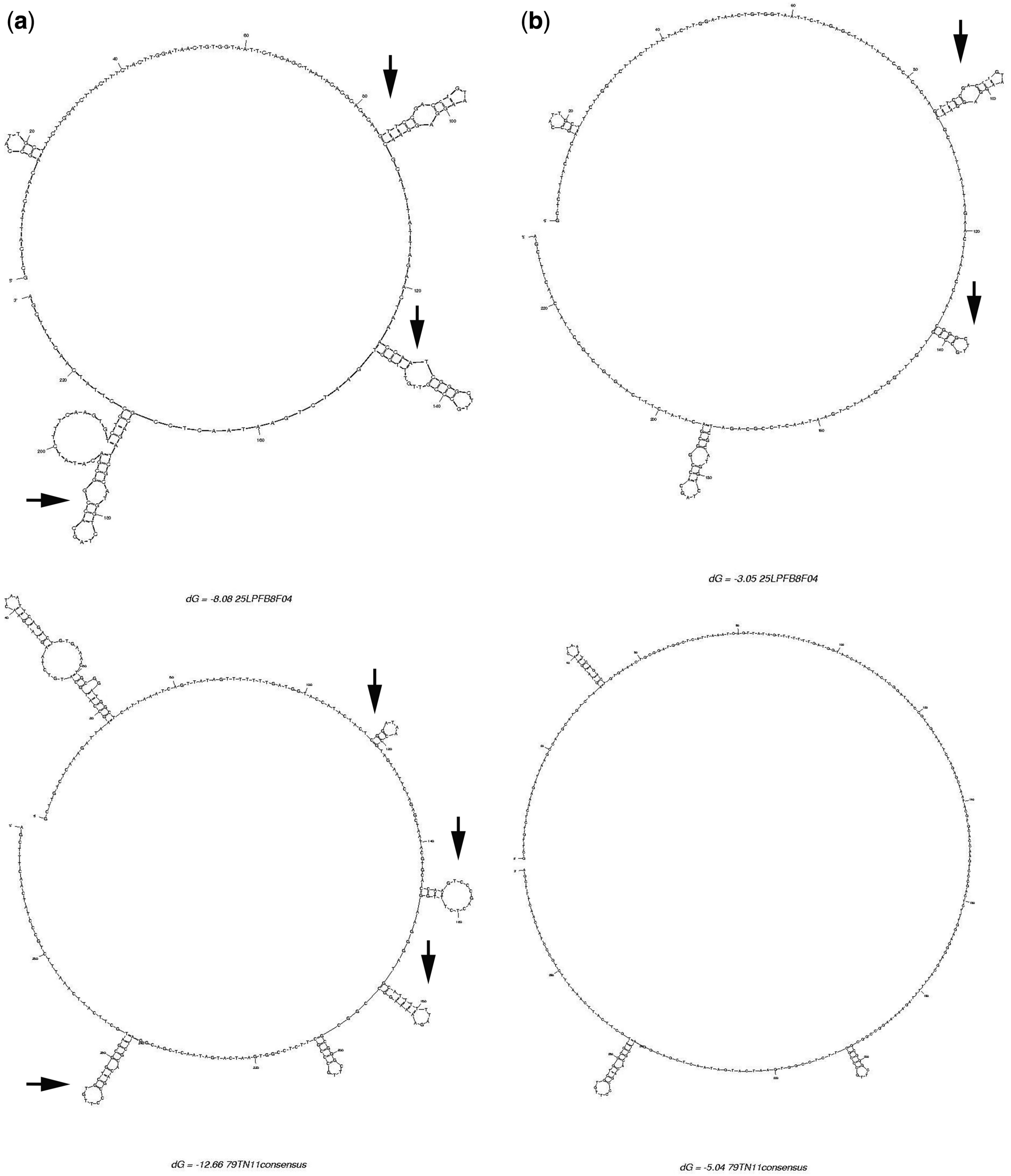

Figure 3. Most frequent predicted secondary structures found on the 18S rDNA amplicon at (**a**) 55°C and (**b**) 65°C folding temperatures, using as an example two single nematodes. Arrows indicate where the most frequent breakpoints occur generally matching hairpin-loops. 79TN11 consensus and 25LBFB8F04 indicate the labels given for each nematode. dG: free energy necessary for sequence stability at a given temperature.

their maintenance will make it more difficult for DNA polymerases to read through. Modeling *in silico* revealed that the nSSU amplicon region in the present study retained some secondary structure at the primer annealing temperature (55°C) and may have been one of the causes of premature termination of DNA synthesis. This is corroborated by results of multiple sequence alignments of PCR-induced chimeras (40) that reveal the recombinant regions were correlated with DNA template secondary structures. Fewer secondary structures of the nSSU amplicon were found at 65°C, suggesting that (i) primers should be designed with a high annealing temperature and/or (ii) genes chosen for environmental metagenetic analyses should be selected for a low tendency to secondary structure formation which should reduce the disposition of complex samples to form chimeras. Nonetheless, other factors are known to minimize frequency of PCR recombination such as adding betaine and dimethylsulfoxide (41); dNTPs should never be limited so that amplification bias is reduced and the use of shorter amplicons will reduce prematurely terminated extension. Further to this, the choice of selectively amplifying target loci from genomic libraries or designing genome-specific amplification primers that will selectively amplify a single homologue (42) also narrow the opportunity for chimera formation.

The investigative use of higher annealing temperatures in our study was not possible since the optimal thermocycling conditions used to amplify meiofaunal representatives (9,43,44) preclude more stringent annealing temperatures. In fact, an intuitive general rule for metagenetic studies would be to avoid high annealing temperatures to ensure the co-amplification of large ranges of taxa from disparate phyla. This implies not only having very high quality DNA samples within uncontaminated laboratory environments but also compulsory stringent analyses by using algorithms to remove artifacts and/ or putative chimeras after sequencing. In addition, the use of reference databases to detect chimeric molecules in environmental data sets is complicated by their unpredictable diversity, meaning that reference data may not be representative of the true diversity. On the other hand, the existence of chimeric sequences in public DNA databases is well known (24,45) and the risk of classifying chimeras as new organisms is becoming higher than the risk of neglecting non-chimeric ones.

Experience from recent studies (4,9) where ~65% of the sequences generated from a 454 Roche environmental data set were discarded leads us to suggest that metagenetic analyses are the ideal 'breeding ground' for recombinant DNA molecules. DNA amplification by PCR has become the main crucial step used for next-generation sequencing technologies in the analysis of environmental samples and so PCR-derived artifacts are continuously increasing. Based on our analyses, the theory of chimera formation having a stochastic distribution (46) should probably be re-evaluated because their occurrence can be influenced by several factors, namely PCR conditions, amplicon nucleotide diversity, molecule folding structure and sequencing strategies. In fact, almost all steps of the molecular approach can introduce biases or errors, including the target DNA template concentration (6,19), DNA polymerases (20,47) and thermal cycling conditions (16,19,26,48). Moreover, the interaction between nucleic acids, DNA polymerases and thermal cycling are likely to be dynamic, suggesting that the identification of optimal molecular biological parameters that reduce chimera formation are likely to be locus/taxon specific.

Overall, we anticipate that our findings will enhance our understanding of chimera formation and associated optimization of pyrosequencing strategies when conducting PCR-based DNA amplification of homologous loci or multigene families. As technologies evolve (49), sequencing is likely to be employed to *de novo* analyse gene partitions from a range of genomic loci in order to address questions that have been hitherto intractable using chain termination sequencing (50). Thus, insights gained here have relevance to biodiversity identification and associated fields such as gene–environment interactions in host parasite co-evolution (51); pathogen recognition (52); genes underpinning immune response (53) and predator–prey arms races (54,55).

ACCESSION NUMBERS

Genbank/EMBL/DDBJ/SRA043810.1, Genbank/EMBL/BankIt/JN968213-JN968286.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figure 1.

ACKNOWLEDGEMENTS

We would like to thank Andrew Cossins, Margaret Hughes and Christiane Hertz-Fowler of the Liverpool Advanced Molecular Genetic Facility for Roche 454 sequencing support. Further thanks go to Alex Dudarenko and Lucy Price, whose undergraduate and MSc thesis work, respectively, contributed to the collection of nematode worms.

FUNDING

Natural Environment Research Council (NERC) New Investigator Grant (NE/E001505/1 to S.C.); Post Genomic and Proteomics Grant (NE/F001266/1 to S.C.); Molecular Genetics Facility Grant (MGF-167 to S.C.); Portuguese Foundation for Science and Technology (FCT) Grant (SFRH/BD/27413/2006 to V.G.F.); EPSRC Career Acceleration Fellowship EP/H003851/1 (to C.Q.) and BBSRC CASE studentship supported by Unilever (to B.N.). Funding for open access charge: Bangor University School of Biological Sciences and BBSRC CASE studentship supported by Unilever (to B.N.).

Conflict of interest statement. None declared.

REFERENCES

1. Huber,J.A., Mark Welch,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.
2. Sogin,M.L., Morrison,H.G., Huber,J.A., Mark Welch,D., Huse,S.M., Neal,P.R., Arrieta,J.M. and Herndl,G.J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl Acad. Sci.*, **103**, 12115–12120.
3. Massana,R. and Pedros-Alio,C. (2008) Unveiling new microbial eukaryotes in the surface ocean. *Curr. Opin. Microbiol.*, **11**, 213–218.
4. 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., Thomas,W.K. *et al.* (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nat. Commun.*, **1**, 98.
5. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
6. Haas,B.J., Gevers,D., Earl,A.M., Feldgarden,M., Ward,D.V., Giannoukos,G., Ciulla,D., Tabbaa,D., Highlander,S.K.,

Sodergren,E. *et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.*, **21**, 494–504.

7. Kunin,V., Engelbrektson,A., Ochman,H. and Hugenholtz,P. (2009) Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. *Environ. Microbiol.*, **12**, 118–123.
8. Quince,C., Lanzen,A., Davenport,R.J. and Turnbaugh,P.J. (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.
9. Creer,S., Fonseca,V.G., Porazinska,D.L., Giblin-Davis,R.M., Sung,W., Power,D.M., Packer,M., Carvalho,G.R., Blaxter,M.L., Lambshead,P.J.D. *et al.* (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Mol. Ecol.*, **19**, 4–20.
10. Pruesse,E., Quast,C., Knittel,K., Fuchs,B., Ludwig,W., Peplies,J. and Glockner,F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.
11. Pawlowski,J., Christen,R., Lecroq,B., Bachar,D., Shahbazkia,H.R., Amaral-Zettler,L. and Guillou,L. (2011) Eukaryotic richness in the abyss: insights from pyrotag sequencing. *PLoS ONE*, **6**, e18169.
12. Porazinska,D., Giblin-Davis,R., Faller,L., Farmerie,W., Kanzaki,N., Morris,K., Powers,T., Tucker,A., Sung,W. and Thomas,W. (2009) Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Mol. Ecol. Resources*, **9**, 1439–1450.
13. Huber,T., Faulkner,G. and Hugenholtz,P. (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*, **20**, 2317–2319.
14. 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. *Nat. Methods*, **6**, 639–641.
15. Reeder,J. and Knight,R. (2009) The 'rare biosphere': a reality check. *Nat. Methods*, **6**, 636–637.
16. von Wintzingerode,F., Gobel,U.B. and Stackebrandt,E. (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol. Rev.*, **21**, 213–229.
17. 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. *Appl. Environ. Microbiol.*, **72**, 5734–5741.
18. Wang,G.C. and Wang,Y. (1997) Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl. Environ. Microbiol.*, **63**, 4645–4650.
19. Qiu,X., Wu,L., Huang,H., McDonel,P.E., Palumbo,A.V., Tiedje,J.M. and Zhou,J. (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl. Environ. Microbiol.*, **67**, 880–887.
20. Lahr,D.J. 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**, 857–866.
21. Engelbrektson,A., Kunin,V., Wrighton,K.C., Zvenigorodsky,N., Chen,F., Ochman,H. and Hugenholtz,P. (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J.*, **4**, 642–647.
22. Cole,J.R., Chai,B., Marsh,T.L., Farris,R.J., Wang,Q., Kulam,S.A., Chandra,S., McGarrell,D.M., Schmidt,T.M., Garrity,G.M. *et al.* (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res.*, **31**, 442–443.
23. Gonzalez,J.M., Zimmermann,J. and Saiz-Jimenez,C. (2005) Evaluating putative chimeric sequences from PCR-amplified products. *Bioinformatics*, **21**, 333–337.
24. Ashelford,K.E., Chuzhanova,N.A., Fry,J.C., Jones,A.J. and Weightman,A.J. (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.*, **71**, 7724–7736.
25. Smyth,R.P., Schlub,T.E., Grimm,A., Venturi,V., Chopra,A., Mallal,S., Davenport,M.P. and Mak,J. (2010) Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene*, **469**, 45–51.
26. Wang,G.C. and Wang,Y. (1996) The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology*, **142**, 1107–1114.
27. Meldal,B.H., Debenham,N.J., De Ley,P., De Ley,I.T., Vanfleteren,J.R., Vierstraete,A.R., Bert,W., Borgonie,G., Moens,T., Tyler,P.A. *et al.* (2007) An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. *Mol. Phylogenet. Evol.*, **42**, 622–636.
28. Tamura,K., Dudley,J., Nei,M. and Kumar,S. (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.*, **24**, 1596–1599.
29. Yoder,M., De Ley,I.T., King,I.W., Mundo-Ocampo,M., Mann,J., Blaxter,M., Poiras,L. and De Ley,P. (2006) DESS: a versatile solution for preserving morphology and extractable DNA of nematodes. *Nematology*, **8**, 367–376.
30. Blaxter,M.L., De Ley,P., Garey,J.R., Liu,L.X., Scheldeman,P., Vierstraete,A., Vanfleteren,J.R., Mackey,L.Y., Dorris,M., Frisse,L.M. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71–75.
31. Huse,S.M., Huber,J.A., Morrison,H.G., Sogin,M.L. and Welch,D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, **8**, R143.
32. Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, **453**, 3–31.
33. Ricotta,C. and Szeidl,L. (2006) Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. *Theor. Popul. Biol.*, **70**, 237–243.
34. Dixon,P. and Palmer,M.W. (2003) VEGAN, a package of R functions for community ecology. *J. Veg. Sci.*, **14**, 927–930.
35. Snelgrove,P.V.R. (1999) Getting to the bottom of marine biodiversity: Sedimentary habitats – Ocean bottoms are the most widespread habitat on Earth and support high biodiversity and key ecosystem services. *Bioscience*, **49**, 129–138.
36. Blaxter,M., Mann,J., Chapman,T., Thomas,F., Whitton,C., Floyd,R. and Abebe,E. (2005) Defining operational taxonomic units using DNA barcode data. *Philos. Trans. R. Soc. Lond., B. Biol. Sci.*, **360**, 1935–1943.
37. Creer,S. (2010) Second-generation sequencing derived insights into the temporal biodiversity dynamics of freshwater protists. *Mol. Ecol.*, **19**, 2829–2831.
38. Clayton,R.A., Sutton,G., Hinkle,P.S. Jr, Bult,C. and Fields,C. (1995) Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. *Int. J. Syst. Bacteriol.*, **45**, 595–599.
39. Chen,G., Znosko,B.M., Jiao,X. and Turner,D.H. (2004) Factors affecting thermodynamic stabilities of RNA 3 x 3 internal loops. *Biochemistry*, **43**, 12865–12876.
40. Wu,L., Tang,T., Zhou,R. and Shi,S. (2007) PCR-mediated recombination of the amplification products of the Hibiscus tiliaceus cytosolic glyceraldehyde-3-phosphate dehydrogenase gene. *J. Biochem. Mol. Biol.*, **40**, 172–179.
41. Shammas,F.V., Heikkila,R. and Osland,A. (2001) Fluorescence-based method for measuring and determining the mechanisms of recombination in quantitative PCR. *Clin. Chim. Acta*, **304**, 19–28.
42. Small,R.L., Ryburn,J.A. and Wendel,J.F. (1999) Low levels of nucleotide diversity at homoeologous Adh loci in allotetraploid cotton (*Gossypium* L.). *Mol. Biol. Evol.*, **16**, 491–501.
43. Floyd,R.M., Rogers,A.D., Lambshead,P.J.D. and Smith,C.R. (2005) Nematode-specific PCR primers for the 18S small subunit rRNA gene. *Mol. Ecol. Notes*, **5**, 611–612.
44. De Ley,P., De Ley,I.T., Morris,K., Abebe,E., Mundo-Ocampo,M., Yoder,M., Heras,J., Waumann,D., Rocha-

Olivares,A., Jay Burr,A.H. *et al*. (2005) An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. *Philos. Trans. R. Soc. Lond., B. Biol. Sci.*, **360**, 1945–1958.

45. Hugenholtzt,P. and Huber,T. (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int. J. Syst. Evol. Microbiol.*, **53**, 289–293.
46. Jumpponen,A. (2007) Soil fungal communities underneath willow canopies on a primary successional glacier forefront: rDNA sequence results can be affected by primer selection and chimeric data. *Microbiol. Ecol.*, **53**, 233–246.
47. Judo,M.S., Wedel,A.B. and Wilson,C. (1998) Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Res.*, **26**, 1819–1825.
48. Meyerhans,A., Vartanian,J.P. and Wain-Hobson,S. (1990) DNA recombination during PCR. *Nucleic Acids Res.*, **18**, 1687–1691.
49. Glenn,T.C. (2011) Field guide to next-generation DNA sequencers. *Mol. Ecol. Resources*, **11**, 759–769.
50. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, **74**, 5463–5467.
51. Little,T.J. (2002) The evolutionary significance of parasitism: do parasite-driven genetic dynamics occur ex silico? *J. Evol. Biol.*, **15**, 1–9.
52. Bishop,J.G., Dean,A.M. and Mitchell-Olds,T. (2000) Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution. *Proc. Natl Acad. Sci. USA*, **97**, 5322–5327.
53. Wegner,K.M. (2009) Massive parallel MHC genotyping: titanium that shines. *Mol. Ecol.*, **18**, 1818–1820.
54. Duda,T.F.J. and Palumbi,S.R. (1999) Molecular genetics of ecological diversification: Duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proc. Natl Acad. Sci. USA*, **96**, 6820–6823.
55. Mebs,D. (2001) Toxicity in animals Trends in evolution? *Toxicon*, **39**, 87–96.