Vol. 66: 13–21, 2012
doi: 10.3354/ame01510

AQUATIC MICROBIAL ECOLOGY
Aquat Microb Ecol

Published online March 14

OPEN ACCESS

Distinct approaches for the detection and removal of chimeric 16S rRNA sequences can significantly affect the outcome of between-site comparisons

Nika Zajec, Blaž Stres, Gorazd Avguštin*

University of Ljubljana, Biotechnical Faculty, Animal Science Department, Chair for Microbiology and Microbial Biotechnology, Groblje 3, 1230 Domžale, Slovenia

ABSTRACT: Comparative analyses of 16S rRNA clone libraries represent a standard tool in microbial ecology. Chimeric sequences are generally removed prior to such comparisons. A literature survey revealed a general pattern: (1) most commonly a single chimera identification approach (CIA) has been used; (2) putative chimeras have routinely been discarded without manual examination; (3) chimera filtered datasets have been submitted to repositories. To explore the effects of various CIAs on the study of microbial β-diversity relationships using complete primary data, 4 bacterial and 4 archaeal clone libraries were generated from a submarine spring and analyzed together with 3 bacterial and 3 archaeal published primary datasets. The primary datasets were compared with their 8 different CIA filtered datasets using Chimera_check, CCODE, Pintail, Chimera Slayer and Bellerophon, the last with 4 different settings. When CIA filtered datasets were pooled according to the CIA used, no significant differences between them could be observed, although there was not complete congruency between the different CIAs. When CIA filtered datasets of the same clone library were compared, generally no significant differences could be observed. In contrast, when CIA filtered datasets of different clone libraries were compared, the statistical significance of the relationships shifted from significant to insignificant or vice-versa in many cases depending on the CIA used. This precludes a correct identification of β-diversity. To solve this problem, we treated all CIA filtered datasets and primary data of a single clone library as CIA replicates in non-parametric MANOVA. This enabled unambiguous delineation of environmental samples by taking into account all CIA introduced data modifications.

KEY WORDS: 16S rRNA · Chimera · Beta diversity · Clustering · Pairwise significance · UniFrac · Classifier

INTRODUCTION

Comparative analysis of bacterial and archaeal 16S rRNA gene sequences has been routinely adopted as a means for the identification of the dominant populations of microbial communities, to infer the within (α-) and between site (β-) relationships and to design tools for monitoring their responses to environmental perturbations (Lozupone & Knight 2005, Lozupone et al. 2007, Ley et al. 2008, Auguet et al. 2010, Barberán & Casamayor 2010). Since rRNA genes from environmental samples are most frequently amplified in such studies via PCR, the resulting clone libraries may contain chimeric sequences, heteroduplexes and mutations (von Wintzingerode et al. 1997, Qiu et al. 2001, Hugenholz & Huber 2003, Ashelford et al. 2005, DeSantis et al. 2006). As a result of public awareness, filtering of putative chimeric sequences before submission to public databases has been reported in the majority of published studies. Several bioinformatic tools have been developed to eliminate chimeras from clone libraries commonly obtained by Sanger sequencing. However, these tools depend on the quality of the public sequence databases. The best known and most common chimera identification approaches (CIA) used so far have been Chimera_check (Maidak et al. 2001) and Bellerophon (Huber et al. 2004), while Pintail (Ashelford et al. 2005) and Mallard (Ashelford et al. 2006), CCODE (Gonzalez et al. 2005) and Chimera Slayer (Schloss et al. 2009) have been less frequently used. As recommended by several

*Corresponding author. Email: gorazd.avgustin@bf.uni-lj.si

authors, all putative chimeras should be manually verified to identify the breakpoint or conversion point at which the chimeric sequence signature shifts from one parent to the next (Wang & Wang 1997, Hugenholtz & Huber 2003, Ashelford et al. 2005, 2006). Despite easy access to 16S rRNA sequences and sequence libraries as well as common availability of the tools for the identification of putative chimeras, numerous corrupted 16S rRNA sequences deposited in the public databases have been reported. It has been reported that proportions of corrupted 16S rRNA sequences in these public data bases range from 3% to 20% (Qiu et al. 2001, Hugenholtz & Huber 2003, Ashelford et al. 2005, DeSantis et al. 2006). In addition, the next generation sequencing methods based on pyrosequencing approaches are not immune to the same problem (Quince et al. 2009). Although numerous approaches exist, there is no universally accepted method to detect chimera in 16S rRNA gene sequence datasets and the results of various CIA are only intended to assist users in making their own decision before submitting their datasets to public repositories. In a recent study, Ley et al. (2008) reported that chimera removal using Bellerophon did not affect sample clustering based on principal coordinate analysis (PCoA) of UniFrac distances. However, the effects of various CIAs on microbial β-diversity have not yet been investigated systematically.

In this study, we surveyed the literature to identify the general practice for chimera identification and removal adopted by researchers and to reanalyze publicly available clone libraries containing primary data for chimera. The following approaches were used: Chimera_check (Maidak et al. 2001), CCODE (Gonzalez et al. 2005), Pintail (Ashelford et al. 2005), Chimera Slayer (Schloss et al. 2009) and Bellerophon (Huber et al. 2004), with 4 different settings. Thus, in total 8 chimera removal procedures were tested. As a baseline, to complement the clone libraries from published literature, we prepared 8 clone libraries, 4 archaeal and 4 bacterial, from a submarine spring located in the northern Adriatic Sea (Izola32). Our null hypothesis was that chimera identification and removal would not effect the phylogenetic signal of the individual microbial community, and thus would not significantly influence the between-community relationship.

MATERIALS AND METHODS

Identification of relevant datasets and general practices

Three separate PubMed and 'Web of Science' searches were performed. The objectives were (1) to quantify the frequency of the use of various CIAs, and (2) to identify the generally adopted strategy for verification and removal of putative chimeric sequences. For our comparative study we also retrieved 3 bacterial (Sievert et al. 2000, Orphan et al. 2001, Garcia-Martinez et al. 2009) and 3 archaeal (Knittel et al. 2005, Oline et al. 2006, Chaudhary et al. 2009) clone libraries of 16S rRNA gene sequences that provided suitable primary data that had not been prescreened by a CIA (see Table S1 in the supplement at www.int-res.com/articles/suppl/a066p013_supp.pdf; the authors' names shown in Table S1 are used hereafter to identify the clone libraries). To allow comparison with our own Izola32 clone libraries, criteria for the selection of studies were that (1) the primary (quality checked) sequence data of the clone library had been made publicly available and (2) that these were obtained from marine and spring sediments.

Izola32 clone library construction

Sediments from 4 locations (denominated Spring out, Spring wall, Spring up and Spring down) at the Izola32 warm spring (45° 32.9′ N, 13° 38.7′ W) (Faganeli et al. 2005) were sampled with plexiglass corers (length = 100 cm, diam. = 6 cm) by a SCUBA diver. Corers were capped on both sides by rubber stoppers, placed on ice and swiftly transported to the laboratory, for total microbial DNA isolation. The sediment cores were aseptically extracted from corers and sliced into 2 cm longitudinal sections. DNA was extracted in triplicates from 0.5 g sediment portions using UltraClean Soil DNA kit (MoBio), according to manufacturer's instructions for maximum DNA yields. 16S rRNA genes were amplified using Bacteria and Archaea specific primer pair fd1-1401R (Weisburg et al. 1991, Nübel et al. 1996) and F109Arch-1386Rarch (Wright & Pimm, 2003), respectively. After heating to 95°C for 3 min, the reaction was cycled as follows: 30 cycles of 30 s at 95°C, followed by 45 s at 56°C for bacterial primers and 53°C for archaeal primers, and finally 100 s at 72°C. The cycles were followed by final elongation for 15 min at 72°C. Each 50 µl reaction contained 0.25 µM of each primer, 1× *Taq* buffer, 2mM $MgCl_2$, 0.2 mM dNTP mix and 1 U *Taq* Polymerase (Fermentas Life Sciences).

PCR products were purified by High Pure PCR Purification Kit (Roche) and cloned using the CloneJET Cloning Kit (Fermentas), according to the manufacturer's instructions. Clones from the obtained bacterial and archaeal clone libraries were randomly selected and sequenced at Macrogen (Korea). The resulting sequences were edited using the base calling program for DNA sequence analysis (PHRED), aided by the Codon Aligner (www.phrap.com). Basecalling, end-clipping and vector trimming were performed according to default settings.

Identification of chimeric sequences

Orientation Checker (www.bioinformatics-toolkit.org/Squirrel/index.html) was used to group 16S rRNA sequences into size classes covering 5′, central or 3′ sections (Ashelford et al. 2006). The Silva-based alignment of the template file for Chimera Slayer was aligned to standard Greengenes alignment in mothur and a full-length *Escherichia coli* 16S rRNA gene sequence (U00096) was also included in analyses as suggested by Ashelford et al. (2006). Chimera_check (Maidak et al. 2001), CCODE (Gonzalez et al. 2005), Pintail (Ashelford et al. 2005), Chimera Slayer (Schloss et al. 2009) and Bellerophon (Huber et al. 2004), the last of these with 4 different correction settings (Huber-Hugenholz, Kimura, Jukes-Cantor and 'none'), were used for identification of putative chimeric sequences in all libraries. Following the general practice that emerged from our literature search (e.g. Lopez-Garcia et al. 2003, Ley et al. 2006, 2008), the identified putative chimeric sequences in clone libraries were removed and the resulting datasets were organized according to the method of chimera identification (Table S1 in the supplement). In total, 63 archaeal and 63 bacterial datasets (7 clone libraries × 9 treatments) were prepared for further analysis. These were produced from primary data (no chimera removed) from the 3 retrieved clone libraries and the 4 clone libraries from Izola32 and, in each case, following the removal of chimera identified by the 8 different approaches.

Building the phylogenetic tree

Silva (Web)Aligner (www.arb-silva.de/aligner/) was used to align sequences from the clone libraries to the standard Arb alignment of Silva100 database (Ludwig et al. 2004, Pruesse et al. 2007). The imported sequences were added to the Arb guide tree using the Arb parsimony insertion tool as previously described (Ludwig et al. 2004). This is essential, as there was very little overlap in sequenced regions of the 16S rRNA genes when comparing data from all of the studies. The tree made in Arb was exported and each sequence was annotated with designations, relating sequences to particular clone library CIA filtered datasets based on the corresponding chimera-removal approach.

Statistical analyses

First, hierarchical clustering and significance tests in the UniFrac web interface (Lozupone & Knight 2005, Lozupone et al. 2007) or mothur (Schloss et al. 2009) were performed, using the produced Arb tree and a file mapping sequence labels of the original clone libraries and following the removal of chimera identified by the 8 different procedures. Datasets generated through different methods of chimera identification were treated as distinct environmental samples. We used UniFrac to test for significant differences between pairs of environmental samples (i.e. sequence collections produced). A qualitative measure of community β-diversity based on presence-absence of particular lineages (unweighted UniFrac) and a quantitative measure, based on how many sequences from each lineage are present (weighted UniFrac), were used. It is known that the 2 approaches measure different aspects of microbial diversity (Lozupone et al. 2006). UniFrac tests were performed using 1000 permutations. PCoA was used to find clusters and the most important axes of variation among samples. Ley et al. (2008) showed that despite the presence of chimeric sequences identified by Bellerophon (one setting) the basic groupings in PCoA remained the same. In addition, this technique was shown to be more successful than cluster recovery by jackknifing to detect similarities in the data (Liu et al. 2007).

The p-values were corrected for multiple comparisons using a modified Benjamini-Yekutieli correction (Benjamini & Yekutieli 2001) and experimentwise error rates where $p < 0.05$ were considered significant (Schloss 2008). This procedure can accommodate the significance testing of large numbers of potentially dependent tests while balancing risks of Type I (the probability of falsely rejecting the null hypothesis) and Type II errors (not rejecting the null hypothesis when it is false) (García 2003, Nakagawa 2004, Narum 2006).

Second, an independent phylogenetic analysis of each clone library and its CIA filtered datasets was conducted using nomenclatural taxonomy and Bergey's Manual together with the sequence Classifier at Ribosomal Database Project II (http://rdp.cme.msu.edu/) (Wang et al. 2007, Cole et al. 2009). Non-metric multidimensional scaling (NMDS) using Bray-Curtis distance, and 250 runs with real and randomized data, was applied for visualizing the differences in CIA filtered datasets of the original clone libraries using PC-ORD V5.0 (MjM Software Design). To assess the significance of correlations, stress decomposition (Monte Carlo [MC] Scree plot) and the number of dimension axes to retain, the MC technique was used (McCune et al. 2002).

To identify which CIAs produced most congruent results and to quantify the overall effect of various CIAs, we used both weighted and unweighted UniFrac calculations in combination with quantitative and presence/absence Classifier hierarchy NMDS approaches. These were repeated using a reorganized set of data, treating all CIA filtered datasets of a single CIA as a novel sequence collection. A novel mapping file for

Archaea and *Bacteria* was prepared, resulting in an additional 9 separate bacterial and 9 archaeal datasets. Classifier hierarchy files of distinct clone libraries analyzed with a particular CIA were merged in the same order. To explore the effects of clone library preparation, a subset of the complete dataset was prepared, retaining only the 4 archaeal and 4 bacterial clone libraries generated in this study.

Last, to identify which environments truly differed significantly, all CIA filtered datasets of a single clone library plus its original dataset were treated as replicate measurements of microbial community and were subjected to non-parametric MANOVA (NP-MANOVA) (Anderson 2001) using 1000 permutations as implemented in PAST (Hammer et al. 2001). This non-parametric test of significant differences between 2 or more groups was based on Bray-Curtis distance measure and is normally used for ecological taxa-in-samples data. The presence/absence for qualitative data was produced in PC-ORD V5.0 (MjM Software Design), whereas quantitative data were first normalized and arc sin square root transformed before analysis. This resulted in an additional 7 separate bacterial and 7 archaeal datasets that contained 9 CIA-filtered datasets next to original datasets that were tested for significant pairwise differences at $p < 0.05$. Experimentwise error rates were corrected using either classical Bonferroni or modified Benjamini-Yekutieli corrections described above.

RESULTS AND DISCUSSION

Chimera identification approaches: general practice and comparison

A double literature survey by PubMed and 'Web of Science' was performed on September 1, 2010. The principal results are outlined in the following paragraphs.

First, in the majority of cases (2037 and 2003 in PubMed and Web of Science, respectively) the clone libraries had been screened prior to publication using either Chimera_check (831/771 screenings in PubMed and Web of Science, respectively), Bellerophon (448/385), Pintail (169/156), Mallard (165/144), CCODE (33/31) or Chimera Slayer (12/10). There were only a few studies where clone libraries had been screened using more than one chimera identification software (e.g. Chimera_Check and Bellerophon) as described by Ashelford et al. (2006).

Second, manual inspection of 'Materials and methods' and 'Results' sections in published papers elucidated the general approach taken. The sequences identified as putative chimera were generally discarded without further manual examination, contrary to the recommended procedure (e.g. Cole et al. 2007, Ashelford et al. 2005).

Third, our literature survey confirmed that the primary data (all sequences of sufficient quality produced within a single study) (Montgomery 1996) are difficult to obtain from public databases (Lozupone & Knight 2007). Many of the studies reported prescreening of clones by restriction fragment length polymorphism (RFLP), temperature gradient gel electrophoresis (TGGE), denaturing gradient gel electrophoresis (DGGE), terminal restriction fragment length polymorphism (T-RFLP) and single-strand conformation polymorphism (SSCP) and other typing approaches or submitted only the clones representing operational taxonomic units at various cut-off values without information on their relative abundance. We were able to identify 6 clone libraries, 3 archaeal and 3 bacterial (Table S1 in the supplement), which did not report identification of putative chimera and were assumed to contain original primary data for the purpose of this study. These datasets were screened in conjunction with the 8 Izola32 sequence collections.

Fourth, following the general procedure we observed in published literature on our model datasets, each of the CIAs identified a different set of sequences as putative chimeric sequences, with little overlap between approaches (Table S1 in the supplement). There was no complete congruency between chimera removal approaches using either Classifier NMDS or UniFrac PCoA weighted or unweighted analyses (Fig. 1; Fig. S1 in the supplement at www.int-res.com/articles/suppl/a066p013_supp.pdf). However, both UniFrac significance tests showed that pooled datasets (according to chimera identification protocol) did not differ significantly from one another. The generalization of this result suggests that either UniFrac did not effectively capture existing differences introduced by chimera-identification approaches in the pooled datasets or that the differences introduced by the same chimera identification protocol in distinct clone libraries counterbalanced or masked potential differences at the level of comparing pooled datasets. Thus, we were not able to identify which were the most congruent CIAs. This suggests that chimera removal approach produced no significant difference between datasets pooled across various clone libraries. Using a smaller dataset, comprised of only 4 bacterial and 4 archaeal clone libraries generated in this study and their CIA filtered datasets, the same result was obtained (results not shown). Therefore, one cannot reject our null hypothesis at the level of $p < 0.05$ that there is no difference between community structures of the 8 pooled CIA filtered datasets and the original pooled data. Thus, it appears that the overall structure of the pooled datasets under analysis was not significantly altered by the CIA of choice.

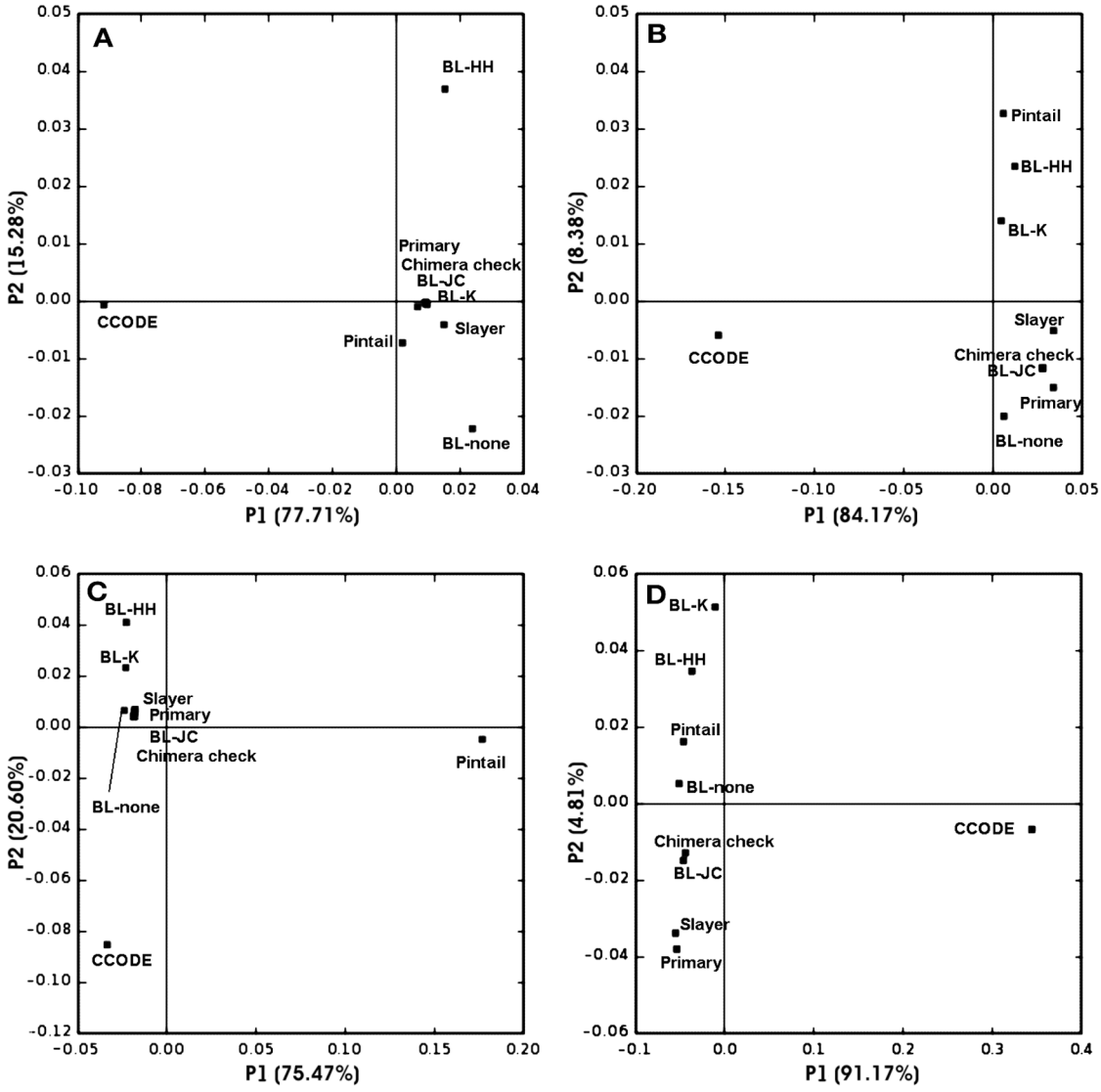

Fig. 1. Principal coordinate analysis (PCoA) of the chimera identification approach congruency with pooled datasets conducted in UniFrac, applied to (A) unweighted archaeal, (B) weighted archaeal, (C) unweighted bacterial and (D) weighted bacterial clone libraries and their chimera identification approach (CIA) filtered datasets. Each dot represents the composite phylogenetic signal of primary data from 7 clone libraries or of their CIA filtered datasets following the application of 8 different CIAs. See also Fig. S1 in the supplement where the results of Classifier hierarchy files NMDS are presented. Percent variation explained is given in brackets. Abbreviations describe the different correction settings used in Bellerophon program. BL-HH: Bellerophon with Huber-Hugenholz correction, BL-K: Bellerophon with Kimura correction, BL-JC: Bellerophon with Jukes-Kantor correction, BL-none: Bellerophon without correction

Pairwise significance of clone library CIA filtered datasets

In the next step we performed the pairwise comparisons of all 64 datasets comprised of primary data and CIA filtered datasets of each clone library. Sequences in these datasets are in the same range as described in the literature (Lozupone et al. 2006, Ley et al. 2008). Both NMDS ordination and PCoA generally grouped CIA filtered datasets of a single clone library into congruent clusters (Fig. 2; see Fig. S2 in the supplement at www.int-res.com/articles/suppl/a066p013_supp.pdf), as was also observed before using only one CIA (Ley et al. 2008). However, exceptions to the general clustering were observed in weighted UniFrac for 3 archaeal (Spring out, Oline, Spring down) and 4 bacterial clone libraries (Sievert, Spring wall, Spring up and Spring down). The significance of pairwise relationships of clustered CIA filtered datasets (Fig. 2) is presented in Fig. 3. In general, pairwise testing using the same CIA-

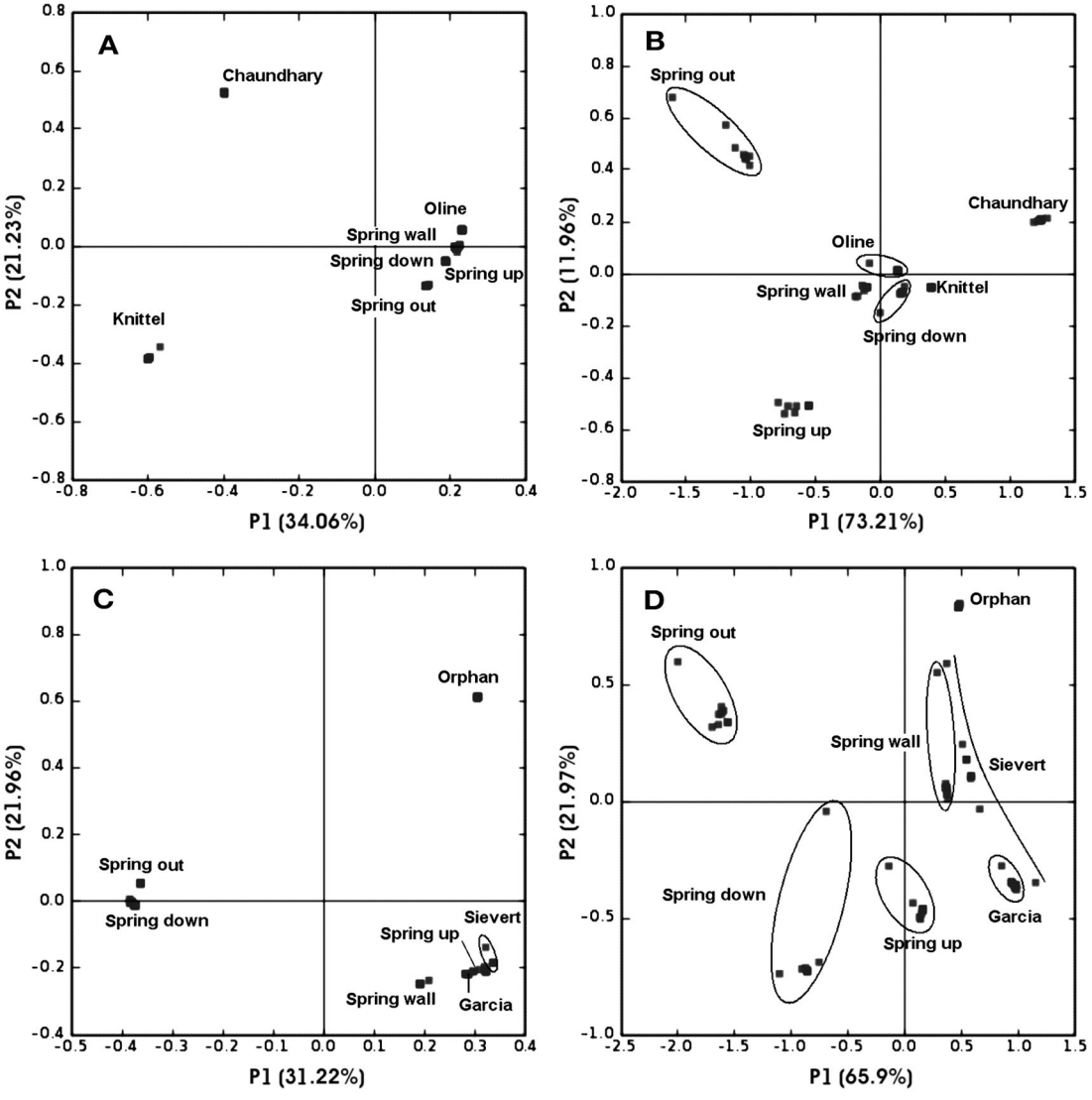

Fig. 2. Principal coordinate analysis (PCoA) of pairwise comparisons of datasets comprised of primary data and CIA filtered datasets of each clone library conducted in UniFrac, applied to (A) unweighted archaeal, (B) weighted archaeal, (C) unweighted bacterial and (D) weighted bacterial clone libraries and their chimera identification approach (CIA) filtered dataset datasets. Each dot represents the phylogenetic signal of a single clone library or its CIA filtered dataset. See also Fig. S2 in the supplement where the results of Classifier hierarchy files NMDS are presented. Percent variation explainedis given in brackets

filtered dataset of a single clone library did not reveal statistically significant differences in either form of the UniFrac test (the diagonal blocks of squares in Fig. 3). This indicates that the fraction of identified putative chimeric sequences and their contributed branch length was in general not large enough to result in significant differences between the CIA filtered datasets of sequences derived from single clone libraries used in this study (Fig. 3). There were 2 exceptions, however: the bacterial libraries Sievert and Spring wall (Fig. 3D).

When different CIA filtered datasets were pairwise compared to each other, the effective outcome of hypothesis testing remained the same for the majority of comparisons between clone libraries irrespective of the CIAs applied (blocks of squares of uniform color in Fig. 3). However, in many cases, the use of different CIAs gave rise to remarkable shifts in the significance levels (blocks of squares of non-uniform color in Fig. 3). These shifts appear to follow the major clustering observed in Fig. 2 and Fig. S2 in the supplement. The exclusion of the putative chimeras changed the significance of the pairwise relationship from significant ($p < 0.05$) to not significant, or vice-versa, thus reversing the hypothesis test outcome depending on the CIA used. UniFrac significance tests resulted in 10, 14, 18 and 20 blocks out of 49 in unweighted archaeal,

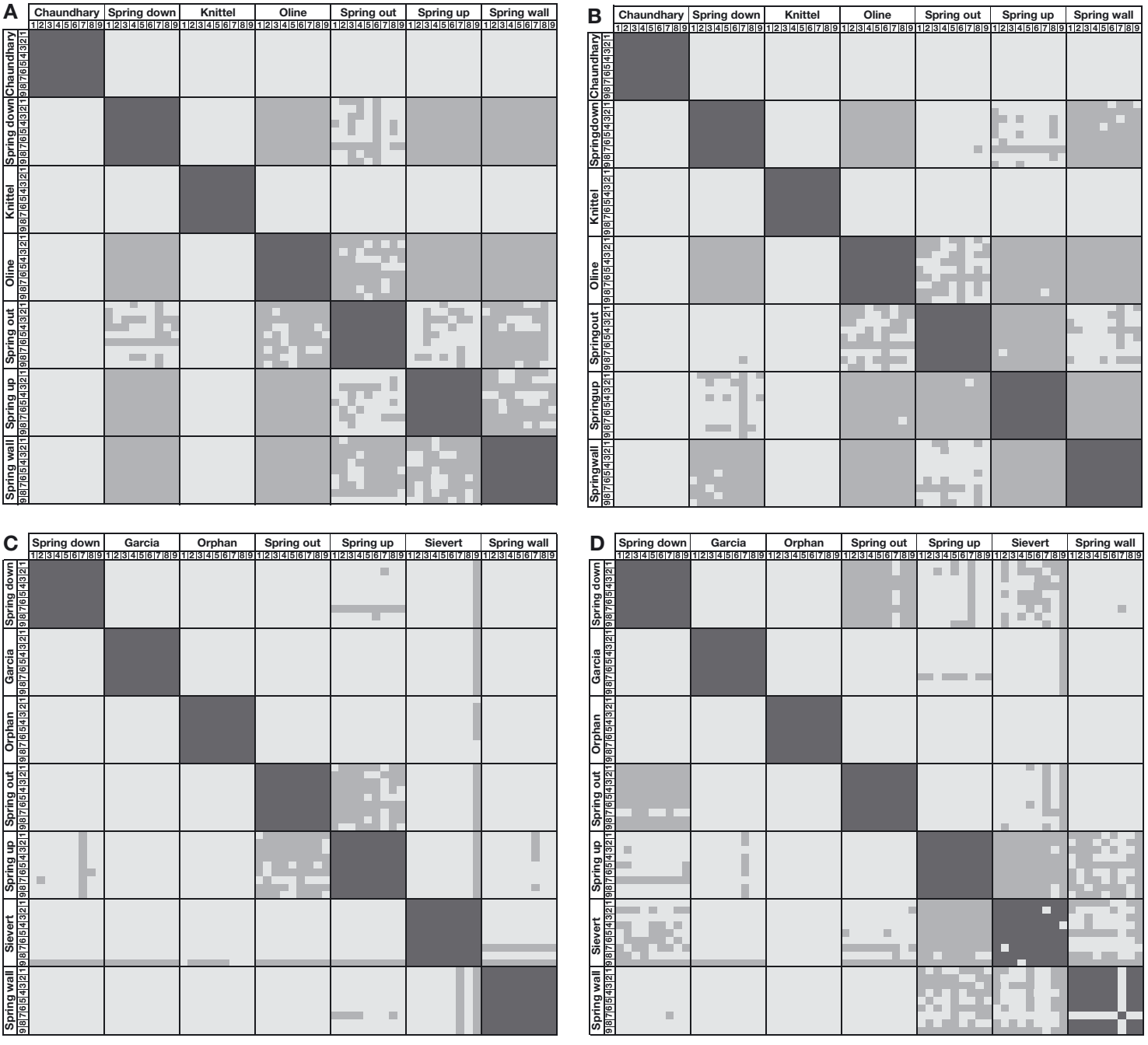

Fig. 3. A schematic matrix showing the results of pairwise UniFrac significance tests of (A) unweighted archaeal, (B) weighted archaeal, (C) unweighted bacterial and (D) weighted bacterial clone libraries and their filtered datasets after the application of 8 chimera removal approaches (CIAs). Entries in the matrix are organized in blocks according to primary data and shaded according to the significance of the difference between pairs of datasets: significant (■) $p < 0.05$, non-significant (■) $p > 0.05$. Diagonal blocks represent comparisons of CIA filtered datasets of the same clone libraries (■). All p-values were corrected for multiple comparisons using the modified Benjamini-Yekutieli correction (see 'Materials and methods'). Designations: 1, original dataset; 2, Chimera_check; 3, Bellerophon-Huber Hugenholtz; 4, Bellerophon-Jukes Cantor; 5, Bellerophon-Kimura; 6, Bellerophon-'none'; 7, CCODE; 8, Chimera Slayer; 9, Pintail

weighted archaeal, unweighted bacterial and weighted bacterial clone libraries and their CIA filtered datasets, respectively, in which shifts in the significance of community relationships were identified as a function of the chimera identification approach selection (Fig. 3). In addition, the use of the same CIA with 4 different settings as tested with Bellerophon resulted in shifts of relationship significance in some of the blocks (Fig. 3)

(4 settings in Bellerophon blocks of squares with varying color, for example the blocks showing comparisons between CIA datasets of clone libraries Spring down–Spring up or Spring up–Sievert). This suggests that even the use of a single CIA and varying settings is not immune to generating datasets that exhibit significant differences in the pairwise tests of community structure.

As multiple statistical tests were conducted in this study comparing 63 bacterial and 63 archaeal clone library CIA filtered datasets, the resulting p-values had to be corrected for false discovery in multiple comparisons using Benjamini-Yekutieli approach. This made the significance tests more conservative. However, if one limited comparisons to just 2 clone libraries, such corrections would not be necessary and the raw scores could be used instead, making shifts in pairwise significance between the 2 distinct CIA filtered datasets of distinct clone libraries even more evident.

In summary, the generation of distinctly processed datasets by CIAs resulted in contradictory pairwise significance tests in 4 to 7 clone libraries out of 7. On the other hand, treating all CIA filtered datasets of a single clone library (n = 8) and primary data (n = 1) as 9 replicates of a sample in NP-MANOVA accounted for all variability in data produced by distinct approaches to chimera removal, not just single pairwise comparison of 2 particular-user defined CIA filtered datasets. In our case, this resulted in unambiguous delineation of real environmental samples for archaeal ($p(same) < 0.0001$; total sum of squares = 2.971; within-group sum of squares = 0. 2244) and bacterial datasets ($p(same) < 0.0001$; total sum of squares = 12.86; within-group sum of squares = 1.316). A non-parametric multivariate analysis of clone library groups of CIA filtered datasets could provide an additional verification step as it takes into account all variability in data produced by distinct approaches to chimera removal, not just single pairwise comparison of 2 CIA filtered datasets.

To conclude, the removal of all putatively chimeric sequences from the datasets using distinct CIAs has resulted in unevenly processed and filtered datasets deposited in public databases. At least a portion of such datasets has the potential to give rise to shifts in pairwise significance and clustering found in our study, but they are nevertheless used for the inference of relationships between microbial communities and the key environmental factors (Lozupone et al. 2006, Ley et al. 2008, Auguet et al. 2010).

The use of unevenly filtered datasets is in sharp contrast to generally practiced strategy in high throughput sequencing (Middelbos et al. 2010, Caporaso et al. 2011). These archives require that all reads are deposited, enabling extensive and independent downstream reanalysis (Wheeler et al. 2008, Kaminuma et al. 2010).

As a consequence, it is impossible to compare the effectiveness of available chimera removal approaches, as the quality of the available data in public databases (Ashelford et al. 2005, Gonzalez et al. 2005), the sequence length in ecological survey studies, the intragenomic heterogeneity of 16S rRNA (Ashelford et al. 2006), the personal experience and judgment of researcher (Lozupone et al. 2006, Ley et al. 2008) and possibly also other unidentified factors, can all potentially affect the accuracy of available CIAs. The release of primary data following the next generation sequencing standards would provide a greater level of control over the large-scale Sanger data contained in public databases.

Acknowledgements. We acknowledge Sarah Westcott and Catherine Lozupone for helpful discussions and the 3 anonymous reviewers for stimulating comments that improved the manuscript. We are indebted to SCUBA diver Alfred Zajic from Koper, Slovenia, for sediment sampling.

LITERATURE CITED

Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. Austral Ecol 26:32–46

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (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

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2006) New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. Appl Environ Microbiol 72:5734–5741

Auguet JC, Barberan A, Casamayor EO (2010) Global ecological patterns in uncultured Archaea. ISME J 4:182–190

Barberán A, Casamayor EO (2010) Global phylogenetic community structure and β-diversity patterns in surface bacterioplankton metacommunities. Aquat Microb Ecol 59:1–10

Benjamini Y, Yekutieli D (2001) The control of false discovery rate under dependency. Ann Stat 29:1165–1188

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D and others (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA 108(Suppl):4516–4522

Chaudhary A, Haack SK, Duris JW, Marsh TL (2009) Bacterial and archaeal phylogenetic diversity of a cold sulfur-rich spring on the shoreline of Lake Erie, Michigan. Appl Environ Microbiol 75:5025–5036

Cole JR, Wang Q, Cardenas E, Fish J and others (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37 (Suppl 1):D141–D145

Cole JR, Chai B, Farris RJ, Wang Q and others (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res 35(SI):D169–D172

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M and others (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072

Faganeli J, Ogrinc N, Walter LM, Žumer J (2005) Geochemical characterization of the submarine spring of Izola (Gulf of Trieste, N Adriatic Sea). Mater Geoenvir 52:35–39

García LV (2003) Controlling the false discovery rate in ecological research. Trends Ecol Evol 18:553–554

García-Martínez M, López-López A, Calleja ML, Marba N, Duarte CM (2009) Bacterial community dynamics in a seagrass (*Posidonia oceanica*) meadow sediment. Estuaries Coasts 32:276–286

Gonzalez JM, Zimmermann J, Saiz-Jimenez C (2005) Evaluating putative chimeric sequences from PCR-amplified products. Bioinformatics 21:333–337

Hammer Ø , Harper DAT, Ryan PD (2001) PAST: paleontological statistics software package for education and data analysis. Palaeontol Electronica 4:1–9

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

Hugenholtz P, Huber T (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. Int J Syst Evol Microbiol 53:289–293

Kaminuma E, Mashima J, Kodama Y, Gojobori T and others (2010) DDBJ launches a new archive database with analytical tools for next-generation sequence data. Nucleic Acids Res 38:D33–D38

Knittel K, Lösekann T, Boetius A, Kort R, Amann R (2005) Diversity and distribution of methanotrophic archaea at cold seeps. Appl Environ Microbiol 71:467–479

Ley RE, Harris JK, Wilcox J, Spear JR and others (2006) Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. Appl Environ Microbiol 72: 3685–3695

Ley RE, Hamady M, Lozupone C, Turnbaugh PJ and others (2008) Evolution of mammals and their gut microbes. Science 320:1647–1651

Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res 35:e120

López-García P, Duperron PP, Foriel J, Susini J, Moreira D (2003) Bacterial diversity in hydrothermal sediment and epsilonproteobacterial dominance in experimental microcolonizers at the Mid-Atlantic Ridge. Environ Microbiol 5: 961–976

Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71:8228–8235

Lozupone CA, Knight R (2007) Global patterns in bacterial diversity. Proc Natl Acad Sci USA 104:11436–11440

Lozupone C, Hamady M, Knight R (2006) UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. Bioinformatics 7:371–384

Ludwig W, Strunk O, Westram R, Richter L and others (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32:1363–1371

Maidak BL, Cole JR, Lilburn TG, Parker CT and others (2001) The RDP-II (Ribosomal Database Project). Nucleic Acids Res 29:173–174

McCune B, Grace JB, Urban DL (2002) Analysis of ecological communities. MjM Software Design, Gleneden Beach, OR

Middelbos IS, Vester Boler BM, Qu A, White BA, Swanson KS, Fahey GC Jr (2010) Phylogenetic characterization of fecal microbial communities of dogs fed diets with or without supplemental dietary fiber using 454 pyrosequencing. PLoS ONE 5:e9768

Montgomery DC (1996) Introduction to statistical quality control, 3rd edn. John Wiley & Sons, New York, NY

Nakagawa S (2004) A farewell to Bonferroni: the problems of low statistical power and publication bias. Behav Ecol 15: 1044–1045

Narum SR (2006) Beyond Bonferroni: less conservative analyses for conservation genetics. Conserv Genet 7: 783–787

Nübel U, Engelen B, Felske A, Snaidr J and others (1996) Sequence heterogeneties of genes encoding 16S rRNAs in *Paenibacillus polymyxa* detected by temperature gradient gel electrophoresis. J Bacteriol 178:5636–5643

Oline DK, Schmidt SK, Grant MC (2006) Biogeography and landscape-scale diversity of the dominant Crenarchaeota of soil. Microb Ecol 52:480–490

Orphan VJ, Hinrichs KU, Ussler V III, Pauli CK and others (2001) Comparative analysis of methane-oxidizing archaea and sulfate reducing bacteria in anoxic marine sediment. Appl Environ Microbiol 67:1922–1934

Pruesse E, Quast C, Knittel K, Fuchs B, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35:7188–7196

Qiu XY, Wu LY, Huang HS, McDonel PE, Palumbo AV, Tiedje JM, Zhou J (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. Appl Environ Microbiol 67:880–887

Quince C, Lanzén A, Curtis TP, Davenport RJ and others (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods 6:639–641

Schloss PD (2008) Evaluating different approaches that test whether microbial communities have the same structure. ISME J 2:265–275

Schloss PD, Westcott SL, Ryabin T, Hall JR and others (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537–7541

Sievert SM, Kuever J, Muyzer G (2000) Identification of 16S ribosomal DNA-defined bacterial populations at a shallow submarine hydrothermal vent near Milos Island (Greece). Appl Environ Microbiol 66:3102–3109

von Wintzingerode FV, Göbel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analyses. FEMS Microbiol Rev 21:213–229

Wang GCJ, Wang Y (1997) Frequency of formation of chimeric molecules is consequence of PCR coamplification of 16S RNA genes from mixed bacterial genomes. Appl Environ Microbiol 63:4645–4650

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267

Weisburg WG, Barns SM, Pelletier DA, Lane DJ (1991) 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol 173:697–703

Wheeler DL, Barrett T, Benson DA, Bryant SH and others (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33(Suppl 1): D39–D45

Wright ADG, Pimm C (2003) Improved strategy for presumptive identification of methanogens using 16S riboprinting. J Microbiol Methods 55:337–349

Editorial responsibility: Rutger de Wit, Montpellier, France

Submitted: June 28, 2010; Accepted: April 7, 2011
Proofs received from author(s): March 2, 2012