doi:10.1111/j.1462-2920.2007.01253.x

TOPO TA is A-OK: a test of phylogenetic bias in fungal environmental clone library construction

D. Lee Taylor,[1]* Ian C. Herriott,[1] James Long[1] and Keith O'Neill[2]

[1]*University of Alaska, Institute of Arctic Biology, 311 Irving I Building, Fairbanks, AK 99775, USA.*
[2]*Broad Institute – Massachusetts Institute for Technology, 320 Charles St., Cambridge, MA 02141, USA.*

Summary

TA cloning methods are widely used in analyses of environmental microbial diversity, yet the potential of TA methods to yield phylogenetically biased results has received little attention. To test for a TA bias, we constructed clone libraries of fungal amplicons spanning the ribosomal internally transcribed spacer (ITS) and partial large subunit (LSU) from 92 boreal forest soil DNA extracts using two contrasting methods: the Invitrogen TOPO-TA system and the Lucigen PCR-SMART system. The Lucigen system utilizes blunt-ended rather than TA cloning and transcription terminators to reduce biases due to toxicity of expressed inserts. We analysed 588 clone sequences from the two libraries. Species diversity estimators applied to operational taxonomical units (OTUs) were slightly higher for Invitrogen than Lucigen, but confidence intervals for accumulation curves overlapped. Abundances of OTUs were correlated between the libraries ($r^2 = 0.5$, $P < 0.0001$), but certain OTUs had contrasting abundances in the two libraries and a likelihood ratio test rejected homogeneity of the OTU counts. We constructed parsimony and Bayesian trees from aligned LSU regions, and the 'phylogenetic test' revealed that lineage representation was not significantly different between the two libraries. We conclude that characterization of this fungal community was fairly robust to cloning method and no biases due to TA cloning were found.

Introduction

Direct characterization of DNA from environmental samples has revolutionized our understandings of microbial evolution and ecology (e.g. Stahl *et al*., 1984; Schadt *et al*., 2003). Efforts to enumerate the abundance and diversity of microbial taxa typically involve PCR amplification of diagnostic gene regions which results in a heterogeneous pool of PCR products. Sense can be made of this taxonomic slurry using fingerprinting methods such as denaturing gradient gel electrophoresis or terminal restriction fragment length polymorphism or by construction of clone libraries followed by sequence analysis. Clone library construction offers the benefit of phylogenetically explicit data. Much effort has been directed to characterizing and minimizing PCR artefacts and biases (Reysenbach *et al*., 1992; Suzuki and Giovannoni, 1996; Polz and Cavanaugh, 1998), but little attention has been paid to potential biases associated with cloning. The most widely utilized cloning strategies take advantage of non-template adenine additions by standard DNA polymerases to create one-base sticky ends complementary to the T vector. TA cloning systems could introduce biases through several mechanisms. First, it is known that the probability of non-template-dependent adenine addition is related to juxtaposed template sequence (Brownstein *et al*., 1996; Magnuson *et al*., 1996). Thus, some taxa might be over- or under-represented in a TA-based clone library depending on their sequences. Second, most commercial TA vectors include a promoter upstream of the insertion site which allows for selection of insert-containing colonies by various strategies such as interruption of the β-galactosidase gene fragment. A transcribed insert might interfere with growth of the transformed *Escherichia coli* cells such that these insert sequences will not be recovered in the resulting library. An obvious example would be an insert encoding a toxin gene. Transcribed rRNA inserts could also potentially interfere with ribosome function in the host *E. coli*. Third, it is known that reaction kinetics favour insertion of shorter DNA fragments over longer DNA fragments (Sambrook and Russell, 2001), and ligation efficiency varies among vector systems.

Our interest is in characterization of fungal communities in soil. Here, we have constructed a library of fungal community ribosomal PCR products with the widely used Invitrogen TOPO-TA topoisomerase-based cloning kit and compared this with a library constructed from the same samples using a blunt-ended cloning system.

Received 8 June, 2006; accepted 11 December, 2006. *For correspondence. E-mail fflt@uaf.edu; Tel. (+1) 907 474 6982; Fax (+1) 907 474 6967.

Results and discussion

We amplified fungal ribosomal internally transcribed spacer – large subunit (ITS-LSU) fragments from 92 black spruce humic horizon soil DNA extracts and pooled the resulting amplicons prior to cloning with (i) the Invitrogen pCR4.0-TOPO vector and (ii) the Lucigen pcrSMART-HC Kan vector. The Lucigen 'low-bias' system was chosen because ligation to blunt-ended vectors should avoid the potential bias related to variable non-template adenine addition. In addition, the Lucigen vector incorporates hairpin loops on both ends of the insertion site designed to block insert transcription and eliminate this potential source of bias.

Read and assembly qualities were similar in the two libraries with 308 passing assemblies in the Lucigen library and 312 passing assemblies in the Invitrogen library. More short inserts were encountered in the Lucigen library, which could be related to the end-repair step used to remove overhanging adenines. A larger number of non-fungal clones, many of which had the TW13 primer at both ends, were encountered in the Invitrogen library and excluded. We used two contrasting methods for chimera detection, but detected only a few chimeras. Of the 620 assemblies, 32 were excluded (for breakdown, see Table S1). Considering only full-length sequences, the mean Lucigen insert size was 1342 while the mean Invitrogen insert size was 1353, which were not significantly different (Kruskal–Wallis rank sum test, $P = 0.96$).

Fungal diversity in terms of total species richness based on OTU (operational taxonomical unit) presence/absence was similar in the two libraries. Sequence clustering using Cap3 (Huang and Madan, 1999) at 90% sequence identity yielded 148 OTUs, 55 of which occurred in both libraries. These sequence data have been submitted to the GenBank database under accession numbers EF433956–EF434155. The Invitrogen library contained 110 OTUs, 46 of which were singletons. The Lucigen library contained 93 OTUs, 28 of which were singletons. Despite a slightly higher total OTU richness in the Invitrogen library, the 95% confidence intervals of the species accumulation curves for the two libraries overlapped (Fig. S1), suggesting similar levels of alpha diversity.

The fungal communities from the black spruce forest humic soil horizon appear to be extremely diverse. Different fungal species belonging to the same genus often display less than 10% sequence divergence in the ITS region (Horton, 2002; Geml *et al*., 2006). Hence, the OTU groupings reported here certainly underestimate true species diversity. Despite this conservative approach, the species accumulation curves for neither individual library, nor the two libraries combined, approach an asymptote (Fig. S1). Non-parametric estimators of total OTU diversity, taking into account undetected types, including Chao 1, ICE, ACE, Jackknife 1 and Jackknife 2, gave estimates of 221–280 total OTUs in the community (Table S2), but the 95% confidence intervals for these estimates were large, ranging from 214 to 380, and did not stabilize with increasing numbers of subsamples in rarefaction analyses. In addition to OTU richness, the deep phylogenetic diversity we encountered is astounding: much of the known deep diversity of fungi is represented in the large-subunit tree, including the Chytridiomycota, Zygomycota, Ascomycota, Basidiomycota and numerous orders and families within them (data not shown). Ecological implications of the community structure of fungi at this site will be discussed elsewhere.

If the two cloning methods operate equivalently, we would expect them to have high community similarity scores. We calculated the Jaccard and Sorenson indices, which range from 0 when there is no species overlap to 1.0 when identical sets of species are present in two communities, using EstimateS 7.0 (Colwell and Coddington, 1994). The Sorenson index was 0.542 and Jaccard index was 0.372; these low values would normally be interpreted to indicate significant differences in the two communities/libraries. However, these measures only take into account species presence/absence, not abundance; hence, the numerous singletons in both libraries have as much impact on the indices as do the more abundant, shared OTUs. Furthermore, these measures assume that all species in each community have been detected, which is clearly not the case for this data set, nor for most clone library studies of microbial diversity. Chao and colleagues (2005) have shown that such measures are consistently biased towards underestimation of similarity in hyper-diverse, undersampled communities, and have proposed modified community similarity indices which take into account abundance data, and also incorporate estimates of the numbers of undetected taxa. These new indices gave similarity estimates much closer to the predicted value of 1.0 when comparing our two libraries: 0.822 for Chao-Jaccard-Est Abundance and 0.902 for Chao-Sorenson Est Abundance (Chao *et al*., 2005).

If the occurrence of an OTU in a clone library is related to the abundance of the corresponding organism in the community, and the two cloning methods give similar representations of the community, we would expect a correlation between the number of clones of a particular OTU which occur in each library. Considering only OTUs which were represented by two or more clones, a significant positive correlation ($r^2 = 0.5$, $F = 75.9447$, $P < 0.0001$) was found between the number clones belonging to a particular OTU in the Lucigen library and the number of clones of that OTU in the Invitrogen library. Rank abundances of the dominant OTUs for the two libraries were similar (Fig. 1). However, several OTUs abundant in the

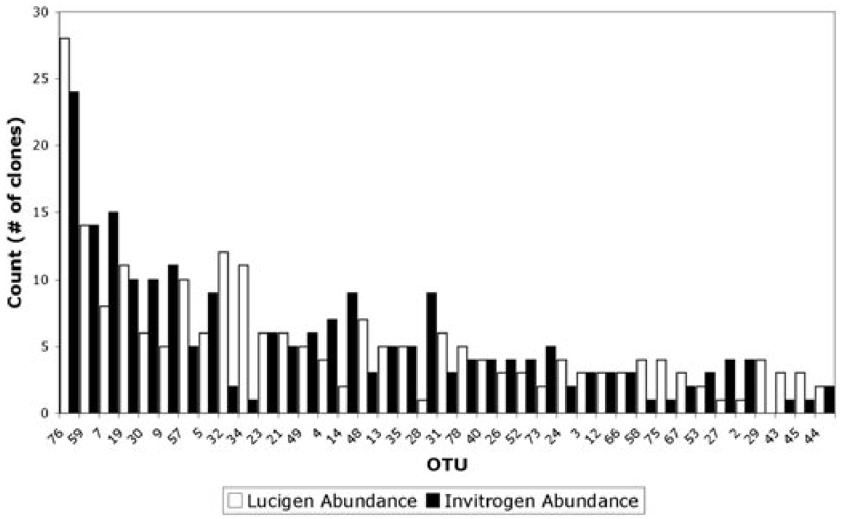


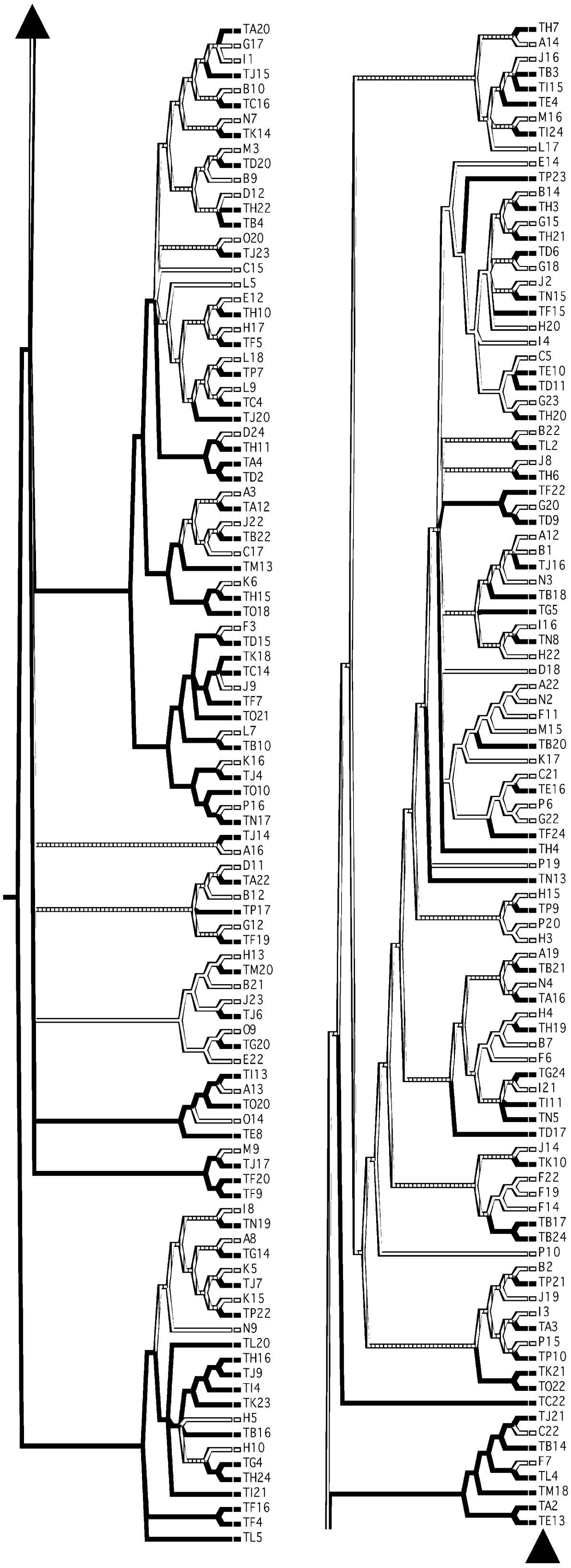
Fig. 1. Comparison of rank abundances of dominant OTUs in the two libraries.

Lucigen library (e.g. OTUs 29, 32 and 34) were rare in the Invitrogen library. The reverse was true in fewer instances (OTU 28). It is conceivable that phylogenetic biases against these OTUs operated during cloning. Upon cataloguing the phylogenetic affinities of all OTUs (data not shown), we found that both OTUs 32 and 34 belong to the Sebacinaceae (Basidiomycota). Across all OTUs, 23 clones associated with the Sebacinaceae were found in the Lucigen library compared with five such clones in the Invitrogen library. Further experiments would be required to determine whether this pattern was due to sampling error or to a systematic bias. However, it is notable that none of the dominant types were completely missing from either library. If abundances of OTUs between the libraries are homogeneous, contingency table tests should give non-significant results. However, when we compared only OTUs with a frequency of 10 or more clones, in order to avoid low expected cell counts, homogeneity of the counts was rejected ($r^2 = 0.024$, $P = 0.0012$), using the likelihood ratio test. The low r^2 suggests the frequencies are not far different from those expected with random sampling from a multinomial distribution. Indeed, if two of the OTUs with contrasting frequencies are excluded, such as 28 and 32, the likelihood ratio becomes non-significant. In other words, the clone counts in the two libraries are quite close to expectations for random samples from a single population. While the use of clone counts as a proxy for organismal abundances is tenuous due to artefacts, biases and sampling error inherent to PCR (Suzuki and Giovannoni, 1996; Polz and Cavanaugh, 1998), and perhaps also the cloning process, our results suggest that moderately consistent species abundance results can be obtained even for hyper-diverse fungal communities.

Of greater importance than species counts, *per se*, are the phylogenetic distributions of taxa represented by the two cloning methods. To evaluate representation of evolutionary lineages in the two libraries, a LSU sequence alignment of 203 taxa and 910 characters was constructed which included a representative of each OTU encountered in each library. In Bayesian tree evaluation, likelihoods and parameter estimates stabilized after 500 000 generations. A consensus was computed from 9500 trees after discarding the first 5000 in the burnin. The best tree had a log likelihood of –27787.07. There were 5093 most parsimonious trees of 6216 steps with consistency indices of 0.255. A 50% majority rule consensus tree was computed from these trees for subsequent analyses. Overall topologies of the Bayesian and parsimony consensus trees were largely congruent.

The parsimony and Bayesian trees were then exported to MacClade 4 (Maddison and Maddison, 1989). The phylogenetic test, or *P*-test, developed by Maddison and Slatkin (1991) and applied to microbial communities by Martin (2002), was carried out. This test asks whether significantly different sets of lineages occur in two samples. The deeper and more diverse the lineage that is unique to a sample, the more significant is the difference. An additional character was added to the data matrix to indicate the source library for each clone, mapped onto the trees, and the minimum number of steps required to explain the distribution of source library on the Bayesian tree was determined to be 81. Mapping of the character onto 1000 equiprobable random trees gave a range from 56 to 80 steps; hence the observed number of steps is outside of the null distribution. For the parsimony tree, source library required 80 steps to map onto the tree, which was again beyond the right-hand tail of the null distribution of 54–79 steps derived from 1000 randomizations of the parsimony tree. These results indicate that the arrays of lineages represented in the two libraries are more similar than would be expected by chance at $P < 0.001$ (Martin, 2002). When this test has been used to

Fig. 2. Bayesian tree of the ribosomal LSU region for representatives of each OTU from each library. There is similar phylogenetic representation across the two libraries. The number of 'evolutionary transitions' from Invitrogen to Lucigen has been traced on the tree. Black terminal branches and clone names beginning in 'T' are from the Invitrogen library. White terminal branches and clone names beginning with letters other than 'T' are from the Lucigen library. To prevent low-quality sequences from confounding the phylogenetic analyses, all consensus base calls with quality scores below 20 were converted to Ns. The ITS regions of divergent OTUs could not be aligned, so we focused only on the LSU region for these analyses. Large subunit regions representing each OTU found in each library were aligned with CLUSTALX (Thompson *et al.*, 1997). The resulting alignment was improved by eye. The GTR model with a portion of invariant sites and gamma distributed rate variation across other sites, was then utilized for tree searches with a heating parameter of 0.3 and eight chains run for 2 000 000 generations in MrBayes 3.1 (Huelsenbeck and Ronquist, 2001). Burnin was set to 500 000 generations based on output from an initial run.



compare microbial communities, the opposite result is more often observed, namely that fewer steps are found for the real tree than the random trees, indicating significant dissimilarity of the communities (e.g. Martin, 2002; Schadt *et al.*, 2003; Stach *et al.*, 2003; Acinas *et al.*, 2004; Dunfield and King, 2005; Eckburg *et al.*, 2005; Papineau *et al.*, 2005). The overlapping lineage representation of the Lucigen and Invitrogen libraries is apparent in the nLSU tree shown in Fig. 2. Note that this analysis takes phylogenetic scale into account, but is based on presence/absence, and hence does not take OTU abundance into account.

Based on the potential filtering effects of insert size, adenine addition and expression of toxic inserts, we predicted that compared with an unbiased method the Invitrogen TOPO-TA cloning method would reveal less fungal diversity and might miss entire lineages of fungi. However, our results suggest that the contrasting cloning strategies behave similarly under our conditions. Despite our emphasis on potential biases of TA cloning methods, it is perhaps not surprising that little difference was found. Non-template adenine addition, a requisite of TA cloning, is known to depend on neighbouring sequences, which could bias adenine addition towards or against particular taxa. However, this effect would only be important if it extended beyond the 18–22 base pair (bp) primers used for the PCR, which has not been shown in previous studies (Brownstein *et al.*, 1996). Size bias, though clearly demonstrated and known to vary among cloning vectors, is also unlikely to exert a major bias in the size range of our inserts (1100–2000 bp). Of greatest concern, *a priori*, was the possible filtering effect of ribosomal inserts expressed from the TA vector that might interfere with host *E. coli*. We found no strong evidence for such a bias, as both libraries revealed similar and largely overlapping phylogenetic diversity (Fig. 2). It is possible that our designation of OTUs based on pairwise divergence values contributed error to the comparison of libraries. While a full phylogenetic analysis of all clones would have been a preferable approach to OTU discrimination, the ITS region cannot be aligned across such diverse fungi. However, we did construct trees of all clones using only the nLSU portion of the sequences, and found that the OTUs corresponded very well with clades in the LSU tree. The differences in abundances of a few of the dominant OTUs (Fig. 1) could be due to stochastic events during PCR, ligation and colony picking steps or to a systematic bias such as insert toxicity. Without repeating library construction and sequencing from numerous PCR replicates, these potential sources of variance cannot be quantified. Such a study would be cost-prohibitive if applied to a hyper-diverse community due to the massive sequencing effort required. The extreme diversity and concomitant lack of replication of clone library construction is a weakness of the experimental design used here. However, use of a less diverse community would reduce the phylogenetic breadth and therefore generality of the comparison. Even without replication of clone library construction, our analysis clearly shows that the contrasting cloning methods provide similar pictures of fungal phylogenetic diversity and community composition at our study site.

In conclusion, our study has demonstrated largely congruent pictures of OTU presence/absence, OTU abundance and deeper phylogenetic diversity using two contrasting cloning methods. Our results should be welcome news to microbial ecologists attempting to document diversity using TA cloning methods.

Acknowledgements

We thank Roger Ruess and Jack McFarland for assistance with soil collection and Tom Marr and Shawn Huston for computational help. Supported by NSF EF 0333308 and the Bonanza Creek LTER (Long-Term Ecological Research) program (funded jointly by NSF grant DEB-0423442 and USDA Forest Service, Pacific North-west Research Station Grant PNW01-JV11261952-231).

References

Acinas, S.G., Klepac-Ceraj, V., Hunt, D.E., Pharino, C., Ceraj, I., Distel, D.L., and Polz, M.F. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430:** 551–554.

Brownstein, M.J., Carpten, J.D., and Smith, J.R. (1996) Modulation of non-templated nucleotide addition by taq DNA polymerase: primer modifications that facilitate genotyping. *Biotechniques* **20:** 1004–1006, 1008–1010.

Chao, A., Chazdon, R.L., Colwell, R.K., and Shen, T.J. (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* **8:** 148–159.

Colwell, R.K., and Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philos Trans R Soc Lond B Biol Sci* **345:** 101–118.

Dunfield, K.E., and King, G.M. (2005) Analysis of the distribution and diversity in recent Hawaiian volcanic deposits of a putative carbon monoxide dehydrogenase large subunit gene. *Environ Microbiol* **7:** 1405–1412.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., *et al.* (2005) Diversity of the human intestinal microbial flora. *Science* **308:** 1635–1638.

Geml, J., Laursen, G.A., O'Neill, K., Nusbaum, H.C., and Taylor, D.L. (2006) Beringian origins and cryptic speciation events in the fly agaric (*Amanita muscaria*). *Mol Ecol* **15:** 225–239.

Horton, T.R. (2002) Molecular approaches to ectomycorrhizal diversity studies: variation in ITS at a local scale. *Plant Soil* **244:** 29–39.

Huang, X.Q., and Madan, A. (1999) Cap3: a DNA sequence assembly program. *Genome Res* **9:** 868–877.

Huelsenbeck, J.P., and Ronquist, F. (2001) MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics (Oxford)* **17:** 754–755.

Maddison, W.P., and Maddison, D.R. (1989) Interactive analysis of phylogeny and character evolution using the computer program MacClade. *Folia Primatol* **53:** 190–202.

Maddison, W.P., and Slatkin, M. (1991) Null models for the number of evolutionary steps in a character on a phylogenetic tree. *Evolution* **45:** 1184–1197.

Magnuson, V.L., Ally, D.S., Nylund, S.J., Karanjawala, Z.E., Rayman, J.B., Knapp, J.I., *et al.* (1996) Substrate nucleotide-determined non-templated addition of adenine by Taq DNA polymerase: implications for PCR-based genotyping and cloning. *Biotechniques* **21:** 700–709.

Martin, A.P. (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol* **68:** 3673–3682.

Papineau, D., Walker, J.J., Mojzsis, S.J., and Pace, N.R. (2005) Composition and structure of microbial communities from stromatolites of Hamelin Pool in Shark Bay, Western Australia. *Appl Environ Microbiol* **71:** 4822–4832.

Polz, M.F., and Cavanaugh, C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* **64:** 3724–3730.

Reysenbach, A.L., Giver, L.J., Wickham, G.S., and Pace, N.R. (1992) Differential amplification of ribosomal-RNA genes by polymerase chain-reaction. *Appl Environ Microbiol* **58:** 3417–3418.

Sambrook, J., and Russell, D.W. (2001) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press.

Schadt, C.W., Martin, A.P., Lipson, D.A., and Schmidt, S.K. (2003) Seasonal dynamics of previously unknown fungal lineages in tundra soils. *Science* **301:** 1359–1361.

Stach, J.E.M., Maldonado, L.A., Masson, D.G., Ward, A.C., Goodfellow, M., and Bull, A.T. (2003) Statistical approaches for estimating Actinobacterial diversity in marine sediments. *Appl Environ Microbiol* **69:** 6189–6200.

Stahl, D., Lane, D., Olsen, G., and Pace, N. (1984) Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science* **224:** 409–411.

Suzuki, M.T., and Giovannoni, S.J. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* **62:** 625–630.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997) The CLUSTAL-X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876–4882.

Supplementary material

The following supplementary material is available for this article online:

Fig. S1. Species accumulation curves. The numbers of OTUs detected relative to the number of clones sequenced are statistically equivalent, showing that the two methods detect similar levels of diversity. Error bars represent 95% confidence intervals for numbers of species (Mao Tao) generated by EstimateS 7.0 (Colwell and Coddington, 1994). To assess the phylogenetic representation across the two libraries, sequences were clustered into OTUs using the program Cap3 (Huang and Madan, 1999). Quality scores exported from Aligner were used in Cap3 to clip bases with scores below 10, and to consider only bases with combined qualities above 40 in determining mismatches between sequences. All parameters were set to lenient values favouring assembly of pairs of sequences except the minimum per cent identity in the overlapping region, which was set to 90%, and the maximum overhang per cent length, which was set to 20%. To create species (OTU) accumulation curves for each library, the observed individuals were randomly distributed to create 50 mock samples for each library, with 50 repetitions of this randomization process.

Table S1. Summary of clones used for analyses.

Table S2. Diversity and similarity statistics for the Invitrogen and Lucigen clone libraries.

This material is available as part of the online article from http://www.blackwell-synergy.com