DIVERSITY AND PHYLOGENETIC STRUCTURE OF TWO COMPLEX MARINE MICROBIAL COMMUNITIES

by

Vanja Klepac-Ceraj
B.S. Molecular Biology and Mathematics
Beloit College, 1998

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
and the
WOODS HOLE OCEANOGRAPHIC INSTITUTION
September 2004

Signature of Author:
Joint Program in Biological Oceanography
Massachusetts Institute of Technology and
Woods Hole Oceanographic Institution
August 02, 2004

Certified by:
Martin F. Polz
Associate Professor, Civil and Environmental Engineering
Massachusetts Institute of Technology
Thesis Advisor

Accepted by:
Heidi M. Nepf
Chairman, Department Committee on Graduate Studies
Massachusetts Institute of Technology

Accepted by:
John B. Waterbury
Chair, Joint Committee for Biological Oceanography
Massachusetts Institute of Technology and
Woods Hole Oceanographic Institution

DIVERSITY AND PHYLOGENETIC STRUCTURE OF TWO COMPLEX MARINE MICROBIAL COMMUNITIES

by

Vanja Klepac-Ceraj

Submitted to the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology and the Woods Hole Oceanographic Institution, August 02, 2004 in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the field of Biological Oceanography

ABSTRACT

Molecular surveys have revealed that microbial communities are extraordinarily diverse. Yet, two important questions remain unanswered: how many bacterial types co-exist, and do such types form phylogenetically discrete units of potential ecological relevance? This thesis explores these questions by investigating bacterial diversity in two complex marine communities (coastal bacterioplankton and sediment sulfate-reducing bacteria) by (i) comprehensive analysis of large 16S rRNA clone libraries, and (ii) refinement and application of parametric diversity estimators. Identification and correction of sequence artifacts demonstrated their potentially significant contribution to diversity estimates. Still, hundreds of unique rRNA sequences (ribotypes) were detected in the corrected libraries, and extrapolation to community diversity with commonly used non-parametric diversity estimators suggested at least thousands of co-existing ribotypes in the two communities. However, close inspection revealed that the non-parametric estimators likely lead to underestimation of ribotype diversity in the clone libraries. Thus, an improved parametric method was developed and shown to closely fit the data. The extrapolated total ribotype diversity in the sample by the improved method was up to one order of magnitude higher than estimated with common non-parametric approaches. Most significantly, the compensation for artifacts and improved estimation revealed that the vast majority of ribotypes fall into microdiverse clusters containing <1% sequence divergence. It is proposed that the observed microdiverse clusters form important units of differentiation in microbial communities. They are hypothesized to arise by selective sweeps and contain high diversity because competitive mechanisms are too weak to purge diversity from within them.

Thesis Advisor: Martin F. Polz
Title: Associate Professor of Civil and Environmental Engineering, M.I.T.

ACKNOWLEDGMENTS

I could not have developed into a scientist or a person that I am now as much as I have without the advice, excitement, and intellectual atmosphere provided by my advisor, Martin Polz. His mentorship as a scientist and advisor has clearly been a tremendous inspiration. I can only hope that some of Martin's ability to teach and treat students so kindly, respectfully, and passionately has rubbed off on me, and that if I ever take upon an advisorship role that I have the ability to as successfully guide students as he has guided me.

I am indebted to my committee: Penny Chisholm, Katrina Edwards, Andreas Teske, and Daniele Veneziano. Their critical thinking, encouragement, and suggestions greatly contributed to the final form of this thesis. Special thanks to Daniele for being patient, kind and having so much energy to introduce me to the world of statistics. Many thanks to Harry Hemond and Dan Distel for helpful discussions and suggestions throughout the past five years. Also, it was my great pleasure to collaborate with John Hobbie, Byron Crump, and Michele Bahr on a salt-marsh project and I gratefully acknowledge their support, friendship and guidance.

I feel extremely lucky to have spent my tenure in the Polz lab with so many dear and amazing friends: Ramahi Sarma-Rupavtarm, Janelle Thompson, Dana Hunt, Luisa Marcelino, and Silvia Acinas. They have edited drafts of my papers, listened to numerous practice talks, been my collaborators and excellent work-out buddies, helped me develop ideas and been always ready to party. Thanks to present and past Polz and Chisholm lab members who made the lab so enjoyable: Sarah Pacocha, Yvette Luyten, Stefan Bertilsson, Eelin Lim, Aoy Tomita-Mitchell, Chanathip Pharino, Linping Kuai, Zackary Johnson, Erik Zinser, Maureen Coleman, and many others.

I have to thank my officemates and Parsonites Matt Sullivan, Holly Michael, and Frédéric Chagnon not only for their help, wit, and generosity, but also for making the process of thesis writing so much more fun and intellectually stimulating with intermittent coffee and dinner breaks. My WHOI companions and my very dear friends Bridget Bergquist, Lara Gulmann, Elle Lilly, Ana Lima, Susan Parks and Oscar Pizzaro have always given me so much support and encouragement in everything I did. Thanks to Joshua Weitz for inspiring conversations on various niches of science.

Sheila Frankel, Vicky Murphy and Sheila Anderson not only were able to cheer me up any time I saw them in the hallway, but also managed to solve odd problems that I seemed to come up with on a rather frequent basis. Ronni Schwartz, Julia Westwater, Marsha Gomes, and Judy McDowell solved so many problems often before I even had a chance to realize that I had them, and they simply made the Joint Program a wonderful experience.

My sister Petra has been my 911 Matlab help, read many of my drafts and made my life so happy here. My immediate and extended family, and host families have been incredibly supportive, helpful and encouraging these last six years. Thanks to them for never doubting that I could do it.

Finally, thanks to Ivan for who he is, for everything above and so much more.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF TABLES

CHAPTER 2

CHAPTER 4

CHAPTER ONE

Introduction

INTRODUCTION

We are only beginning to understand the extent of microbial diversity and principles controlling the distribution and abundance of microorganisms in natural environments. Although cultivation-independent techniques have revealed that microbial communities are extraordinarily diverse (Pace, 1997; Hugenholz et al., 1998), the number of co-existing bacterial types in a natural microbial community and whether these are organized into some natural units of differentiation remain unknown.

Most microbes defy cultivation by standard methods. Therefore, the only reliable way to estimate microbial diversity is by using sequence data from genes obtained directly from environmental samples. The rRNA gene has been adopted as the most commonly used tool to determine evolutionary relationships and diversity (Woese, 1987; Pace, 1997). Most commonly, this is done by the extraction of environmental DNA, polymerase chain reaction (PCR) amplification of target genes, clone library construction from amplified gene fragments, and gene sequencing (Head et al., 1998).

The development of molecular methods has provided a powerful tool for the study of microbial diversity. In 1998, GenBank contained only approximately 8,500 16S rRNA sequences, a majority of which belonged to cultured prokaryotes (Rappe and Giovannoni, 2003). Within a period of six years, this number has grown to over 183,000 entries. Most of the recently added sequences are from environmental clones derived from sediments and aquatic environments. It has been observed that a high number of these are closely related, but not identical to the sequences already deposited in GenBank, suggesting that these closely related sequences may form discrete groups (Rappe and Giovannoni, 2003)

Microdiverse sequences have been observed in rRNA genes retrieved from various environments (Field et al., 1997; Garcia-Martinez and Rodriguez-Valera, 2000;

Casamayor et al., 2002). Although the nucleotide divergence observed in these cloned sequences could be explained by evolution, it remains unclear how much of this variation stems from experimental errors and small-scale variation in sequences among rRNA operons. Since the contribution of experimental errors to diversity is difficult to establish, most studies have clustered sequences into 95-99% sequence similarity groups (Hughes et al., 2001; Martin, 2002; Torsvik et al., 2002). However, in order to elucidate relationships and diversity at finer scales of sequence divergence, methods that minimize and account for contribution of sequence artifacts need to be developed. The development of such methods is addressed in this thesis.

Microbial diversity estimates

Statistical methods hold promise for describing the diversity of a given environment, wherein the observed sequence diversity is extrapolated to that of a sampled clone library or environment (Moyer et al., 1998; Dunbar et al., 1999; Hughes et al., 2001). Typically, the diversity of microorganisms is assessed using sequence data from genes obtained directly from the environment. The diversity of ribotypes or clusters in a clone library is then estimated using statistical approaches (Dunbar et al., 1999; Hughes et al., 2001; Stach et al., 2003). These methods were only recently introduced to microbial ecology (Hughes et al., 2001) and the critical evaluation of their accuracy has remained a challenge because molecular surveys have not produced large enough data sets.

Most of the statistical methods were developed for estimating macroorganismal diversity and these fall into two general categories: (i) parametric methods, where the abundance distribution of taxa is assumed to have a specified parametric form, and (ii) non-parametric methods, where no abundance distribution model is assumed (May, 1975; Colwell and Coddington, 1994; Krebs, 1999). In microbial ecology, the most commonly applied richness estimator is the non-parametric Chao1 estimator (Chao, 1984, 1987; Hughes et al., 2001; Bohannan and Hughes, 2003). It has been noted that the Chao1 estimator gives biased (low) estimates of diversity, especially for very heterogeneous communities (Mao and Lindsay, 2001; Bohannan and Hughes, 2003). Thus, other approaches such as parametric analyses, although computationally more demanding, should be considered in microbial diversity analyses. Especially for large data sets, parametric methods, rather than non-parametric alternatives, would be expected to more accurately estimate the diversity of highly diverse communities. The application of both parametric and non-parametric methods to large microbial data sets is evaluated in this thesis.

Goals of this thesis

This thesis addresses two important questions: (i) How many distinct types of bacteria co-exist in a microbial community, and (ii) do these types form discrete phylogenetic clusters of potential ecological significance? These questions were explored by comprehensively sampling two large clone libraries obtained from two complex marine communities (coastal bacterioplankton and sediment sulfate-reducing bacteria) by (i) removal of sequence artifacts generated by library construction techniques that may confound accurate diversity estimation, (ii) detailed phylogenetic analysis of large 16S rRNA clone libraries, and (iii) refinement and application of parametric diversity estimators.

One objective of this thesis was to explore the genetic diversity and fine-scale phylogenetic structure of two complex microbial communities. We chose to sample communities from two environments displaying extremes in structural difference: mixed coastal ocean and highly structured salt-marsh sediment (Chapter 2 and 3). Therefore, one may expect that the underlying community composition of the two communities would be different. For example, it may be hypothesized that the efficient mixing in the pelagic environment may allow for more efficient selective sweeps within the community. These would serve to purge diversity leading to a more simple overall community composition. Since sediment communities are highly diverse (Ravenschlag et al., 1998; Whitman et al., 1998) we focused on investigating one metabolic guild – sulfate-reducing bacteria (SRB). However, for the coastal ocean sample we investigated the entire bacterioplankton community.

Large libraries based on amplified 16S rRNA gene fragments were constructed from both communities, and were corrected for chimeric sequences and amplification errors to allow phylogenetic interpretation at all levels of sequence divergence. The corrected set of sequences was used to estimate total diversity and patterns of relationships by grouping sequences into similarity clusters (100%, 99%, 98% etc).

Large clone libraries were used to critically evaluate the existing statistical approaches in microbial ecology and to develop new ones. We observed that the most commonly applied diversity estimator, Chao1, significantly underestimated diversity of the complex bacterioplankton library. This was established by evaluating the Chao1 using simulated communities whose species abundances were based on the dataset presented in Chapter 3. Thus, we modified parametric estimators to better account for the way microbial communities are sampled. The modified parametric estimator was applied to the bacterioplankton library and simulated communities, and compared to the commonly used diversity estimator Chao1 (Chapter 4). The modified parametric approach may ultimately provide more reliable estimates of microbial diversity.

References

1. Bohannan, B.J.M., and Hughes, J. (2003) New approaches to analyzing microbial biodiversity data. *Current Opinion in Microbiology* **6**: 282-287.

2. Casamayor, E.O., Pedros-Alio, C., Muyzer, G., and Amann, R. (2002) Microheterogeneity in 16S ribosomal DNA-defined bacterial populations from a stratified planktonic environment is related to temporal changes and to ecological adaptations. *Applied and Environmental Microbiology* **68**: 1706-1714.

3. Chao, A. (1984) Nonparametric-Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics* **11**: 265-270.

4. Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**: 783-791.

5. Colwell, R.K., and Coddington, J.A. (1994) Estimating Terrestrial Biodiversity through Extrapolation. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **345**: 101-118.

6. Dunbar, J., Takala, S., Barns, S.M., Davis, J.A., and Kuske, C.R. (1999) Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. *Applied and Environmental Microbiology* **65**: 1662-1669.

7. Field, K.G., Gordon, D., Wright, T., Rappe, M., Urbach, E., Vergin, K., and Giovannoni, S.J. (1997) Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Appl. Environ. Microbiol.* **63**: 63-70.

8. Garcia-Martinez, J., and Rodriguez-Valera, F. (2000) Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine Archaea of Group I. *Molecular Ecology* **9**: 935-948.

9. Head, I.M., Saunders, J.R., and Pickup, R.W. (1998) Microbial evolution, diversity, and ecology: A decade of ribosomal RNA analysis of uncultivated microorganisms. *Microbial Ecology* **35**: 1-21.

10. Hugenholz, P., Goebel, B.M., and Pace, N.R. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**: 4765-4774.

11. Hughes, J.B., Hellmann, J.J., Ricketts, T.H., and Bohannan, B.J.M. (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* **67**: 4399-4406.

12. Krebs, C.J. (1999) *Ecological methodology.* Menlo Park, CA: Benjamin/Cummings.

13. Mao, C.M., and Lindsay, B.G. (2001) Moment-based nonparametric estimators for the number of classes in a population. In. University Park: The Pennsylvania State University, pp. 1-44.

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

15. May, R.M. (1975) Patterns of species abundance and diversity. In *Ecology and evolution of communities.* Cody, M.L., and Diamond, J.M. (eds). Cambridge, Massachusetts, and London, England: The Belknap Press of Harvard University Press, pp. 81-120.

16. Moyer, C.L., Tiedje, J.M., Dobbs, F.C., and Karl, D.M. (1998) Diversity of deep-sea hydrothermal vent Archaea from Loihi seamount, Hawaii. *Deep-Sea Research Part Ii-Topical Studies in Oceanography* **45**: 303-317.

17. Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science* **276**: 734-740.

18. Rappe, M.S., and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annual Review of Microbiology* **57**: 369-394.

19. Ravenschlag, K., Sahm, K., Pernthaler, J., and Amann, R. (1998) High bacterial diversity in permanetnly cold marine sediments. *Appl. Environ. Microbiol.* **65**: 3982-3989.

20. 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. *Applied and Environmental Microbiology* **69**: 6189-6200.

21. Torsvik, V., Øvreås, L., and Thingstad, T.F. (2002) Prokaryotic diversity -- magnitude, dynamics, and controlling factors. *Science* **296**: 1064-1066.

22. Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998) Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 6578-6583.

23. Woese, C.R. (1987) Bacterial evolution. *Microb. Rev.* **51**: 221-271.

CHAPTER TWO

High overall diversity and dominance of microdiverse relationships in salt marsh sulphate-reducing bacteria

Vanja Klepac-Ceraj, Michele Bahr, Byron C. Crump, Andreas P. Teske, John E. Hobbie and Martin F. Polz

2004 Klepac-Ceraj, V., M. Bahr, B. C. Crump, A. P. Teske, J. E. Hobbie & M. F. Polz. High Overall Diversity and Dominance of Microdiverse Relationships in Salt Marsh Sulphate-Reducing Bacteria. *Env. Microbiol.* 6:686-698.

Environmental Microbiology (2004) 6(7), 686–698 doi:10.1111/j.1462-2920.2004.00600.x

High overall diversity and dominance of microdiverse relationships in salt marsh sulphate-reducing bacteria

Vanja Klepac-Ceraj,[1] Michele Bahr,[2] Byron C. Crump,[2†] Andreas P. Teske,[3‡] John E. Hobbie[2] and Martin F. Polz[1*]

[1] *Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Bldg 48-421, 77 Massachusetts Ave., Cambridge, MA 02139, USA.*
[2] *The Ecosystems Center, Marine Biological Laboratories, and* [3] Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA.

Summary

The biogeochemistry of North Atlantic salt marshes is characterized by the interplay between the marsh grass *Spartina* and sulphate-reducing bacteria (SRB), which mineralize the diverse carbon substrates provided by the plants. It was hypothesized that SRB populations display high diversity within the sediment as a result of the rich spatial and chemical structuring provided by *Spartina* roots. A 2000-member 16S rRNA gene library, prepared with delta-proteobacterial SRB-selective primers, was analysed for diversity patterns and phylogenetic relationships. Sequence clustering detected 348 16S rRNA sequence types (ribotypes) related to delta-proteobacterial SRB, and it was estimated that a total of 623 ribotypes were present in the library. Similarity clustering showed that ª 46% of these sequences fell into groups with <1% divergence; thus, microheterogeneity accounts for a large portion of the observable genetic diversity. Phylogenetic comparison revealed that sequences most frequently recovered were associated with the Desulfobacteriaceae and Desulfobulbaceae families. Sequences from the Desulfovibrionaceae family were also observed, but were infrequent. Over 80% of the delta-proteobacterial ribotypes clustered with cultured representatives of *Desulfosarcina*, *Desulfococcus* and *Desulfobacterium* genera, suggesting that complete oxidizers with high substrate versatility dominate. The large-scale approach demonstrates the co-existence of numerous SRB-like sequences and reveals an unexpected amount of microdiversity.

Received 28 August, 2003; revised 5 January, 2004; accepted 9 January, 2004. *For correspondence. E-mail mpolz@mit.edu; Tel. (+1) 617 253 7128; Fax (+1) 617 258 8850. Present addresses: †Horn Point Laboratory, University of Maryland, Cambridge, MD, USA. ‡Department of Marine Sciences, University of North Carolina, Chapel Hill, NC, USA.

Introduction

Salt marshes are among the most productive environments, with primary production rates ranging from 460 to 3700 g cm^{-2} $year^{-1}$ (Gallagher *et al.*, 1980; Wiegert and Pomeroy, 1981). Formation and persistence of marshes is determined by the growth of marsh grasses because their dense stands can trap and stabilize sediment in the face of the erosive power of tides and waves. Along the Atlantic coast of the United States, marshes are dominated by the smooth cord grasses *Spartina alterniflora* and *S. patens*, which permeate the sediment with a complex rhizome system and reach high production and turnover rates. Some of the plant-derived organic matter is exported to coastal waters (Teal, 1962; Howes and Goehringer, 1994), but a large portion remains within marsh sediments where it is decomposed by fermentation and anaerobic respiration. Sulphate reduction, mediated by sulphate-reducing bacteria (SRB), is typically the prevailing carbon mineralization process in marine anoxic sediments and exceeds respiration using other electron acceptors, including oxygen, nitrate and metal oxides (Jørgensen, 1982; Canfield *et al.*, 1993). Although in some marsh sediments, iron(III) has been suggested to be important to respiration (Canfield and Des Marais, 1993; Joye *et al.*, 1996; Lowe *et al.*, 2000), sulphate reduction usually accounts for 67–80% of all respiration processes (Howarth and Teal, 1979; Howarth and Giblin, 1983; Howarth and Hobbie, 1985).

Marsh grasses and SRB appear to maintain a complex relationship, which displays elements of both antagonism and dependence. On the one hand, sulphide, the end-product of sulphate reduction, may have negative effects on the plant because of its toxicity, and oxygen leaking into the sediment from the aerenchyma may inhibit sulphate reduction. On the other hand, sulphate reduction rates can be tightly correlated to marsh grass production. For example, a fivefold increase in sulphate reduction rates has been observed during the above-ground elongation of the tall form of *S. alterniflora* in a New England marsh. Between early June and August, *Spartina* roots grow rapidly and leak large amounts of exudates, which serve as substrates for SRB growth (Hines *et al.*, 1989; 1999). This is consistent with the observation that sulphate-reducing activity follows seasonal patterns of vegetational changes (Currin *et al.*, 1995) and that rRNA abundance of the SRB genera *Desulfonema*, *Desulfococcus* and *Desulfosarcina* peaks during the early summer root growth of *Spartina* (Rooney-Varga *et al.*, 1998; Hines *et al.*, 1999). Thus, it is likely that the SRB community is tightly coupled to marsh grass activity and that seasonal and spatial differentiation of root activity has a strong influence on the diversity of niche spaces available to SRB.

Spartina plants provide a large variety of potential carbon sources to the SRB community, yet the extent to which these different compounds are used by diverse types of SRB is only beginning to be understood. Roots directly exude simple fatty acids and alcohols, such as malate, ethanol (Mendelssohn *et al.*, 1981) and acetate (Hines *et al.*, 1994). These may be important substrates for SRB, which may take them up directly from the plant as suggested by an increase in SRB populations associated with the rhizosphere during the growth season (Hines *et al.*, 1999). However, the quantitative importance of plant exudates within the total sediment community remains unknown, and acetate, which is regarded as one of the central metabolites in anaerobic communities, was found to support only about 10% of sulphate reduction in marsh sediments (Howarth, 1993). This points to the importance of other plant-derived substrates released during the decay of *Spartina* litter and includes complex carbohydrates (Opsahl and Benner, 1999), phenolics and humic acids (Wilson *et al.*, 1986). Recent isolation of SRB capable of degrading diverse and previously unsuspected compounds has suggested that they are metabolized by SRB in the environment. For example, SRB capable of utilization of long-chain fatty acids and alcohols, glycolate (Friedrich and Schink, 1995), hydrocarbons (Aeckersberg *et al.*, 1991) and aromatic compounds (Beller and Spormann, 1997; Phelps *et al.*, 1998; Galushko *et al.*, 1999; Harms *et al.*, 1999) have been described. High metabolic versatility is found particularly in the genera *Desulfosarcina*, *Desulfococcus* and *Desulfobacterium* (Widdel and Bak, 1992), and it is thus hypothesized that these play an important role in the marsh.

Among the major phylogenetic groups of SRB, delta-proteobacterial SRB have been shown to be important in salt marsh sediments by both culture-dependent and independent studies. For example, using rRNA-targeted, quantitative oligonucleotide hybridization, *Desulfovibrio*, Desulfobacteriaceae and *Desulfobulbus* accounted for ª 30% of Bacteria rRNA, and Desulfobacteriaceae alone accounted for ª 20%, probably making it the dominant group in the marsh sediment (Devereux *et al.*, 1996). The metabolic, physiological and phylogenetic diversity of delta-proteobacterial SRB has been studied extensively, and it appears that a number of metabolic properties are confined to specific phylogenetic groups. Indeed, the traditional classification of SRB into complete and incomplete oxidizers has largely been confirmed by rRNA-based phylogeny. Complete oxidizers, capable of acetate mineralization, are mainly represented by the genera *Desulfobacter*, *Desulfobacterium*, *Desulfosarcina* and *Desulfococcus*, whereas incomplete oxidizers, which oxidize carbon substrates to acetate, are mainly represented by *Desulfovibrio* and *Desulfobulbus*. The links between physiology and phylogeny of the delta-proteobacterial SRB enable some prediction of community properties, based on quantitative 16S rRNA hybridization and some sequence surveys (Devereux and Stahl, 1993; Loy *et al.*, 2002).

Here, we investigated the 16S rRNA diversity of the delta-proteobacterial SRB community associated with the sediments of a New England *S. alterniflora* salt marsh (Fig. 1). We hypothesize that the plant rhizomes structure the environment into numerous microniches, which are reflected in high overall diversity of the SRB community. Furthermore, we explored the question of whether the high substrate diversity created by the plant is reflected in the presence of phylogenetic groups of metabolically differentiated SRB. For example, it may be expected that complete oxidizers dominate the SRB community, because of their broad substrate spectra. Our approach is based on a large-scale survey of a 16S rRNA gene library generated using a polymerase chain reaction (PCR) primer set specific for delta-proteobacterial sequences. The library was constructed from monthly

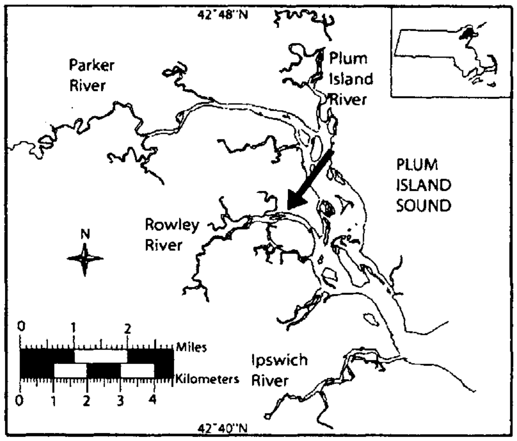


Fig. 1. Location of the sampling site (arrow) in the Plum Island salt marsh, MA, USA.

samples collected over an entire growth cycle of the marsh grass *S. alterniflora* and was analysed by diversity estimators and phylogenetic methods.

Results

Clone library analysis

Approximately 47% of the ≈1650 positive 16S rRNA gene sequences obtained from the 2000-member clone library were associated with the delta-proteobacterial subclass. Non-delta-proteobacterial sequences related to epsilon-Proteobacteria, Firmicutes and *Cytophaga/Flexibacter* were also amplified because of the broad specificity of primer 385F, which was designed to cover all known SRB groups within delta-Proteobacteria. A large majority of the 774 delta-proteobacterial sequences clearly fell within previously identified SRB families: Desulfobacteriaceae, Desulfobulbaceae and Desulfovibrionaceae. The only exceptions were eight sequences for which the association with other delta-proteobacterial groups *Bdellovibrio*, *Syntrophobacter* and *Geobacter/Pelobacter* was ambiguous, i.e. the sequences were less than 85% similar to these groups. These sequences were included in further analysis of delta-proteobacterial SRB relationships because of their close relationship to the SRB. Furthermore, the *Geobacter/Pelobacter* group harbours representatives capable of sulphur reduction and are thus ecologically related to SRB (Lonergan *et al.*, 1996). A second library skewed towards Gram-positive bacteria was constructed; however, despite sequencing of 1000 clones, no sequence related to SRBs was detected (data not shown).

The delta-proteobacterial sequences were subjected to a detailed evaluation of *Taq* errors and chimera formation. A large fraction of the sequences displayed a very high similarity to each other, a phenomenon potentially caused by base misincorporation during amplification. It was thus decided to conduct a detailed estimation of the potential contribution of *Taq* error to sequence diversity. This was done by fitting nucleotide positions of amplified sequences to 16S rRNA secondary structure models (Cannone *et al.*, 2002). We considered that there was a base misincorporation if nucleotide changes occurred in positions that (i) are >98% conserved in the entire bacterial 16S rRNA secondary structure data set; and (ii) lead to non-canonical basepairing in stem structures and were absent in closely related sequences. The first of these rules is expected to lead to a slight overestimation of *Taq* error while the second is likely to result in underestimation. However, the determined *Taq* error rate agreed remarkably well with theoretically predicted rates based on reported values of *Taq* misincorporation rates of 2×10^{-5} nucleotides per cycle (Tindall and Kunkel, 1988). The rate based on the above rationale was 1.7×10^{-5} and 2.5×10^{-5} nucleotides per cycle for regions analysed with rules (i) and (ii) respectively. The data were also checked for putative chimeras by the RDP CHIMERA_CHECK (Maidak *et al.*, 1999) and the CHIMERABUSTER algorithm, which was newly developed specifically to analyse clone libraries with high coverage. This identified 32 sequences deemed likely chimeras by one of the chimera identification methods. Based on these analyses, an additional, corrected data set of delta-proteobacterial sequences was created in which putative *Taq* errors were corrected and from which putative chimeras were removed.

Overall features of the sequences

Both the original and the corrected data set indicated that very high numbers of delta-proteobacterial SRB sequences co-existed in the marsh sediment samples. In the corrected data set, a total of 348 ribotypes (groups of identical sequences) were identified. This is ≈ 23% lower than in the uncorrected data set primarily because of the removal of putative *Taq* errors and chimeras. Rarefaction analysis and the Chao-1 non-parametric diversity estimator were applied to both data sets to estimate how completely the library had been sampled and to extrapolate to total sequence diversity (Hughes *et al.*, 2001). Rarefaction, which plots the number of clones screened versus the number of ribotypes detected, showed that neither data set reached an asymptote, indicating that the diversity in the clone library is even higher (Fig. 2). This was confirmed by the Chao-1 estimator, which computed 623

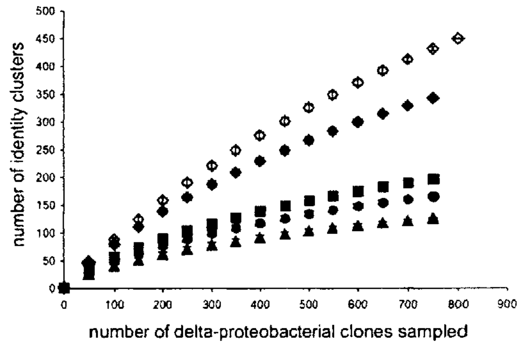


Fig. 2. Rarefaction analysis of similarity groups within delta-proteobacterial SRB sequences with different sequence identity cut-off. Curves were calculated using the algorithm described by Hurlbert (1971) and are plotted as the number of identity clusters versus the number of clones. Identity clusters were identified for uncorrected 100% (open diamonds) and corrected 100% (filled diamonds) 99% (filled squares), 98% (filled circles) and 97% (filled triangles) nucleotide identity; bars represent standard deviation of the statistical resampling process.

ribotypes for the total sequence diversity in the clone library.

Further analysis suggested that the initial observation of close relationships among large numbers of sequences is preserved even after correction of putative *Taq* errors, but that deep phylogenetic lineages were well sampled. Rarefaction analysis of the corrected data set, using 100% and 99% sequence identity to define taxonomic units, indicated that ª 46% of the sequences fell into clusters in which members differed by <1% nucleotide difference (Fig. 2). At 99% identity, only 200 sequence groups were detected, and Chao-1 yielded an estimate of 332 groups. Decreasing the sequence identity cut-off to 98% and 97% produced 168 and 127 sequence types respectively. The total sequence diversity based on Chao-1 was 261 for the 98% group and 191 for the 97% identity groups.

Phylogeny of delta-proteobacterial SRB-like sequences

The delta-proteobacterial SRB-like community clustered into three large and several smaller clades based on the distance and parsimony analysis using one representative sequence of each 98% similarity group (Fig. 3). Over 80% were associated with the family Desulfobacteriaceae (*Desulfosarcina*, *Desulfobacterium*, *Desulfococcus* and *Desulfonema*) (Fig. 3), suggesting that complete oxidizers with high substrate versatility dominate. Clades I, II and V fell into the Desulfobacteriaceae (Fig. 3). Clade I, described in detail below, was by far the largest containing ª 55% of the total sequences (424). Clade II, although phylogenetically diverse, comprised only about 15% of sequences (113) and represented the third largest cluster. Closely related sequences in clade II were environmental clones recovered from benzene-degrading enrichments (Phelps *et al.*, 1998), hydrocarbon seeps (AF154102) and wetlands (AY216442). The only cultured representative was *Desulfobacterium anilinii* (Fig. 3). Clade III represented 17% of the sequences and fell into the incomplete-oxidizing Desulfobulbaceae. Cultured relatives were members of the genera *Desulforhopalus*, *Desulfofustis*, *Desulfocapsa* and *Desulfotalea*, but none closely matched the sequences recovered from the marsh sediment (Fig. 3). For example, one of the two largest 98% consensus groups of sequences (17 members) is only 93% similar to its closest cultured relative *Desulfocapsa* sp. La4.1 (AF228119). Members of the *Desulfovibrio* genus are assigned to clade IV and are only represented by 10 sequences. Eight *Desulfovibrio* sequences were 99% similar to *Desulfovibrio* BG-6, isolated from salt marsh sediments in New Hampshire (Rooney-Varga *et al.*, 1998). Finally, at least two small, deep branching clades (Fig. 3, clades V and VI), containing 10 and 19 sequences, represent novel lineages with no published sequence matching with >90% similarity.

Clade I contained the three largest monophyletic subclades, which all had >96% sequence identity. Two are shown in detail in Fig. 4 to illustrate relationships among closely related sequences. The largest subclade, IA (125 clones), fell within the Desulfosarcinales and displayed >96% similarity to two *Desulfosarcina variabilis* strains and to a *Desulfobacterium cetonicum* strain (Fig. 4A). Some sequences had <1% nucleotide difference from clones recovered from geographically disparate environments. Most notably, the most abundant ribotype with 42 sequences (ª 6% of total delta-proteobacterial clones) was only 2 bases different from clone SB 4.53 obtained from Antarctic shallow-marine sediments (Purdy *et al.*, 2003) (Fig. 4A). The second subclade, IB (Fig. 4B), contained 71 sequences closely related to cloned sequences from permanently cold marine sediments (Sva0081) (Ravenschlag *et al.*, 1998) and an oligochaete endosymbiont (*Olavius algarvensis* sulphate-reducing endosymbiont) (Dubilier *et al.*, 2001). The third subclade (IC; not shown in detail) contained 57 sequences closely related to uncultured SRB (95% similar to the clone Eel-3G12) from anoxic methane-oxidizing consortia recovered from continental shelf sediments (Orphan *et al.*, 2001). Subclades IB and IC had *Desulfobacterium indolicum* and *Desulfonema magnum* as the only distantly related cultured relatives (< 93% similar) respectively. A number of sequences that fell into clade I but were not associated with any of the large subclades were closely related to strains isolated on crude oil extracts and aromatic hydrocarbons, in particular strain NaphS2 (Galushko *et al.*, 1999). These sequences also matched clones SB10 and SB29 recovered from a benzene-mineralizing consortium (Phelps *et al.*, 1998) (Fig. 4).

Discussion

Large-scale analysis of delta-proteobacterial 16S rRNA genes revealed surprising structural features of the microbial community within the marsh sediments. Although it was hypothesized that the marsh grass *Spartina* subdivides the sediment into a large number of microenvironments, the extremely high diversity of co-existing SRB-like sequences was unexpected. Our, to date unprecedented, sampling effort of this phylogenetically defined group comprised 1650 sequences, of which ª 47% were identified as delta-proteobacterial SRB-like sequences. Nonetheless, only ª 55% of the total delta-proteobacterial SRB-like sequence diversity was captured, comprising 623 ribotypes based on the Chao-1 diversity estimator. Most of this diversity resulted from almost 50% of the ribotypes displaying <1% nucleotide difference from each other. Deeper lineages were well sampled, and additional sequencing effort would not have yielded a significant number of new types. Comparison with published

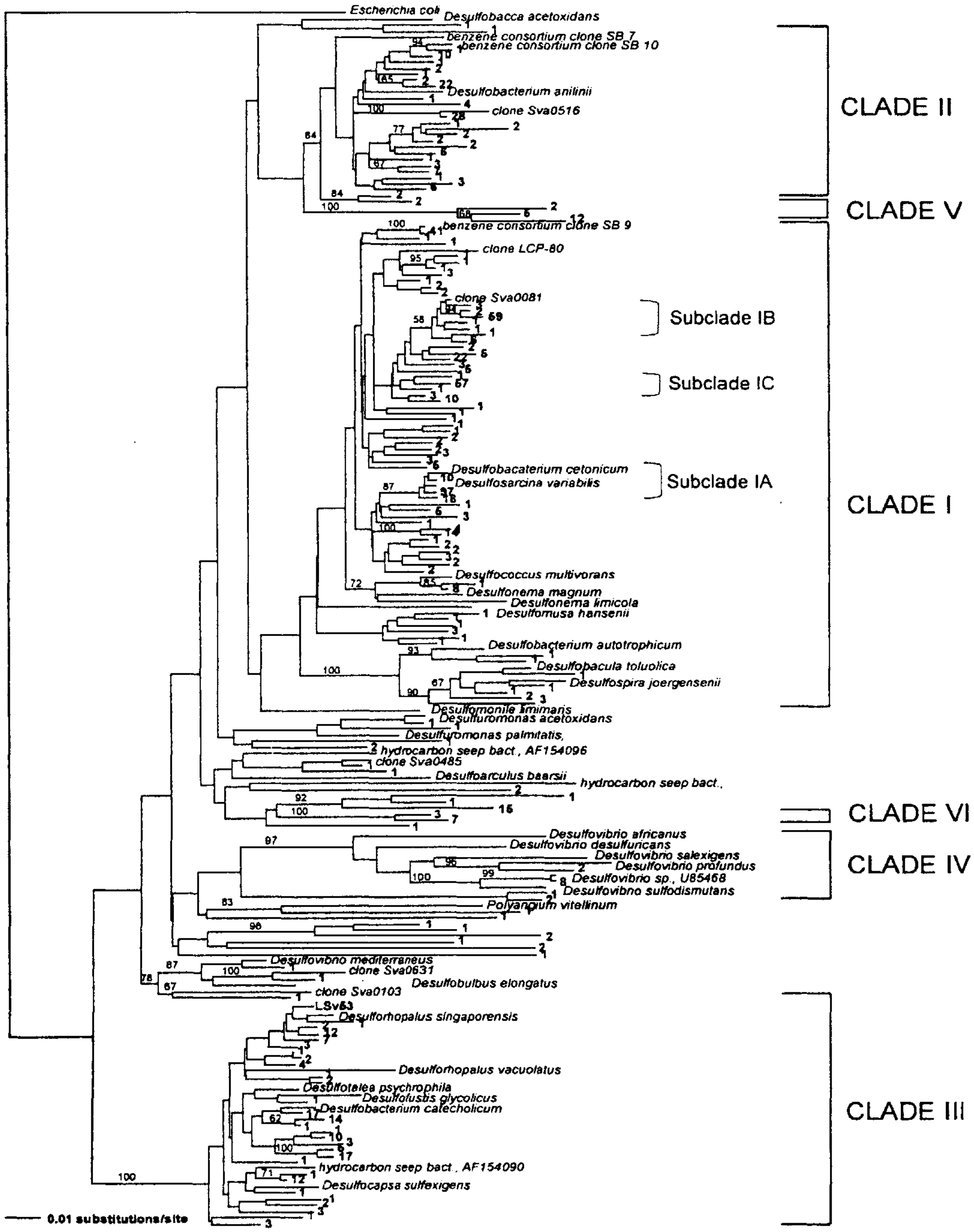

Escherichia coli
Desulfobacca acetoxidans
benzene consortium clone SB 7
benzene consortium clone SB 10
Desulfobacterium anilinii
clone Sva0516
benzene consortium clone SB 9
clone LCP-80
clone Sva0081
Desulfobacaterium cetonicum
Desulfosarcina variabilis
Desulfococcus multivorans
Desulfonema magnum
Desulfonema limicola
Desulfomusa hansenii
Desulfobacterium autotrophicum
Desulfobacula toluolica
Desulfospira joergensenii
Desulfomonile limimaris
Desulfuromonas acetoxidans
Desulfuromonas palmitatis,
hydrocarbon seep bact., AF154096
clone Sva0485
Desulfoarculus baarsii
hydrocarbon seep bact.,
Desulfovibrio africanus
Desulfovibrio desulfuricans
Desulfovibrio salexigens
Desulfovibrio profundus
Desulfovibrio sp., U85468
Desulfovibrio sulfodismutans
Polyangium vitellinum
Desulfovibrio mediterraneus
clone Sva0631
Desulfobulbus elongatus
clone Sva0103
LSv53
Desulforhopalus singaporensis
Desulforhopalus vacuolatus
Desulfotalea psychrophila
Desulfofustis glycolicus
Desulfobacterium catecholicum
hydrocarbon seep bact., AF154090
Desulfocapsa sulfexigens
0.01 substitutions/site
CLADE II
CLADE V
Subclade IB
Subclade IC
Subclade IA
CLADE I
CLADE VI
CLADE IV
CLADE III

Fig. 3. Phylogenetic relationships based on partial 16S rDNA sequences of delta-proteobacterial clones from Plum Island marsh sediment. Each sequence represents a 98% identity cluster, and numbers at terminal nodes represent how many clones fell into the 98% cluster. Tree construction was by neighbour joining using the Jukes–Cantor correction; bootstrap values are based on 100 replicates and are shown for branches with >50% support.

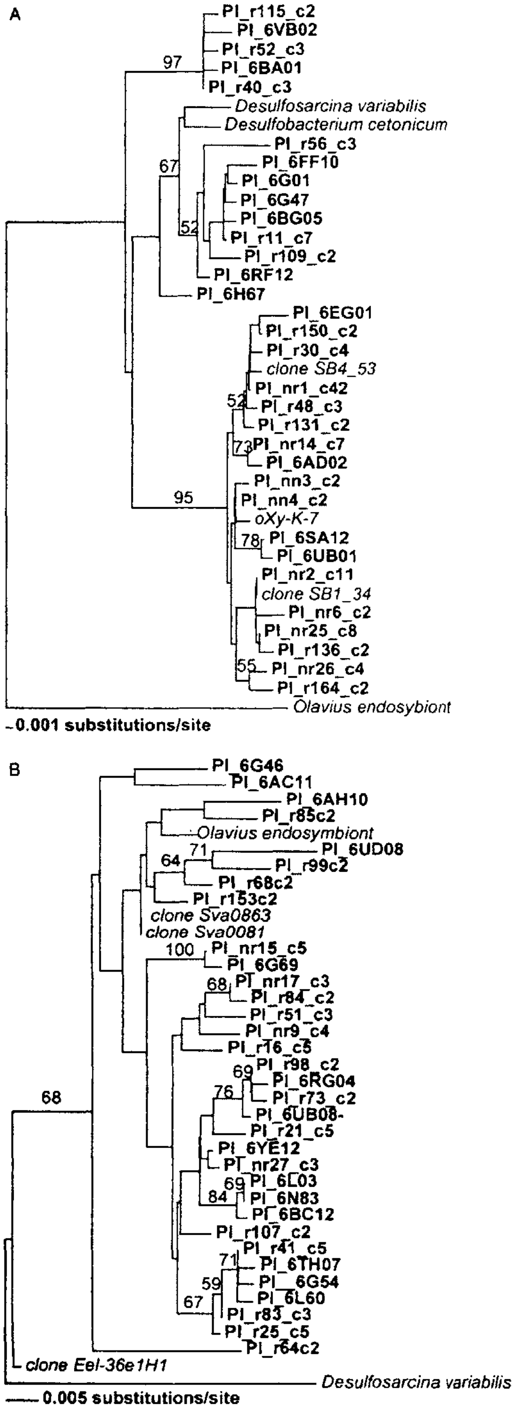


Fig. 4. Phylogenetic relationships among two dominant subclades. Identical sequences were removed from the analysis, and their number is shown at the terminal nodes. Trees were constructed by neighbour joining using the Jukes–Cantor correction; bootstrap values are based on 100 replicates and are shown for branches with >50% support. Subclade IA (A); subclade IB (B).

sequences suggests that deep phylogenetic relationships of delta-proteobacterial SRB are beginning to be well sampled as no novel lineages with <90% identity to known SRB were detected. Approximately 80% of the sequences were associated with lineages of metabolically versatile Desulfobacteriaceae. Thus, the data suggest that diverse carbon substrates produced by marsh plants, or released during plant decay, have a strong effect on structuring the community, but the data also pose questions about the ecological significance of the observed microdiversity.

The surprising observation in the initial data set that over 50% of the sequences fell into 1% consensus groups was confirmed overall by detailed evaluation of potential sources of error. The amplification protocol was designed to minimize PCR artifacts that may cause small-scale sequence divergence. The genes were amplified to minimize PCR bias (Polz and Cavanaugh, 1998) and errors (Thompson *et al.*, 2002). Chimeras, analysed by two methods CHIMERA_CHECK and CHIMERABUSTER, were determined to be relatively insignificant. However, 1% divergent sequences remain nearly impossible to evaluate. Nonetheless, the 32 chimeras identified among sequences with a 1% similarity cut-off represent such a small number that, even if one allows for a significant increase in chimera formation among near-identical sequences, the effect on overall diversity estimates would be small. The evaluation of *Taq* errors suggested that only ≈ 23% of the initially observed diversity resulted from *Taq* error. This indicates that, indeed, a large number of microdiverse delta-proteobacterial sequences co-exist in the marsh sediments.

Sequence microdiversity, described previously in clone libraries (Ferris and Ward, 1997; Field *et al.*, 1997; Amann, 2000; Garcia-Martinez and Rodriguez-Valera, 2000; Casamayor *et al.*, 2002; Ferris *et al.*, 2003), has been ignored in recent estimates of microbial diversity because of the assumption that, in addition to *Taq* error, small-scale variation in sequences among rRNA operons is responsible for this pattern (Hughes *et al.*, 2001; Curtis *et al.*, 2002; Torsvik *et al.*, 2002). Indeed, bacteria can harbour up to 15 rRNA operons (Rainey *et al.*, 1996), and 16S rRNA sequences commonly differ among operons, but differences are typically <1% (Klappenbach *et al.*, 2001). Although interoperon variation may be responsible for a considerable fraction of microdiversity in clone libraries, several lines of evidence suggest that microdiversity is a feature of co-existing bacterial strains. We recently conducted a detailed examination of 16S rRNA divergence within 78 published whole bacterial genomes with multiple operons (Acinas *et al.*, 2004). These genomes contained a total of 397 operons but only 220 different sequences, showing that a large portion of the operons harbour identical sequences. If these genomes were treated as a microbial community, 16S rRNA cloning and sequencing would only result in roughly threefold overestimation of strain diversity. However, this estimate neglects genomes with single rRNA operons. Taking these into account, overestimation is closer to 2. Furthermore, association of 16S rRNA microvariation with different co-existing cells has been documented by *in situ* hybridization (Amann, 2000). Finally, physiologically distinct bacterial strains with identical16S rRNA have been isolated from the same environment (Sass *et al.*, 1998). These considerations suggest that a portion of the observed microdiversity in the delta-proteobacterial sequences may be caused by closely related co-existing strains.

To what extent the observed microdiversity represents ecologically differentiated populations is difficult to ascertain at this point. On the one hand, such variation may represent ecologically undifferentiated populations that have simply arisen by accumulation of mutations during clonal diversification. On the other hand, several lines of evidence suggest that at least some of the 16S rRNA microdiversity represents populations occupying differentiated niche spaces. The rRNAs are slowly evolving genes, and it has been hypothesized that protein-coding genes show evidence of selective sweeps based on adaptive mutations before variation would be seen at the 16S rRNA level (Palys *et al.*, 1997). Furthermore, isolates with identical 16S rRNA need not be metabolically or physiologically identical (Fox *et al.*, 1992). This has been confirmed by comparison of genome sequences of closely related organisms, which show quite extensive differences in gene arrangement, number and sequence (Alm *et al.*, 1999; Welch *et al.*, 2002; Ivanova *et al.*, 2003; Read *et al.*, 2003). That genomic variation in strains with identical 16S rRNA sequences can co-exist in the same environment has been demonstrated by two environmental genomics studies (Schleper *et al.*, 1997; Béja *et al.*, 2002) and by isolation of SRB strains displaying some physiological differences from the same sample (Sass *et al.*, 1998). Thus, it appears likely that microdiversity among salt marsh delta-Proteobacteria indicates some level of ecological adaptation and shows the need for more detailed studies.

Despite high microdiversity, the sequences fell into well-delineated phylogenetic groups. For some of these groups, inference of likely biogeochemical functions is possible. Over two-thirds of the delta-proteobacterial clones were most similar to representatives of the genera *Desulfosarcina* and *Desulfobacterium*, suggesting that complete oxidizers with high substrate versatility dominate the marsh sediments. For example, *Desulfosarcina variabilis*, which was associated with the largest single subclade (Fig. 4A), has metabolic capabilities that are well matched to carbon substrates such as acetate, lactate and ethanol exuded by *Spartina* roots. This is corroborated by previous detection of *Desulfosarcina*-like organisms by quantitative slot-blot hybridization of rRNA extracted from *Spartina* rhizosphere (Rooney-Varga *et al.*, 1997). Furthermore, plants and decomposing plant litter release hydrocarbons and aromatics, which can also be used directly by *D. variabilis*-like organisms such as strains oXyS1 and mXyS1 (Harms *et al.*, 1999). The utilization of the complex substrates such as plant phenolics and flavonoids may be an overall important property of the salt marsh SRB community.

SRB diversity based on 16S rRNA sequences correlated with a recent exploration of diversity in dissimilatory sulphite reductase (dsr) genes conducted on exactly the same sediment samples (M. Bahr *et al.*, unpublished). Both 16S rRNA and dsr clone libraries were dominated by sequences associated with the family Desulfobacteriaceae. In addition, in both libraries, incomplete-oxidizing *Desulfobulbus* and *Desulforhopalus* genera were detected. Moreover, both studies also failed to detect members of the completely oxidizing but nutritionally restricted genus *Desulfobacter* (Widdel and Bak, 1992). This suggests that *Desulfobacter*, which almost exclusively uses acetate as an energy source (Widdel and Bak, 1992), may be at a competitive disadvantage in the rhizosphere where diverse carbon substrates predominate and acetate has been found to support only ª 10% of SRB activity (Howarth, 1993). The *Desulfovibrio* and *Desulfofrigus/Desulfotalea* groups, which were at low abundance in the 16S rRNA gene library, were not detected in the smaller, dsr library (M. Bahr *et al.*, unpublished).

Other molecular diversity studies have detected both differences and similarities in SRB-like community composition. Desulfobacteriaceae dominate SRB communities in other salt marsh sediments based on cloning and quantitative hybridization studies (Devereux *et al.*, 1996; Rooney-Varga *et al.*, 1997). *Desulfosarcina* and *Desulfonema* have been detected in marine sediments (Llobet-Brossa *et al.*, 2002; Purdy *et al.*, 2003), microbial mats (Risatti *et al.*, 1994; Teske *et al.*, 1998; Minz *et al.*, 1999) and hydrocarbon seeps (Orphan *et al.*, 2001). *Desulfobacterium*-like sequences dominated the delta-proteobacterial portion of 16S rRNA clone libraries from the hydrocarbon-rich hydrothermal sediments of the Guaymas Basin (Dhillon *et al.*, 2003). The genus *Desulfobacterium* is nutritionally versatile, as are the genera *Desulfosarcina* and *Desulfococcus* (Widdel and Bak, 1992), and they may thrive in habitats with a similarly diverse substrate spectrum. However, groups that were not abundant in the *Spartina* marsh library can be important in other environments. Cultivation surveys often result in a predominance of *Desulfovibrio* strains, for example in water columns (Teske *et al.*, 1996) and in freshwater lake sediments (Sass *et al.*, 1998). Although cultivation bias favours quickly growing and robust *Desulfovibrio* strains, some molecular surveys also indicate a significant abundance of incompletely oxidizing sulphate reducers in some environments (Trimmer *et al.*, 1997; Llobet-Brossa *et al.*, 2002).

This study represents, to our knowledge, the most extensive sampling of a specific metabolic guild within a microbial community so far. The large-scale approach yielded several surprising results, which pose important questions for future research. The clone library demonstrated the co-existence of a high diversity of SRB organisms with similar overall metabolism and revealed high amounts of microdiversity. One of the most important questions will be to determine at what level of genetic differentiation these co-existing organisms are ecologically differentiated. This may be approached by a combination of targeted isolation of closely related organisms followed by extensive physiological and population biological studies. In addition, new techniques, which allow simultaneous detection of metabolic activity and molecular identification of microorganisms (Boschker *et al.*, 1998; Ouverney and Fuhrman, 1999; Radajewski *et al.*, 2000; Adamczyk *et al.*, 2003; Polz *et al.*, 2003), may provide insights into niche differentiation among both closely and distantly related organisms. A further challenge will be to determine what specific environmental factors select for the presence of one SRB group over another. For example, this study showed, in agreement with previous investigations, a clear dominance of *Desulfosarcina*-like sequences. It will be important to carry out comparative environmental studies, perhaps combined with genomics, to elucidate relevant factors that govern the distribution of microorganisms among different types of environments.

Experimental procedures

Study site and sampling

Samples were collected monthly from March to October 1998 from the bulk sediment of a monotypic stand of the tall form (2 m) of the marsh grass *Spartina alterniflora* at the mouth of the Rowley River in Plum Island Sound salt marsh (northeastern Massachusetts) (Fig. 1). The creekside sampling site had a continuous and dense cover of *S. alterniflora*, and the sediments showed no evidence of macrofaunal activity. The mean tidal range was 2.6 m, and the site was flooded during high tide although it stayed ≈ 1 m above water level during low tide. Salinity was measured in a small tidal pool near the site and was between 20‰ and 34‰ during the sampling period. Triplicate cores (5 cm diameter) were taken within a few metres of each other at mid-tide, immediately cooled to 4°C and transported to the laboratory. The top 4 cm of each core was collected using a sterile scalpel, pooled, placed in sterile 50 ml polypropylene tubes and stored frozen until further processing.

DNA extraction and purification

DNA was extracted by a modified version of the bead beating extraction protocol (Lin and Stahl, 1995). One gram of sample was combined with 0.5 g of sterilized 0.1-mm-diameter zirconium–silica beads (BioSpec Products) with 500 μl of equilibrated phenol (pH 7.0) and 35 μl of 10× buffer (500 mM sodium acetate and 100 mM EDTA buffer, pH 7.0), vortexed for ≈20 s and homogenized four times for 30 s in a reciprocal shaker (Mini-bead beater; BioSpec Products) with intermittent cooling on ice. The sample was then incubated for 10 min at 60°C, homogenized for an additional 30 s and centrifuged at 10 000 r.p.m. at 4°C for 10 min to pellet the beads and separate the phases. The supernatant was transferred to a clean 2 ml tube. The remaining beads were amended with 100 μl of 1× buffer, subjected to additional homogenization for 30 s, and the supernatant was collected after centrifugation for 10 min. Both supernatants were combined and recentrifuged at 14 000 r.p.m. at 4°C for 10 min to separate the remaining phenol. The upper aqueous phase was transferred to a clean tube and extracted twice with an equal volume of buffer-equilibrated phenol (pH 7.0), followed by additional extractions with an equal volume of phenol–chloroform and chloroform respectively. Nucleic acid was precipitated overnight at −20°C after the addition of ammonium acetate (2.5 M final concentration), $MgCl_2$ (2 mM final concentration) and 0.7 volumes of isopropanol. Nucleic acids were recovered by 10 min centrifugation at 14 000 r.p.m., followed by washing twice with 1 ml of 80% ethanol and resuspension in 100 μl of milliQ water. RNA was removed from a subsample of 30 μl by incubation at 37°C for 30 min with 20 U of RNase I (NE BioLabs). Final purification was performed using a Qiagen spin column PCR purification kit according to the manufacturer's instructions.

16S rRNA gene amplification, cloning and sequencing

The delta-proteobacterial SRB species-specific primer 385F was developed by combination of previously published SRB-specific oligonucleotide hybridization probes (Amann, 1995; Rabus *et al.*, 1996), and was used in combination with the bacterial primer 1492R for PCR amplification of 16S rRNA (Table 1). Each of the six monthly samples was amplified in 10 replicate reactions to minimize stochastic PCR bias (Polz and Cavanaugh, 1998). Each 20 μl reaction contained 0.2 mM each dNTP, 2 mM $MgCl_2$, 0.1 μM each primer, 1 μl of template DNA (5–10 ng), 1× PCR buffer and 0.1 U of *Taq* polymerase (Invitrogen) and was carried out in a Robocycler (Stratagene) using the following conditions: initial denaturation at 94°C for 3 min, followed by 15 cycles of denaturation at 94°C for 1 min, primer annealing at 50°C for 1 min, elongation at 72°C for 2 min with a final extension step at 72°C for 5 min. The amplification was carried out for only 15 cycles to decrease PCR bias (Polz and Cavanaugh, 1998) and the formation of *Taq* error and chimeric sequences (Qiu *et al.*,

Table 1. 16S rRNA gene-targeted primers.

Primer[a]	Used in:	Sequence	Specificity	Reference
385F[b]	PCR amplification	CTG ACG CAG CRA CGC CG	Most delta-proteobacterial SRB	Amann (1995); Rabus *et al.* (1996)
907R	Sequencing	CCG TCA ATT CMT TTR AGT TT	Most Bacteria	Lane (1991)
1492R[c]	PCR amplification	TAC GGY TAC CTT GTT AYG ACT T	Most Bacteria and Archaea	Lane (1991); Vergin *et al.* (1998)

a. 16S rRNA positions; *E. coli* numbering.
b. Designed from the commonly used 385 and 385b SRB probes by combining all degeneracies.
c. Modified according to information provided by Vergin *et al.* (1998) by incorporating a degeneracy (T/C) at position 1508 (*E. coli* numbering).

2001). The replicates of PCR amplifications were combined, precipitated using a QIAquick PCR purification kit (Qiagen), resuspended in 40 µl of milliQ water and purified additionally using the QIAquick gel extraction kit (Qiagen). The combined products were reamplified with five additional PCR cycles to minimize the formation of heteroduplex molecules (Thompson *et al.*, 2002) and purified using a QIAquick gel extraction kit. Subsequently, all 6 month samples were combined in equal amounts of DNA and used for cloning.

Four microlitres of the combined PCR products (final concentration 9.5 µg ml^{-1}) were ligated into PCR 2.1-TOPO vector and transformed into One Shot TOP10 chemically competent *Escherichia coli* cells (Invitrogen). Cells containing plasmid inserts were selected by growing on LB agar plates (Difco) in the presence of ampicillin and Xgal according to the manufacturer's specifications (Invitrogen). White colonies were transferred to 96-well deep blocks containing in each well 1.2 ml of Super Broth (32 g of tryptone, 20 g of Bacto yeast extract, 5 g of NaCl per litre) and ampicillin (50 mg l^{-1}). After overnight growth at 37°C with shaking at 250 r.p.m., cells were harvested by centrifugation at 2800 r.p.m. for 8 min at 4°C, and plasmids were extracted using the RevPrep Orbit™ workstation. Purified plasmids served as templates for partial 16S rRNA sequence determination using the bacterial primer 907R (Table 1) and the BigDye Termination kit version 3.0 (Applied Biosystems). Completed reactions were run on a 96-capillary 3730xl DNA analyser (Applied Biosystems).

Sequence analysis

The SEQUENCHER software package (Gene Codes) was used to remove vector and primer sequence and to check each sequence visually for ambiguities not scored by the automated sequence analysis program. Subsequently, sequences affiliated with the delta-proteobacterial subclass were identified by BLASTN (Altschul *et al.*, 1990) and preliminary phylogenetic tree construction using the neighbour-joining method within the ARB sequence analysis package (Ludwig *et al.*, 2004). These putative delta-proteobacterial sequences were kept for further analysis and subjected to a robust screening to score potential PCR-induced errors. First, putative *Taq* errors were identified by mapping each sequence manually to a secondary structure model of *Desulfovibrio desulphuricans* 16S rRNA (Cannone *et al.*, 2002). A sequence position was scored as a *Taq* error if (i) the nucleotide differed from universally conserved positions in the >98% consensus sequence assembled for all bacterial 16S rRNAs (Cannone *et al.*, 2002) or (ii) a non-canonical basepairing occurred in a stem region of the secondary structure and was absent in other closely related sequences. Secondly, sequences were tested for indication of chimera formation during the amplification. Initially, CHIMERA_CHECK implemented in the RDP (Maidak *et al.*, 2001) was used. However, for a large number of the sequences, it was difficult to conclude with high probability that they had originated from distinct parental sequences because no sufficiently close relatives were present in the RDP. Thus, the CHIMERABUSTER analysis tool (http://web.mit.edu/polz/seqtools/chimera.html) was developed. Briefly, the rationale for CHIMERABUSTER is derived from the fact that chimeras are combinations of sequences present in the sample and that well-sampled clone libraries should have a high incidence of co-occurrence of chimeras and their parental sequences. The program CHIMERABUSTER uses the two highly variable regions at each end of the molecule as *in silico* probes with adjustable specificity cut-offs. The program flags all sequences in which each probe matches two or more distinct sequences with >1% sequence difference. Thus, three sequences are identified, of which two are parental and one the potential chimera. Of these three, the sequence with the lowest incidence in the clone library was identified as more likely to be chimeric because chimeras form at later stages in the amplification when the parental sequences are already abundant. These putative chimeras, in addition to those identified by CHIMERA_CHECK, were excluded from the data set for phylogenetic analysis.

To determine how well the clone library was sampled at different sequence similarity levels, the sequences were first grouped into 100%, 99%, 98% and 97% similarity groups and rarefied. A clustering tool, which uses the nearest neighbour approach, adds a sequence to a cluster if there is at least one sequence that is within the similarity threshold set for the clustering (http://web.mit.edu/polz/seqtools/clusters.html). Rarefaction was carried through the Rarefaction Calculator (http://www2.biology.ualberta.ca/jbrzusto/rarefact.php). To estimate the total number of similarity clusters in the clone library at the different cut-offs, the Chao-1 non-parametric species richness estimator was calculated (Chao, 1987; Hughes *et al.*, 2001).

Phylogenetic analyses were carried out in PAUP*, version 4.0b10 (Swofford, 1993). For determination of the relationship of deeply divergent groups, a data set containing a single representative from each 98% identity cluster was assembled. Relationships were determined using the neighbour-joining method with Jukes–Cantor correction and checked for consistency using parsimony. The most variable regions (*E. coli* positions 452–463 and 849–850) were excluded from phylogenetic analyses of single representative sequences of each 98% identity cluster. For the analyses of sequences within 100% identity clusters, no length variation was observed, and all sequence positions were included. For each analysis, the robustness was tested by bootstrap resampling with the minimum evolution method with 100 replicates.

Nucleotide sequence data

The sequences of the cloned 16S rRNA SRB-like genes were deposited in GenBank under accession numbers AY374653–AY374982.

Acknowledgements

We would like to thank Ivica Ceraj for writing the chimera and clustering programs. This study was supported by the NSF Microbial Observatories programme (grant 9977897). Plasmid preps and sequencing were carried out at the W. M. Keck Ecological and Evolutionary Genetic Facility in the Josephine Bay Paul Center for Comparative Molecular Biology and Evolution at the Marine Biological Laboratory, Woods Hole, MA. USA.

References

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., and Polz, M.F. (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J Bacteriol* (in press).

Adamczyk, J., Hesselsoe, M., Iversen, N., Horn, M., Legner, A., Halkjaer Nielsen, P., *et al.* (2003) The isotope array, a new tool that employs substrate-mediated labeling of rRNA for determination of microbial community structure and function. *Appl Environ Microbiol* **69:** 6875–6887.

Aeckersberg, F., Bak, F., and Widdel, F. (1991) Anaerobic oxidation of saturated hydrocarbons by a new type of sulfate-reducing bacterium. *Arch Microbiol* **156:** 5–14.

Alm, R.A., Ing, L.-S.L., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., *et al.* (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori. Nature* **397:** 176–180.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Amann, R. (1995) Fluorescently labelled, rRNA-targeted oligonucleotide probes in the study of microbial ecology. *Mol Ecol* **4:** 543–554.

Amann, R. (2000) Who is out there? Microbial aspects of biodiversity. *Syst Appl Microbiol* **23:** 1–8.

Béja, O., Koonin, E.V., Aravind, L., Taylor, L.T., Seitz, H., Stein, J.L., *et al.* (2002) Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl Environ Microbiol* **68:** 355–345.

Beller, H.R., and Spormann, A.M. (1997) Benzylsuccinate formation as a means of anaerobic toluene activation by sulfate-reducing strain PRTOL1. *Appl Environ Microbiol* **63:** 3729–3731.

Boschker, H.R.S., Nold, S.C., Wellsburt, P., Bos, D., de Graaf, W., Pel, R., *et al.* (1998) Direct linking of microbial populations to specific biogeochemical processes by ^{13}C-labelling of biomarkers. *Nature* **392:** 801–805.

Canfield, D.E., and Des Marais, D.J. (1993) Biogeochemical cycles of carbon, sulfur, and free oxygen in a microbial mat. *Geochim Cosmochim Acta* **57:** 3971–3984.

Canfield, D.E., Jørgensen, B.B., Fossing, H., Glud, R., Gundersen, J., Ramsing, N.B., *et al.* (1993) Pathways of organic carbon oxidation in three continental margin sediments. *Mar Geol* **113:** 27–40.

Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *Biomed Central Bioinf* **3:** 2.

Casamayor, E.O., Pedros-Alio, C., Muyzer, G., and Amann, R. (2002) Microheterogeneity in 16S ribosomal DNA-defined bacterial populations from a stratified planktonic environment is related to temporal changes and to ecological adaptations. *Appl Environ Microbiol* **68:** 1706–1714.

Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43:** 783–791.

Currin, C.A., Newell, S.Y., and Paerl, H.W. (1995) The role of standing dead *Spartina alterniflora* and benthic microalgae in salt marsh food webs: considerations based on multiple stable isotope analysis. *Mar Ecol Prog Series* **121:** 99–116.

Curtis, T.P., Sloan, W.T., and Scanell, J.W. (2002) Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA* **99:** 10494–19499.

Devereux, R., and Stahl, D.A. (1993) Phylogeny of sulfate-reducing bacteria and a perspective for analyzing their natural communities. In *The Sulfate-Reducing Bacteria: Contemporary Perspectives.* Odom, J.M., and Singleton, R.J. (eds). New York: Springer-Verlag, pp. 131–160.

Devereux, R., Hines, M.E., and Stahl, D.A. (1996) S cycling: characterization of natural communities of sulfate-reducing bacteria by 16S rRNA sequence comparisons. *Microb Ecol* **32:** 283–292.

Dhillon, A., Teske, A., Dillon, J., Stahl, D.A., and Sogin, M.L. (2003) Molecular characterization of sulfate-reducing bacteria in the Guaymas Basin. *Appl Environ Microbiol* **69:** 2765–2772.

Dubilier, N., Mulders, C., Ferdelman, T., de Beer, D., Pernthaler, A., Klein, M., *et al.* (2001) Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. *Nature* **411:** 298–302.

Ferris, M.J., and Ward, D.M. (1997) Seasonal distributions of dominant 16S rRNA-defined populations in a hot spring microbial mat examined by denaturing gradient gel electrophoresis. *Appl Environ Microbiol* **63:** 1375–1381.

Ferris, M.J., Kuhl, M., Wieland, A., and Ward, D.M. (2003) Cyanobacterial ecotypes in different optical microenvironments of a 68°C hot spring mat community revealed by 16S-23S rRNA internal transcribed spacer region variation. *Appl Environ Microbiol* **69:** 2893–2898.

Field, K.G., Gordon, D., Wright, T., Rappe, M., Urbach, E., Vergin, K., and Giovannoni, S.J. (1997) Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Appl Environ Microbiol* **63:** 63–70.

Fox, G.E., Wisotzkey, J.D., and Jurtshuk, P.J. (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol* **42:** 166–170.

Friedrich, M., and Schink, B. (1995) Isolation and characterization of a desulforubidin-containing sulfate-reducing bacterium growing with glycolate. *Arch Microbiol* **164:** 271–279.

Gallagher, J.L., Reimold, R.J., Linthurst, R.A., and Pfeiffer, W.J. (1980) Aerial production, mortality, and mineral accumulation export dynamics in *Spartina alterniflora* and *Juncus roemerianus* plant stands in a Georgia salt-marsh. *Ecology* **61:** 303–312.

Galushko, A., Minz, D., Schink, B., and Widdel, F. (1999) Anaerobic degradation of naphthalene by a pure culture of a novel type of marine sulphate-reducing bacterium. *Environ Microbiol* **1:** 415–420.

Garcia-Martinez, J., and Rodriguez-Valera, F. (2000) Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine Archaea of group I. *Mol Ecol* **9:** 935–948.

Harms, G., Zengler, K., Rabus, R., Aeckersberg, F., Minz, D., Rossello-Mora, R., and Widdel, F. (1999) Anaerobic oxidation of o-xylene, m-xylene, and homologous alkylbenzenes by new types of sulfate-reducing bacteria. *Appl Environ Microbiol* **65:** 999–1004.

Hines, M.E., Knollmeyer, S.L., and Tugel, J.T. (1989) Sulfate reduction and other sedimentary biogeochemistry in a northern New England salt marsh. *Limnol Oceanogr* **34:** 578–590.

Hines, M.E., Banta, G., Giblin, A.E., Hobbie, J.E., and Tugel, J.T. (1994) Acetate concentration and oxidation in salt marsh sediments. *Limnol Oceanogr* **39:** 140–148.

Hines, M.E., Evans, R.S., Genthner, B.R.S., Willis, S.G., Friedman, S., Rooney-Varga, J.N., and Devereux, R. (1999) Molecular phylogenetic and biogeochemical studies of sulfate-reducing bacteria in the rhizosphere of Spartina alterniflora. *Appl Environ Microbiol* **65:** 2209–2216.

Howarth, R.W. (1993) Microbial processes in salt marshes. In *Aquatic Microbiology, an Ecological Approach*. Ford, T.E. (ed.). Cambridge, MA: Blackwell Scientific Publications, pp. 239–259.

Howarth, R.W., and Giblin, A. (1983) Sulfate reduction in the salt marshes at Sapelo Island, Georgia. *Limnol Oceanogr* **28:** 70–82.

Howarth, R.W., and Hobbie, J.E. (1985) Annual carbon mineralization and belowground production of *Spartina alterniflora* in a New England salt marsh. *Ecology* **66:** 595–605.

Howarth, R.W., and Teal, J.M. (1979) Sulfate reduction in a New England salt-marsh. *Limnol Oceanogr* **24:** 999–1013.

Howes, B.L., and Goehringer, D.D. (1994) Porewater drainage and dissolved organic-carbon and nutrient losses through the intertidal creekbanks of a New England saltmarsh. *Mar Ecol Prog Series* **114:** 289–301.

Hughes, J.B., Hellmann, J.J., Ricketts, T.H., and Bohannan, B.J.M. (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67:** 4399–4406.

Hurlbert, H.S. (1971) The non-concept of species diversity: a critique and alternative parameters. *Ecology* **52:** 577–586.

Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., Kapatral, V., *et al.* (2003) Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature* **423:** 87–91.

Jørgensen, B.B. (1982) Mineralization of organic matter in the sea bed – the role of sulfate reduction. *Nature* **296:** 643–645.

Joye, S.B., Mazzotta, M.L., and Hollibaugh, J.T. (1996) Community metabolism in microbial mats: The occurrence of biologically-mediated iron and manganese reduction. *Estuar Coastal Shelf Sci* **43:** 747–766.

Klappenbach, J.A., Saxman, P.R., Cole, J.R., and Schmidt, T.M. (2001) rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res* **29:** 181–184.

Lane, D.J. (1991) 16S/23S rRNA sequencing. In *Nucleic Acid Techniques in Bacterial Systematics*. Stackebrandt, E., and Goodfellow, M. (eds). New York: John Wiley and Sons, pp. 115–148.

Lin, C.Z., and Stahl, D.A. (1995) Taxon-specific probes for the cellulolytic genus *Fibrobacter* reveal abundant and novel equine-associated populations. *Appl Environ Microbiol* **61:** 1348–1351.

Llobet-Brossa, E., Rabus, R., Bottcher, M.E., Konneke, M., Finke, N., Schramm, A., *et al.* (2002) Community structure and activity of sulfate-reducing bacteria in an intertidal surface sediment: a multi-method approach. *Aquat Microb Ecol* **29:** 211–226.

Lonergan, D.J., Lenter, H.L., Coates, J.D., Phillips, E.J.P., Schmidt, T.M., and Lovley, D.R. (1996) Phylogenetic analysis of dissimilatory Fe(III)-reducing bacteria. *J Bacteriol* **178:** 2402–2408.

Lowe, K.L., Dichristina, T.J., Roychoudhury, A.N., and Van Cappellen, P. (2000) Microbiological and geochemical characterization of microbial Fe(III) reduction in salt marsh sediments. *Geomicrobiol J* **17:** 163–176.

Loy, A., Lehner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J., *et al.* (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl Environ Microbiol* **68:** 5064–5081.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* **32:** 1363–1371.

Maidak, B.L., Olsen, G.J., Larsen, N., Overbeek, R., McCaughey, M.J., and Woese, C.R. (1999) The RDP (Ribosomal Database Project). *Nucleic Acids Res* **25:** 109–110.

Maidak, B.L., Cole, J.R., Lilburn, T.G., Parker, C.T., Saxman, P.R., Farris, R.J., *et al.* (2001) The RDP-II (Ribosomal Database Project). *Nucleic Acids Res* **29:** 173–174.

Mendelssohn, I.A., McKee, K.L., and Patrick, W.H.J. (1981) Oxygen deficiency in *Spartina alterniflora* roots: metabolic adaptation to anoxia. *Science* **214:** 439–441.

Minz, D., Flax, J.L., Green, S.J., Muyzer, G., and Cohen, Y. (1999) Diversity of sulfate-reducing bacteria in oxic and anoxic regions of a microbial mat characterized by comparative analysis of dissimilatory sulfite reductase genes. *Appl Environ Microbiol* **65:** 4666–4671.

Opsahl, S., and Benner, R. (1999) Characterization of carbohydrates during early diagenesis of five vascular plant tissues. *Org Geochem* **30:** 83–94.

Orphan, V.J., Hinrichs, K.U., Ussler, W., Paull, C.K., Taylor, L.T., Sylva, S.P., *et al.* (2001) Comparative analysis of methane-oxidizing archaea and sulfate-reducing bacteria in anoxic marine sediments. *Appl Environ Microbiol* **67:** 1922–1934.

Ouverney, C.C., and Fuhrman, J.A. (1999) Combined microautoradiography-16S rRNA probe technique for determination of radioistope uptake by specific microbial cell types *in situ*. *Appl Environ Microbiol* **65:** 1746–1752.

Palys, T., Nakamura, L.K., and Cohan, F.M. (1997) Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. *Int J Syst Bacteriol* **47:** 1145–1156.

Phelps, C., Kerkhof, I., and Young, I. (1998) Molecular characterization of a sulfate-reducing consortium which mineralizes benzene. *FEMS Microbial Ecol* **27:** 269–279.

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

Polz, M.F., Bertilsson, S., Acinas, S.G., and Hunt, D. (2003) A(r)ray of hope in analysis diversity of microbial of the function and communities. *Biol Bull* **204:** 196–199.

Purdy, K.J., Nedwell, D.B., and Embley, T.M. (2003) Analysis of the sulfate-reducing bacterial and methanogenic archaeal populations in contrasting Antarctic sediments. *Appl Environ Microbiol* **69:** 3181–3191.

Qiu, X.Y., Wu, L.Y., Huang, H.S., McDonel, P.E., Palumbo, A.V., Tiedje, J.M., and Zhou, J.Z. (2001) Evaluation of PCR-generated chimeras: mutations and heteroduplexes with 16S rRNA gene-based cloning. *Appl Environ Microbiol* **67:** 880–887.

Rabus, R., Fukui, M., Wilkes, H., and Widdle, F. (1996) Degradative capacities and 16S rRNA-targeted whole-cell hybridization of sulfate-reducing bacteria in anaerobic enrichment culture utilizing alkylbenzenes from crude oil. *Appl Environ Microbiol* **62:** 3605–3613.

Radajewski, S., Ineson, P., Parekh, N., and Murrell, J. (2000) Stable-isotope probing as a tool in microbial ecology. *Nature* **403:** 646–649.

Rainey, F.A., Ward-Rainey, N.L., Janssen, P.H., Hippe, E., and Stackebrandt, E. (1996) *Clostridium paradoxum* contains multiple 16S rRNA genes with heterogeneous intervening sequences. *Microbiology* **142:** 2087–2095.

Ravenschlag, K., Sahm, K., Pernthaler, J., and Amann, R. (1998) High bacterial diversity in permanently cold marine sediments. *Appl Environ Microbiol* **65:** 3982–3989.

Read, T.D., Peterson, S.N., Tourasse, N., Baillie, L.W., Paulsen, I.T., Nelson, K.E., *et al.* (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* **423:** 81–86.

Risatti, J.B., Capman, W.C., and Stahl, D.A. (1994) Community structure of a microbial mat: the phylogenetic dimension. *Proc Natl Acad Sci USA* **91:** 10173–10177.

Rooney-Varga, J.N., Devereux, R., Evans, R.S., and Hines, M.E. (1997) Seasonal changes in the relative abundance of uncultivated sulfate-reducing bacteria in a salt marsh sediment and in the rhizosphere of *Spartina alterniflora*. *Appl Environ Microbiol* **63:** 3895–3901.

Rooney-Varga, J.N., Genthner, B.R.S., Devereux, R., Willis, S.G., Friedman, S.D., and Hines, M.E. (1998) Phylogenetic and physiological diversity of sulphate-reducing bacteria isolated from a salt marsh sediment. *Syst Appl Microbiol* **21:** 557–568.

Sass, H., Wieringa, E.B.A., Cypionka, H., Babenzien, H.D., and Overmann, J. (1998) High genetic and physiological diversity of sulfate-reducing bacteria isolated from an oligotrophic lake sediment. *Arch Microbiol* **170:** 243–251.

Schleper, C.A., Holben, W., and Klenk, H.-P. (1997) Recovery of crenarchaeotal ribosomal DNA sequences from freshwater-lake sediments. *Appl Environ Microbiol* **63:** 321–323.

Swofford, D.L. (1993) PAUP – a computer program for phylogenetic inference using maximum parsimony. *J Gen Physiol* **102:** A9–A9.

Teal, J.M. (1962) Energy flow in the salt marsh ecosystem of Georgia. *Ecology* **43:** 614–624.

Teske, A., Waver, C., Muyzer, G., and Ramsing, N.B. (1996) Distribution of sulfate-reducing bacteria in a stratified fjord (Mariager Fjord, Denmark) as evaluated by most-probable-number counts and denaturing gradient gel electrophoresis of PCR-amplified ribosomal DNA fragments. *Appl Environ Microbiol* **62:** 1405–1415.

Teske, A., Ramsing, N.B., Habicht, K., Fukui, M., Kuver, J., Jorgensen, B.B., and Cohen, Y. (1998) Sulfate-reducing bacteria and their activities in cyanobacterial mats of Solar Lake (Sinai, Egypt). *Appl Environ Microbiol* **64:** 2943–2951.

Thompson, J.R., Marcelino, L.A., and Polz, M.F. (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. *Nucleic Acids Res* **30:** 2083–2088.

Tindall, B.J., and Kunkel, T.A. (1988) Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase. *Biochemistry* **27:** 6008–6013.

Torsvik, V., Øvreås, L., and Thingstad, T.F. (2002) Prokaryotic diversity – magnitude, dynamics, and controlling factors. *Science* **296:** 1064–1066.

Trimmer, M., Purdy, K.J., and Nedwell, D.B. (1997) Process measurement and phylogenetic analysis of the sulfate reducing bacterial communities of two contrasting benthic sites in the upper estuary of the Great Ouse, Norfolk, UK. *FEMS Microbiol Ecol* **24:** 333–342.

Vergin, K.L., Urbach, E., Stein, J.L., DeLong, E.F., Lanoil, B.D., and Giovannoni, S.J. (1998) Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order *Planctomycetales*. *Appl Environ Microbiol* **64:** 3075–3078.

Welch, R.A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci USA* **99:** 17020–17024.

Widdel, F., and Bak, F. (1992) Gram-negative mesophilic sulfate-reducing bacteria. In *The Prokaryotes*. Balows, A., Trüper, H.G., Dworkin, M., Harder, W., and Schleifer, K.-H. (eds). New York: Springer, pp. 3352–3378.

Wiegert, R.G., and Pomeroy, L.R. (1981) Ecology of salt marshes: an introduction. In *The Ecology of a Salt Marsh*. Pomeroy, L.R., and Wiegert, R.G. (eds). New York: Springer Verlag, pp. 3–20.

Wilson, J.O., Buchsbaum, R., Valiela, I., and Swain, T. (1986) Decomposition in salt-marsh ecosystems – phenolic dynamics during decay of litter of *Spartina alterniflora*. *Mar Ecol Prog Series* **29**: 177–187.

Erratum Chapter 2

Page 23 “Finally, at least two small, deep branching clades (Fig. 3, clades V and VI), containing 10 and 19 sequences, ...” should read “Finally, at least two small, deep branching clades (Fig. 3, clades V and VI), containing 20 and 10 sequences, ...”

Page 24, Fig. 3: “*Desulfovibrio mediterraneus*” exported from ARB package where it was incorrectly referenced. It should read “*Desulfobulbus mediterraneus.*”

Chapter Three

Fine-scale phylogenetic architecture of a complex bacterial community

Silvia G. Acinas*, Vanja Klepac-Ceraj*, Dana E. Hunt, Chanathip Pharino, Ivica Ceraj, Daniel L. Distal and Martin F. Polz

* co-first authors

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

16. Pollitz, F. F. Transient rheology of the uppermost mantle beneath the Mojave Desert, California. *Earth Planet. Sci. Lett.* **215,** 89–104 (2003).
17. Wald, D. J. & Heaton, T. H. Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake. *Bull. Seismol. Soc. Am.* **84,** 668–691 (1994).
18. Kaverina, A., Dreger, D. & Price, E. The combined inversion of seismic and geodetic data for the source process of the 16 October 1999 Mw 7.1 Hector Mine, California, earthquake. *Bull. Seismol. Soc. Am.* **92,** 1266–1280 (2002).
19. Masterlark, T. & Wang, H. F. Transient stress-coupling between the 1992 Landers and 1999 Hector Mine, California, earthquakes. *Bull. Seismol. Am.* **92,** 1470–1486 (2002).
20. Peltzer, G., Rosen, P. & Rogez, F. Poroelastic rebound along the Landers 1992 earthquake surface rupture. *J. Geophys. Res.* **103,** 30131–30145 (1998).
21. Williams, C. F. Temperature and the seismic/aseismic transition; observations from the 1992 Landers earthquake. *Geophys. Res. Lett.* **23,** 2029–2032 (1996).
22. Melbourne, T. & Helmberger, D. Mantle control of plate boundary deformation. *Geophys. Res. Lett.* **28,** 4003–4006 (2001).
23. Goes, S. & van der Lee, S. Thermal structure of the North American uppermost mantle inferred from seismic tomography. *J. Geophys. Res.* **107,** doi:2000JB000049 (2002).
24. Farmer, G. L. *et al.* Origin of late Cenozoic basalts at the Cima volcanic field, Mojave Desert, California. *J. Geophys. Res.* **100,** 8399–8415 (1995).
25. Kronenberg, A. K. & Tullis, J. A. Flow strengths of quartz aggregates; grain size and pressure effects due to hydrolytic weakening. *J. Geophys. Res.* **89,** 4281–4297 (1984).
26. Shelton, G. & Tullis, J. A. Experimental flow laws for crustal rocks. *Trans. Am. Geophys. Union* **62,** 396 (1981).
27. Jaoul, O., Tullis, J. A. & Kronenberg, A. K. The effect of varying water contents on the creep behaviour of Heavitree Quartzite. *J. Geophys. Res.* **89,** 4298–4312 (1984).
28. Hansen, F. D. & Carter, N. L. Creep of selected crustal rocks at 1000 MPa. *Trans. Am. Geophys. Union* **63,** 437 (1982).
29. Hirth, G. & Kohlstedt, D. Rheology of the upper mantle and the mantle wedge: A view from the experimentalists. In *The Subduction Factory* (ed. Eiler, J.) (American Geophysical Union 2004).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements Funding was provided by the National Science Foundation and the Southern California Earthquake Center. GPS data were provided by the US Geological Survey and the Southern California Integrated GPS Network..

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.M.F. (freed@purdue.edu).

Fine-scale phylogenetic architecture of a complex bacterial community

Silvia G. Acinas[1]*, Vanja Klepac-Ceraj[1]*, Dana E. Hunt[1], Chanathip Pharino[1], Ivica Ceraj[2], Daniel L. Distel[3] & Martin F. Polz[1]

[1] *Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[2] *Bugaco, Somerville, Massachusetts 02144, USA*
[3] *Department of Biochemistry, Microbiology and Molecular Biology, University of Maine, Orono, Maine 04469, USA*

* These authors contributed equally to this work

Although molecular data have revealed the vast scope of microbial diversity[1], two fundamental questions remain unanswered even for well-defined natural microbial communities: how many bacterial types co-exist, and are such types naturally organized into phylogenetically discrete units of potential ecological significance? It has been argued that without such information, the environmental function, population biology and biogeography of microorganisms cannot be rigorously explored[2]. Here we address these questions by comprehensive sampling of two large 16S ribosomal RNA clone libraries from a coastal bacterioplankton community. We show that compensation for artefacts generated by common library construction techniques reveals fine-scale patterns of community composition. At least 516 ribotypes (unique rRNA sequences) were detected in the sample and, by statistical extrapolation, at least 1,633 co-existing ribotypes in the sampled population. More than 50% of the ribotypes fall into discrete clusters containing less than 1% sequence divergence. This pattern cannot be accounted for by interoperon variation, indicating a large predominance of closely related taxa in this community. We propose that such microdiverse clusters arise by selective sweeps and persist because competitive mechanisms are too weak to purge diversity from within them.

Traditional species concepts have largely been concessions to the need to identify bacteria reproducibly, but none adequately describe natural units of microbial diversity[3]. It has recently been proposed that natural taxa are distinct groups of strains that arise by periodic selection—a process of continuing, selectionally neutral, diversification punctuated by adaptive mutations leading to selective sweeps[4]. The latter events purge all sequence variants except those associated with the genome carrying the adaptive mutation[4]. One of the attractive features of this concept is that it should be applicable to molecular surveys of microbial diversity because taxa would be identifiable in phylogenetic trees as distinct clusters of closely related sequences[1]. Moreover, such clusters should be detectable independently of the gene used to construct these trees as long as the accumulation of variation is commensurate with the occurrence of sweeps[5]. However, this theory has not been applied to broad-scale studies of bacterial diversity in the environment. Over the past 20 years, diversity studies have primarily been based on analyses of 16S rRNA clone libraries but it has remained uncertain to what extent fine-scale patterns of variation are due to sequence artefacts, to heterogeneity among paralogous operons or to the co-existence of similar but differentiated taxa[1]. Furthermore, it has not been explored whether naturally defined units of differentiation emerge from recently released shotgun sequence data from the Sargasso Sea[6].

We deduced that the discovery of ecologically significant patterns of relationships between co-existing ribotypes requires, first, an examination of clone libraries large enough to elucidate relationships at all levels of differentiation, and second, methods that minimize and account for the contribution of sequence artefacts and paralogous variation to diversity estimates. We sequenced about 1,000 clones from each of two polymerase chain reaction (PCR)-derived 16S rRNA libraries constructed from the same coastal bacterioplankton sample. The first library employed common (standard) amplification protocols. For the second, a modified protocol was designed to minimize artefacts and to identify Taq errors and chimaeric molecules through extensive sequence analyses (see Methods). This approach allowed the most comprehensive analysis of any single gene from co-occurring populations so far, even in view of the recently released Sargasso Sea study, which in aggregate sampled a similar number of rRNA genes but from several locations, dates and diverse biogeochemical conditions[6]. Our overall rationale was to achieve high coverage of rRNA genes from a single community while estimating and compensating for the influence of artefacts on ribotype diversity, potentially revealing emergent patterns.

Comparison of the two libraries showed that changes to the amplification protocol alone decreased the incidence of unique sequences from 76% (692 of 909) in the standard to 61% (686 of 1,131) in the modified library. Correction for chimaeras and Taq error lowered the percentage to 48% (516 of 1,067) unique sequences (Fig. 1a), demonstrating a potentially significant contribution of PCR-induced artefacts to (micro)diversity estimates. Consequently, these corrections yield a significantly lower estimate of total ribotype diversity for the sampled community when compared with the unmodified standard library (1,633 versus 3,881) with the use of the Chao-1 estimator[7]. A novel estimator (N_T/N_{max}) (ref. 2) yielded a similar value of 2,236 sequences for the corrected data set. This good agreement, combined with the low incidence of chimaeras and the observation that corrections account for most expected Taq errors (see Methods), provides confidence in the corrected estimate of ribotype diversity.

A vast and previously unrecognized predominance of microdiverse ribotypes was revealed by further analysis of relationships. More than half of the observed sequences in the modified library fell into clusters sharing at least 99% sequence consensus (Fig. 1a). This result is still more marked when the Chao-1 diversity estimator is applied to the data, indicating that more than two-thirds of ribotypes might be members of 99% sequence clusters in the sampled bacterioplankton. Defining such 99% clusters, rather than unique ribotypes, as operational taxonomic units (OTUs) decreases diversity estimates from 1,633 to 520 OTUs, a decline of about 70%. However, further clustering into 98% and 97% consensus groups decreases the number of OTUs by only 3% (507 OTUs) and 11% (450 OTUs), respectively (Fig. 1a). In fact, a remarkably consistent exponential decline was observed in the number of OTUs as cluster cut-off values were decreased from 99% to 75% (Fig. 1b). In stark contrast, the number of OTUs greatly exceeds this exponential trend for values above 99% (Fig. 1b). An essentially identical relationship emerged from a phylogenetic (maximum-likelihood) analysis, in which the accumulation of lineages per arbitrary time unit was inferred under a molecular clock model[8] (data not shown). This exponential accumulation of clusters or lineages is expected if the creation and removal of taxa are on average constant over time[8]. The sharp discontinuity observed above 99% similarity therefore suggests increased diversification or decreased removal of diversity within microdiverse clusters.

The overall predominance of extremely closely related ribotypes also emerges from phylogenetic analyses as large clusters of closely related taxa (Fig. 2, and Supplementary Information). These are typically well separated from other clusters, as indicated by a comparison of average within-cluster and between-cluster sequence divergence (data not shown)[5]. The most sequence-rich microdiverse clusters are formed within the Pelagibacter (SAR11) group of the alpha-Proteobacteria (Fig. 2) but all highly represented lineages contain such clusters, including the gamma-Proteobacteria and the Bacteriodetes group (Supplementary Information). Nevertheless, microdiverse clusters are not uniformly distributed between lineages in the modified clone library. For example, the Cytophaga group contains more deeply divergent lineages and fewer microdiverse clusters than the Pelagibacter group (Supplementary Information). However, such differences might be due to incomplete sampling because rarefaction (Fig. 1a) suggests that deeper branching lineages in this library are well sampled and that additional sequencing should therefore primarily reveal microdiverse ribotypes.

To what extent can the observed ribotype microdiversity be explained by variation among paralogous operons within single genomes[1]? We have recently explored this question by an analysis of 97 available complete bacterial genomes[9]. These contain, because of multiple non-identical operons, a total of 242 different 16S rRNA sequences[9]; that is, the number of ribotypes exceeds the number of genomes about 2.5-fold. Remarkably, interoperon sequence difference remains within about 1% among these genomes. Only five genomes deviate from this rule, four of which were thermophilic bacteria all with a single operon with higher sequence divergence[9]. Therefore, if one accepts that the distribution of operons among free-living bacteria is similar to that of the 97 sequenced genomes, a conservative correction factor of about 2.5 (ref. 9) can be applied to

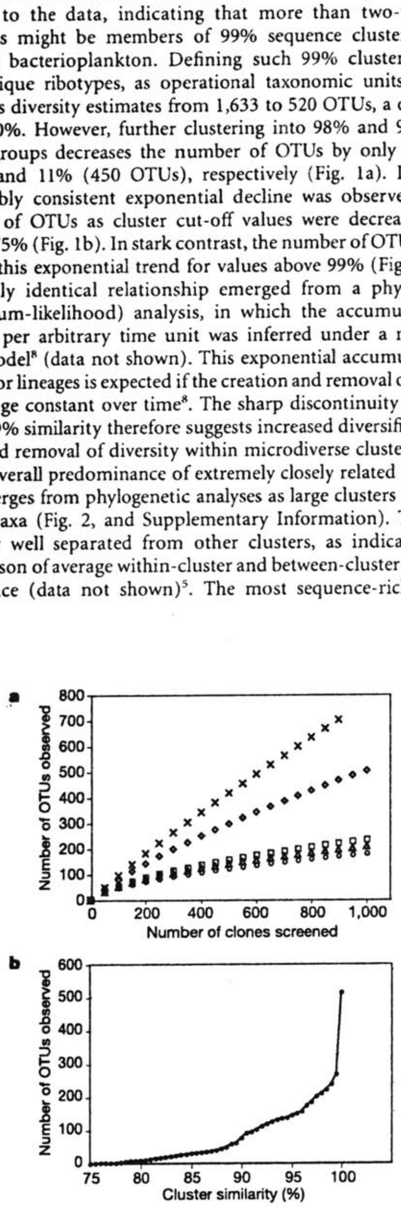


Figure 1 Compositional pattern of the coastal bacterioplankton sample. **a**, Rarefaction curves of the number of OTUs in a 16S rRNA library constructed with standard (crosses, 100% sequence similarity cluster) and modified (diamonds, 100% sequence similarity clusters; squares, 99%; triangles, 98%; circles, 97%) amplification protocols. Standard deviations fall within the symbols and are not shown. **b**, Number of OTUs plotted against changing degrees of cutoffs in 0.5% increments for grouping of sequences into similarity clusters.

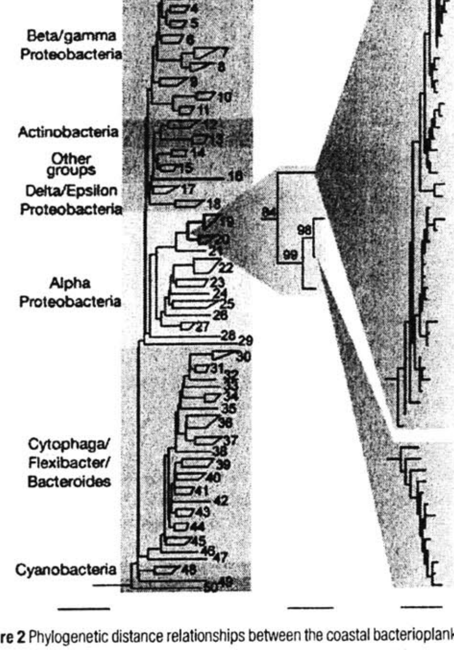


Figure 2 Phylogenetic distance relationships between the coastal bacterioplankton based on partial 16S rRNA sequencing. **a**, Summary of groups represented in the sample, in which each number denotes a phylogenetic cluster of sequences (for identification key see Supplementary Information). **b**, Relationships between *Pelagibacter* (SAR11) clusters represented by one sequence of each 99% similarity cluster. Numbers associated with nodes represent bootstrap values. **c**, Examples of microdiverse relationships between SAR11 ribotypes. Scale bars, 0.1 (**a**), 0.05 (**b**) and 0.01 (**c**) substitutions per site. Arrows connecting trees point to expanded nodes.

the estimated number of sequences (1,113) in the 99% similarity group to yield a revised estimate of at least 446 closely related genomes co-existing in the sample. However, this is probably an overcorrection because opportunistic bacteria with multiple operons are thought to predominate in culture collections and among sequenced genomes[9,10]. Moreover, *Pelagibacter ubique* HTCC1062, which is identical in sequence to clones within the largest SAR11 99% similarity clusters, seems to contain a single rRNA operon (S. J. Giovannoni, personal communication). Given that operon numbers vary little between closely related bacteria[9] it is unlikely that the observed SAR11 microdiversity can be explained by operon differences. Finally, shotgun sequencing of Sargasso Sea prokaryotes revealed a total of about 1,400 rRNA and about 600 RecA sequences[6]. The latter is a single-copy gene in all currently published genomes. Their frequency in the sample therefore provides an independent and almost identical estimate of 2.3 rRNA operons per genome. Thus, we conclude that, even after conservative correction, genomes denoted by microdiverse ribotypes represent by far the dominant fraction of bacterial diversity in this coastal bacterioplankton sample.

The observed pattern raises the question: what level of similarity should be expected between genomes carrying microvariant ribotypes? Comparative genomics has shown that genomes can be divided into stable and variable sets of genes, termed the core and flexible/auxiliary genome, respectively[11,12]. The latter arises primarily by means of phage and transposon-mediated lateral gene transfer and comprises between 1% and 18% of genes[11] but possibly as much as 60% (ref. 13). The core genome, in contrast, is a stable complement of genes that includes rRNA and housekeeping genes. This core reflects the overall evolutionary history of the lineage because little lateral gene transfer is detectable[9,12,14]. Microdiverse ribotype clusters should therefore also be apparent in comparisons of other housekeeping genes, possibly more so because of the higher substitution rates typical of protein coding genes[5].

Do microdiverse sequences denote co-existing, ecologically differentiated genomes? Among free-living bacteria of very similar ribotypes, correlation of genomic variation with ecological parameters has been demonstrated convincingly in only a single case involving two strains of *Prochlorococcus*[15], but these would not fall within a single microdiverse cluster as defined here. In contrast, no evidence of functional differentiation was detected in several environmental BAC clones with microdiverse 16S rRNA, despite considerable polymorphisms in protein-coding genes[16,17]. This is consistent with recently advanced theories for the interpretation of microdiverse sequence clusters. It has been shown[5] that clustering of housekeeping genes, resulting from periodic selection, predicts ecologically differentiated strains within cultivated bacterial taxa. If microdiverse ribotype clusters in the environment arise by the same mechanism[1] their very existence implies that intracluster competition is too weak to sweep members from within their ranks. However, this does not require that these genomes are functionally identical. Subdifferentiation within the flexible genome might provide increased fitness under episodic or spatially confined environmental conditions, but not sufficient growth advantage to sweep competing microdiverse genomes[12]. Furthermore, ecological factors might decrease effective competition. Particularly, predation has been suggested to promote the coexistence of diverse lineages by 'killing the winner' of competitive events[18]. Finally, recombination might be important in delineating and preserving genetic diversity among members of clusters by allowing sweeps of adaptive alleles without removing selectionally neutral variation[19].

The above considerations lead us to suggest that microdiverse ribotype clusters are important units of differentiation in natural bacterial communities. Indeed, such clusters might be widespread. We have recently detected numerous microdiverse ribotypes in saltmarsh sulphate-reducing bacteria[20], and ribotype clusters have previously been tentatively suggested for some open-ocean microbial groups[1]. To determine whether microdiverse ribotype clusters described here represent ecotypes—that is, ecologically cohesive populations—will require a detailed comparison of their encompassed genomic variation. Indeed, high-throughput sequencing[6] and cultivation[21] provide the means for rigorous testing of the hypothesized ecological importance of ribotype clusters. Most importantly, such inquiries challenge us to re-examine concepts of microbial diversity and invigorate the search for ecologically and evolutionarily defined species concepts. □

Methods

Study site and sampling

A 2.2-litre water sample was collected on 6 October 2001 from the marine end of the Plum Island Sound estuary (northeastern Massachusetts), and bacterioplankton was concentrated on a 0.22-μm filter (Supor; Gelman), which was stored at −80 °C until DNA extraction. Measured water parameters were as follows: temperature 16 °C, pH 8.0, prokaryotic cell numbers $0.99 \times 10^9\,\mathrm{l}^{-1}$, dissolved organic carbon $0.4\,\mathrm{mg\,C\,l}^{-1}$, chlorophyll *a* $5.94\,\mu\mathrm{g\,l}^{-1}$.

DNA extraction and clone library construction

Cells on filters were lysed and nucleic acids were extracted with a modified version of a bead-beating method[20] followed by treatment with RNase I and purification on Qiagen DNA purification spin columns. Two 16S rRNA clone libraries were constructed from the same sample to estimate the total coexisting sequence diversity and the effect of PCR-induced artefacts. For both, the bacteria-specific primers 27F and 1492R as modified in ref. 22 were used. Each PCR reaction contained genomic DNA equivalent to 4.9×10^6 cells. Ten replicate reactions were combined and gel-purified, and the same amount of amplicons were cloned with the PCR 2.1-TOPO kit (Invitrogen). The PCR amplification for the first (standard) library used 35 cycles mimicking commonly used protocols (typically between 30 and 40 cycles). The second (modified) library was constructed to minimize the accumulation of the three known PCR artefacts (Taq errors, chimaeras and heteroduplex molecules)[20]. In brief, the sample was amplified for 15 cycles followed by a 3-cycle 'reconditioning step', which eliminated heteroduplex molecules[23] and decreased the incidence of Taq errors and chimaeras. Purified plasmids served as templates for partial 16S rRNA sequence determination with the bacterial primer 27F (ref. 20).

Sequence analysis

Sequences (position 68 to 805, *E. coli* numbering) from the corrected library were further analysed for evidence of PCR artefacts. About 3% of the sequences were removed as putative chimaeras on the basis of identification by a combination of the three bioinformatics tools Chimera_Check[24], Bellerophon[25] and ChimeraBuster[20]. Taq errors were identified by manual reconstruction of 16S rRNA secondary structures as pioneered in ref. 26 and detailed in ref. 20. In brief, the method scores as Taq errors sequence changes that violate either the sequence-conservation rule (nucleotides that are different in positions more than 98% conserved in all bacterial sequences) or the secondary-structure-conservation rule (apparent changes resulting in mismatches in stem structures that are not detected in related sequences). The Taq error rate determined from these rules was 3.3×10^{-5} per nucleotide per duplication, which agrees remarkably well with the experimentally determined value of 2×10^{-5} per nucleotide per duplication for the Taq polymerase used[20]. Further confidence that the large majority of Taq errors are captured is lent by several simple considerations. First, inspection of alignments showed the remaining variation clustered in regions known to be highly variable, which is inconsistent with the expected random distribution of Taq errors. Second, separate quantification of Taq error rate for positions falling under the conservation and the secondary structure rule previously gave highly similar rates of 1.7×10^{-5} and 2.5×10^{-5}, respectively[20]. Third, after 18 cycles, the inferred Taq error rate would lead to a misincorporation of bases at a rate of 3.6×10^{-4} per nucleotide. Because about 800 base pairs of sequence reads were obtained, this would translate into an average of 0.3 errors per sequence. This is close to the fraction of sequences (0.26; 181 of 686) removed from the modified library owing to identification of Taq errors. Last, potentially undetectable Taq errors by the secondary structure rule are those that change one allowed base pairing into another (for example, A-U to G-U). Although this can happen in two-thirds of all positions in rRNA, only 30% of the time will random replacement of one nucleotide by another due to Taq error result in another allowed base pairing. Therefore, at worst 20% (0.67 × 0.3) of Taq errors are missed but since only about 64% of the nucleotide positions fell under the secondary-structure rule this number translates into about 13%. Considering the probable incidence of errors per sequence based on a Taq error rate of 2×10^{-5} the calculation again results in an estimated 4% of sequences (0.13 × 0.3) that carry errors missed by the applied corrections.

The corrected set of sequences was used to estimate total diversity in the bacterioplankton community and patterns of relationships. An algorithm was developed[20] to group sequences into percentage similarity clusters (100%, 99%, 98%, and so on). This formed the basis for statistical extrapolation of total sequence diversity with the Chao-1 (ref. 7) and N_T/N_{MAX} (ref. 2) estimators. Accumulation of lineages through time was calculated with GENIE[27] from a non-optimized tree inferred by maximum likelihood with the molecular clock assumption enforced[8]. Identification of phylogenetic affiliation of the sequences was performed with the neighbour-joining method implemented in ARB[28] and followed by an analysis of more restricted groups of sequences by using distance and parsimony methods in PAUP* (ref. 29).

Received 10 February; accepted 11 May 2004; doi:10.1038/nature02649.

1. Rappé, M. S. & Giovannoni, S. J. The uncultured microbial majority. *Annu. Rev. Microbiol.* **57,** 369–394 (2003).
2. Curtis, T. P., Sloan, W. T. & Scannell, J. W. Estimating prokaryotic diversity and its limits. *Proc. Natl Acad. Sci. USA* **99,** 10494–10499 (2002).
3. Rosselló-Mora, R. & Amann, R. The species concept for prokaryotes. *FEMS Microbiol. Rev.* **25,** 39–67 (2001).
4. Cohan, F. M. What are bacterial species. *Annu. Rev. Microbiol.* **56,** 457–487 (2002).
5. Palys, T., Nakamura, L. K. & Cohan, F. M. Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. *Int. J. Syst. Bacteriol.* **47,** 1145–1156 (1997).
6. Venter, C. J. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
7. Hughes, J. B., Hellmann, J. J., Ricketts, T. H. & Bohannan, B. J. M. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* **67,** 4399–4406 (2001).
8. Martin, A. P. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl. Environ. Microbiol.* **68,** 3673–3682 (2002).
9. Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. & Polz, M. F. Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J. Bacteriol.* **186,** 2629–2635 (2004).
10. Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* **66,** 1328–1333 (2000).
11. Hacker, J. & Carniel, E. Ecological fitness, genomic islands and bacterial pathogenicity—a Darwinian view of the evolution of microbes. *EMBO Rep.* **2,** 376–381 (2001).
12. Lan, R. T. & Reeves, P. R. When does a clone deserve a name? A perspective on bacterial species based on population genetics. *Trends Microbiol.* **9,** 419–424 (2001).
13. Welch, R. A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc. Natl Acad. Sci. USA* **99,** 17020–17024 (2002).
14. Daubin, V., Moran, N. A. & Ochman, H. Phylogenetics and the cohesion of bacterial genomes. *Science* **301,** 829–832 (2003).
15. Rocap, G. *et al.* Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424,** 1042–1047 (2003).
16. Schleper, C. A. *et al.* Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Crenarchaeum symbiosum*. *J. Bacteriol.* **180,** 5003–5009 (1998).
17. Béja, O. *et al.* Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl. Environ. Microbiol.* **68,** 335–345 (2002).
18. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* **45,** 1320–1328 (2000).
19. Lan, R. T. & Reeves, P. R. Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol.* **8,** 396–401 (2000).
20. Klepac-Ceraj, V. *et al.* High overall diversity and dominance of microdiverse relationships in salt marsh sulfate-reducing bacteria. *Environ. Microbiol.* doi: 10.1111/j.1462-2920.2004.00600.x (2004).
21. Connon, S. A. & Giovannoni, S. J. High-throughput method for culturing microorganisms in very low nutrient media yield diverse new marine isolates. *Appl. Environ. Microbiol.* **768,** 3878–3885 (2002).
22. Vergin, K. L. *et al.* Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order *Planctomycetales*. *Appl. Environ. Microbiol.* **64,** 3075–3078 (1998).
23. Thompson, J. R., Marcelino, L. A. & Polz, M. F. Heteroduplexes in mixed-template amplifications: formation, consequences and elimination by 'reconditioning PCR'. *Nucleic Acids Res.* **30,** 2083–2088 (2002).
24. Cole, J. R. *et al.* 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 (2003).
25. Hugenholtz, P. & Huber, T. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int. J. Syst. Evol. Microbiol.* **53,** 289–293 (2003).
26. Field, K. G. *et al.* Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Appl. Environ. Microbiol.* **63,** 63–70 (1997).
27. Pybus, O. G. & Rambaut, A. GENIE: Estimating demographic history from molecular phylogenies. *Bioinformatics* **18,** 1404–1405 (2003).
28. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32,** 1363–1371 (2004).
29. Swofford, D. L. *PAUP*. Phylogenetic Analysis Using Parsimony (*And Other Methods) Version 4* (Sinauer Associates, Sunderland, Massachusetts, 2002).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank S. Bertilsson (Uppsala University) for sample collection, D. Veneziano (MIT) for discussions on statistical extrapolation of sequence diversity, P. Chisholm (MIT) for numerous suggestions, and researchers and staff of the PIE-LTER (Plum Island Ecosystem-Long Term Ecological Research) for logistical support. This work was supported by grants from the National Science Foundation, the Department of Energy Genomes to Life program and a postdoctoral fellowship from the Spanish Ministry of Education to S.G.A.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.F.P. (mpolz@mit.edu). The sequences of the cloned partial 16S rRNA genes have been deposited in the GenBank database under accession numbers AY580336–AY580882.

Cambrian origins and affinities of an enigmatic fossil group of arthropods

N. E. Vaccari[1], **G. D. Edgecombe**[2] **& C. Escudero**[3]

[1] *Instituto de Geología y Minería, Universidad Nacional de Jujuy, Avenida Bolivia 1661, San Salvador de Jujuy (4600), Argentina*
[2] *Australian Museum, 6 College Street, Sydney, NSW 2010, Australia*
[3] *O'Higgins y Zarate, 108, Palpalá (4612), Jujuy, Argentina*

Euthycarcinoids are one of the most enigmatic arthropod groups, having been assigned to nearly all major clades of Arthropoda. Recent work has endorsed closest relationships with crustaceans[1] or a myriapod–hexapod assemblage[2], a basal position in the Euarthropoda[3], or a placement in the Hexapoda[4] or hexapod stem group[5]. Euthycarcinoids are known from 13 species ranging in age from Late Ordovician or Early Silurian to Middle Triassic, all in freshwater or brackish water environments[6]. Here we describe a euthycarcinoid from marine strata in Argentina dating from the latest Cambrian period, extending the group's record back as much as 50 million years. Despite its antiquity and marine occurrence, the Cambrian species demonstrates that morphological details were conserved in the transition to fresh water. Trackways in the same unit as the euthycarcinoid strengthen arguments that similar traces of subaerial origin from Cambro-Ordovician rocks were made by euthycarcinoids[7,8]. Large mandibles in euthycarcinoids[6,9] are confirmed by the Cambrian species. A morphology-based phylogeny resolves euthycarcinoids as stem-group Mandibulata, sister to the Myriapoda and Crustacea plus Hexapoda.

Mandibulata Snodgrass, 1938
Euthycarcinoidea Gall and Grauvogel, 1964
Euthycarciniformes Starobogatov, 1988
Apankura gen. nov.

Etymology. *Apankura* (Quechua), meaning crab.

Type species. *Apankura machu* gen. et sp. nov.

Diagnosis. Euthycarciniform with large mandibles that occupy most of the space beneath the posterior cephalic tergite; anterior two pairs of pre-abdominal limbs smaller than the posterior nine pairs; limbs markedly taper distally, composed of about ten podomeres, distal podomeres are shorter, large setae are absent; at least six post-abdominal segments; post-abdominal tergites are each about 2.5-times wider than they are long.

Apankura machu sp. nov.

Etymology. Genus as above; *machu* (Quechua), meaning grandfather.

Holotype. Museo de Geología, Mineralogía y Paleontología, Universidad Nacional de Jujuy (JUY-P 24; Fig. 1).

Locality and horizon. Bed of Río Huasamayo, Garganta del Diablo, near Tilcara, Jujuy Province, Argentina. The holotype (the only known specimen) is in greenish-grey mudstone from the Casa Colorada Member, Santa Rosita Formation. The trilobites *Neoparabolina frequens argentina* and *Plicatolina scalpta* on the same slab indicate a latest Cambrian age (lower part of *Neoparabolina frequens argentina* zone)[10]. Green shales of the Casa Colorada Member represent lower offshore deposition in an open marine facies[11].

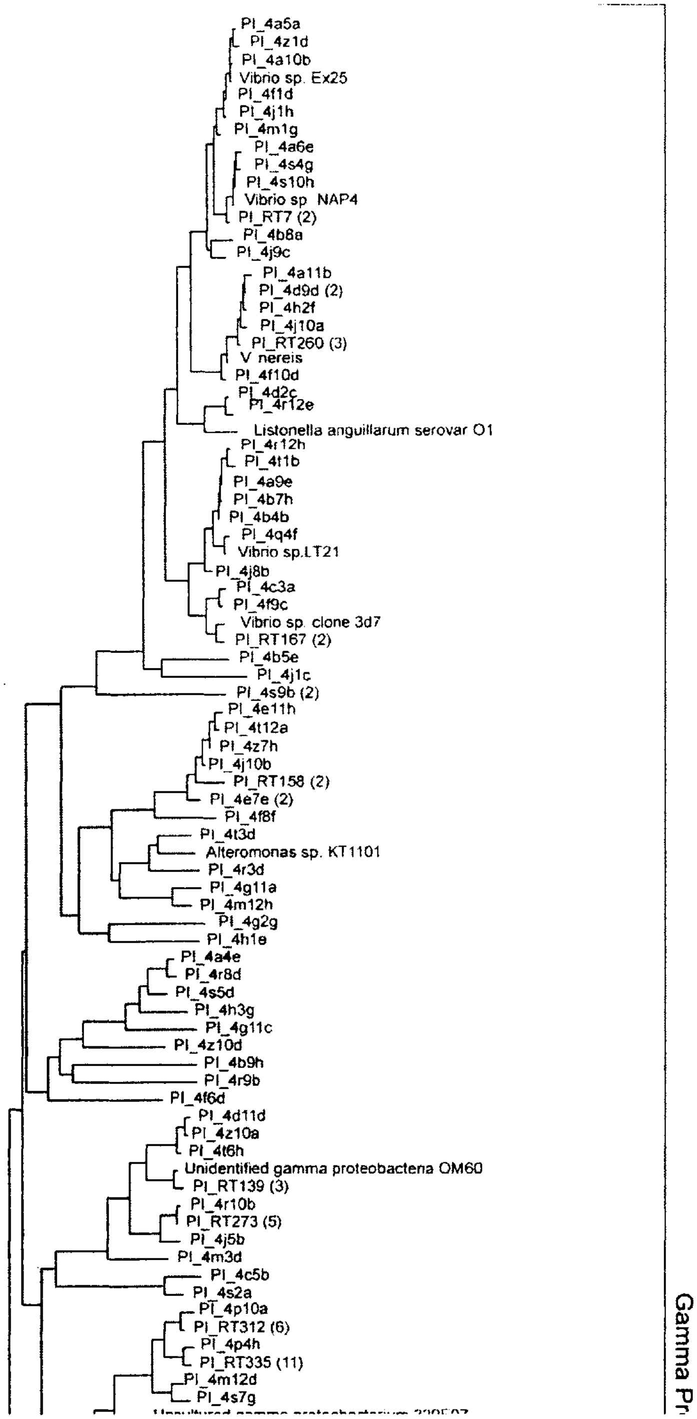
Diagnosis. As for genus.

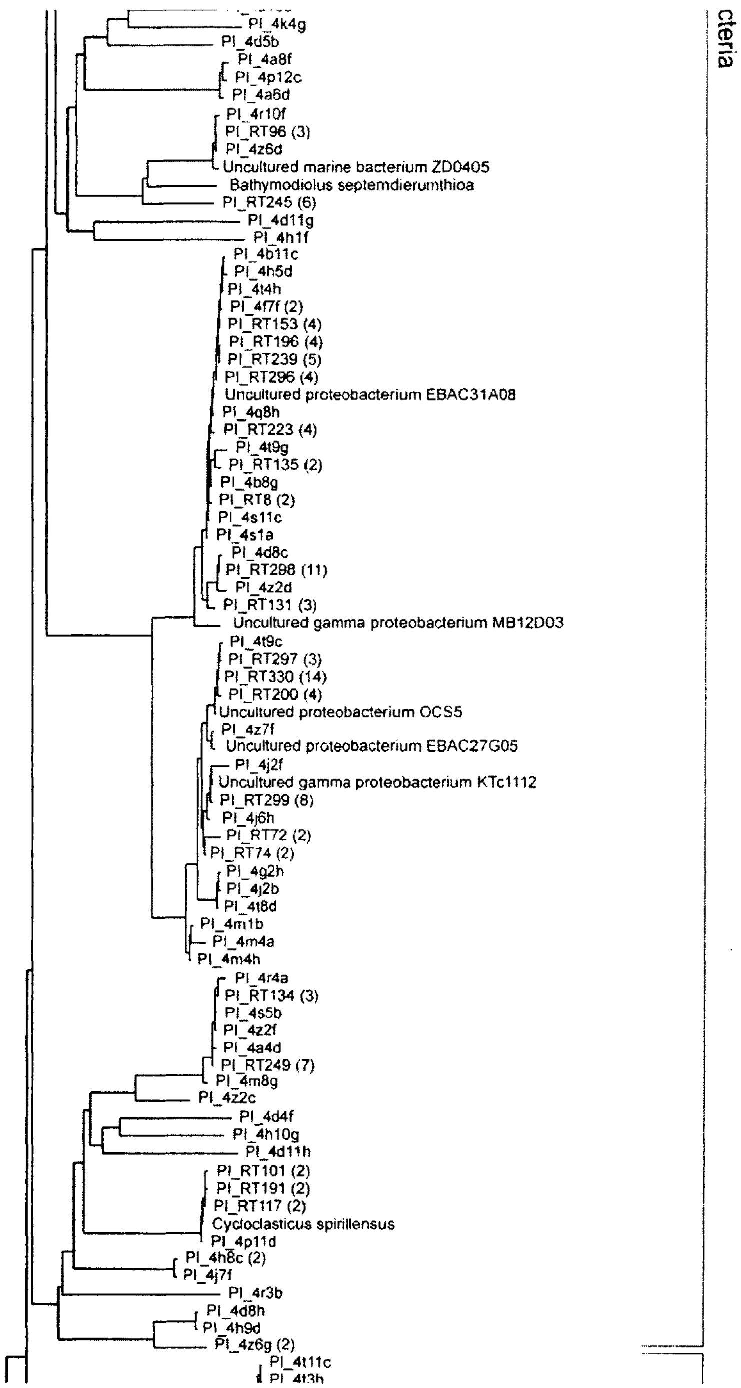
The holotype is 38 mm long, including the head, pre-abdomen and six segments of the post-abdomen. The maximum width of the pre-abdominal tergites is 16 mm. As in other euthycarcinoids[2,12], the head is composed of a short anterior tergite and a longer, wider posterior tergite. The latter is trapezoidal, with gently curved lateral margins. The antenna is uniramous, with at least nine short articles. Large, well-defined spheroidal processes[13] are at the lateral margin

Key to cluster numbers from Figure 2a.

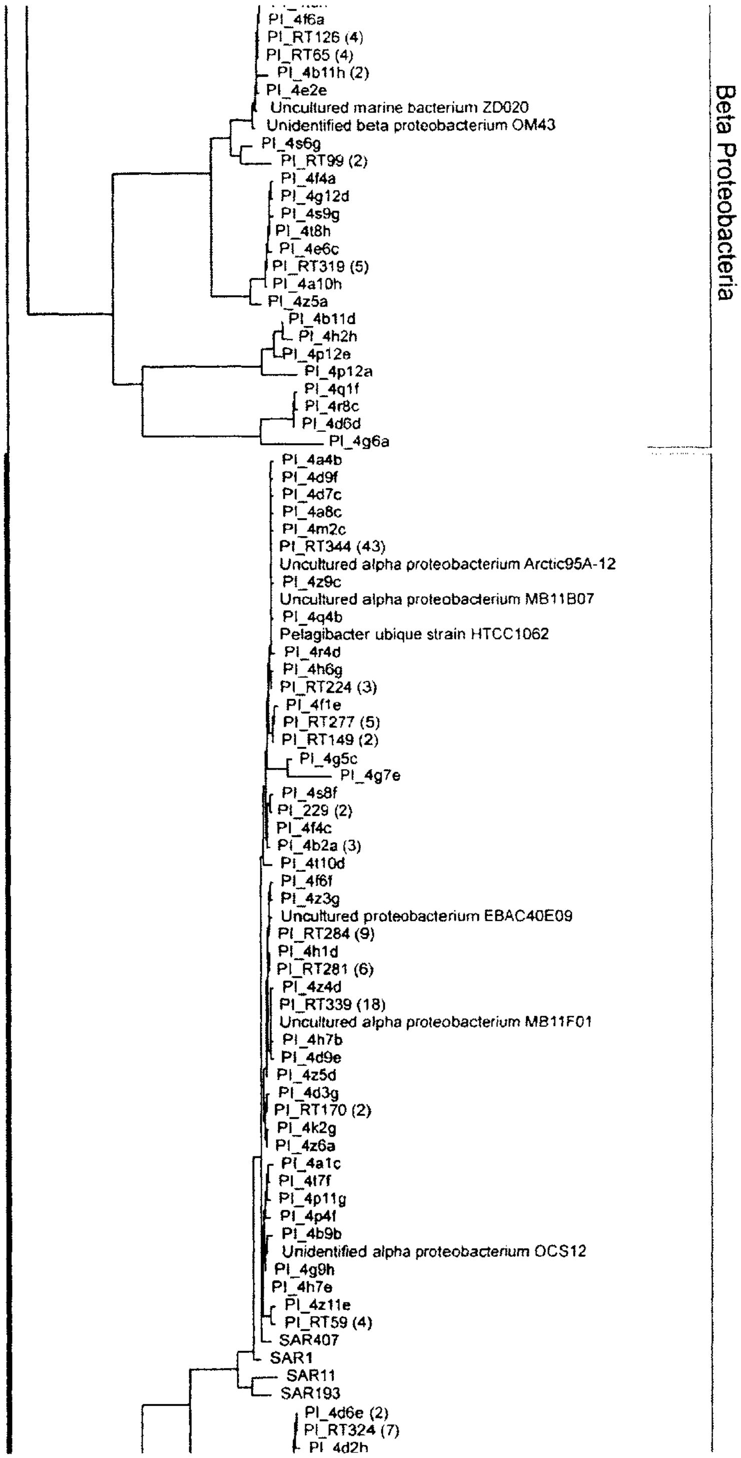
Numbers in parentheses represent ribotypes and sequences associated with each cluster, respectively. 1, Vibrionale*s* (33/40); 2, *Pseudoalteromonas/Shewanella* (12/14); 3, Uncultured Gamma1 (UGAMMA1) (9/9); 4, Uncultured Gamma2 (UGAMMA2) (8/12); 5, *Methylomonas/Methylomicrobium* (8/24); 6, Uncultured Gamma3 (UGAMMA3)/Symbionts (13/20); 7, Uncultured Gamma4 (UGAMMA4) (20/51); 8, Uncultured Gamma5 (UGAMMA5) (13/33); 9, *Cyclocasticus*/Symbionts (20/33); 10, Methylophylaceae (Beta Proteobacteria) (17/29); 11, Comamonadaceae (Beta Proteobacteria) (7/7); 12, Uncultured Actinomycetes1 (UACTINO1) (12/45); 13, *Cryobacterium* (8/11); 14, Uncultured Verrucomicrobiales (7/10); 15, *Planctomyces* (6/7); 16, *Fusibacter* (Clostridiales) (2/2); 17, *Bacteriovorax* (Delta Proteobacteria) (8/8); 18, *Arcobacter* (Epsilon Proteobacteria) (8/11); 19, SAR_surface group (Alpha Proteobacteria) (43/129); 20, SAR_deep group (Alpha Proteobacteria) (17/25); 21, Uncultured Alpha1 (UALPHA1) (2/5); 22, Roseobacter/Roseovarius (29/87); 23, Uncultured Alpha2 (UALPHA2) (8/10); 24, Uncultured Alpha3 (UALPHA3) (1/1); 25, Uncultured Alpha4 (UALPHA4) (18/28); 26, Uncultured Alpha5 (UALPHA5) (3/3); 27, Uncultured Alpha6 (UALPHA6) (6/9); 28, Unknown1 group (UNK1) (5/17); 29, Unknown2 group UNK2 (3/4); 30, Uncultured CFB1 group (UCFB1) (10/13); 31, Uncultured CFB2 group (UCFB2) (7/22); 32, Uncultured CFB3 group (UCFB3) (5/5); 33, Uncultured CFB4 group (UCFB4) (4/5); 34, Cytophagales1 (13/55); 35, *Polaribacter* (5/8); 36, Uncultured CFB5 group (UCFB5) (28/56); 37, Uncultured CFB6 group

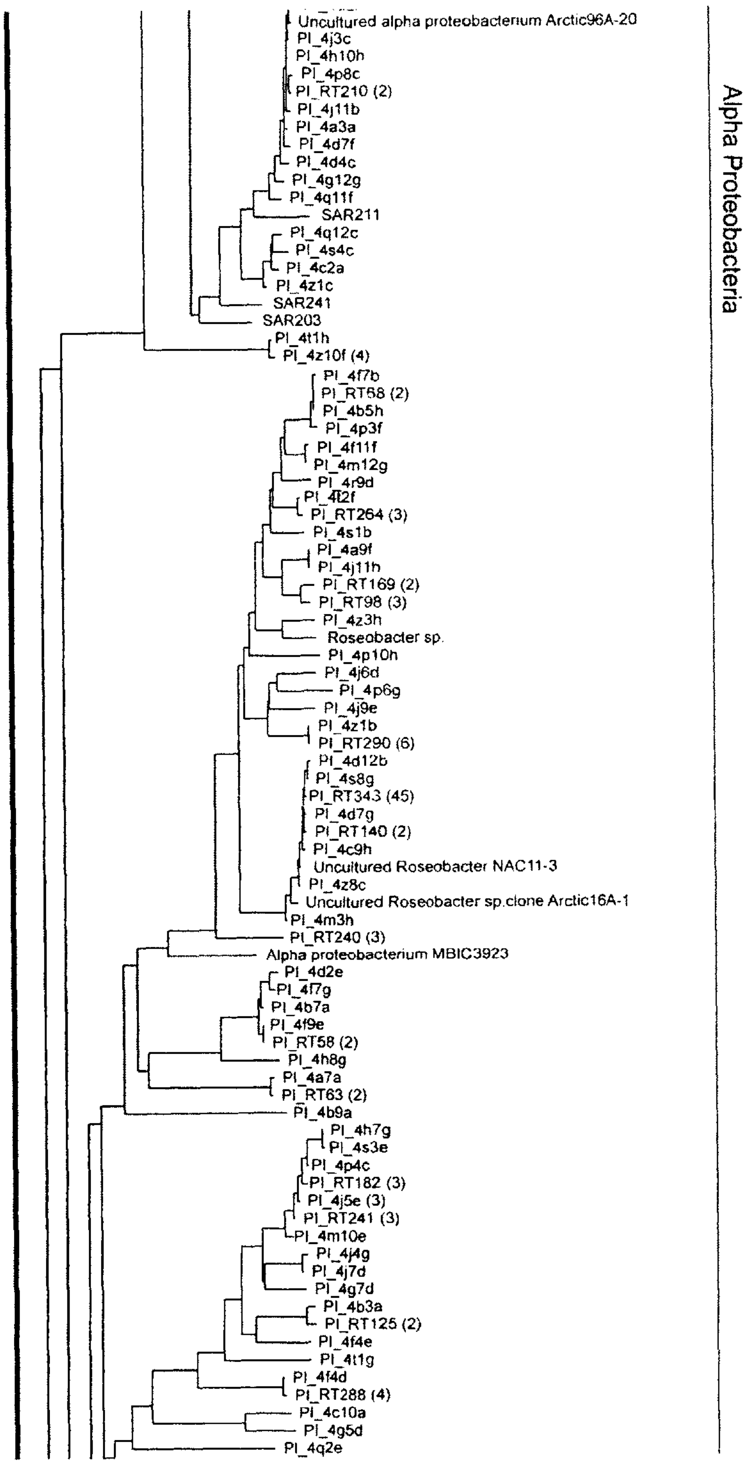
16S rRNA gene sequences (ribotypes) retrieved from the Plum Island bacterioplankton sample and representative reference sequences.

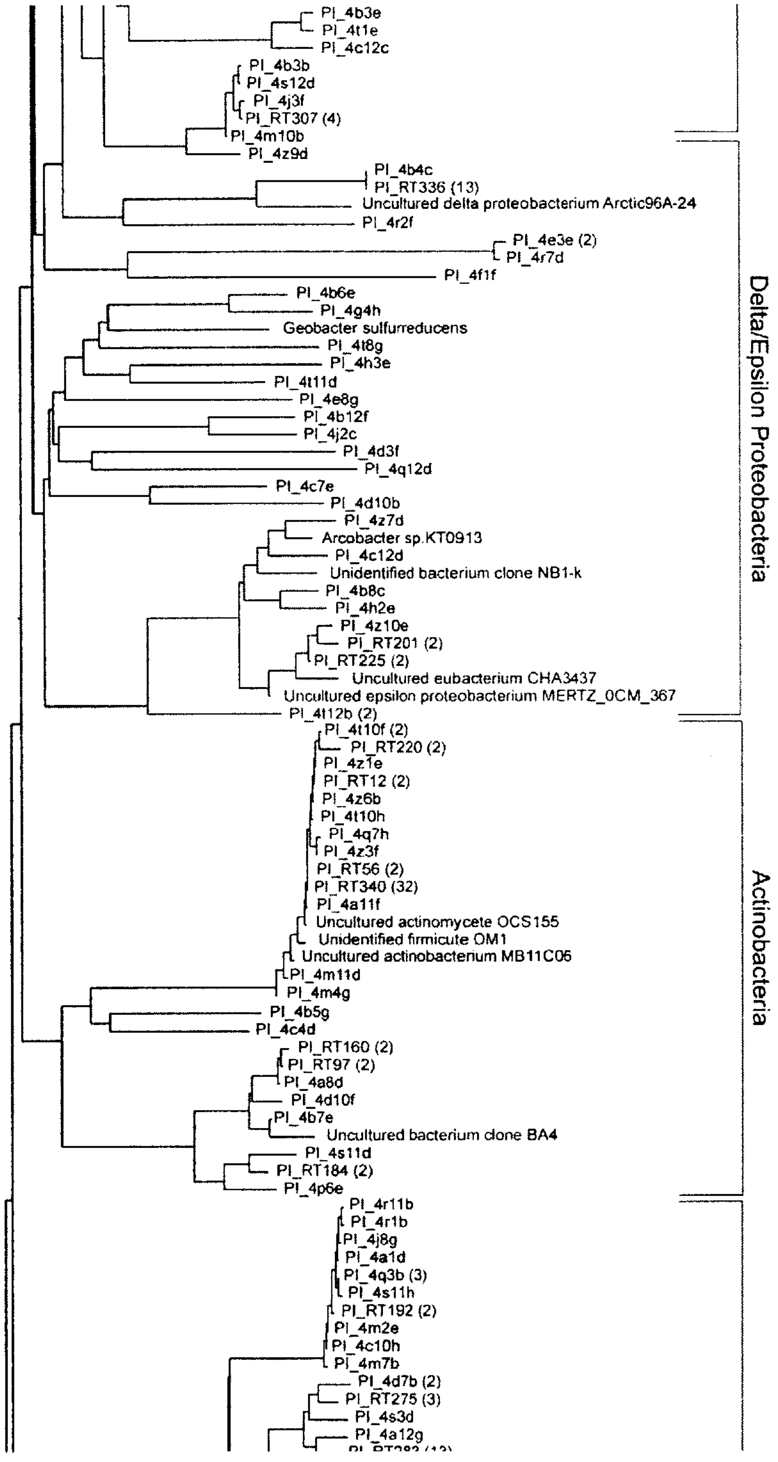

Pl_4a5a
Pl_4z1d
Pl_4a10b
Vibrio sp. Ex25
Pl_4f1d
Pl_4j1h
Pl_4m1g
Pl_4a6e
Pl_4s4g
Pl_4s10h
Vibrio sp NAP4
Pl_RT7 (2)
Pl_4b8a
Pl_4j9c
Pl_4a11b
Pl_4d9d (2)
Pl_4h2f
Pl_4j10a
Pl_RT260 (3)
V nereis
Pl_4f10d
Pl_4d2c
Pl_4r12e
Listonella anguillarum serovar O1
Pl_4r12h
Pl_4t1b
Pl_4a9e
Pl_4b7h
Pl_4b4b
Pl_4q4f
Vibrio sp.LT21
Pl_4j8b
Pl_4c3a
Pl_4f9c
Vibrio sp. clone 3d7
Pl_RT167 (2)
Pl_4b5e
Pl_4j1c
Pl_4s9b (2)
Pl_4e11h
Pl_4t12a
Pl_4z7h
Pl_4j10b
Pl_RT158 (2)
Pl_4e7e (2)
Pl_4f8f
Pl_4t3d
Alteromonas sp. KT1101
Pl_4r3d
Pl_4g11a
Pl_4m12h
Pl_4g2g
Pl_4h1e
Pl_4a4e
Pl_4r8d
Pl_4s5d
Pl_4h3g
Pl_4g11c
Pl_4z10d
Pl_4b9h
Pl_4r9b
Pl_4f6d
Pl_4d11d
Pl_4z10a
Pl_4t6h
Unidentified gamma proteobactena OM60
Pl_RT139 (3)
Pl_4r10b
Pl_RT273 (5)
Pl_4j5b
Pl_4m3d
Pl_4c5b
Pl_4s2a
Pl_4p10a
Pl_RT312 (6)
Pl_4p4h
Pl_RT335 (11)
Pl_4m12d
Pl_4s7g
Uncultured gamma proteobacterium
Gamma Pr

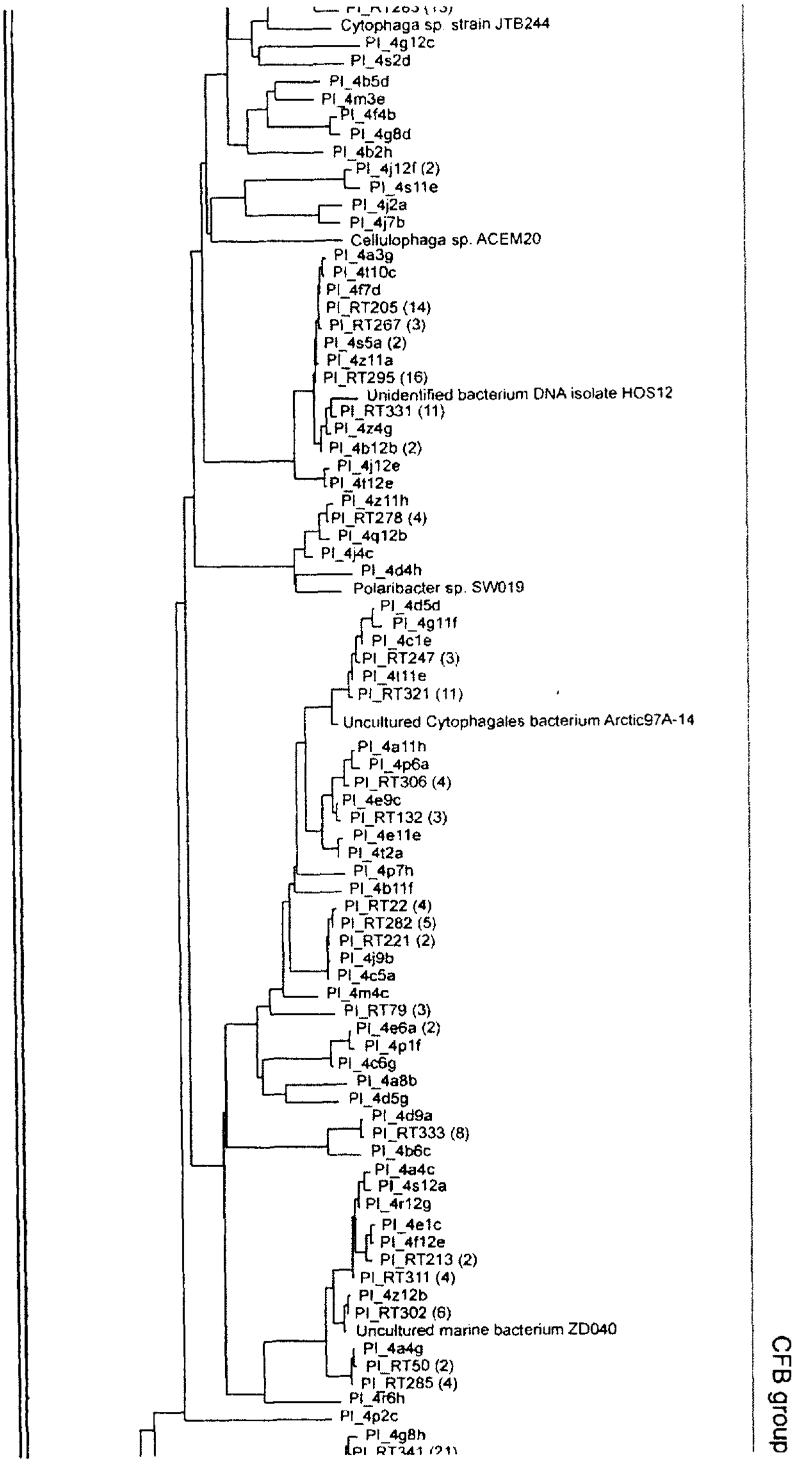


Pl_4k4g
Pl_4d5b
Pl_4a8f
Pl_4p12c
Pl_4a6d
Pl_4r10f
Pl_RT96 (3)
Pl_4z6d
Uncultured marine bacterium ZD0405
Bathymodiolus septemdierumthioa
Pl_RT245 (6)
Pl_4d11g
Pl_4h1f
Pl_4b11c
Pl_4h5d
Pl_4t4h
Pl_4f7f (2)
Pl_RT153 (4)
Pl_RT196 (4)
Pl_RT239 (5)
Pl_RT296 (4)
Uncultured proteobacterium EBAC31A08
Pl_4q8h
Pl_RT223 (4)
Pl_4t9g
Pl_RT135 (2)
Pl_4b8g
Pl_RT8 (2)
Pl_4s11c
Pl_4s1a
Pl_4d8c
Pl_RT298 (11)
Pl_4z2d
Pl_RT131 (3)
Uncultured gamma proteobacterium MB12D03
Pl_4t9c
Pl_RT297 (3)
Pl_RT330 (14)
Pl_RT200 (4)
Uncultured proteobacterium OCS5
Pl_4z7f
Uncultured proteobacterium EBAC27G05
Pl_4j2f
Uncultured gamma proteobacterium KTc1112
Pl_RT299 (8)
Pl_4j6h
Pl_RT72 (2)
Pl_RT74 (2)
Pl_4g2h
Pl_4j2b
Pl_4t8d
Pl_4m1b
Pl_4m4a
Pl_4m4h
Pl_4r4a
Pl_RT134 (3)
Pl_4s5b
Pl_4z2f
Pl_4a4d
Pl_RT249 (7)
Pl_4m8g
Pl_4z2c
Pl_4d4f
Pl_4h10g
Pl_4d11h
Pl_RT101 (2)
Pl_RT191 (2)
Pl_RT117 (2)
Cycloclasticus spirillensus
Pl_4p11d
Pl_4h8c (2)
Pl_4j7f
Pl_4r3b
Pl_4d8h
Pl_4h9d
Pl_4z6g (2)
Pl_4t11c
Pl_4t3h
cteria

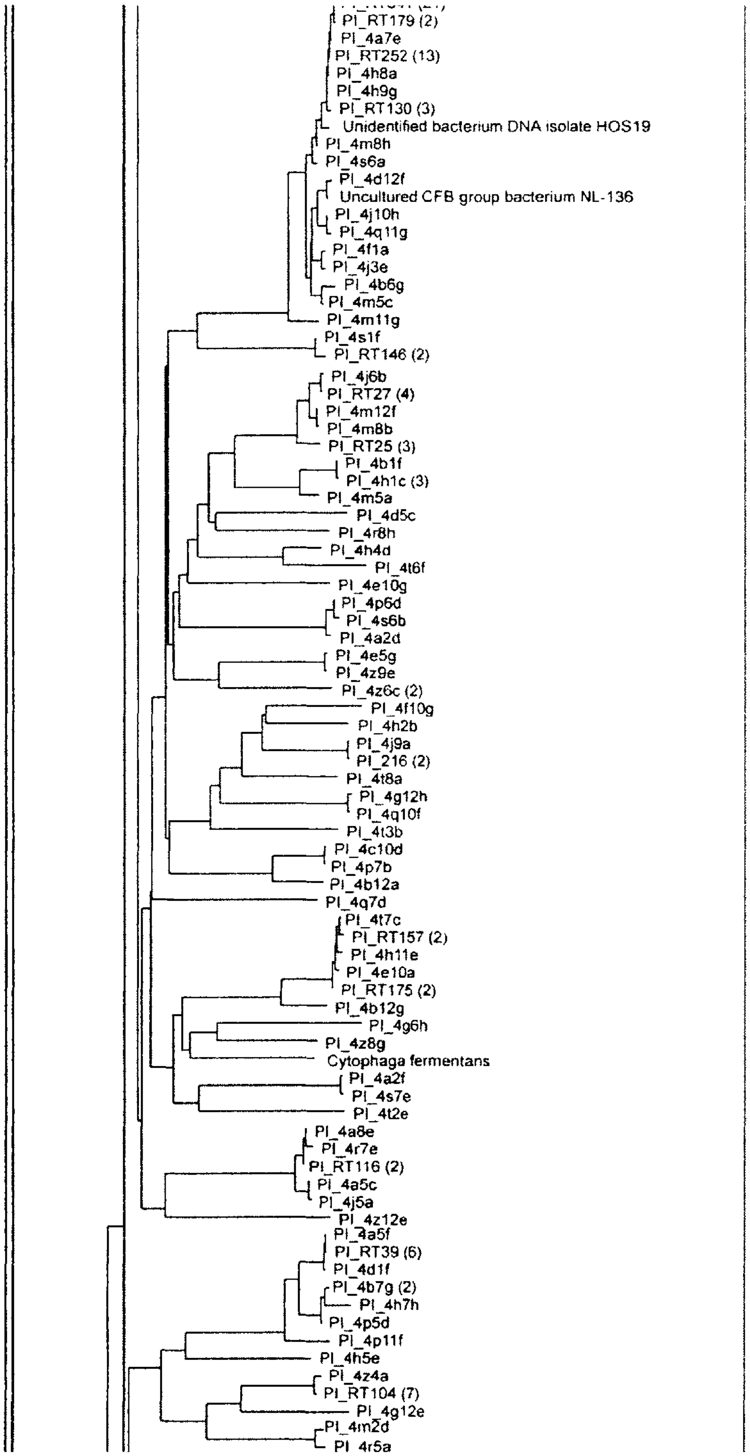

PI_4f6a
PI_RT126 (4)
PI_RT65 (4)
PI_4b11h (2)
PI_4e2e
Uncultured marine bacterium ZD020
Unidentified beta proteobacterium OM43
PI_4s6g
PI_RT99 (2)
PI_4f4a
PI_4g12d
PI_4s9g
PI_4t8h
PI_4e6c
PI_RT319 (5)
PI_4a10h
PI_4z5a
PI_4b11d
PI_4h2h
PI_4p12e
PI_4p12a
PI_4q1f
PI_4r8c
PI_4d6d
PI_4g6a
PI_4a4b
PI_4d9f
PI_4d7c
PI_4a8c
PI_4m2c
PI_RT344 (43)
Uncultured alpha proteobacterium Arctic95A-12
PI_4z9c
Uncultured alpha proteobacterium MB11B07
PI_4q4b
Pelagibacter ubique strain HTCC1062
PI_4r4d
PI_4h6g
PI_RT224 (3)
PI_4f1e
PI_RT277 (5)
PI_RT149 (2)
PI_4g5c
PI_4g7e
PI_4s8f
PI_229 (2)
PI_4f4c
PI_4b2a (3)
PI_4t10d
PI_4f6f
PI_4z3g
Uncultured proteobacterium EBAC40E09
PI_RT284 (9)
PI_4h1d
PI_RT281 (6)
PI_4z4d
PI_RT339 (18)
Uncultured alpha proteobacterium MB11F01
PI_4h7b
PI_4d9e
PI_4z5d
PI_4d3g
PI_RT170 (2)
PI_4k2g
PI_4z6a
PI_4a1c
PI_4t7f
PI_4p11g
PI_4p4f
PI_4b9b
Unidentified alpha proteobacterium OCS12
PI_4g9h
PI_4h7e
PI_4z11e
PI_RT59 (4)
SAR407
SAR1
SAR11
SAR193
PI_4d6e (2)
PI_RT324 (7)
PI_4d2h
Beta Proteobacteria


Uncultured alpha proteobacterium Arctic96A-20
Pl_4j3c
Pl_4h10h
Pl_4p8c
Pl_RT210 (2)
Pl_4j11b
Pl_4a3a
Pl_4d7f
Pl_4d4c
Pl_4g12g
Pl_4q11f
SAR211
Pl_4q12c
Pl_4s4c
Pl_4c2a
Pl_4z1c
SAR241
SAR203
Pl_4t1h
Pl_4z10f (4)
Pl_4f7b
Pl_RT68 (2)
Pl_4b5h
Pl_4p3f
Pl_4f11f
Pl_4m12g
Pl_4r9d
Pl_4t2f
Pl_RT264 (3)
Pl_4s1b
Pl_4a9f
Pl_4j11h
Pl_RT169 (2)
Pl_RT98 (3)
Pl_4z3h
Roseobacter sp.
Pl_4p10h
Pl_4j6d
Pl_4p6g
Pl_4j9e
Pl_4z1b
Pl_RT290 (6)
Pl_4d12b
Pl_4s8g
Pl_RT343 (45)
Pl_4d7g
Pl_RT140 (2)
Pl_4c9h
Uncultured Roseobacter NAC11-3
Pl_4z8c
Uncultured Roseobacter sp.clone Arctic16A-1
Pl_4m3h
Pl_RT240 (3)
Alpha proteobacterium MBIC3923
Pl_4d2e
Pl_4f7g
Pl_4b7a
Pl_4f9e
Pl_RT58 (2)
Pl_4h8g
Pl_4a7a
Pl_RT63 (2)
Pl_4b9a
Pl_4h7g
Pl_4s3e
Pl_4p4c
Pl_RT182 (3)
Pl_4j5e (3)
Pl_RT241 (3)
Pl_4m10e
Pl_4j4g
Pl_4j7d
Pl_4g7d
Pl_4b3a
Pl_RT125 (2)
Pl_4f4e
Pl_4t1g
Pl_4f4d
Pl_RT288 (4)
Pl_4c10a
Pl_4g5d
Pl_4q2e
Alpha Proteobacteria


Pl_4b3e
Pl_4t1e
Pl_4c12c
Pl_4b3b
Pl_4s12d
Pl_4j3f
Pl_RT307 (4)
Pl_4m10b
Pl_4z9d
Pl_4b4c
Pl_RT336 (13)
Uncultured delta proteobacterium Arctic96A-24
Pl_4r2f
Pl_4e3e (2)
Pl_4r7d
Pl_4f1f
Pl_4b6e
Pl_4g4h
Geobacter sulfurreducens
Pl_4t8g
Pl_4h3e
Pl_4t11d
Pl_4e8g
Pl_4b12f
Pl_4j2c
Pl_4d3f
Pl_4q12d
Pl_4c7e
Pl_4d10b
Pl_4z7d
Arcobacter sp.KT0913
Pl_4c12d
Unidentified bacterium clone NB1-k
Pl_4b8c
Pl_4h2e
Pl_4z10e
Pl_RT201 (2)
Pl_RT225 (2)
Uncultured eubacterium CHA3437
Uncultured epsilon proteobacterium MERTZ_0CM_367
Pl_4t12b (2)
Pl_4t10f (2)
Pl_RT220 (2)
Pl_4z1e
Pl_RT12 (2)
Pl_4z6b
Pl_4t10h
Pl_4q7h
Pl_4z3f
Pl_RT56 (2)
Pl_RT340 (32)
Pl_4a11f
Uncultured actinomycete OCS155
Unidentified firmicute OM1
Uncultured actinobacterium MB11C06
Pl_4m11d
Pl_4m4g
Pl_4b5g
Pl_4c4d
Pl_RT160 (2)
Pl_RT97 (2)
Pl_4a8d
Pl_4d10f
Pl_4b7e
Uncultured bacterium clone BA4
Pl_4s11d
Pl_RT184 (2)
Pl_4p6e
Pl_4r11b
Pl_4r1b
Pl_4j8g
Pl_4a1d
Pl_4q3b (3)
Pl_4s11h
Pl_RT192 (2)
Pl_4m2e
Pl_4c10h
Pl_4m7b
Pl_4d7b (2)
Pl_RT275 (3)
Pl_4s3d
Pl_4a12g
Delta/Epsilon Proteobacteria
Actinobacteria



Cytophaga sp. strain JTB244
PI_4g12c
PI_4s2d
PI_4b5d
PI_4m3e
PI_4f4b
PI_4g8d
PI_4b2h
PI_4j12f (2)
PI_4s11e
PI_4j2a
PI_4j7b
Cellulophaga sp. ACEM20
PI_4a3g
PI_4t10c
PI_4f7d
PI_RT205 (14)
PI_RT267 (3)
PI_4s5a (2)
PI_4z11a
PI_RT295 (16)
Unidentified bacterium DNA isolate HOS12
PI_RT331 (11)
PI_4z4g
PI_4b12b (2)
PI_4j12e
PI_4t12e
PI_4z11h
PI_RT278 (4)
PI_4q12b
PI_4j4c
PI_4d4h
Polaribacter sp. SW019
PI_4d5d
PI_4g11f
PI_4c1e
PI_RT247 (3)
PI_4t11e
PI_RT321 (11)
Uncultured Cytophagales bacterium Arctic97A-14
PI_4a11h
PI_4p6a
PI_RT306 (4)
PI_4e9c
PI_RT132 (3)
PI_4e11e
PI_4t2a
PI_4p7h
PI_4b11f
PI_RT22 (4)
PI_RT282 (5)
PI_RT221 (2)
PI_4j9b
PI_4c5a
PI_4m4c
PI_RT79 (3)
PI_4e6a (2)
PI_4p1f
PI_4c6g
PI_4a8b
PI_4d5g
PI_4d9a
PI_RT333 (8)
PI_4b6c
PI_4a4c
PI_4s12a
PI_4r12g
PI_4e1c
PI_4f12e
PI_RT213 (2)
PI_RT311 (4)
PI_4z12b
PI_RT302 (6)
Uncultured marine bacterium ZD040
PI_4a4g
PI_RT50 (2)
PI_RT285 (4)
PI_4r6h
PI_4p2c
PI_4g8h
PI_RT341 (21)
CFB group


PI_RT179 (2)
PI_4a7e
PI_RT252 (13)
PI_4h8a
PI_4h9g
PI_RT130 (3)
Unidentified bacterium DNA isolate HOS19
PI_4m8h
PI_4s6a
PI_4d12f
Uncultured CFB group bacterium NL-136
PI_4j10h
PI_4q11g
PI_4f1a
PI_4j3e
PI_4b6g
PI_4m5c
PI_4m11g
PI_4s1f
PI_RT146 (2)
PI_4j6b
PI_RT27 (4)
PI_4m12f
PI_4m8b
PI_RT25 (3)
PI_4b1f
PI_4h1c (3)
PI_4m5a
PI_4d5c
PI_4r8h
PI_4h4d
PI_4t6f
PI_4e10g
PI_4p6d
PI_4s6b
PI_4a2d
PI_4e5g
PI_4z9e
PI_4z6c (2)
PI_4f10g
PI_4h2b
PI_4j9a
PI_216 (2)
PI_4t8a
PI_4g12h
PI_4q10f
PI_4t3b
PI_4c10d
PI_4p7b
PI_4b12a
PI_4q7d
PI_4t7c
PI_RT157 (2)
PI_4h11e
PI_4e10a
PI_RT175 (2)
PI_4b12g
PI_4g6h
PI_4z8g
Cytophaga fermentans
PI_4a2f
PI_4s7e
PI_4t2e
PI_4a8e
PI_4r7e
PI_RT116 (2)
PI_4a5c
PI_4j5a
PI_4z12e
PI_4a5f
PI_RT39 (6)
PI_4d1f
PI_4b7g (2)
PI_4h7h
PI_4p5d
PI_4p11f
PI_4h5e
PI_4z4a
PI_RT104 (7)
PI_4g12e
PI_4m2d
PI 4r5a

CHAPTER FOUR

Estimating the diversity of bacterial communities

To be submitted with Daniele Veneziano and Martin F. Polz as co-authors.

Abstract

Reliable estimates of diversity are a prerequisite for many studies of community assembly, environmental function, and biogeography. However, for microorganisms, diversity assessments have only recently become possible through the advancement of molecular techniques and application of statistical methods generally developed for estimating diversity of macroorganisms. Currently, the most commonly used diversity estimator in microbial ecology is Chao1, however, theoretical studies have suggested it underestimates the diversity of complex microbial communities. This is confirmed by our analysis of a large dataset of 16S rRNA sequences derived from a single microbial sample. Here, we consider existing parametric approaches, as these should perform better for very diverse communities. We modify them to better account for the specific way microbial communities are sampled. The number of taxa and other model parameters are estimated using maximum likelihood (ML). Of the tested distributions for abundances of bacterial types, the lognormal distribution, which is commonly used for microbial communities, results in the best fit to our data. The number of taxa estimated using our parametric approach and the lognormal distribution is one order of magnitude higher than the value given by the Chao1 estimator. Through simulation, we show that the difference is due to bias of the latter estimator and that the diversity within a complex marine microbial community is considerably higher than previously observed.

Introduction

Prokaryotes are by far the most abundant and diverse organisms. Their global abundance is estimated to be 4-6 x 10^{30} cells (Whitman et al., 1998) and molecular surveys have revealed an astounding microbial diversity that was previously undetected by culture-dependent methods (Head et al., 1998; Hugenholz et al., 1998; Rappe and Giovannoni, 2003). However, the quantification of microbial diversity has remained a technical challenge because molecular surveys have not produced data sets large enough to critically evaluate statistical methods that were only recently introduced to microbial ecology (Hughes et al., 2001).

As most bacteria remain unculturable, the most accurate way to estimate microbial diversity is to use sequence data from genes obtained directly from environmental samples. This involves extraction of environmental DNA, polymerase chain reaction (PCR) amplification of target genes (most commonly ribosomal 16S rRNA genes), clone library construction from amplified gene fragments, and gene sequencing (Head et al., 1998). Typically, sequences are clustered into unique rRNA sequences (ribotypes) or clusters of ribotypes based on 1-5% sequence divergence (Martin, 2002). The diversity of ribotypes or clusters in a clone library is then estimated using statistical approaches (Dunbar et al., 1999; Hughes et al., 2001; Stach et al., 2003).

The many existing methods for estimating organismal diversity fall into two general categories: (i) parametric methods, where the abundance distribution of taxa is assumed to have a specified parametric form, and (ii) non-parametric methods, where no abundance distribution model is assumed (May, 1975; Colwell and Coddington, 1994;

Krebs, 1999). In microbial ecology, the most commonly applied richness estimator is the non-parametric Chao1 estimator (Hughes et al., 2001; Bohannan and Hughes, 2003). It calculates diversity of organisms based on the number of taxa represented in the sample by one individual (singletons) and two individuals (doubletons) (Chao, 1984, 1987). Since the Chao1 estimator depends only on the number of singletons and doubletons, it is very simple to use. However, it has been noted that the Chao1 estimator gives biased (low) estimates of diversity, especially for very heterogeneous communities (Mao and Lindsay, 2001; Bohannan and Hughes, 2003).

Parametric analysis has only rarely been used to estimate microbial diversity. A major reason is that parametric methods are computationally more demanding than non-parametric alternatives. In addition, it has been argued that the need to assume an underlying abundance distribution of taxa in a clone library is disadvantageous because existing data sets are not large enough to support the choice of a particular abundance model (Bohannan and Hughes, 2003). However, if a distribution of abundances can be chosen based on theoretical or empirical observations, parametric models should estimate the diversity more accurately than non-parametric approaches.

In this paper we show that for complex microbial communities the commonly used Chao1 estimator is highly biased and severely underestimates diversity. Since our data set is the largest derived from a single environmental sample to date (sample size = 1,033 16S rRNA sequences), we consider as an alternative, various parametric maximum-likelihood estimators to better account for the way microbial communities are sampled. We compare the modified parametric estimators with existing ones and the

Chao1 estimator, and evaluate the performance of different estimators by sampling from the bacterioplankton data set and simulated communities.

Modification of a parametric model and ML formulation

Here we consider (i) the formulation of parametric abundance models and sampling methods appropriate for microbial communities, and (ii) estimation of the model parameters using maximum likelihood. We first describe abundance distribution models for taxa in the population and the sample. This is necessary because, the distribution of taxon abundance in the sample depends on the distribution of taxon abundance in the population and the sampling procedure used. Second, we provide a detailed description of the maximum likelihood formulation. Lastly, we evaluate how well different parametric models fit the data and compare the results with diversity estimates obtained by non-parametric approaches.

Distribution of taxon abundances in a population

Clone libraries constructed from complex natural environments are typically highly heterogeneous, being comprised of a few dominant and many rare taxa (Dunbar et al., 1999; Stach et al., 2003; Acinas et al., 2004a; Venter and al., 2004). In nature, abundances vary significantly from taxon to taxon and can be represented by a random variable with a particular distribution form. Lognormal and gamma distributions have been widely used for highly heterogeneous communities of macroorganisms and are reviewed below (Fisher et al., 1943; Preston, 1948; Kempton and Taylor, 1974; May, 1975).

The lognormal distribution has probability density function:

$$q(r) = \frac{1}{r\sigma\sqrt{2\pi}} e^{\frac{-(\ln r - \mu)^2}{2\sigma^2}}, \qquad [1]$$

where $\ln(r)$ is normally distributed with mean value μ and variance σ^2.

It has been argued that the abundances of microbial taxa should be lognormally distributed (Curtis et al., 2002). For example, May (1975) proposed that the highly dynamic and random growth, such as that commonly observed for microorganisms in natural communities, should lead to a lognormal distribution of taxa. Furthermore, multiple biotic and abiotic interactions, such as nutrient composition and availability, light, temperature, and competition, should also lead to such a distribution (May, 1975).

Another commonly used abundance model is the gamma distribution, with density:

$$q(r) = \frac{1}{\beta_R^{\alpha}\Gamma(\alpha)} r^{\alpha-1} e^{-\frac{r}{\beta_R}}, \qquad [2]$$

where $\Gamma(\alpha)$ is the gamma function, $\alpha>0$ is a shape parameter, and $\beta>0$ is a scale parameter. The variable r in Eq. [2] has mean value $\alpha\beta_R$ and variance $\alpha\beta_R^2$ (Engen and Lande, 1996; Diserud and Engen, 2000).

There are two special cases of the gamma distribution, which have been proposed for highly heterogeneous communities: the broken stick distribution (MacArthur, 1957) and the logarithmic distribution (Fisher et al., 1943). When the shape parameter α

equals one, the gamma distribution is exponential (Plotkin and Muller-Landau, 2002), which is a continuous version of MacArthur's broken stick model (Cohen, 1968). MacArthur's broken stick distribution is to be expected when there is a fixed amount of some governing resource that ecologically homogeneous taxa divide up among themselves in an independent way (May, 1975). When $\alpha \rightarrow 0$, the gamma distribution becomes a logseries distribution (Plotkin and Muller-Landau, 2002), which is a continuous version of Fisher's logarithmic (logseries) model (Fisher et al., 1943). Fisher's logarithmic model is appropriate for taxa that compete for a single resource, as in the broken stick model, but ecologically homogeneous taxa divide up the resource in some dependent fashion (May, 1975).

Distribution of taxon abundances in a sample and ML parameter estimation

The number of individuals of a given taxon observed in a sample is a random variable whose distribution depends on the total abundance of that taxon in the sampled community, the sample size, and the method of sampling. Given the abundances of the taxa in the community, the corresponding abundances in the sample are typically assumed to have independent Poisson distributions (Preston, 1948; Bulmer, 1974). While often not exact, the independent Poisson assumption leads to relatively simple analysis and gives a good approximation to more accurate, but more complex sampling distributions.

Sampling methods for macroorganisms are generally different from those for microorganisms. There are two main methods of sampling: exhaustive or complete (in time or space) and random with fixed size (number of sampled individuals). Plants and insects are typically sampled exhaustively inside quadrats and from traps, respectively. In this kind of sampling, the total sample size is not known *a priori* as everything within a given space or time volume is collected. In contrast, clone libraries are often sampled using a fixed sample size. As we shall see, main difference between the two sampling strategies is in the mean value of the sample abundances.

Compound Poisson model for exhaustive sampling

We first describe the maximum-likelihood formulation for exhaustive sampling and then show how the formulation is modified when the sample size is controlled.

A likelihood method for exhaustive independent Poisson sampling was recently described by Mao and Lindsay (2001). This model applies when the space-time distributions of individuals belonging to different taxa are independent Poisson variables and sampling is exhaustive inside a certain region. Let N be the unknown total number of taxa to be estimated and denoted by m_i the sample abundance for taxon i with possible values 0, 1, 2, ... The number of taxa with sample abundance m_i=m is denoted by n_m. Hence $\sum_{m=1}^{\infty} n_m = N$. In this case, that n_1, n_2, ..., are observed but n_0 is not, and that $\sum_{m=1}^{\infty} n_m = n$, where n is the total number of taxa observed in the sample. In the exhaustive sampling model, the m_i' s are taken to be Poisson random variables with mean values λ_i, i = 1,...N. The mean values λ_i are proportional to the population abundances and are treated as independent random variables with the same probability distributions (usually lognormal or gamma) with $Q(\lambda)$. Given Q, the probability $f(m|Q)$ that a generic taxon i has sample abundance $m_i = m$ is:

$$f(m \mid Q) = \int_0^\infty \frac{1}{m!} \lambda^m e^{-\lambda} dQ(\lambda). \qquad [3]$$

In this case $f(m|Q)$ does not depend on the number of taxa in the community, N. The likelihood function of (N,Q) has the form:

$$\begin{aligned} l(N,Q \mid \{n_m\}) &\propto f(n_0 = N - n, \{n_m, m > 0\} \mid N,Q) = \\ &= \frac{N!}{(N-n)! \prod_{m=1} n_m!} f(0 \mid Q)^{N-n} \prod_{m=1} f(m \mid Q)^{n_m} \end{aligned} \qquad [4]$$

The factor $\frac{N!}{(N-n)! \prod_{m=1} n_m!}$ in eqn [4] is a multinomial coefficient, which gives the number of ways in which sample abundances m_i, $i = 1,\ldots,N$ can be ordered. The likelihood function in eqn [4] can be factored into a binomial probability for the number of distinct taxa observed in the sample (n) and a multinomial probability for the observed frequency counts $\{n_m, m > 0\}$ given n. These two factors are

$$f(n \mid N,Q) = \frac{N!}{n!(N-n)!} f(0 \mid Q)^{N-n} [1 - f(0 \mid Q)]^n \qquad [5]$$

and

$$f(\{n_m, m > 0\} \mid N,Q,n) = \frac{n!}{\prod_{m=1} n_m!} \prod_{m=1} \left(\frac{f(m \mid Q,n)}{1 - f(0 \mid Q,n)} \right)^{n_m}, \qquad [6]$$

respectively. Note that, in the likelihood function, the factor $\frac{n!}{\prod_{m=1} n_m!}$ in eqn [6] may be omitted.

The probabilities in eqn [6] do not depend on the total number of taxa in the community *N*. However, the distribution *Q* appears in both likelihood components, and thus inference of *Q* must be based on the entire likelihood function (eqn [4]).

Compound Poisson model for fixed sample size

Since clone libraries contain a finite number of individuals and are sampled without replacement, for given population abundances $M_1, M_2,\ldots, M_N$, the abundances $m_1, m_2,\ldots, m_N$, in a sample of fixed size *s* have hypergeometric distribution. However, because of the large number of individuals in each taxon, the hypergeometric distribution may be replaced with a multinomial distribution, which describes sampling with replacement or sampling from an infinite number of individuals. Furthermore, if the sample size *s* is large, the multinomial distribution of the m_i's may be approximated with an independent Poisson distribution with the same mean values $\lambda_i = \frac{M_i}{\sum_{j=1}^{N} M_j} s = R_i \frac{s}{N}$ where $R_i = \frac{M_i}{\overline{M}}$ is the relative abundance of taxon *i* in the population and $\overline{M} = \frac{1}{N}\sum_{j=1}^{N} M_j$. In the Poisson approximation, the sample size is not fixed but is a random variable with Poisson distribution and mean value equal to the actual sample sizes. The coefficient of variation is $1/\sqrt{s}$ and is therefore small for sample sizes of the order of 1,000 individuals, used here to characterize microbial diversity. If the M_i's are independent and identically distributed, the R_i's are also identically distributed, but are dependent. For simplicity, we assume that the R_i's are also independent with some distribution Q, inherited from the distribution of the abundance M_i.

Under the above modeling assumptions and approximations, the probability that a taxon is observed m times is given by:

$$f(m \mid N,Q) = \frac{s^m}{m!} \int_0^{\infty} (\frac{r}{N})^m e^{-\frac{sr}{N}} dQ(r) \quad . \qquad [7]$$

Notice that, contrary to the exhaustive sampling, *f(m)* now depends not only on Q but also on N. In this case, the likelihood function for N and Q is given by:

$$l(N,Q \mid \{n_m\}) = \frac{N!}{(N-n)! \prod_{m=1} n_m!} f(0 \mid N,Q)^{N-n} \prod_{m=1} f(m \mid N,Q)^{n_m} \qquad [8]$$

The likelihood function may be factored in analogy with eqns [5] and [6]. To avoid calculation of the binomial coefficient $\frac{N!}{n!(N-n)!}$ in eqn. [6] one can approximate the binomial probability of *f(n|N,Q)* with the density at n of the normal distribution having the same mean and variance

$$\begin{cases} E[n \mid N,Q] = N[1 - f(0 \mid N,Q)] \\ Var[n \mid N,Q] = Nf(0 \mid N,Q)[1 - f(0 \mid N,Q)] \end{cases} \quad . \qquad [9]$$

This gives:

$$l(N,Q \mid \{n_m\}) \propto f(n_0 = N - n, \{n_m, m > 0\} \mid N,Q) = f(n \mid N,Q) \times f(\{n_m, m > 0\} \mid N,Q,n) =$$

$$= \frac{1}{\{Nf(0 \mid N,Q)[1 - f(0 \mid N,Q)]\}^{1/2}} e^{-\frac{(n - N[1 - f(0|N,Q)])^2}{2Nf(0|N,Q)[1 - f(0|N,Q)]}} \times \prod_{m=1} \left(\frac{f(m \mid N,Q)}{1 - f(0 \mid N,Q)} \right)^{n_m} \quad .[10]$$

Eqn. [10] is similar to the likelihood explained by Bulmer (1974). However, Bulmer's formulation does not contain the first likelihood term $f(n \mid N,Q)$. Bulmer (1974) stated that the first part of the distribution is unnecessary, since n_0, the number of taxa not represented in the sample, is unknown. A comparison of the likelihood functions and the associated ML estimates of N and σ will be made later in this chapter.

Next, we specialize the present formulation for the cases when Q is a lognormal or a gamma distribution.

$f(m|N,Q)$ for Poisson-lognormal model

Suppose that the relative abundances R_i's have lognormal distribution. Specifically, $\ln R_i$ has normal distribution with variance σ^2 and mean $\mu = -\frac{1}{2}\sigma^2$, so that $E[R_i]=1$. Then R has probability density function:

$$q(r) = \frac{1}{r\sigma\sqrt{2\pi}} e^{\frac{-(\ln r + \frac{\sigma^2}{2})^2}{2\sigma^2}}. \qquad [11]$$

and eqn. [7] becomes:

$$f(m \mid N,Q) = \frac{s^m}{m!}\int_0^\infty (\frac{r}{N})^m e^{-\frac{sr}{N}} \frac{1}{r\sigma\sqrt{2\pi}} e^{\frac{-(\ln r + \frac{\sigma^2}{2})^2}{2\sigma^2}} dr. \qquad [12]$$

f(m|N,Q) for Poisson-Gamma model

Suppose now that the relative abundances R_i's have gamma(α,β_R) distribution with unit mean value and variance σ_R^2. Since this distribution has mean value $\alpha\beta_R$ and variance $\alpha\beta_R^2$, it follows that

$$\begin{cases} \alpha = \dfrac{1}{\beta_R} = \dfrac{1}{\sigma_R^2} \\ \beta_R = \sigma_R^2 \end{cases} . \qquad [13]$$

It also follows that the mean sample abundances $\lambda_i = R_i \frac{s}{N}$ are iid with gamma(α,β) distribution and parameters

$$\begin{cases} \alpha = \dfrac{1}{\beta_R} = \dfrac{1}{\sigma_R^2} \\ \beta = \dfrac{s}{N}\beta_R = \dfrac{s}{N}\sigma_R^2 \end{cases} . \qquad [14]$$

In this case, the compound Poisson-gamma distribution in eqn [7] is a negative binomial distribution (Fisher et al., 1943). Specifically,

$$f(m\,|N,\sigma_R^2) = \binom{\alpha+m-1}{\alpha-1}\left(\frac{\beta}{\beta+1}\right)^m\left(\frac{1}{\beta+1}\right)^\alpha, \qquad [15]$$

where $\beta = \beta_R \frac{s}{N}$, and $\binom{\alpha+m-1}{\alpha-1} = \frac{\Gamma(\alpha+m)}{\Gamma(\alpha)m!} = \frac{\alpha+m-1}{m} \times \frac{\alpha+m-2}{m-1} \times ... \times \frac{\alpha}{1}$.

Results

Analysis of a large microbial data set and validation of the modified parametric ML estimators

We apply the diversity estimates derived above to the largest available 16S rRNA clone library from a single microbial sample (Acinas et al., 2004b). The library was constructed from a complex marine bacterioplankton community collected on 6 October 2001 from the marine end of the Parker River Estuary, MA. The sample consists of 1,033 16S rRNA gene sequences. Among these, 516 are unique rRNA sequences (ribotypes) and approximately 50% of these sequences occurred only once in the sample. The observed values of ribotype abundances are given in Table 1. Construction of the 16S rRNA library, corrections for sequence artifacts, and a detailed phylogenetic analysis of this bacterioplankton data set is described elsewhere (see chapter 3).

The maximum likelihood of the bacterioplankton data set applying the Poisson-lognormal model estimated 25,000 ribotypes with the standard deviation of the taxa log-abundances of 2.7. The likelihood for our Poisson-lognormal model, eqn [8], is shown in Figure 1A as a function of number of distinct taxa in the library, N, and the standard deviation of the taxa log-abundances, σ. This likelihood is the product of two terms, which are given by eqns [6] and [5] and plotted separately in Figures 1B and 1C, respectively. Figure 1B corresponds to Bulmer's (1974) likelihood formulation and ignores the fact that $(N\text{-}n)$ taxa were not observed in the sample. This likelihood component constrains well the total number of taxa N, but is less informative on $(\sigma|N)$. This can be seen from the separation of the contour lines in the vertical direction (Fig.

1B). The second component, eqn [5] and Figure 1C, is also important: it provides little additional constraint on N, but imposes a clear relation between σ and N. Thus, both components are necessary to describe maximum-likelihood formulation for random sampling with fixed number of sampled individuals.

The bias and variance of ML estimators in our likelihood formulation of N and σ and the ML estimators in Bulmer's formulation were evaluated using simulated communities and samples. We simulated 10 communities with N and σ set to 25,000 and 2.7, which are the ML estimates obtained assuming the lognormal distribution (Fig. 1A). From each simulated community, 10 samples of 1,033 individuals were drawn at random, as in the original bacterioplankton data set. For each sample, the ML estimates using eqn [6] (Bulmer's likelihood formulations) and eqn [10] (our likelihood formulation) were obtained (Fig. 1D). The average value and the standard deviation of the estimates of the 10 populations obtained for both parametric ML estimators are shown in Figure 1D. The average values of N and σ of 100 ML estimates were $23{,}061 \pm 2.47 \times 10^8 (1\sigma^2)$ and $2.56 \pm 0.041 (1\sigma^2)$ for our likelihood, and $29{,}301 \pm 4.60 \times 10^8 (1\sigma^2)$ and $2.69 \pm 0.075 (1\sigma^2)$ for Bulmer's likelihood, respectively. Thus, the variances were smaller for the our likelihood. Even the average variance of N and σ of the 10 population variances is lower for our likelihood (1.32×10^8 and 0.024) than for the Bulmer's likelihood formulation (2.81×10^8 and 0.052). Thus, although the two functions are comparable, the modified likelihood results in lower uncertainty of the maximum likelihood estimate of N and σ.

The bias and variance was also evaluated for the Chao1 estimator using the same simulated communities and samples as with the parametric ML estimators (Fig 1D). For each sample, the Chao1 estimates were obtained, averaged and compared with the values obtained by ML for the bacterioplankton sample (Fig. 1D). The average of the 10 Chao1 values obtained from simulated lognormal communities equals 1673 ±133($1\sigma^2$) ribotypes. Thus, the Chao1 values are more than 5 standard deviations away from the mean of the MLE for the bacterioplankton data and simulated communities. Overall, this suggests that the ribotype abundances in the library are well explained by the lognormal distribution and that the Chao1 significantly underestimates this diversity.

To validate the lognormal abundance model, we compared the empirical expected rarefaction curve with the theoretical mean curve for the lognormal model and the gamma model using the modified parametric ML approach. The Poisson-lognormal model (Fig. 1A) has the best fit to our data set for N equals 25,000 ribotypes and σ of the lognormal distribution equals 2.7 (ML value equals –561.86). The ±1 standard deviation (σ) away from the MLE mean ranges from (N=13,500, ln(σ)=2.4) to (N=49,000, ln(σ)=2.9). For the Poisson-gamma model (Fig. 2) the MLE is 5.91×10^{10} ribotypes and the ln(σ) equals nine (ML value > –630). The ±1σ ranges from (N=1.36×10^5, ln(σ)=2.9) to some unbounded value of N. Using the underlying distribution (lognormal or gamma) and the corresponding MLE value, one can generate a rarefaction curve by calculating the observed taxa in the sample given the sample size. Comparing such rarefaction curves to a rarefaction curve of the data as well as the ML values, we determined that the lognormal abundance model fits the clone library data most closely (Fig. 3A).

Poisson-gamma model fits the ribotype data more poorly than the Poisson-lognormal model. The MLE for the Poisson-gamma model (Fig. 2) gives a very large number of ribotypes, $>10^{15}$, which exceeds the number of individuals present in the sample and therefore is not believable. Interestingly, with essentially equal confidence (within one σ away from the mean of the MLE), N can have values between10^5 and $>10^{15}$ (Fig 2.). These estimates lie on the ridge defined by the $\frac{N}{\sigma^2}$ =constant, where N is the total number of taxa in a sampled community and σ^2 is the variance of the gamma distribution. However, the Poisson-gamma fits the rarefaction curve badly (Fig. 3A), and thus, we can discount these results. The two special cases of the Poisson-gamma distributions, Fisher's logarithmic (Fisher et al., 1943) and MacArthur's broken stick (MacArthur, 1957) distributions, fit the rarefaction curve even more poorly (Fig. 3A).

A comparison of the Chao1 and Curtis estimators to the Poisson-lognormal model

The best estimate of the number of ribotypes in the library using the Poisson-lognormal model (25,000 ribotypes) is over one order of magnitude higher than the value of 1,633 obtained by applying the Chao1 estimator (Chao, 1984, 1987). We fixed the value of the total number of ribotypes to the Chao1 value of 1,633 and under this constraint determined the most likely value of the associated standard deviation of the lognormal distribution (σ). Using the lognormal distribution and the corresponding MLE value, we generated a rarefaction curve for the Chao1 by calculating the observed taxa in the sample given the sample size. The expected rarefaction curve for the Chao1 for the lognormal community is given in the Figure 3A. The shape of the Chao1 rarefaction curve is more concave down compared to the shape of the ribotype accumulation curve because it has a lower asymptote (Fig. 3A). Also, in Figure 3A, we show the fit to the rarefaction curve of another estimator, recently developed by Curtis et al. (Curtis et al., 2002), which is a parametric estimator based on the lognormal distribution. This estimator uses only the highest observed abundance and the total number of individuals in a sampled population. It resulted in an estimate of 2,236 ribotypes for our data set. This value is still an order of magnitude lower that that of the ML value. Although the Curtis estimator uses an underlying distribution (lognormal), it may suffer from the same problem as the Chao1. Since both of these estimators use only partial information from the sample, this may suggest one reason for the unreliable estimates.

Analysis of clustered data

The MLE values and comparison of the fits for the Poisson-lognormal and the Poisson-gamma models were also obtained for the data set after clustering ribotypes into groups of sequences that were ≥97% identical, i.e. 97% similarity groups. Using the Poisson-lognormal model, the number of 97% similarity groups is estimated to be about 1,500 taxa. The standard deviation of log-abundance distribution is estimated to be 2.7. Although the Poisson-lognormal model resulted in the best fit also for this data set, the fit to the 97% rarefaction curve is not as tight as that to the 100% (ribotype) rarefaction curve (Figs. 3A and 3B).

Effect of sample size

The sample size of 1,033 is larger than commonly used in microbial community analysis. It is thus of interest to investigate whether the ML estimator for the Poisson-lognormal model performs adequately for smaller sample sizes. Thus, we randomly chose 500 sequences from the actual sample. This resulted in similar maximum likelihood estimates of around 25,000 ribotypes, but displayed higher uncertainty due to smaller sample size (data not shown).

We were further interested in how sensitive the Chao1 estimators are to sample size and for what sample sizes the Chao1 estimators eventually become unbiased. For this purpose, we sampled the same simulated community, with sample sizes ranging from 25 to 7.5x10^6 individuals. Two Chao1 estimators were applied (Fig. 4): the commonly-used uncorrected (Chao, 1984) and the bias corrected (Colwell, 1997). Since bias-corrected Chao1 estimator has been rarely used, we used the uncorrected Chao1 estimator in all of our other analyses. The formulas for the two estimators are given in the legend of Figure 4. Although both estimators significantly underestimate diversity for sample sizes smaller that 10^5 individuals, the bias-corrected Chao1 estimate gives higher estimates than the uncorrected Chao1 for sample sizes smaller than 10^3 individuals (Fig. 4). However, for sample sizes larger than 10^3, the uncorrected Chao1 gives higher estimates than the bias-corrected Chao1 (Fig. 4), thus displaying less bias than the bias-corrected Chao1. Surprisingly, only for the largest sample sizes (>10^6) the Chao1 estimates the sample size within the same order of magnitude. Therefore, especially for the small sample size, both versions of the Chao1 estimator, uncorrected or bias-corrected, greatly underestimate the ribotype diversity of diverse clone libraries.

Discussion

Estimating diversity is a basic task in ecology, yet for microbial communities it has remained elusive. Non-parametric estimators (e.g., Chao 1) have started to be used to evaluate microbial diversity because of reported low bias (Hughes et al., 2001; Hill et al., 2003; Kemp and Aller, 2004). However, we have observed from simulated communities whose species abundances were based on our data that the Chao1 is usually negatively biased. Although Mao and Lindsay (2002) have shown that Chao1 is generally biased for heterogeneous communities, the extent of under-estimation has been overlooked. To obtain more accurate estimates of microbial diversity, we have turned to parametric models. Given our large sample, parametric models have the advantage over non-parametric models, that the underlying distribution of taxa abundances can be inferred from the sample and this information can be used to more accurately estimate the total number of taxa. Specifically, we have used a parametric approach based on maximum likelihood, which consists of (i) a distribution model of the abundances of taxa in a community, and (ii) a sampling model appropriate for microbial communities.

We have concentrated our investigations on two commonly used abundance distribution models: the lognormal (Preston, 1948) and the gamma distribution (Fisher et al., 1943). It has been suggested that abundances of taxa in microbial communities are lognormally distributed (Dunbar et al., 1999; Curtis et al., 2002). Our data set is indeed best described by the Poisson-lognormal distribution, suggesting that the lognormal distribution is the underlying distribution of the sampled clone library (Fig. 3A). Compared to the Poisson-lognormal, the Poisson-gamma distributions have resulted in inferior fit (Fig. 3A). We have tested the gamma model because of its suggested good fit to heterogeneous communities (Kempton and Taylor, 1974). Two special cases of the Poisson-gamma, Fisher's logarithmic (Fisher et al., 1943) and MacArthur's broken stick (MacArthur, 1957), may be viewed as distributions characteristic of relatively simple communities whose dynamics are dominated by a single factor (May, 1975). Fisher's logarithmic model has been criticized because fitting the log-series distribution presumes an infinite pool of species available for sampling (Kempton and Taylor, 1974). It is unlikely that complex microbial communities are governed by a single factor, so Fisher's and MacArthur's models are likely inappropriate for modeling abundances of taxa and in turn, for estimating diversity from a sample. Thus, based on our results and theoretical grounds, we can confidently discount the gamma-based distributions for complex microbial communities.

We tested the sensitivity of the estimates from the Poisson-lognormal diversity model relative to (i) sub-sampling of simulated communities, and (ii) sub-sampling of the actual bacterioplankton data. The results from the simulated populations were comparable to those observed for the bacterioplankton library data set, falling within two standard deviations from the mean of the maximum likelihood estimate for the data (Fig. 4). Similar values of ~25,000 ribotypes were obtained for random sub-sampling of 500 individuals from the bacterioplankton data set, but yielded higher uncertainty (data not shown). Therefore, the tests of reproducibility suggest that our model keeps performing adequately for the smaller sample size.

Based on our findings, the bias of the Chao1 estimator for diverse data sets is much higher than previously recorded. The extent of bias associated with diverse libraries was measured using the Chao1 estimator for samples ranging from 50 to 7.5 x 10^7 individuals randomly sampled from simulated communities (Fig. 4). Bias by a factor smaller than one order of magnitude was observed only for very large sample sizes ($>10^6$) and only for the commonly-used uncorrected Chao1 estimator (Chao, 1984, 1987). Furthermore, we showed that the bias-corrected Chao1 does not correct for the biased (low) estimates of the diversity. Although the Chao1 has been cited as giving a lower bound estimate for heterogeneous environments (Mao and Lindsay, 2001; Bohannan and Hughes, 2003) and a small sample size (Kemp and Aller, 2004), we find it 5 standard deviations away from the mean of the MLE of lognormal distribution (Fig. 1D). Furthermore, the Chao1 estimates of the number of ribotypes for the sample size of 1,033 are narrowly distributed around a mean of 1,673, with the largest standard deviation of only 133 (Fig. 1D). By contrast, the ML estimates were essentially unbiased with a variance consistent with the curvature of the likelihood function. In actuality, the non-parametric methods should give a higher uncertainty than parametric methods because they are not constrained by the assumed underlying distribution of taxon abundances. However, the variance of Chao1 is surprisingly small and as such is a very misleading measure of uncertainty (Fig. 1D).

We observed that the Poisson-lognormal model results in a tighter fit to the 100% rarefaction curve (Fig. 1A) in comparison to the 97% rarefaction curve (Fig. 1B). It is possible that the ribotype and cluster data may contain two different underlying distributions and that these result in the differential fit between the two data sets. Lunn et al. (2004) pointed out that differences in the underlying abundance distribution of taxa may be a function of a chosen taxonomic resolution, e.g. species versus genera. However, closer inspection of the data set is required before we can determine whether the difference of the underlying distributions is indeed due to taxonomic resolution.

Although distribution of ribotypes in the bacterioplankton clone library was best described with the Poisson-lognormal model, it does not necessarily mean that lognormal is the only distribution available to explain the given data set. From the abundance data we only have the information on the upper portion of the underlying distribution. Thus, we can only provide a good fit for this portion of the curve. While some distributions can be rejected with authority (i.e. gamma distributions), the data could be fitted equally well with the power-law distribution. However, this has not yet been tested with the present bacterioplankton library. The data of the bacterioplankton sample in Figures 3C and 3D are plotted together with the power trend line as well as a sample taken from a simulated lognormal community. Indeed, from the observations of the data alone, one cannot conclude which distribution (power-law or lognormal) would produce a better fit (Figs. 3C and 3D). However, the power-law model is likely to produce a similar order of magnitude estimate as the lognormal model.

Independent of the statistical approaches used to estimate diversity, an important question remains: to what extent can the diversity observed for a given clone library be extrapolated to a sampled microbial community? The construction of the libraries themselves and PCR amplification may introduce artifacts and biases, which may alter the estimated diversity and abundance distribution (Suzuki and Giovannoni, 1996; Polz and Cavanaugh, 1998; Speksnijder et al., 2001; Thompson et al., 2002). The bacterioplankton data set was corrected for PCR artifacts such as Taq errors, chimeras and heteroduplexes and it can be reasonably assumed that no additional taxa were added to the libraries. However, it is likely that universal primers used in the PCR failed to amplify some organisms from the sampled community. It has been shown that a number of 16S rRNA primers previously thought to be universally conserved are in fact not conserved (Vergin et al., 1998; Daims et al., 1999). Therefore, the estimates for the library should be regarded as minimum diversity estimates for the sampled environment.

In addition, the bias introduced due to preferential amplification of some templates (Suzuki and Giovannoni, 1996) or due to relative abundances of multiple rRNA operons remains an open question. These biases can potentially distort the inferred distribution of abundances of taxa in the sampled community. However, the extent to which they contribute to the standard deviation of the taxon log-abundances, σ, observed for the clone library cannot be entirely resolved at this point. We do, however, know that the ML estimator is essentially unbiased for both total number of taxa, N, and σ, and that the value of σ is large for our library. This suggests that (i) the experimental bias would have to be very large for the distribution of ribotypes in the library to be different from the underlying distribution of the community and (ii) reducing this bias would not change significantly N, but only result in the more accurate estimates of N.

The large diversity of ~25,000 ribotypes suggested to co-exist in the coastal bacterioplankton sample raises several questions. Firstly, how many individual genomes underlie the ribotype diversity? Secondly, to what extent can all these types actually be ecologically differentiated? We have previously suggested, based on extensive analysis of all available completely sequenced genomes, that on average the number of ribotypes exceeds the number of genomes by ~2.5 fold (Acinas et al., 2004a). Thus, correcting the value of 25,000 ribotypes for the contribution of multiple operons, a value of 10,000 genomes is obtained. If these were ecologically differentiated, the functional diversity within our sample would indeed be striking. We previously determined that the number of individuals from which our clone library was constructed was ~10^6 bacterial cells. Therefore, if the genomes were uniformly distributed in the sample each would only be represented by ~100 individuals, but under the lognormal distribution, which is suggested for this sample, most genomes would be present at a very small fraction. However, it is highly unlikely that individual ribotypes represent functional units, because functional units would be then composed of only a few individuals. It appears more likely that the functional units are formed of microdiverse clusters of ribotypes, implying that the functional diversity may be far lower than the observed ribotype diversity.

In fact, we have previously suggested that individual genomes may not actually represent distinct functional units within the community but that such units are represented as microdiverse ribotype clusters (Chapter 2 and 3). These clusters may arise by selective sweeps and persist because competitive mechanisms are too weak to purge diversity from within them (Acinas et al., 2004b). Indeed, the large diversity estimate of ribotypes supports these previous suggestions. In addition, we observed that the Poisson-lognormal MLE value for the 97% similarity groups data set is ~1,500 taxa suggesting that the vast majority of the ribotype diversity (85%) is contained within 3% sequence divergence and therefore very likely organized into microdiverse clusters.

In summary, this study reveals previously unsuspected diversity within a complex marine bacterial community. This is evident from development and application of parametric methods based on maximum likelihood. Using these methods, we are able, for the first time, to constrain the value of diversity estimates and critically evaluate the commonly applied Chao1 non-parametric estimator. We are confident that the Chao1 estimator significantly underestimates diversity of complex microbial communities. Most importantly we show that the diversity of complex microbial communities could be much greater than previously thought.

References

1. Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., and Polz, M.F. (2004a) Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J. Bacteriol.* **186**: 2629-2635.

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

3. Bohannan, B.J.M., and Hughes, J. (2003) New approaches to analyzing microbial biodiversity data. *Current Opinion in Microbiology* **6**: 282-287.

4. Bulmer, M.G. (1974) Fitting Poisson Lognormal Distribution to Species-Abundance Data. *Biometrics* **30**: 101-110.

5. Chao, A. (1984) Nonparametric-Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics* **11**: 265-270.

6. Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**: 783-791.

7. Cohen, J.E. (1968) Alternative derivation of a species abundance relation. *American Naturalist* **102**: 165-171.

8. Colwell, R.K. (1997) Estimates: statistical estimation of species richness and shared species from samples. Version 5. User's Guide and application. - Published at: http://viceroy.eeb.uconn.edu/estimates.

9. Colwell, R.K., and Coddington, J.A. (1994) Estimating Terrestrial Biodiversity through Extrapolation. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **345**: 101-118.

10. Curtis, T.P., Sloan, W.T., and Scannell, J.W. (2002) Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 10494-10499.

11. Daims, H., Bruhl, A., Amann, R., Schleifer, K.-H., and Wagner, M. (1999) The domain-specific probe EUB338 is insufficient for the detection of all bacteria: development and evaluation of a more comprehensive probe set. *Syst. Appl. Microbiol.* **22**: 434-444.

12. Diserud, O.H., and Engen, S. (2000) A general and dynamic species abundance model, embracing the lognormal and the gamma models. *American Naturalist* **155**: 497-511.

13. Dunbar, J., Takala, S., Barns, S.M., Davis, J.A., and Kuske, C.R. (1999) Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. *Applied and Environmental Microbiology* **65**: 1662-1669.

14. Engen, S., and Lande, R. (1996) Population dynamic models generating species abundance distributions of the gamma type. *Journal of Theoretical Biology* **178**: 325-331.

15. Fisher, R.A., Corbet, A.S., and Williams, C.B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**: 42-58.

16. Head, I.M., Saunders, J.R., and Pickup, R.W. (1998) Microbial evolution, diversity, and ecology: A decade of ribosomal RNA analysis of uncultivated microorganisms. *Microbial Ecology* **35**: 1-21.

17. Hill, T.C.J., Walsh, K.A., Harris, J.A., and Moffett, B.F. (2003) Using ecological diversity measures with bacterial communities. *Fems Microbiology Ecology* **43**: 1-11.

18. Hugenholz, P., Goebel, B.M., and Pace, N.R. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**: 4765-4774.

19. Hughes, J.B., Hellmann, J.J., Ricketts, T.H., and Bohannan, B.J.M. (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* **67**: 4399-4406.

20. Kemp, P.F., and Aller, R.C. (2004) Estimating prokaryotic diversity: When are 16S rDNA libraries large enough? *Limnol. Oceanogr. Methods* **2**: 114-125.

21. Kempton, R.A., and Taylor, L.R. (1974) Log-Series and Log-Normal Parameters as Diversity Discriminants for Lepidoptera. *Journal of Animal Ecology* **43**: 381-399.

22. Krebs, C.J. (1999) *Ecological methodology.* Menlo Park, CA: Benjamin/Cummings.

23. Lunn, M., Sloan, W.T., and Curtis, T.P. (2004) Estimating bacterial diversity from clone libraries with flat rank abundance distributions. *Environ. Microbiol.*: doi:10.1111/j.1462-2920.2004.00641.x.

24. MacArthur, R.H. (1957) On the relative abundance of bird species. *Proc. Natl. Acad. Sci. USA* **43**: 293-295.

25. Mao, C.M., and Lindsay, B.G. (2001) Moment-based nonparametric estimators for the number of classes in a population. In. University Park: The Pennsylvania State University, pp. 1-44.

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

27. May, R.M. (1975) Patterns of species abundance and diversity. In *Ecology and evolution of communities.* Cody, M.L., and Diamond, J.M. (eds). Cambridge, Massachusetts, and London, England: The Belknap Press of Harvard University Press, pp. 81-120.

28. Plotkin, J.B., and Muller-Landau, H.C. (2002) Sampling the species composition of a landscape. *Ecology* **83**: 3344-3356.

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

30. Preston, F.W. (1948) The commonness and rarity of species. *Ecology* **41**: 611-627.

31. Rappe, M.S., and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annual Review of Microbiology* **57**: 369-394.

32. Speksnijder, A.G.C.L., Kowalchuk, G.A., De Jong, S., Kline, E., Stephen, J.R., and Laanbroek, H.J. (2001) Microvariation artifacts introduced by PCR and cloning of closely related 16S rRNA gene sequences. *Applied and Environmental Microbiology* **67**: 469-472.

33. 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. *Applied and Environmental Microbiology* **69**: 6189-6200.

34. 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.

35. Thompson, J.R., Marcelino, L.A., and Polz, M.F. (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. *Nucleic Acids Research* **30**: 2083-2088.

36. Venter, C.J., and al., e. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.

37. Vergin, K.L., Urbach, E., Stein, J.L., DeLong, E.F., Lanoil, B.D., and Giovannoni, S.J. (1998) Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order *Planctomycetales. Appl. Environ. Microbiol.* **64**: 3075-3078.

38. Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998) Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 6578-6583.

Table 1. Ribotype abundances observed in the bacterioplankton sample.

Abundance of ribotypes	Number of ribotypes
1	381
2	65
3	23
4	18
5	4
6	6
7	3
9	1
11	4
13	3
14	2
16	1
21	1
27	1
32	1
43	1
45	1

Figure 1. Likelihood surface for the compound Poisson-Lognormal model. The variables on the axes are the total number of taxa *N* and the standard deviation of the lognormal distribution σ. In all four parts of the figure, black contours represent estimates of the parameter pair values, *N* and σ, 1-5 standard deviations away from the mean (the most likely estimate for the total number of types). **A.** Likelihood surface of the modified Poisson-lognormal model; generated as a product of the two likelihoods shown in B and C. **B.** Likelihood surface for a likelihood function obtained from the information on sample abundances in the sample *s* (equivalent to the Bulmer's likelihood function (Bulmer, 1974)). **C.** Likelihood surface for a likelihood function obtained from the information on the number of types observed in the sample *s*. **D.** Likelihood surface for the compound Poisson-Lognormal model and the most likely estimate of the total number of types in a bacterioplankton community sample (black cross); averages of 10 samples from each of the 10 simulated communities estimated for the Poisson-lognormal model using the modified likelihood function (red solid circles); averages of 10 samples from each of the 10 simulated communities estimated for the Poisson-lognormal model using the Bulmer's likelihood function (blue open circles); averages of the 10 Chao1 values of the 10 samples from the each of the 10 simulated communities (green solid diamonds).

Figure 2. Likelihood surface for the Poisson-gamma model showing relative support for each pair of parameter estimates of the number of types *N* and *ln(σ)*. Black contours (1-

5) represent estimates of the parameter pair values, N and σ, 1-5 standard deviations away from the mean (the most likely estimate for the total number of types).

Figure 3. Lognormal and gamma distributions under the Poisson sampling fitted to data by the method of maximum likelihood and compared to the rarefaction curve of the actual data. **A.** Rarefaction curve of the bacterioplankton sample – taxa constructed from 100% sequence similarity clustering (blue solid diamonds); Poisson-lognormal distribution (red solid circles); Chao1 estimate under the Poisson-Lognormal (orange open circles); Curtis estimate under the Poisson-Lognormal (green open circles); Poisson-Gamma distribution (black star); and MacArthur's broken stick distribution (black cross). **B.** Rarefaction curve of the bacterioplantkon sample - taxa constructed from 97% sequence similarity clustering (blue solid diamonds); Poisson-lognormal distribution (red solid circles); Chao1 estimate under the Poisson-Lognormal (orange open circles); and Poisson-Gamma distribution (black star). **C.** Probability exceedance plot of the bacterioplantkon sample - 100% sequence similarity clustering OTUs (blue solid diamonds) and a sample taken from a simulated lognormal community (black open triangles); and power trendline (black solid line). **D.** Probability exceedance plot of the bacterioplantkon sample - 97% sequence similarity clustering OTUs (blue solid diamonds) and a sample taken from a simulated lognormal community (black open triangles); and power trendline (black solid line).

Figure 4. Chao1 estimates of the total number of different types calculated from a simulated community of 25,000 different types and σ of 2.7 using "bias-corrected" (green open squares) and the approximate "uncorrected" formula (blue solid diamonds) as a function of sample size. Error bars are one standard deviation and were calculated with the variance formula derived by Chao (1987).

Bias-corrected: $N_{Chao1} = S_{obs} + \frac{S_1^2}{2(S_2+1)} - \frac{S_1 S_2}{2(S_2+1)^2}$; uncorrected: $N_{Chao1} = S_{obs} + \frac{S_1^2}{2S_2}$.

Standard deviation: $\sigma = \sqrt{\mathrm{var}(N_{Chao1})} = \sqrt{S_2\left[\frac{G^4}{4} + G^3 + \frac{G^2}{2}\right]}, G = \frac{S_1}{S_2}$.

Estimated population size average from 10 simulated communities calculated from Poisson-lognormal model using maximum likelihood is shown in red solid circles. The red lines are error bars showing one standard deviation for the Poisson-lognormal model.

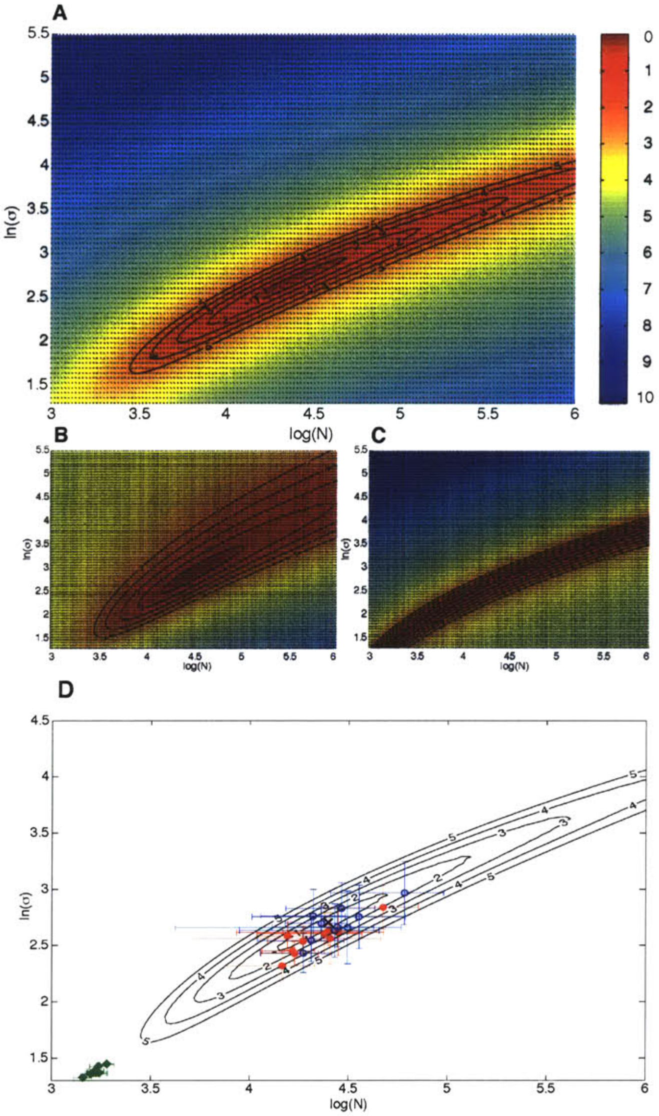


Figure 1.

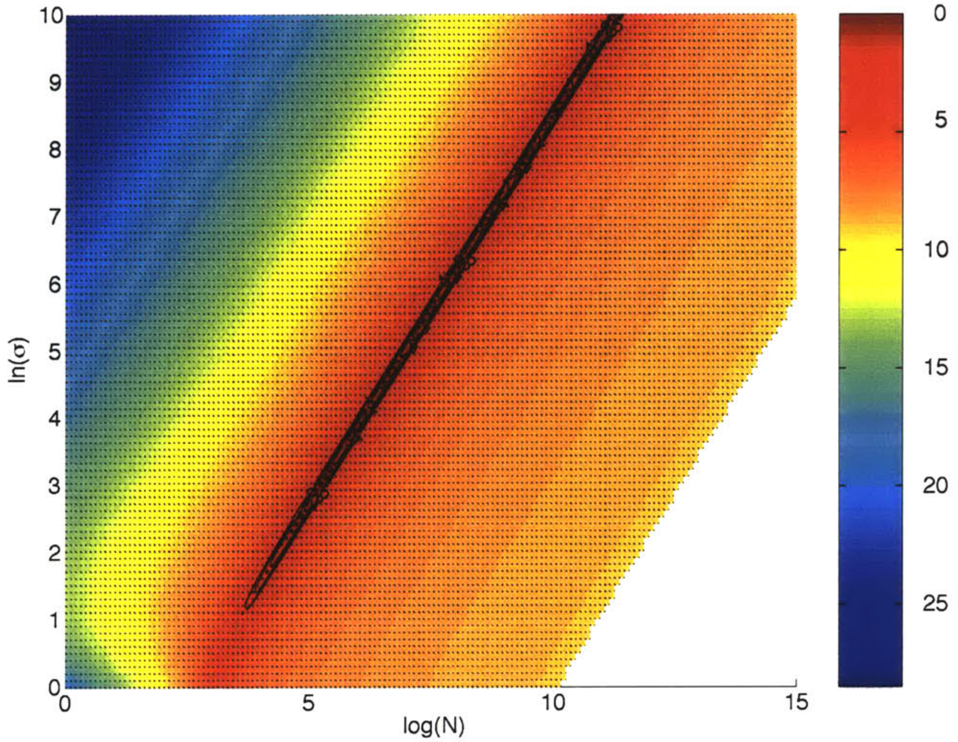


Figure 2.

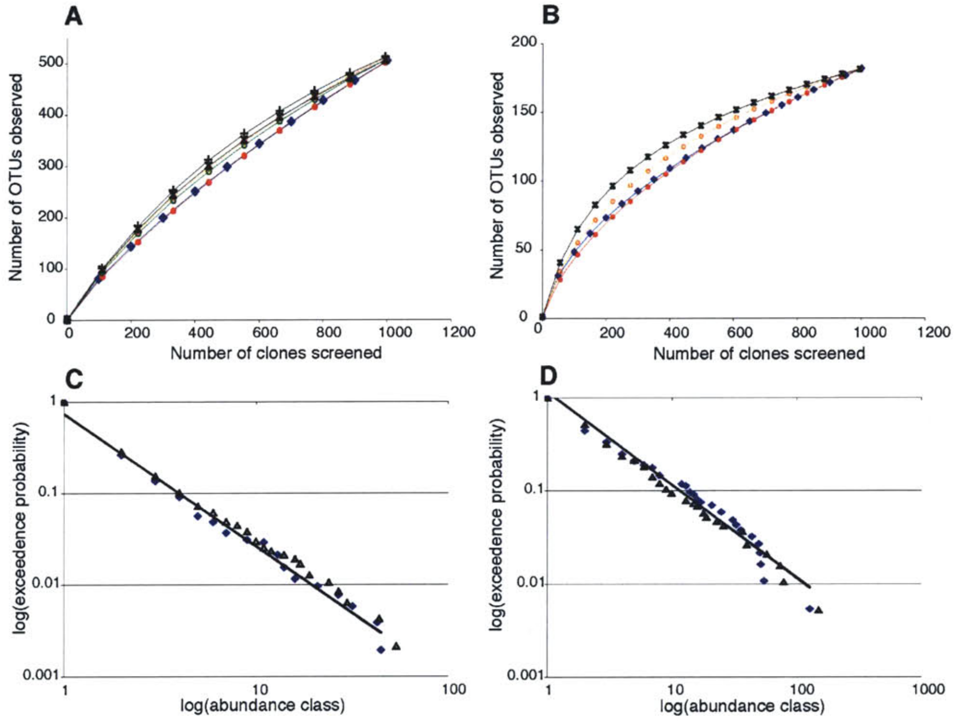

A
500
400
300
200
100
0
Number of OTUs observed
0
200
400
600
800
1000
1200
Number of clones screened
B
200
150
100
50
0
Number of OTUs observed
0
200
400
600
800
1000
1200
Number of clones screened
C
1
0.1
0.01
0.001
log(exceedence probability)
1
10
100
log(abundance class)
D
1
0.1
0.01
0.001
log(exceedence probability)
1
10
100
1000
log(abundance class)

Figure 3.

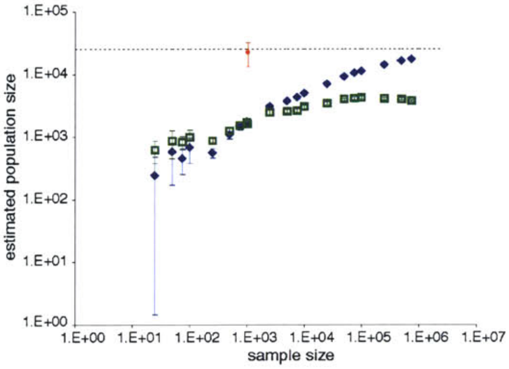

1.E+05
1.E+04
1.E+03
1.E+02
1.E+01
1.E+00
estimated population size
1.E+00
1.E+01
1.E+02
1.E+03
1.E+04
1.E+05
1.E+06
1.E+07
sample size

Figure 4.

CHAPTER FIVE

Summary and Future Directions

SUMMARY AND FUTURE DIRECTIONS

This thesis explores questions of microbial community structure in the marine environment. Specifically, two fundamental questions are investigated: (i) how many bacterial types co-exist, and (ii) does a phylogenetic structure exist that would suggest units of differentiation among natural microbial communities? Bacterial diversity in two complex marine communities (coastal bacterioplankton and sediment sulfate-reducing bacteria) was investigated by (i) comprehensive analysis of large 16S rRNA clone libraries, and (ii) refinement and application of parametric diversity estimators.

Several major results are presented:

- development of analysis protocols and tools to constrain artifacts that may lead to overestimation of diversity and, more significantly, obscure patterns of community organization (Chapters 2 and 3).
- refinement and application of statistical methods to estimate microbial diversity (Chapter 4).
- demonstration of co-existence of previously undetected high diversity within marine bacterial communities (Chapters 2, 3, and 4).
- identification of patterns of community organization revealing predominance of microdiverse ribotype clusters that are hypothesized to correspond to ecologically distinct units (Chapters 2 and 3).

Chapter 2 explores the extent of diversity and phylogenetic structure of a bacterial sediment community. The clone library investigated in this chapter was constructed from 16S rRNA genes of delta-proteobacterial sulfate-reducing bacteria (SRB) from salt-marsh sediment samples. We observed unexpected high diversity of ribotypes and predominance of microdiverse relationships among the co-existing SRB. This high diversity is indeed surprising as the SRB overall possess very similar metabolism; however, as detailed in Chapter 3 the actual functional units within this community may not be individual ribotypes but microdiverse clusters of ribotypes, so that the functional diversity within the sediment SRB may be far lower than the observed ribotype diversity. It may be possible in the future to critically test hypotheses on structure and function using this community as a model system. First, the SRB community was well sampled so that it is likely that all major groups were detected. This good coverage may serve as a foundation to develop tools for monitoring specific SRB populations in order to correlate their growth with prevalence of specific conditions within the sediment. For example, this may be achieved by application of DNA microarrays specifically designed to differentiate individual ribotypes and ribotype clusters within the community. Second, SRB communities are overall well studied and therefore an extensive dataset exists whose comparison with the current study may reveal patterns in occurrence of specific SRB types in different environments. For example, in the marsh sediment previously unidentified completely oxidizing SRB dominated but other studies have also found predominance of the same group of SRB in similar environments.

Diversity and phylogenetic structure of a coastal bacterioplankton community was investigated in Chapter 3. This community was chosen because it differs in its overall ecological features from the sediment community and would therefore be a test case for the general applicability of the findings in Chapter 2. The pelagic environment is well mixed, while sediments are highly structured environments. Therefore, one may expect that the underlying community composition of the two communities would be different. For example, it may be hypothesized that the efficient mixing in the pelagic environment may allow for more efficient selective sweeps within the community. These would serve to purge diversity leading to a more simple overall community composition. However, similar fine-scale phylogenetic structures were observed for the coastal community as well. Although microdiversity has been previously suggested by analysis of specific microbial groups in PCR-generated clone libraries (Field et al., 1997; Garcia-Martinez and Rodriguez-Valera, 2000; Casamayor et al., 2002), it had remained unclear to what extent microdiversity arises by PCR induced artifacts or is the result of paralogous rRNA operons within the same genome (Suzuki and Giovannoni, 1996; Polz and Cavanaugh, 1998; Speksnijder et al., 2001; Thompson et al., 2002; Rappe and Giovannoni, 2003).

Application of PCR-based approaches on the 16S rRNA gene can potentially affect the estimation of diversity in several ways: by (1) formation of sequence artifacts (Taq errors, chimeras, and heteroduplexes), (2) preferential amplification of some templates over others, which can skew sequence abundances, (3) missing some of the sequence diversity due to the primer selection, and (4) the incidence of multiple rRNA operons within a single genome.

(1) We have developed methods that minimize and account for chimeras, Taq errors and heteroduplex errors. It has been shown empirically that the methods we employed removed all error due to heteroduplex formation (Thompson et al., 2002). The remaining types of error cannot be detected directly and must be inferred. We have employed the two most widely used chimera-checking programs and have written our own software specifically designed for clone libraries that have been sampled to a high degree (Chapter 2 and 3). Finally, we have developed methods to identify and eliminate polymerase errors based on well-known patterns of primary and secondary structure conservation and have provided several independent estimates of their effectiveness (Chapter 2 and 3). It is important to note that one important component of this thesis is the development, for the first time of a means to estimate the number of rRNA sequences affected by TAQ errors and some guidelines on how to identify them.

(2 & 4) At this point we cannot ascertain the extent of bias resulting from the preferential amplification of templates (Suzuki and Giovannoni, 1996) or from different abundances of multiple rRNA operons. However, preliminary evidence suggests that this bias may not be large. We have conducted and published a detailed investigation of operon heterogeneity among published genomes (Acinas et al., 2004). About 40% of genomes have 1 or 2 operons and a majority of their sequences are identical, indicating that the number of operons per genome is highly skewed towards the lower spectrum. Also, an extensive survey of 97 published bacterial genomes showed that a correction factor of 0.4 can be applied to estimate diversity of genomes from unique rRNA sequences (Acinas et al., 2004). Thus, after applying this correction, the bias stemming

from the preferential amplification of some templates and different number of identical operons may only distort the distribution of ribotype abundances of the sampled community, but should not change their total number.

(3) PCR using universal primers may fail to amplify some organisms from a sampled community. It has been shown that several 16S rRNA primers previously thought to be universally conserved are in fact not (Vergin et al., 1998; Daims et al., 1999). Therefore, as we cannot exclude the possibility that the primers may fail to amplify some members of the domain *Bacteria*, the estimates of the diversity should be regarded as a minimum diversity for a sampled community. This implies that the diversity of the sampled community may be even higher than diversity reported for the library. It is important to note that prior to the amplification of the bacterioplankton community, we carefully evaluated the existing universal primers 27F and 1492R and modified these to include members from the order *Planctomycetales* based on the information provided by Vergin et al. (1998). The mismatch of universal primers to the 16S rRNA sequences was inferred from a survey of *Planctomycetales* clones recovered from a marine fosmid library (Vergin et al., 1998). Although it is possible that these modified primers may have a mismatch with some unknown bacterial 16S rRNA sequences, a marine molecular survey conducted without prior PCR amplification indicated that no novel clades of the domain *Bacteria* could be observed (Venter and al., 2004). Thus, it is very likely that we detected at least the dominant members of the bacterioplankton community.

Overall, by (i) developing methods that minimize and account for the contribution of sequence artifacts, (ii) accounting for variation in multiple operons within single genomes, and (iii) improving the existing primers, we were able to reduce the difference of the total number of ribotypes between the sampled communities and the constructed large clone libraries.

The large size of the clone libraries enabled comparison of statistical approaches used in estimating microbial diversity, as well as development and application of parametric methods based on maximum likelihood (Chapter 4). With our dataset and simulations we were able to critically evaluate the estimates obtained by the commonly applied Chao1 non-parametric estimator and the bias associated with this estimator. In addition, we evaluated existing parametric methods as they should perform better for diverse microbial communities, and modified them to better account for the specific way microbial communities are sampled. The diversity estimated using the parametric approach revealed an even higher number of co-existing microdiverse sequences, and the estimated diversity was over one order of magnitude higher than that suggested by common non-parametric approaches. However, when sequences were clustered into 97% sequence similarity groups, diversity estimates increased by a much smaller factor compared to the Chao1. This suggests that the overall pattern of predominance of microdiverse clusters is strongly confirmed by the new analysis tools.

Overall, the compensation for artifacts and improved estimation revealed that the vast majority of ribotypes fall into microdiverse clusters containing <1% sequence divergence. Whether the observed ribotype clusters represent ecotypes, i.e. ecologically cohesive populations, will have to be determined by detailed examination of the environmental dynamics of genomic variants. It is proposed that the observed microdiverse clusters form important units of differentiation in microbial communities. They are hypothesized to arise by selective sweeps and contain high diversity because competitive mechanisms are too weak to purge diversity from within them.

References

1. Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., and Polz, M.F. (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J. Bacteriol.* **186**: 2629-2635.

2. Casamayor, E.O., Pedros-Alio, C., Muyzer, G., and Amann, R. (2002) Microheterogeneity in 16S ribosomal DNA-defined bacterial populations from a stratified planktonic environment is related to temporal changes and to ecological adaptations. *Applied and Environmental Microbiology* **68**: 1706-1714.

3. Daims, H., Bruhl, A., Amann, R., Schleifer, K.-H., and Wagner, M. (1999) The domain-specific probe EUB338 is insufficient for the detection of all bacteria: development and evaluation of a more comprehensive probe set. *Syst. Appl. Microbiol.* **22**: 434-444.

4. Field, K.G., Gordon, D., Wright, T., Rappe, M., Urbach, E., Vergin, K., and Giovannoni, S.J. (1997) Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Appl. Environ. Microbiol.* **63**: 63-70.

5. Garcia-Martinez, J., and Rodriguez-Valera, F. (2000) Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine Archaea of Group I. *Molecular Ecology* **9**: 935-948.

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

7. Rappe, M.S., and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annual Review of Microbiology* **57**: 369-394.

8. Speksnijder, A.G.C.L., Kowalchuk, G.A., De Jong, S., Kline, E., Stephen, J.R., and Laanbroek, H.J. (2001) Microvariation artifacts introduced by PCR and cloning of closely related 16S rRNA gene sequences. *Applied and Environmental Microbiology* **67**: 469-472.

9. 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.

10. Thompson, J.R., Marcelino, L.A., and Polz, M.F. (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. *Nucleic Acids Research* **30**: 2083-2088.

11. Venter, C.J., and al., e. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.

12. Vergin, K.L., Urbach, E., Stein, J.L., DeLong, E.F., Lanoil, B.D., and Giovannoni, S.J. (1998) Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order *Planctomycetales. Appl. Environ. Microbiol.* **64**: 3075-3078.

GLOSSARY

Bacterioplankton – all bacteria that passively drift in lakes and ocean

Chimera – a sequence formed from two different sequences.

Clone – a lineage of individuals produced asexually.

Cluster – a set of sequences grouped by some sequence similarity cutoff.

Diversity – the heterogeneity of a system; the variety of different types of organisms occurring together in a biological community.

Ecological niche – the functional role of an organism within an ecosystem; the combined description of the physical habitat, functional role, and interactions of the microorganisms.

Guild – populations within a community, which use the same resources.

m – sample abundances (Chapter 4).

M – population abundances (Chapter 4).

n – a number of taxa in a sample (Chapter 4).

N – the total number of taxa in a population (Chapter 4).

Non-parametric diversity estimators – diversity estimators that assume no models of distribution of taxon abundances.

OTU – operational taxonomic unit.

Parametric diversity estimators – diversity estimators that assume abundance distribution of taxa.

Phylogeny – the line or lines, of direct descent in a given group of organisms; also the study or the history of such relationships.

Population – the set of data from which a statistical sample is taken (Chapter 4).

Population abundance – taxon abundance in a population (Chapter 4).

R – relative abundance of taxon (Chapter 4).

Ribotype – unique rRNA sequence.

σ - standard deviation of the taxa log-abundances (Chapter 4).

s **–** sample size (Chapter 4).

Sample – a set of sequences from a clone library used in diversity estimates (Chapter 4).

Sample abundance – taxon abundance in a sample.

Taxon – a taxonomic category or a group.

2424- 62