PROTISTAN MICROBIAL OBSERVATORY IN THE CARIACO BASIN: AN EXAMINATION OF SPECIES RICHNESS, HABITAT SPECIALIZATION, AND BIOGEOGRAPHY

A dissertation presented

by

William Orsi

to
The Department of Biology

In partial fulfillment of the requirements for the degree of Doctor of Philosophy

in the field of

Biology

Northeastern University
Boston, Massachusetts
March, 2011
PROTISTAN MICROBIAL OBSERVATORY IN THE CARIACO BASIN: AN EXAMINATION OF SPECIES RICHNESS, HABITAT SPECIALIZATION, AND BIOGEOGRAPHY

by

William Orsi

ABSTRACT OF DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biology in the Graduate School of Arts and Sciences of Northeastern University, March 2011
ABSTRACT

The largest 18S rRNA Sanger sequence dataset from an oceanographic regime is combined with pyrosequencing to interrogate protistan diversity in the Cariaco Basin. A multiple PCR primer approach was applied to a total of 16 samples from 3 locations across the Basin and 4 depths along the basin’s geochemical gradients. Phylogenetic analyses identify new clades at multiple levels of taxonomic hierarchy, revealing increased ecological and geographic distributions. Several clades were detected only in anoxic samples, suggestive of habitat specialization. This is supported by multivariate analyses and parametric richness estimations showing a division between communities present in different geochemical layers of the water column and at the different geographic locations sampled. In situ fixed samples were obtained to enumerate protists from different biogeochemical habitats using fluorescent in situ hybridization and scanning electron microscopy, and show that 90% of cells in deep anoxic layers of the Cariaco Basin do not hybridize with a universal probe. This suggests divergence in one of the most conserved regions of the 18S gene, and thus substantial novelty. Within anoxic in situ samples, a new class-level lineage within the phylum Ciliophora is identified and described here as ‘candidatus Cariacotrichea’.
Acknowledgements

I thank the captain and crew of the B/O Hermano Gines and the staff of the Fundación La Salle de Ciencias Naturales, Margarita Island for their assistance during my field-work in Venezuela. I am deeply grateful to Yrene Astor and Ramon Varela who were instrumental in the transportation of the DEEP-SID to and from Isla Margarita, and provided assistance in many practical aspects of my field work. Also, I am indebted to Tine Hohmann and Jose Faria for their help in post-cruise processing of numerous Bouin’s fixed samples. I appreciate the assistance of Drs. Wilhelm Foissner and Peter Vdacny in the morphological classification of the ‘candidatus Cariacotrichea’. I would like to thank Valentin Ilyin (Northeastern University), Chesley Leslin, (Northeastern University), Nathan Cahoon (Northeastern University), and Richard Christen (University of Nice, France) for their assistance with bioinformatics. Also, thank you to Steven Vollmer for introducing me to multivariate approaches to analyze ecological communities. I would like to thank Gordon Taylor, Virginia Edgcomb, and Slava Epstein for providing helpful guidance throughout the duration of my Ph.D research, and all members of my Ph.D committee for their support. This research was supported by grants from NSF (MCB-0348341 and DEB-0816840 to Slava Epstein [Northeastern University], DEB-0816638 to John Bunge [Cornell University], MCB-0348407 to Virginia Edgcomb [Woods Hole Oceanographic Institution] and OCE 03-26175 and MCB-03-47811 to Gordon Taylor [Stony Brook University]), from ANR-Biodiversité project Aquaparadox and the ERA-net project BiodivERsA BioMarKs program to Richard Christen [University of Nice, France], and from Venezuela FONACIT (nos. 96280221 and 2000001702 to Ramon Varela [Fundación La Salle de Ciencias Naturales]). Finally, I would like to thank my parents for their loving support throughout my life and academic career.
# TABLE OF CONTENTS

Abstract...........................................................................................................2

Acknowledgements..............................................................................................4

Table of Contents..................................................................................................5

Introduction.............................................................................................................6

Chapter 1: Species richness and endemicity.........................................................12

Chapter 2: Habitat specialization...........................................................................56

Chapter 3: ‘\textit{candidatus} Cariacotrichea’.........................................................97

Chapter 4: \textit{In situ} access to marine protists...............................................120

Summary...............................................................................................................132

Appendix...............................................................................................................136
INTRODUCTION

This work is a presentation and discussion of the results of a 5-year protistan Microbial Observatory in the Cariaco Basin (Caribbean Sea, Venezuela). The principal goals of this work, and of the Microbial Observatory, is to study protistan species richness, biogeography, phylogenetic diversity, habitat specialization, and community structure, and by doing so address one of the oldest tenets in microbial ecology: Is everything everywhere (Beijerinck MW (1913). *De infusies en de ontdekking der backteriën. Jaarboek van de Koninklijke Akademie voor Wetenschappen*. Müller: Amsterdam, the Netherlands.)?

Why protists? They are the focal point of this research because of their importance as members of any aquatic microbial community. Through grazing on prokaryotic and other eukaryotic prey, they modify or re-mineralize organic matter and regenerate nutrients. Through direct and indirect effects, protists help determine the metabolic potential of microbial communities. Inorganic nutrients released by grazing increase primary production and make organic carbon available to higher trophic levels, thus influencing the carbon cycle. Despite the importance of protists in aquatic ecosystems, relatively little is known about protistan diversity, biogeography, and the degree to which specific habitats select for unique communities, particularly in the deep sea.

Why Cariaco? The site of this Microbial Observatory is not coincidental: the Cariaco Basin offers a multitude of different biogeochemical niches in its chemically stratified water column. In fact, the Basin is so chemically diverse that it serves as a convenient location to examine marine habitats that are oxic, anoxic, and everything in between. If all microbes are indeed everywhere, one would expect the microbial communities present across the multitude of biogeochemical habitats in the Basin to serve as a proof for that. The following short description of our model environment supports this point. The Cariaco water column progresses from fully oxic to sulfidic across a temporally varying boundary between the depths of 250 and
350 meters. Opposing gradients in oxidants and reductants are present within the redox transition zone where prokaryotic metabolic activity and cell numbers often peak, which typically coincides with peaks in protist cell numbers. Collectively, this creates a diversity of biogeochemical niches representative of many different habitats that otherwise would have to be sampled in different locations.

To address protistan richness, phylogenetic diversity, habitat specialization, and biogeography, we conducted the largest rRNA gene survey of microbial eukaryotes to date. This survey employed RNA analyses, scanning electron microscopy, fluorescence in situ hybridization, as well as a variety of statistical tools.

The body of this research is divided into four chapters. Each has its own specific conclusions of the respective discussion sections. The first chapter describes the protistan 18S rRNA gene sequence diversity surveyed using both Sanger and pyrosequencing-based approaches, employing multiple PCR primers, and state-of-the-art statistical analyses to estimate microbial richness. Sampling the Basin at three stations, two seasons, and four geochemically distinct depths, we obtained over 6,000 nearly full-length protistan rRNA gene sequences. This dataset was complemented with a novel, high-throughput, sequencing approach that was applied to the same Cariaco samples and produced a total of over 80,000 short (~100 bases) sequence reads from a variable portion of the rRNA gene (V9). Both sequence collections represent all major and many minor protistan taxa known today, at globally similar frequencies. This large dataset provides, via a recently developed parametric modeling method, the first statistically sound prediction of the total size of protistan taxon richness in a large and varied environment, such as the Cariaco Basin. The number estimated is far lower than the global known richness, indicating a degree of endemicity in the distribution of the majority of microbial eukaryotes.

Chapter 2 uses phylogenetic information, multivariate community analyses, and statistical richness predictions to test whether protists exhibit habitat specialization within
defined geochemical layers of the water column, and to examine the likelihood that at least some are endemic to these environments. This, coupled to spatio-temporal distributions of protists across two seasons and two geographic sites within the Basin, provides new insights into the spatial partitioning of marine protistan communities, their distributions along geochemical gradients, and the extent of protistan phylogenetic diversity and biogeography. Specifically, these analyses reveal an apparent restriction of some protists to different sites in the Basin, as well as oxygenated, suboxic, or anoxic/sulfidic environments. Phylogenetic and multivariate community analyses along with parametric estimations of richness imply that geographic location, seasonality, and geochemical gradients define the community structure of marine protists in the Cariaco Basin. Thus, it is concluded that substantially different communities can exist in close proximity to one another, which permits speciation to proceed differently at different sites and depths in the Basin. These analyses support the moderate endemcity model (Bass et al., 2007; Foissner, 2006), which views protists as an assemblage of species some of which are cosmopolitan and others endemic, and thus biogeographically similar to macroorganisms.

Chapter 3 describes the detection, quantification, and morphological description of a novel class-level lineage within the protistan phylum Ciliophora. This marks the first discovery of a new class of its kind in decades. By combining ribosomal RNA detection with scanning electron microscopy, a novel morphological feature was identified in this group whose rRNA-based phylogenetic position suggests divergence at the class-level. Based on this specific feature and molecular phylogenies, a novel ciliate class is proposed and described as ‘candidatus Cariacotrichea’.

Chapter 4 focuses on some of the practical lessons learned through the Microbial Observatory. Perhaps the most important one is the value of accessing marine protists in situ. Several findings have been made exclusively because of the use of a remotely operated deep-sea robotic sampler. This device fixed the seawater samples in situ, from four biogeochemically
distinct depths in the water column, and provided access to cells and their RNA that we could not access in any other way. The degree of novelty of protistan assemblages within these samples was assessed by surveying OTUs divergent in one of the most conservative regions of the rRNA gene, the sites routinely used for ‘universal’ fluorescent *in situ* hybridization (FISH) probes. The disparity between total cell numbers and those determined by a universal FISH probe suggest the gross majority of planktonic protists in the deep layers of the Basin have no close relatives in existing databases. Only a small portion of the cells that did not hybridize to the rRNA probe could be determined to be dead or inactive protist cells. The implication is that these waters and possibly other anoxic aquatic ecosystems, contain significant taxonomic novelty, possibly missed by the rRNA approach entirely, making them exciting targets for future microbial discovery efforts.

The appendix of this thesis contains additional studies that I participated in during my Ph.D work. Although these studies are not the focus of my thesis, they are united with the majority of my Ph.D work under the general theme of microbial ecology, with an emphasis on protists. The first of these studies focuses on the protistan community inhabiting Deep-Sea Hypersaline Lakes (DHALs) in the Mediterranean Sea. Applying a ribosomal RNA approach, our team demonstrated that a rich and diverse community of protists exists within these extreme environments and suggested that the environmental conditions defining the limits of eukaryotic life are still far from being established. The second study is a taxonomic revision of two classes within the phylum Ciliophora. Using phylogenetic and morphological analyses we proposed that two classes, which have been traditionally grouped into different infraphyla, are actually sister groups and correspond to a new infraphylum of their own. The third study used high-throughput sequencing to investigate the protistan communities in two different anoxic marine habitats, the Cariaco Basin (Venezuela) and Framvaren Fjord (Norway). The results of this study demonstrated that anaerobic communities of marine protists are vastly complex and also
highlight new high-throughput sequencing technologies as valuable tools for the assessment of microbial diversity.

Chapters 1, 2, and 4 of this dissertation have been accepted for publication as a series of three papers in *The ISME Journal*. Some images that emerged from this work were also chosen for publication in *The ISME Journal* as the cover artwork for 2011, and they are presented on the next page. Chapter 3 is a manuscript that is currently undergoing preparation for submission to *The International Journal of Systematic and Evolutionary Microbiology*. These four manuscripts are interrelated, and thus each manuscript contains numerous cross-references to the other three. To make cross-referencing between them easier in this dissertation, additional text has been added here to the citations. For example, the citation 'see Orsi *et al.*, this issue' has been modified to read 'see Orsi *et al.*, this issue [Chapter 2 of this dissertation]'.
Marine ciliates exhibiting epibiotic bacteria recovered from the deep, permanently anoxic and hypersulfidic waters of the Cariaco Basin (Caribbean Sea, Venezuela). Images were taken with a Hitachi S-4800 scanning electron microscope at Northeastern University, Boston, MA.

Authors: William Orsi¹, Virginia Edgcomb², Bill Fowle¹, Gordon T. Taylor³, and Slava Epstein¹

Affiliations: ¹Department of Biology, Northeastern University, 313 Mugar Building, Boston, MA 02115, USA. ²Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA. ³Marine Sciences Research Center, State University of New York, Stony Brook, New York 11794, USA
Chapter 1: Species Richness and Endemicity
Protistan Microbial Observatory in the Cariaco Basin, Caribbean. I.

Pyrosequencing vs. Sanger Insights into Species Richness

V. Edgcomb1, W. Orsi2, J. Bunge3, S. O. Jeon4, R. Christen5, C. Leslin2, M. Holder6, G. T. Taylor7, P. Suarez8, R. Varela9, S. Epstein2,10*

1. Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA.

2. Department of Biology, Northeastern University, 313 Mugar Building, Boston, MA 02115, USA.

3. Department of Statistical Science, 301 Malott Hall, Cornell University, Ithaca, NY 14853, USA.

4. Department of Environmental Science, Kangwon National University, Kangwon-Do, South Korea.

5. Université de Nice et CNRS UMR 6543, Laboratoire de Biologie Virtuelle, Centre de Biochimie, Parc Valose. F 06108 Nice, France

6. Department of Ecology and Evolutionary Biology, University of Kansas, 6031 Haworth Hall, 1200 Sunnyside Avenue, Lawrence, Kansas 66045, USA.

7. School of Marine & Atmospheric Sciences, Stony Brook University, Stony Brook, New York 11794, USA.

8. Departamento de Biología de Organismos, Universidad Simón Bolívar, Sartenejas, Baruta, Estado Miranda, Venezuela.


10. Marine Sciences Center, Northeastern University, 430 Nahant Rd., Nahant, MA, 01908, USA.

Running title: Diversity, ecology, and biogeography of microbial eukaryotes
Keywords: protists, diversity, species richness, anoxic, endemicity, pyrosequencing, 18S rRNA approach

Subject Categories: Microbial ecology and functional diversity of natural habitats, and/or Microbial population and community ecology

* corresponding author
Abstract

Microbial diversity and distribution are topics of intensive research. In two companion papers in this issue, we describe the results of the Cariaco Microbial Observatory (Caribbean Sea, Venezuela). The Basin contains the largest body of marine anoxic water, and presents an opportunity to study protistan communities across biogeochemical gradients. In the first paper, we survey 18S rRNA gene sequence diversity using both Sanger and pyrosequencing-based approaches, employing multiple PCR primers, and state-of-the-art statistical analyses to estimate microbial richness missed by the survey. Sampling the Basin at three stations, in two seasons, and at four depths with distinct biogeochemical regimes, we obtained the largest, and arguably the least biased collection of over 6,000 nearly full-length protistan rRNA gene sequences from a given oceanographic regime to date, and over 80,000 pyrosequencing tags. These represent all major and many minor protistan taxa, at frequencies globally similar between the two sequence collections. This large dataset provided, via the recently developed parametric modeling, the first statistically sound prediction of the total size of protistan richness in a large and varied environment, such as the Cariaco Basin: over 36,000 species, defined as almost full-length 18S rRNA gene sequence clusters sharing over 99% sequence homology. This richness is a small fraction of the grand total of known protists (over 100,000-500,000 species), suggesting a degree of protistan endemism.
Introduction

The history of research on protistan diversity and taxonomy spans almost two centuries. Protists attained their status as a separate Kingdom even before the role of the nucleus was first proposed (and so Protists included bacteria as a result, (Haeckel 1866, Owen 1860). Centuries later, it is still essentially unknown whether all the major protistan groups have or have not been discovered, whether or not protistan species are globally distributed (Baldauf 2003, Cavalier-Smith 2004, Finlay and Fenchel 1999, Finlay 2002, Foissner 1999, Foissner, 2006), or how many species are present in a given environment (Jeon et al 2006).

One of the main reasons for such a lamentable state of affairs is the extent of protistan diversity. The number of protistan life forms appears to overwhelm our current abilities to completely inventory all species in a single sample from most environments, let alone get a complete representation of the community present in the ecosystem from which this sample was drawn. Over the past 10 years, numerous studies have applied rRNA gene sequencing in attempts to describe the surprising diversity of microbial eukaryotes (e.g., Diez et al 2001, Lopez-Garcia et al 2001, Moon-van der Staay et al 2001, Moreira and Lopez-Garcia 2002). Most of these studies were limited to <1,000 clones sequenced (Christen 2008), forcing researchers to focus on single samples. As shown by the preponderance of sequences registered only once in the library, and by statistical modeling (Hong et al 2006, Jeon et al 2008, Jeon et al 2006), current practices have universally resulted in only partial recovery of any single sample’s diversity and richness, have provided little to no sample replication within a particular environment, and have proved insufficient to settle biogeographic arguments regarding the global scale of microbial eukaryotic diversity, their degree of endemicity, and even the species richness of a single sample. High-throughput pyrosequencing technology promised to address methodological shortcomings by recovering uncommon, perhaps even exceedingly rare species (Huber et al 2007, Sogin et al 2006), but the short read lengths of 454 sequences made it necessary to rely on the existing long rRNA gene sequences to establish taxonomic
identities. Also, concerns remain about the role that sequencing errors may play in producing an
draftbf...time, the
design of PCR primers used for both Sanger and pyrosequencing approaches may be
significantly biased in their recovery of protists, possibly creating a distorted view of the extant
richness and diversity. Under these circumstances comparing two species lists is difficult and
these problems are compounded by uncertainties about nonparametric statistical methods
commonly used to estimate the size of the species pool.

In these companion papers, we attempt to address at least some of the limitations of the
studies published to date. Our experimental plan was to (i) focus on an environment rather than
on a sample, (ii) scale up sequencing efforts proportionally to the large number of samples
required to survey this environment, (iii) employ a dual sequencing (Sanger/pyrosequencing)
strategy, (iv) minimize the biases of both approaches, (v) and analyze the molecular diversity
data using statistical tools we developed for the purpose of estimating microbial richness. With
these goals, we established in 2004 a multi-year Microbial Observatory in the Cariaco Basin off
the coast of Venezuela. The Cariaco Basin is the world’s largest truly marine anoxic system
and it offers a large number of contrasting biogeochemical habitats that can be sampled in a
single hydrographic cast. We sampled the Basin extensively at three stations, in two
contrasting seasons, each time at four depths and across strong biogeochemical gradients
spanning fully oxygenated layers to deep, highly sulfidic habitats. We believe that these
samples represent fairly broad coverage of the Basin’s geochemistry, its protists, their spatial
heterogeneity and seasonal dynamics. In this first paper we present the largest to date set of
protistan gene sequences from an environment >800-nt long each (16,000 total, 6,489
protistan) for the purpose of providing reliable statistical estimates of total protistan richness
within our samples with meaningful standard error. To minimize bias, we employed a multiple
PCR primer strategy, and also complemented our Sanger clone libraries with ~82,000 quality
protistan sequences using massively parallel pyrosequencing of amplicons from the V9 region.
This dual sequencing approach enabled the first direct comparison of environmental 18S rRNA gene sequence diversity obtained by these two methods. The two sequence technologies showed substantial similarity in the types of protists recovered, and in their relative proportions. This tentatively indicated that both approaches represented the target diversity reasonably well.

**Material and Methods**

For expanded details see Supplementary Materials and Methods file.

**Sample Collection.** Samples were collected from three stations in the Cariaco Basin, Venezuela: Station A (10.50°N, 64.66°W), Station B (10.40°N, 64.46°W) and Station C (10.40°N, 65.35°W) (Fig. 1 and 2, Table S1). Samples for DNA extractions were collected at depths corresponding to 40m above the oxic/anoxic interface, the oxic/anoxic interface, 40m below the interface, and at 900m. The oxic/anoxic interface was defined as the depth at which oxygen became undetectable. This depth usually corresponded to a particle density maximum detected by transmissometry. Samples were withdrawn from Niskin bottles under N₂ pressure into sterile, intravenous bags immediately after retrieval and stored at *in situ* temperatures. Samples were then filtered onto 47mm Durapore membranes (0.45 µM pore size) onboard, with little to no exposure of the sample to the atmosphere. Depending on the cell concentration in each sample, 2.5-2.8 liters of seawater were filtered per sample under gentle vacuum (<25 cm Hg) until no passage of water through filters was observed. The variable filtration volume (filtering until passage of water through the filter ceased) does not completely control for differences in population abundances between our samples, but it likely reduces biases that may be introduced by potentially unequal sampling efforts. Filtration was conducted under a gentle vacuum (<25 cm Hg). Membranes were stored in cryovials containing DNA extraction buffer (Stoeck et al 2003) at -20°C until further processing.

**DNA preparation and PCR amplification of 18S rDNA.** High molecular weight DNA was
prepared using previously described methods (Stoeck et al 2003, Stoeck et al 2007). DNA quality was checked by agarose gel electrophoresis (1%), and the DNA yield was quantified using a Nanodrop ND-1000 UV-Vis spectrophotometer (Nanodrop Technologies, Wilmington, DE). The DNA yield from a minimum of two filters per depth at each station was combined prior to PCR. The 18S rRNA gene was amplified using four different primer combinations; Euk528F (5'-CGGTAATTCCAGCTCC-3) paired with either U1492R (5'-GGTTACCTTGTTACGACTT-3'), U1391R (5'-GGGCGGTGTGTACAARGR-3'), or U1517R (5'-ACGGCTACCTTGTTACGACTT-3') and Euk360F (5'-CGGAGARGGMGCTGAGA-3') paired with U1492R. All primer combinations were applied to all samples and PCR products from a minimum of 3 reactions per primer pair were pooled.

Clone Library Construction and Sequencing. To constrain cost, PCR products for stations B and C were pooled from the same depths prior to clone library construction. This admittedly caused loss of some information on distribution of protists but allowed a more in-depth sequencing of samples, better covering the Basin. Separate clone libraries were constructed for each of 4 depths from the 8 pooled B/C samples (4 from May and 4 from January 2005) and from the 8 station A samples, producing 64 different clone libraries. Inserts were sequenced bi-directionally using an Applied Biosystems 3730XL capillary sequencer. Processing of the data used PHRED, PHRAP (Ewing and Green 1998, Ewing et al 1998) and a pipeline script to call bases from chromatograms and perform quality control procedures. The sequences were checked for chimeras using the Bellerophon Chimera Check and the Check_Chimera utilities (Ribosomal Database Project) (Cole et al 2003). After removal of short sequences (< 800 bp), putative chimeras, bacterial, archaeal and metazoan sequences, remaining sequences were grouped into Operational Taxonomic Units (OTUs) based on 90, 95, 98, and 99% rRNA gene sequence similarity levels. This was achieved by first making all possible pair-wise sequence alignments by using ClustalW (Thompson et al 1994), calculating % sequence identities,
followed by clustering the sequences by using the unweighted pair group method with arithmetic mean (UPGMA) as implemented in the OC clustering program (http://www.compbio.dundee.ac.uk/Software/OC/oc.html). The number of OTUs and their frequencies at each cut-off value became the subject of statistical analyses. All protistan sequences have been deposited in GenBank under the accession numbers GU819239-GU825728.

Statistical Analyses of Clone Library Data. For each sample, we obtained "frequency count" data at the 90, 95, 98 and 99% sequence identity levels, i.e., the numbers of OTUs registered in the corresponding clone library only once (the "singletons"), or twice (the "doubletons"), etc. Based on these frequency count data we estimated the total number of OTUs at each % identity level, representing the sum of seen (empirically registered) and unseen OTUs (present but undetected due to limited sequencing effort). To do this we used the program CatchAll (Bunge 2011) to compute each of five nonparametric (Good-Turing, Chao1, ACE, ACE1, and Chao-Bunge) and eight parametric (Poisson; negative binomial; inverse Gaussian, Pareto and lognormal-mixed Poisson; and mixtures of one, two, or three geometrics) estimators, at every right-truncation point of the data (i.e., eliminating outliers at every possible cut-off point in the data), as described previously (Hong et al 2009, Hong et al 2006, Jeon et al 2008, Jeon et al 2006). For the nonparametric analyses we used a fixed right-truncation point (maximum frequency) equal to 10, and selected optimal analyses as recommended in the statistical literature (we also compared nonparametric results at higher truncation points, but these are typically nonsensical (Bunge and Barger 2008). For the parametric analyses, we first eliminated all results with an asymptotically corrected goodness-of-fit hypothesis test p-value less than 0.01. We then selected the analysis at each right-truncation point with a minimum value of AICc (Akaike's information criterion, corrected for sample size). Finally, we eliminated analyses with a standard error greater than 1/2 of the total richness estimate. From the remaining results, we selected the one at the highest right-truncation point (so as to use the maximum amount of
data) such that an uncorrected chi-square goodness-of-fit hypothesis test $p$-value exceeded 0.01 (or if this was not possible, the maximum corresponding $p$-value). This yielded one nonparametric and one parametric estimate of total richness for each sample at each % identity level.

**Massively Parallel 454 Pyrosequencing.** The PCR primers we used for 454 sequencing the V9 hypervariable region were 1,391F (5’-GTACACACCGCCCGTC-3’) and EukB (reverse) (5’-TGATCCTTCTGCAGGTTCACCTAC-3’). For each individual environmental DNA extract (each station, depth, and season combination) we ran three independent PCR reactions, which were pooled and cleaned using the MinElute PCR purification kit (Qiagen, Valencia, CA, USA). Pyrosequencing (Roche GS FLX Sequencing) was performed by MWG Biotech, Huntsville, Alabama. In total, we recovered 251,648 sequence reads that were subjected to quality control leaving us with 82,484 protistan sequences for further consideration. Tag sequences have been deposited in the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) under the accession number SRP003469. Annotated tag sequences are available from the authors.

**Bioinformatic Analysis of Tag Sequences.**
We proceeded with our analyses by following a pipeline developed earlier for inspecting and quality checking pyrosequencing reads (Stoeck et al 2009). Sequences shorter than 100 nucleotides, as well as those with at least one ambiguous position (N), were removed from consideration. To make taxonomic assignments, we first built a reference database containing all SSU rRNA gene sequences in public databases longer than 500 nucleotides. Each sequence tag was compared to these reference sequences using similarity searches with the program BLASTN and requesting up to 30 best hits, using the following BLAST parameters: -m 7 -r 3 -q -2 -G 6 -E 6  (Stoeck et al 2010). We then parsed the BLAST output to extract
taxonomic assignments at a series of thresholds for sequence similarity (70, 75, 80, 85, 90, 92, 95, 96, 97, 98, 99, and 100%). Sequence similarity was calculated as the sum of identities for non-overlapping (if any) HSP (High Scoring Pairs, see the BLAST documentation) divided by the length of the query sequence. For each query, we used only the most similar target sequence for which a good taxonomic assignment was provided (i.e. longer description in the OC field in the EMBL entries). To enable direct comparison between taxonomic assignments of 454- and Sanger produced sequences, the above BLAST analyses were repeated for nearly full-length 18S rRNA gene sequences, and separately for their V9 domains.

Results

1. Overall protistan diversity in the Cariaco Basin.

We examined the extent of protistan diversity in the Basin in two ways. First, we amplified the 18S rRNA gene from community DNA, followed by cloning and sequencing the amplicons. From each of the 64 libraries, we sequenced an average of 250 clones, resulting in over 16,000 rRNA gene sequences, or at a minimum, 672 sequences for each of the 16 samples. Culling of confirmed or suspected metazoan, chimeric, and short (< 800 bp) sequences left 6,498 good quality target 18S protistan sequences for further analyses. Table S2 presents the final number of sequences, grouped into OTUs sharing over 99% of gene sequence identity, and categorized by PCR primer pair, season, depth, and site. Second, we used the same DNA extracts as templates for 16 separate PCR reactions followed by pyrosequencing of the V9 region of the 18S rRNA gene. This produced a collection of over quarter of a million reads, with 173,723 meeting our quality control criteria, 52% of which were metazoan and other non-target sequences.

It is important to demonstrate the agreement between the taxonomic assignments established from the nearly full-length rRNA gene sequences vs the V9 tags contained therein. We achieved this by extracting the V9 region from 3107 of our clone sequences that contained
this region, and comparing the taxonomic assignments made by BLAST. Table S3 shows a general global agreement, as well as several minor differences. The proportion of sequences falling into the unassigned/unclassified categories is larger for the V9 tags and this effect appears stronger in Alveolata. We note that for taxonomic assignment of the pyrosequencing tags, we chose the ≥70% sequence similarity threshold because at this level fewer tags remained unassigned than at higher thresholds.

At the highest level of taxonomic assignment, approximately corresponding to kingdoms and phyla, the frequencies of different rRNA gene sequences appear similar between the conventional clone library/Sanger- and pyrosequencing methods (Fig. 3). Both collections were dominated by Alveolata, Rhizaria, Stramenopiles and unclassified 18S rRNA gene sequences, and in the same order of dominance, followed in both cases by Fungi and Euglenozoa. The most visible difference between Sanger- and pyrosequencing-based methods at this level of taxonomic assignment is the number of rarer taxa constituting <1% of respective sequence collections: 9 vs 14, respectively.

These gross similarities in the overall recovery of protists by the two methods were observed at lower levels of taxonomic assignment as well. For example, in both sequence collections, the Alveolata were dominated by the ciliate subphylum Intramacronucleata and four dinoflagellate orders (Gymnodiniales, Prorocetrales, Syndiniales, and Gonyaulacales), as well as unclassified alveolates (Fig. 4). Similarly, Stramenopiles were dominated by the same taxa regardless of the method used to detect them (data not shown). Expectedly, pyrosequencing detected more of the rare taxa (again defined as those comprising <1% of the entire collection): 11 vs 8 rare stramenopile taxa registered using clone libraries/Sanger sequencing. Classification to order was the lowest taxonomic level possible for the majority of our tag sequences. At this level, Sanger sequencing and pyrosequencing again recovered similar groups at similar frequencies (Robertson and Burke, 1989). For example, within the dinoflagellate class Dinophyceae, almost three quarters of all sequences were represented by
three orders present at essentially the same proportions in our Sanger- and pyrosequencing-based data sets (Gymnodiniales, Peridiniales, and Prorocentrales). The other abundantly represented orders were also in the same proportions (Syndiniales and Gonyaulacales) (Fig. 5). We note that these gross similarities between the two libraries appear independent of the cutoff level of query/hit similarity used to assign V9 sequences to taxa by BLAST, as raising this cutoff to 80% did not affect the overall pattern (data not shown).

2. Biases of the rRNA gene sequence collection.

To generate conventional 18S rRNA gene clone libraries, we used four different PCR primer sets applied to each DNA extract obtained. Table 1 provides a representative subset of these data summarizing the number of occurrences of ciliates by primer set. We chose to present data for this group because we recovered roughly the same number of rRNA gene sequences (≈ 100) for each of the four PCR primer sets, which simplified the comparison of their respective performances. Also, the relatively large size of the ciliate 18S rRNA gene sequence collection minimized the influence of stochastic variation on the pattern of recovery. Differential recovery of several ciliate clades illustrates the substantial biases of the PCR primer pairs used, such as an apparent discrimination against Euplotida by the 360F/1492R pair, or discrimination for Grossglockneriidae and Choreotrichida by the 528F/1391R pair.

3. Total protistan richness in the Cariaco Basin.

We clustered protistan 18S rRNA gene sequences recovered from our clone libraries into OTUs at different levels of sequence identity. The 6,498 individual sequences fell into 2,099, 1,496, 767, and 392 OTUs at 99, 98, 95, and 90% identity levels, respectively. We then reconstructed the OTU frequency distributions (Fig. 6), and modeled these distributions to estimate the size of microbial richness that must exist in our samples, and therefore in the Basin to account for the empirical data collected, using the program CatchAll (Bunge 2011).
accordance with previous experience (Bunge and Barger 2008, Hong et al 2009, Jeon et al 2008), the best-fitting models are finite mixtures of one, two, or three geometric distributions, while the Poisson, negative binomial, inverse Gaussian-, Pareto- and lognormal-mixed Poisson provided inferior fits or presented other statistical anomalies (data not shown). Typically higher-order mixtures are required to produce acceptable richness estimates and goodness-of-fit assessments that satisfy our criteria (see Materials and Methods). For purely comparative purposes, Table 2 also presents the richness predictions based on nonparametric richness lower bound Chao1 and the nonparametric estimator ACE1; these are shown only for comparison with the parametric estimates.

The estimates of the total protistan richness in our samples based on 2- and 3-mixed exponential distributions show that they may harbor thousands of protistan genera (defined as OTUs formed at 95-90% gene sequence similarity), and up to tens of thousands of protistan species (defined as OTUs formed at 98-99% gene sequence similarity), indicating that our empirical collection sampled about 20-30% of all the genera, and about 5-10% of all the species in our samples. By extension, these are minimal estimates of protistan richness in the entire Basin.

**Discussion**

The first of our companion papers offers a general description of the Cariaco Basin, the overall view of protistan diversity we detected, and a comparison of the 18S rRNA gene sequence collections produced by the tandem approach of two alternative sequencing technologies (Sanger and pyrosequencing). The Cariaco Basin is an excellent model environment to study protistan species richness, phylogenetic diversity, habitat specialization, community structure, its dynamics, and global uniqueness. This is because the Basin is a mosaic of dramatically different biogeochemical niches, and has existed as such for millions of years (Robertson and Burke 1989), though it likely went through periods of oxidation (Peterson et al 2000).
Cariaco water column progresses from fully oxic to sulfidic across a temporally varying boundary between 250 and 350m. Within the redox transition zone lie strong gradients in $O_2$, $NO_3^-$, $H_2S$, $NH_4^+$, $NO_2^-$, $PO_4^{3-}$ and $CH_4$, and enrichments in $S_2O_3^{2-}$, $SO_3^{2-}$, $S^0$, $Mn^{2+}$ and $Fe^{2+}$ that select for specific bacterial and archaeal phylotypes (Li et al 2008, Lin et al 2008, Scranton et al 2006, Taylor et al 2001). This transition zone exhibits a peak in prokaryotic metabolic activity and cell numbers, which often coincides with peaks in protist cell numbers.

Protists are important members of aquatic microbial communities due to their autotrophic and heterotrophic activities. Their grazing of prokaryotic and other eukaryotic microbes regenerates nutrients, modifies or re-mineralizes organic matter (e.g. Sherr and Sherr, 2002, Taylor, 1982, Taylor et al., 1985), and can affect the quantity, activity and physiological state of their prey (Frias-Lopez et al., 2009; Madsen et al., 1991). Grazing releases inorganic nutrients that often limit primary production, and makes organic carbon available to higher trophic levels (Berman et al., 1987). In our companion paper (Orsi et al., this issue [Chapter 2 of this dissertation]), we identify many new clades at species to class levels, some of which appear restricted to specific layers of the water column and have a significantly non-random distribution. We also use multivariate community analyses coupled with parametric richness estimates for selected regions of the Basin to demonstrate that distinct communities of protists exist within close proximity (40 m) to each other. These findings suggest many pelagic protists are specialized to specific habitats, and likely diversify, at least in part due to separation by geochemical barriers. For a discussion of the influence of biogeochemical factors on protistan community structure, see Orsi et al. (this issue [Chapter 2 of this dissertation]).

1. Overall protistan diversity in the Cariaco Basin.

Owing to the scale of sequencing efforts, we anticipated and demonstrated recovery of the expectedly large diversity that contains representatives of all major clades (Fig. 3). The remarkable similarity between the results produced by the two sequencing approaches
employed came as a surprise. Indeed, at all three hierarchical levels examined, both sequence collections were dominated by the same taxa, often present at essentially the same frequencies (Figs. 3-5). Such comparisons are rare, and, to the best of our knowledge, ours is the first published report of such a comparison for protists. While similarity at the highest level of taxonomic hierarchy (Fig. 3, 4) is not counterintuitive, we find it remarkable that the overall similarity between compositions of the dominant clades holds in many cases down to the level of protistan orders (Fig. 5). For example, three dinoflagellate orders, Gymnodiniales, Peridiniales, and Prorocentrales, comprised 72% of either sequence collection, at essentially the same frequencies. We did not extend these analyses to the levels of family or genus because this level of assignment is not currently reliable for pyrosequencing data. The differences between the two sequence collections do not appear until we consider relatively rare taxa with frequencies in single percentage points of the total respective databases. Phyla represented by less than 1% of the total number of sequences are more numerous in the collection produced by pyrosequencing than Sanger sequencing (Fig. 3), which we attribute to the sequencing depth afforded by the former method. We note however that the impact of pyrosequencing errors on recovered diversity has not yet been fully established (Kunin et al 2010), and we interpret the rare taxa with caution.

2. Biases of the rRNA gene approach and their minimization.

One of our goals was to produce a minimally biased molecular survey of protists, and we attempted to minimize the PCR primer biases (Hong et al 2009, Jeon et al 2008, Polz and Cavanaugh 1998, Suzuki and Giovannoni 1996) by using a multiple-primer approach (Stoeck et al 2006). Recently, our group showed statistically that no tested pair of conserved primers could provide amplification of all templates present in the DNA extract, and that a combination of PCR reactions each using different primers led to qualitatively richer clone libraries for both protists (Jeon et al 2008) and bacteria (Hong et al 2009). The difference in performance of the
four primer pairs used here is illustrated in Table 1, which gives examples of organisms completely missed from the clone libraries created by all but one PCR primer set, or organisms universally detected by all but one primer pair. Note that for Table 1 we chose to present representatives of the phylum Ciliophora. Such a choice is conservative because PCR primers have been designed using a significant number of ciliate sequences, potentially leading to less discrimination against these organisms compared to less studied groups.

We tentatively conclude that the multiple PCR primer approach provided a more balanced view of microbial composition in the Cariaco Basin, and that this study is less biased than the majority of prior protistan surveys (including, e.g., our own earlier study of the Cariaco Basin, (Stoeck et al 2003). Coupled with the size of the data set, we are well-positioned to model distributions in order to estimate the total number of protists in the Basin, including those missed by our libraries.

3. Total protistan richness in the Cariaco Basin.

We used a new statistical approach (Hong et al 2006) to model OTU frequency distributions and provide estimates of the total protistan richness in our samples, and thus conservatively in the entire Basin*. Our group developed this method as a general tool for richness estimation (Hong

* The estimate of total taxonomic richness for our samples is an estimate of the total richness that would be observed if sampling were carried out using the same procedures, but with infinite effort, i.e., exhaustively. It is also an approximation of the total richness of the Basin, under the assumption that the conditions occurring in the individual samples are "representative," that is, that the conditions of the individual samples represent an approximately random sample within a given biogeochemical regime extant in the Basin. The locations for sample collection were carefully selected based on a decade of prior Cariaco Time Series data showing the three sites and four depths we sampled to be representative of the
et al. 2006) and later applied it to microbial eukaryotes (Edgcomb et al 2009, Stoeck et al 2007, Jeon et al 2006). Here we produced a sequence collection large enough to be modeled by our parametric statistical method, with an appropriate goodness-of-fit, and most importantly to obtain a standard error sufficiently low as to not render the estimate uninformative. We compared essentially all known parametric and nonparametric estimates for the total richness, along with associated statistics including standard errors and goodness-of-fit diagnostics. Our data also illustrated a potential underestimation of total richness by nonparametric estimators. Table 3 shows that, at the level of OTUs based on 99% gene sequence identity, the OTUs empirically observed in our samples constitute about 10% of all predicted to be present, and the difference between richness predictions based on parametric vs nonparametric approaches is substantial. As OTUs become more inclusive, and the portion of the sampled richness increases, such differences become smaller; for OTUs based on 90% gene sequence identity, the parametric modeling- and nonparametric estimators-based richness predictions converge on essentially same values.

Our approach enabled us to predict how many OTUs must be present in our samples, and thus conservatively in the Basin: 35,968 ± 13,123 (SE) and 4,039 ± 798 (SE) for OTUs comprised of 18S rRNA gene sequences sharing >99% and >95% identity, respectively. These OTU cutoffs are roughly comparable to species and genera. We argue that this is the first estimate of total protistan richness that is relatively unbiased, and with realistic standard errors. We note that, while individual clone libraries sampled the respective environments to varying

range of biogeochemical conditions in the Basin. It is, however, possible that distributions of OTUs in the Basin are even patchier than we detected, due to additional variations in conditions that we did not sample. This would cause an unknown number of OTUs to be missed in our samples. If this is the case, our estimates of the total diversity are likely to be conservative.
degrees (i.e., similarly-sized libraries cover diversity differently if the target habitats have different richness), this has no bearing on our estimates. Indeed, the sample size does not influence the value of the sample’s predicted richness, only the standard error of its estimate. Some bias in the estimators of total richness may be present for small samples, but this is not a major factor here. Furthermore, differences in sampling intensity may have been minimized by the variable filtration volumes we applied (filtering until filters clogged -- see above).

It is instructive to compare our results to the previous richness estimates obtained using a single 3L sample from the Cariaco Basin (395 ± 153 (SE) and 164 ± 40 (SE) at 99% and 95% OTU clustering, respectively; Jeon et al 2006). Samples in this study covered temporal, spatial, and biogeochemical variations in the Basin, and resulted in estimates of diversity that are 2 to 3 orders of magnitude higher. This indicates a very patchy distribution of species over the volume of the Basin, and tentatively suggests the possibility of a high degree of habitat specialization.

The debate over the degree to which protist communities exhibit global distribution has not been fully resolved (Bass et al., 2007, Boenigk et al 2006, Caron 2009; Foissner 1999, Foissner, 2006, Lawley et al 2004). We would like to emphasize that all existing knowledge of microbial diversity comes from samples, and, strictly speaking, is limited to the collection of samples analyzed so far – rather than environments from which they were obtained. Under these circumstances, a reasonable analysis would be to compare a proxy for global richness, such as the total number of protistan species known from samples collected to date, to a proxy for richness in a large an environment, such as a collection of samples covering its multitude of habitats. This collection should be reasonably large because, as Caron (2009) and Ramette and Tiedje (2006) argue, a species could be globally distributed even if it is not found in each and every sample. The express purpose of our study was to overcome a common limitation of microbial diversity studies, typically based on one to a few samples per environment. We did so by focusing on a large system, and revisiting multiple locations in different seasons, and sampling across its principal biogeochemical regimes, without sacrificing the per sample depth
of sequence coverage. It should be noted that not all protist populations were sampled in this study, because we intentionally excluded the photic zone. There is an additional important caveat: the total richness, whether in a single sample or collection of samples, can only be estimated statistically, as even these pyrosequencing efforts do not allow for a complete coverage of sample(s) richness. The statistical tools typically used in microbial diversity studies do not necessarily work well for the purpose (Chao and Bunge 2002). The state-of-the-art statistical tool kit we developed, coupled with a large-scale survey of an environment as large and chemically diverse as the Cariaco Basin, meet the above requirements, and provide us with what is needed for a proxy for total protistan richness in the aphotic zone of the Cariaco Basin. Our analyses show that this proxy, or the number of species of microbial eukaryotes predicted to exist in the collection of our samples, appears to be a substantial figure, but it falls far short of the total number of protistan species described to date (100,000 to 500,000 species (Adl et al 2007, Corliss 1982) by being 3 to 15 times lower, suggesting a degree of endemism in this environment.

Conclusions
As part of the Cariaco Microbial Observatory, we collected the largest, and possibly the least biased, collection of protistan molecular signatures to date from a single environment. Two sequencing approaches produced globally similar pictures of protistan diversity, which contain representatives of all major protistan clades known today. Parametric statistical modeling predicted the total number of species in samples spanning different locations, seasons, and a range of diverse habitats within the Cariaco Basin. This number is far lower than the global known richness, indicating a degree of endemism in the distribution of the majority of microbial eukaryotes.
Acknowledgements

We thank the captain and crew of the B/O Hermano Gines and the staff of the Fundación La Salle de Ciencias Naturales, Margarita Island for their assistance during our field work in Venezuela, particularly Yrene Astor. We are grateful to the dedicated researchers of the CARIACO biogeochemical time series program without whom this work would not have been possible. We would like to thank Valentin Ilyin (Northeastern University), David Mark Welch (MBL), and Jed Goldstone (WHOI) for helpful discussions on bioinformatics approaches, Hilary Morrison and Rich Fox (MBL) for the use of their pipeline scripts for sequence data processing, and the Academic Research Computing User Group at Northeastern University for permitting us to use the Opportunity computer cluster. We thank Linda Woodard for supervising the statistical computations. Comments from two anonymous reviewers significantly improved this manuscript. This research was conducted using the resources of the Cornell University Center for Advanced Computing, which receives funding from Cornell University, New York State, the National Science Foundation, and other leading public agencies, foundations, and corporations. This research was supported by grants from NSF (MCB-0348341 and DEB-0816840 to SSE, DEB-0816638 to JB, MCB-0348407 to VE and OCE 03-26175 and MCB-03-47811 to GTT), from ANR-Biodiversité project Aquaparadox and the ERA-net project BiodivERsA BioMarkKs program to RC, and from Venezuela FONACIT (nos. 96280221 and 2000001702 to RV). This is a contribution # 268 from the Marine Science Center, Northeastern University, Nahant MA, U.S.A.
Figures and Tables

Figure 1. Map of study site. Station A represents the site of the U.S.-Venezuelan CARIACO biogeochemical times series program in the eastern sub-basin. Station B is a shallower station south of the Tortuga channel, the source of incoming Caribbean waters. Station C is centered in the less productive western sub-basin.
Figure 2. Vertical profiles of oxygen, hydrogen sulfide, and bacterial and flagellate abundances at the time of sampling for protistan DNA in (a) January and (b) May 2005; modified from (Taylor et al., 2001; Scranton et al., 2006).
Figure 3. Phylum and kingdom-level assignments of the 18S rRNA gene sequence collections obtained by conventional cloning/Sanger sequencing and 454 pyrosequencing approaches.
Figure 4. Subphylum and class level assignments of the alveolate 18S rRNA gene sequence collections obtained by conventional cloning/Sanger sequencing and 454 pyrosequencing approaches.
Figure 5. Order assignments of the Dinophyceae 18S rRNA gene sequence collections obtained by conventional cloning/Sanger sequencing and 454 pyrosequencing approaches.
Figure 6. The fit of the 3-mixed exponential distribution (solid line) to the empirical data points (filled circles; OTUs clustered at 99% 18S rRNA gene sequence identity).
<table>
<thead>
<tr>
<th></th>
<th>3F14R</th>
<th>5F13R</th>
<th>5F14R</th>
<th>5F15R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choreotrichida</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Colpodidae</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cryptocaryon</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Epalxellidae</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Euplotida</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Furgasoniidae</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Grossglockneriidae</td>
<td>9</td>
<td>74</td>
<td>3</td>
<td>46</td>
</tr>
<tr>
<td>Halteriidae</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hausmanniella</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nassulidae</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Parameciidae</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Philasterida</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Pleuronematida</td>
<td>8</td>
<td>0</td>
<td>31</td>
<td>11</td>
</tr>
<tr>
<td>Prorodontidae</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Stichotrichida</td>
<td>79</td>
<td>0</td>
<td>58</td>
<td>32</td>
</tr>
<tr>
<td>Strombidiidae</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>Tintinnida</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>104</td>
<td>107</td>
<td>106</td>
<td>117</td>
</tr>
</tbody>
</table>

Table 1. Differential recovery of ciliate taxa by four different PCR primer sets.
Table 2. Predicted richness of protistan assemblage in the Cariaco Basin, with associated statistics (SE: standard error; GOF: goodness-of-fit (p-value for the corrected Pearson chi-square goodness-of-fit test); CB: 95% Confidence Interval; NA: Not applicable).

<table>
<thead>
<tr>
<th>% Identity of sequences within OTU</th>
<th>Statistic</th>
<th>Best Parametric Model</th>
<th>Nonparametric estimators</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥99</td>
<td>Estimated Total number of OTUs</td>
<td>35968</td>
<td>8960</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>13123</td>
<td>613</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>18373-72596</td>
<td>7860-10269</td>
</tr>
<tr>
<td></td>
<td>GOF</td>
<td>0.1396</td>
<td>NA</td>
</tr>
<tr>
<td>≥98</td>
<td>Estimated Total number of OTUs</td>
<td>12480</td>
<td>4351</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>3489</td>
<td>282</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>7480-21666</td>
<td>3850-4959</td>
</tr>
<tr>
<td></td>
<td>GOF</td>
<td>0.8204</td>
<td>NA</td>
</tr>
<tr>
<td>≥95</td>
<td>Estimated Total Sp</td>
<td>4039</td>
<td>1951</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>798</td>
<td>156</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>2815-6015</td>
<td>1685-2300</td>
</tr>
<tr>
<td></td>
<td>GOF</td>
<td>0.0598</td>
<td>NA</td>
</tr>
<tr>
<td>≥90</td>
<td>Estimated Total number of OTUs</td>
<td>1143</td>
<td>748</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>162</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>CI</td>
<td>887-1532</td>
<td>630-924</td>
</tr>
<tr>
<td></td>
<td>GOF</td>
<td>0.6788</td>
<td>NA</td>
</tr>
<tr>
<td>Sample Site</td>
<td>Sample Depth (m)</td>
<td>Sample Depth (m)</td>
<td>Dissolved CO₂ (µM)</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Station A</td>
<td>240 (A) 185 (A)</td>
<td>280 (I) 225 (I)</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>320 (B) 265 (B)</td>
<td>900 (D) 900 (D)</td>
<td>0.00</td>
</tr>
<tr>
<td>Station B</td>
<td>260 (A) 216 (A)</td>
<td>300 (I) 256 (I)</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>340 (B) 296 (B)</td>
<td>400 (D) 670 (D)</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Station C</td>
<td>240 (A) 230 (A)</td>
<td>280 (I) 270 (I)</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>320 (B) 310 (B)</td>
<td>900 (D) 900 (D)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Supplementary Table 1. Hydrographic conditions at sampling stations and depths. (n/a – not available due to sensor malfunction).
Supplementary Table 2. The number of rRNA gene sequences analyzed in this study classified by season, biogeochemical habitat, station, and PCR primer pair from each sample.

<table>
<thead>
<tr>
<th>Sample/Primer set</th>
<th>Number of protistan sequences (99% OTUs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 2005</td>
<td>3985 (1450)</td>
</tr>
<tr>
<td>January 2005</td>
<td>2513 (719)</td>
</tr>
<tr>
<td>40m above oxic/anoxic interface</td>
<td>2211 (702)</td>
</tr>
<tr>
<td>Within the redoxcline</td>
<td>1737 (702)</td>
</tr>
<tr>
<td>40m below oxic/anoxic interface</td>
<td>1560 (433)</td>
</tr>
<tr>
<td>900m sample from deep anoxic and sulfidic waters</td>
<td>981 (335)</td>
</tr>
<tr>
<td>Station A</td>
<td>3886 (1207)</td>
</tr>
<tr>
<td>Stations B and C pooled</td>
<td>2603 (961)</td>
</tr>
<tr>
<td>5F15R</td>
<td>1724 (577)</td>
</tr>
<tr>
<td>5F14R</td>
<td>1266 (358)</td>
</tr>
<tr>
<td>5F13R</td>
<td>2343 (938)</td>
</tr>
<tr>
<td>3F14R</td>
<td>1165 (296)</td>
</tr>
</tbody>
</table>
Supplementary Table 3. Taxonomic assignments made by a test dataset of 3107 18S rRNA clone sequences and the V9 tag sequences extracted from them (see Materials and Methods).

<table>
<thead>
<tr>
<th>Domain</th>
<th>Phylum</th>
<th>Clone Sequence</th>
<th>Extracted V9 Tag Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>105</td>
<td>387</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Alveolata</td>
<td>1174</td>
<td>854</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Amoebozoa</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Centr heliozoa</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Choano flagellida</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Euglenozoa</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Fungi</td>
<td>162</td>
<td>130</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Haptophysae</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Ichthyosporea</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Metazoa</td>
<td>15</td>
<td>56</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Rhizaria</td>
<td>1084</td>
<td>1041</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Telonemida</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Viridiplantae</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>Unclass. Env</td>
<td>303</td>
<td>417</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>stramenopiles</td>
<td>213</td>
<td>186</td>
</tr>
</tbody>
</table>
Supplemental Material and Methods

Sample Collection. Samples were collected on 17-19 January and 23-25 May 2005 from three stations in the Cariaco Basin, Venezuela: Station A (10.50°N, 64.66°W), Station B (10.40°N, 64.46°W) and Station C (10.40°N, 65.35°W) (Fig. 1 and 2). Sampling was conducted aboard the B/O Hermano Gines, operated by Estación de Investigaciones Marinas (EDIMAR), Fundación la Salle de Ciencias Naturales, located on Margarita Island, Venezuela. Hydrographic conditions (salinity, temperature, pressure, turbidity, chlorophyll fluorescence, dissolved O₂ concentrations) were continuously profiled with a SeaBird CTD instrument package and YSI oxygen sensor mounted on a rosette. Water samples were collected by 12 TFE-lined, 8-12L Niskin bottles mounted on the rosette. O₂ concentrations were validated by Winkler titrations of discrete samples from Niskin bottles. Each station was profiled for additional key biogeochemical variables (H₂S, NH₄⁺, NO₃⁻, NO₂⁻, PO₄³⁻, CH₄, pH, prokaryotic biomass and productivity) at 12 (Stn. B) to 18 (Stns. A and C) discrete depths using Niskin bottle samples (see (Scranton et al 2006, Taylor et al 2001), Fig. 2, and Table S1 for details and typical profiles). Samples for DNA extractions were collected at depths corresponding to 40m above the oxic/anoxic interface, within the oxic/anoxic interface, 40m below the interface, and at 900m. The oxic/anoxic interface was defined as the depth at which oxygen became undetectable. This depth usually corresponded to a particle density maximum as detected by the Sea Tech c-beam Transmissometer on the rosette. Sampling protocols prevented exposure of the water samples to the atmosphere and minimized alterations in the water’s redox potential. Samples were withdrawn from the Niskin bottles on deck under N₂ pressure into sterile, evacuated 2-L hospital intravenous bags (non-DEHP vinyl; Secure Medical Inc., Whitewater, Wis.) immediately
after retrieval and stored in large coolers of seawater at \textit{in situ} temperatures. Samples were then filtered onto 47mm Durapore membranes (0.45 $\mu$M pore size) using a multiplex high volume filtration unit that allowed for simultaneous onboard filtration of eight 2 to 4L samples, with little to no exposure of the sample to the atmosphere. Depending on the cell concentration in each sample, 2.5-2.8 liters of seawater were filtered per sample under gentle vacuum (<25 cm Hg) until no passage of water through filters was observed. The variable filtration volume (filtering until passage of water through the filter ceased) does not completely control for differences in population abundances between our samples, but it likely reduces biases that may be introduced by potentially unequal sampling efforts. Filtration was conducted under a gentle vacuum (<25 cm Hg). Membranes were stored in cryovials containing 1 ml of DNA extraction buffer (100 mM Tris-HCl pH8, 100 mM Na$_2$EDTA pH8, 100 mM NaPO$_4$ pH8, 1.5M NaCl) with 1% CTAB, and Proteinase K (0.1 mg ml$^{-1}$ final) at -20°C until further processing.

**DNA preparation and PCR amplification of 18S rDNA.** High molecular weight DNA was prepared using previously described methods (Stoeck et al 2003, Stoeck et al 2007). Briefly, the frozen filter was shaken horizontally for 30 min. at 37°C, 225 rpm after adding 5 ml DNA extraction buffer and 25 µl proteinase K (20 mg/ml). Filters were then incubated at 65°C after adding 0.8 ml 20% SDS, inverting every 15-20 minutes. The supernatant was transferred to a 50 ml centrifuge tube (PC or PPCO) and the filter was extracted a second time by adding another 2.5 ml of the DNA extraction buffer and 0.25 ml 20% SDS, vortexing for 10 seconds, and incubating for 10 minutes in 65°C water bath. The supernatants were combined and extracted with 1 volume chloroform:isoamyl alcohol (24:1) by mixing and
centrifuging at 6,000 x g for 10 minutes at room temperature. The supernatant was precipitated with 0.7 volume isopropanol at room temperature for one hour, and then centrifuged at 16,000 x g for 20 minutes at room temperature. The pellet was washed with 3 ml 70% ethanol, and centrifuged again for 15 minutes at room temperature. The pellet was dried and re-suspended in 50-100 µl TE buffer (pH 8). DNA quality was checked by agarose gel electrophoresis (1%), and the DNA yield was quantified using a Nanodrop ND-1000 UV-Vis spectrophotometer (Nanodrop Technologies, Wilmington, DE). The DNA yield from minimum 2 filters per depth at each station was combined prior to PCR. A multiple primer approach was used to minimize PCR primer bias. The 18S rRNA gene was amplified using four different primer combinations; Euk528F (5'-CGGTAATTCCAGCTCC-3') paired with either U1492R (5'-GGTTACCTTCTTGTACGACTT-3'), U1391R (5'-GGGCGGTGTGTACAARGR-3'), or U1517R (5'-ACGGCTACCTTGTACGACTT-3') and Euk360F (5'-CGGAGARGGMGCMGTAGA-3') paired with U1492R. All primer combinations were applied to all samples (for January and May 2005 samples were collected from three stations and water was filtered from four depths/station, for a total of 24 samples).

Clone Library Construction and Sequencing. Amplified DNA was checked for quality by agarose gel electrophoresis, and cloned into the vector TopoXL (Invitrogen). PCR products for stations B and C were pooled from the same depths prior to clone library construction. A minimum of two PCR amplifications per primer combination were pooled prior to cloning. Separate clone libraries were constructed for the products of each PCR primer combination for the 8 pooled B/C samples (4 from May and 4 from January 2005) and from the 8 station A samples. This produced 64 different clone libraries. Plasmid DNA from an
average of two to three 96-well plates of clones from each library was prepared using a MWG Biotech RoboPrep2500, and inserts were sequenced bi-directionally using an Applied Biosystems 3730XL capillary sequencer at the Josephine Bay Paul Center at the Marine Biological Laboratory (MBL), Woods Hole, MA (a total of approximately 16,000 clones were sequenced). Processing of the data used PHRED, PHRAP (Ewing and Green 1998, Ewing et al 1998) and a pipeline script to call bases from chromatograms, perform quality control procedures including checks for data consistency, data integrity, and data quality, to trim vector and low quality data and to assemble the sequences into contigs. The sequences were checked for chimeras using the Bellerophon Chimera Check and the Check_Chimera utilities (Ribosomal Database Project) (Cole et al 2003). After removal of short sequences (< 800 bp), putative chimeras, bacterial, archaeal and metazoan sequences (leaving 6498 sequences), the remaining sequences were grouped into Operational Taxonomic Units (OTUs) based on 90, 95, 98, and 99% rRNA gene sequence similarity levels. This grouping was achieved by first making all possible pair-wise sequence alignments by using ClustalW (Thompson et al 1994) at default settings and calculating % sequence identities, followed by clustering the sequences into OTUs by using the unweighted pair group method with arithmetic mean (UPGMA) as implemented in the OC clustering program (http://www.compbio.dundee.ac.uk/Software/OC/oc.html). The OTU grouping was checked manually to verify that all OTUs were assembled at the cut-off level desired. The number of OTUs and their frequencies at each cut-off value became the subject of statistical analyses. All protistan sequences have been deposited in Genbank under the accession numbers GU819239-GU825728.
Statistical Analyses of Clone Library Data. For each sample, we obtained "frequency count" data at the 90, 95, 98 and 99% sequence identity levels, i.e., the numbers of OTUs registered in the corresponding clone library only once (the "singletons"), or twice (the "doubletons"), etc. Based on these frequency count data we estimated the total number of OTUs at each % identity level, representing the sum of seen (empirically registered) and unseen OTUs (present but undetected due to limited sequencing effort). To do this we used the program CatchAll (Bunge 2011) to compute each of five nonparametric (Good-Turing, Chao1, ACE, ACE1, and Chao-Bunge) and eight parametric (Poisson; negative binomial; inverse Gaussian, Pareto and lognormal-mixed Poisson; and mixtures of one, two, or three geometrics) estimators, at every right-truncation point of the data (i.e., eliminating outliers at every possible cut-off point in the data), as described previously (Hong et al 2009, Hong et al 2006, Jeon et al 2008, Jeon et al 2006). For the nonparametric analyses we used a fixed right-truncation point (maximum frequency) equal to 10, and selected optimal analyses as recommended in the statistical literature (we also compared nonparametric results at higher truncation points, but these are typically nonsensical (Bunge and Barger 2008). For the parametric analyses, we first eliminated all results with an asymptotically corrected goodness-of-fit hypothesis test $p$-value less than 0.01. We then selected the analysis at each right-truncation point with a minimum value of AICc (Akaike's information criterion, corrected for sample size). Finally, we eliminated analyses with a standard error greater than 1/2 of the total richness estimate. From the remaining results, we selected the one at the highest right-truncation point (so as to use the maximum amount of data) such that an uncorrected chi-square goodness-of-fit hypothesis test $p$-value exceeded 0.01 (or if this was not possible, the maximum corresponding $p$-value). This yielded one nonparametric
Massively Parallel 454 Pyrosequencing. The PCR primers we used for 454 sequencing flanked the V9 region of eukaryote SSU rRNA genes and produced an average read length of 138 bases. These primers were 1,391F (5'-GTACACACCGCCCGTC-3') and EukB (reverse) (5'-TGATCCTTCTGCAGGTTCACCTAC-3'). For each individual environmental DNA extract (each station, depth, and season combination) we ran three independent 30-µl PCR reactions with reaction mix consisting of 5U of Pfu Turbo polymerase (Stratagene, La Jolla, CA, USA), 1x Pfu reaction buffer, 200 µm dNTPs (Pierce Nucleic Acid Technologies, Milwaukee, WI, USA), a 0.2 µM concentration of each primer in a volume of 100 µl, and 3-10 ng genomic DNA as template. The PCR protocol employed an initial denaturation at 94°C for 3 min; 30 cycles of 94°C 30 s, 57°C for 45 s, and 72°C for 1 min; and a final 2 min extension at 72°C. PCR products from all DNA samples were pooled and cleaned using the MinElute PCR purification kit (Qiagen, Valencia, CA, USA). Pyrosequencing (Roche GS FLX Sequencing) was performed by MWG Biotech, Huntsville, Alabama. In total, we recovered 251,648 sequence reads that were subjected to quality control. Removal of low quality sequences (see below) left us with 173,723 high-quality reads, including 82,484 protistan sequences, for further consideration. Tag sequences have been deposited in the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) under the accession number SRP003469. Annotated tag sequences are available from the authors.
Tag sequences ranged from 52 to 280 nucleotides in length, with 131-nucleotide long reads representing the most abundant size category. Not all sequences reached the distal primer, thus a "loose" dereplication procedure was used in which shorter sequences were considered to be identical to longer sequences when they overlapped with 100% identity. Sequences shorter than 100 nucleotides, as well as those with at least one ambiguous position (N), were removed from consideration. During the follow up process of taxonomic assignment (see below), we further removed non-ribosomal and suspected chimera sequences.

We proceeded with our analyses by following a pipeline developed earlier for inspecting and quality checking pyrosequencing reads (Stoeck et al 2009). To make taxonomic assignments, we first built a reference database containing all SSU rRNA gene sequences in public databases longer than 500 nucleotides, and formatted this database using the formatdb utility of the NCBI BLAST stand-alone package. Each sequence tag was compared to these reference sequences using similarity searches with the program BLASTN and requesting up to 30 best hits, using the following BLAST parameters: -m 7 -r 3 -q -2 -G 6 -E 6 (Stoeck et al 2010). We then parsed the BLAST output to extract taxonomic assignments at a series of thresholds for sequence similarity (70, 75, 80, 85, 90, 92, 95, 96, 97, 98, 99, and 100%). Sequence similarity was calculated as the sum of identities for non-overlapping (if any) HSP (High Scoring Pairs, see the BLAST documentation) divided by the length of the query sequence, which is a superior method than considering solely the first HSP (the latter may include only a short conserved domain thus not reflecting the true similarity between the two sequences). For each query, we retained for the follow up
analyses only the most similar target sequence for which a good taxonomic assignment was provided (i.e. longer description in the OC field in the EMBL entries). The underlying hypothesis is that a well-described taxonomy is often the result of a proper phylogenetic analysis, while a short description is often the result of assignments made using the most similar sequence, often resulting in errors. To enable direct comparison between taxonomic assignments of 454- and Sanger produced sequences, the above BLAST analyses were repeated for nearly full-length 18S rRNA gene sequences, and separately for their V9 domains.
REFERENCES


Chapter 2: Habitat specialization
Protistan Microbial Observatory in the Cariaco Basin, Caribbean. II. Habitat specialization.

W. Orsi¹, V. Edgcomb², S. O. Jeon³, C. Leslin¹, J. Bunge⁴, G. T. Taylor⁵, R. Varela⁶, S. Epstein¹,⁷

¹ Department of Biology, Northeastern University, 313 Mugar Building, Boston, MA 02115, USA.
² Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA.
³ Department of Environmental Science, Kangwon National University, Kangwon-Do, South Korea.
⁴ Department of Statistical Science, 301 Malott Hall, Cornell University, Ithaca, NY 14853, USA.
⁵ School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York 11794, USA.
⁷ Marine Science Center, Northeastern University, 430 Nahant Rd, Nahant MA 01908.

Running title: Diversity, ecology, and distribution of marine microbial eukaryotes

Keywords: protists, diversity, anoxic, 18S rRNA approach

Subject Categories: Microbial ecology and functional diversity of natural habitats, and/or Microbial population and community ecology

* corresponding author
Abstract

This is the second paper in a series of three that investigate eukaryotic microbial diversity and taxon distribution in the Cariaco Basin, Venezuela, the ocean's largest anoxic marine basin. Here, we use phylogenetic information, multivariate community analyses, and statistical richness predictions to test whether protists exhibit habitat specialization within defined geochemical layers of the water column. We also analyze spatio-temporal distributions of protists across two seasons and two geographic sites within the Basin. Nonmetric multidimensional scaling indicates that these two Basin sites are inhabited by distinct protistan assemblages, an observation that is supported by the minimal overlap in observed and predicted richness of sampled sites. A comparison of parametric richness estimations indicates that protistan communities in closely spaced – but geochemically different – habitats are very dissimilar, and may share as few as 5% of total OTUs. This is supported by a canonical correspondence analysis indicating that the empirically observed OTUs are organized along opposing gradients in oxidants and reductants. Our phylogenetic analyses identify many new clades at species to class levels, some of which appear restricted to specific layers of the water column and have a significantly non-random distribution. These findings suggest many pelagic protists are restricted to specific habitats, and likely diversify, at least in part due to separation by geochemical barriers.
Introduction

Protists are pivotal members of aquatic microbial communities. Through grazing on prokaryotic and other eukaryotic prey, they regenerate nutrients and modify or remineralize organic matter (Jumars et al., 1989; Sherr and Sherr, 2002; Taylor, 1982). In addition, they can affect the quantity, activity and physiological state of their prey (Frias-Lopez et al., 2009; Madsen et al., 1991). Through direct and indirect effects, protists help determine metabolic potentials of microbial communities and influence aquatic carbon cycling. Bacterial grazing is principally performed by small flagellated protists and ciliates (Frias-Lopez et al., 2009; Sherr and Sherr, 2002). Grazing releases inorganic nutrients that often limit primary production, and makes organic carbon available to higher trophic levels (Berman et al., 1987; Caron, 1994; Caron, 2000; Jurgens and Massana, 2008; Pernthaler, 2005; Sherr and Sherr, 2002). In recognition of their importance in aquatic communities, protists are now considered in numerical models of carbon cycling and in paradigms of surface and deep-ocean microbial ecology (Aristegui et al., 2009). Yet relatively little is still known about protistan diversity and distributions, particularly in the deep sea, and the degree to which specific habitats select for unique communities.

A century ago Martinus Beijerinck suggested that, in the microbial world, “everything is everywhere” (Beijerinck, 1913), implying that microorganisms are limitless in their dispersal abilities. This idea became a topic of debate (Bass et al., 2007; Finlay, 2002; Finlay and Fenchel, 1999; Martiny et al., 2006; Richards and Bass, 2005). For example, one view holds (Finlay and Clarke, 1999) that a soil sample under 0.1 cm$^2$, can contain at least 78% of all globally found species of examined protists (the Genus Paraphysomonas), implying that liters of sea water examined here might contain the bulk of protistan diversity described to date. A logical extension of this view is a proposition that protistan species that are abundant locally should also be abundant globally, and vice versa (Finlay et al., 2001). An alternative view is the “moderate endemicity” model, proposed by Wilhelm Foissner (Foissner, 2006). The latter posits
that some protistan species have cosmopolitan distributions, while other species have restricted distributions, a hypothesis gaining experimental support (Bass et al., 2007).

Determining whether microbial eukaryotes have distinctive biogeographies is far from trivial. Given the high abundances and reproductive potentials of microorganisms, high dispersal rates could allow many species to physically permeate almost all environments (Cavalier-Smith, 2004; Fenchel and Finlay, 2004; Finlay, 2002), but a number of empirical studies do not support this idea (Boenigk et al., 2006; Foissner, 1999; Fuhrman et al., 2008; Lawley et al., 2004; Richards and Bass, 2005; Rutherford et al., 1999; Soininen and Heino, 2007; Telford et al., 2006). Two examples are studies of longitudinal richness gradients of some marine foraminiferans (Rutherford et al., 1999) and bacteria (Fuhrman et al., 2008) suggesting that temperature explains a large portion of microbial endemcity. A handful of studies have addressed habitat specialization of marine protists but have either focused on specific groups (Guillou et al., 2008) or are based on a relatively small number of samples and sequences (Alexander et al., 2009; Countway et al., 2005; Edgcomb et al., 2009; Edgcomb et al., 2002; Stoeck et al., 2006; Stoeck et al., 2007; Stoeck et al., 2003; Zuendorf et al., 2006).

The Cariaco Basin, Venezuela, harbors the world’s largest truly marine body of permanently anoxic water, and provides a convenient model to study the role of geochemical gradients as possible barriers for active and passive protistan migration and colonization. Temporal changes in the biogeochemistry of the Basin have been well studied (Astor et al., 2003; Muller-Karger et al., 2001a; Scranton et al., 2001) as well as the spatiotemporal dynamics of bacterial populations (Lin et al., 2008). Seasonal shifts in local upwelling intensity, rates of fluvial discharge, trade wind intensity, and lateral intrusions of oxic waters influence primary production, microbial biomass and metabolic rates (Astor et al., 2003; Lin et al., 2008; Muller-Karger et al., 2001a; Taylor et al., 2001a). However, little is known about these effects on protistan populations. The Cariaco’s water column transitions from fully oxic to sulfidic across a temporally varying boundary between 250 and 350m of depth. Within the redox transition zone
lie strong gradients in $O_2$, $NO_3^-$, $H_2S$, $NH_4^+$, $NO_2^-$, $PO_4^{3-}$ and $CH_4$, and enrichments in $S_2O_3^{2-}$, $SO_3^{2-}$, $S^0$, $Mn^{2+}$ and $Fe^{2+}$ that select for specific prokaryotic phylotypes at different depths (Li et al., 2008; Lin et al., 2006; Lin et al., 2008; Lin et al., 2007; Taylor et al., 2001a). Within the redoxcline, peaks in prokaryotic metabolic activity and cell numbers are observed, which often coincide with peaks in protistan cell numbers (Taylor et al., 2006; Taylor et al., 2001a). Lin et al. (2008) showed significant vertical variations in bacterial community structure between oxic, transition, and anoxic zones, as well as horizontally between different sites. Vertical, seasonal, and geographic patterns in prokaryotic community distribution have also been documented in other locations (see Treusch et al., 2009) for recent synopsis). Recently, genetic approaches have begun to document marine protistan diversity (see Edgcomb et al. in this issue for a detailed discussion [Chapter 1 in this dissertation]). However, spatiotemporal analyses of protistan communities have been limited (e.g. (Countway et al., 2007; Not et al., 2007)). Studies have thus far been generally constrained to fewer than 2000 total SSU rRNA gene sequences from any single environment or sample, which in light of observed diversity makes it difficult to assess spatiotemporal differences in protistan communities. Here, we attempt to provide a more comprehensive approach. While we did not collect multiple samples at each sampling point, we went substantially beyond earlier investigations by collecting samples spanning the major biogeochemical habitats of the Cariaco Basin, revisiting these habitats in different seasons, and undertaking the largest Sanger-based 18S rRNA gene sequencing effort within a given oceanographic regime to date. Our data provide a unique opportunity to examine species richness and distribution within an oceanographic regime.
**Materials and Methods**

More extensive information is in the Supplementary Materials and Methods in the companion paper in this issue by Edgcomb et al. (Chapter 1 of this dissertation).

**Sample collection**

Samples were collected from three sites in the Cariaco Basin, Venezuela: Site A (10.50°N, 64.66°W), Site B (10.40°N, 64.46°W) and Site C (10.40°N, 65.35°W) during January and May of 2005 (Figure 1). Samples for DNA extractions were collected at depths corresponding to 40 m above the oxic/anoxic interface, within the oxic/anoxic interface, 40 m below the interface, and at 900 m. The oxic/anoxic interface was defined as the depth at which oxygen became undetectable. Thus a total of 24 water samples were collected (two months x four depths x three sites). Samples were withdrawn from Niskin bottles under N\textsubscript{2} pressure and captured on 47mm Durapore membranes (0.45 µM pore size) onboard, with little to no sample exposure to the atmosphere. Depending on cell concentration in each sample, 2.5-2.8L samples were filtered under gentle vacuum (<25 cm Hg) until no passage of water through filters was observed. At least two filters from each sample were used for DNA extractions. Membranes were stored in cryovials containing DNA extraction buffer (Stoeck *et al.*, 2003) at -20°C until further processing.

**DNA preparation and PCR amplification of 18S rDNA**

DNA was prepared using previously described methods (Stoeck *et al.*, 2007; Stoeck *et al.*, 2003). Depending on yield, DNA from 2-3 filters per depth at each site was combined prior to PCR. The 18S rRNA gene was amplified using four different primer combinations; Euk528F (5'-CGGTAATTCCAGCTCC-3') either paired with U1492R (5'-GGTTACCTTGTTACGACTT-3'), U1391R (5'-GGGCCGTGTGTACAARGR-3'), or U1517R (5'-ACGGCTACCTTGTTACGACTT-3') and Euk360F (5'-CGGAGARGGMGCMTGAGA-3') paired with U1492R. All primer combinations were applied to all samples, and PCR products from ≥3 reactions were pooled per primer pair for each sample.
Clone library construction and sequencing

To constrain costs and allow for deeper sequencing of individual libraries, PCR products for sites B and C were pooled from the same depths prior to clone library construction (reducing the number of analyzed sites to two, and the total number of analyzed samples from 24 to 16). Separate clone libraries were constructed for each of the 4 depths from the 8 site B/C samples (4 from May 2005 and 4 from January 2005) and from the 8 site A samples. Processing of the data used PHRED and PHRAP (Ewing and Green, 1998; Ewing et al., 1998) and a pipeline script to call bases from chromatograms and perform quality control procedures. The sequences were checked for chimeras using the Bellerophon Chimera Check and the Check_Chimera utilities (Ribosomal Database Project) (Cole et al., 2003). After removal of short sequences (< 800 bp), putative chimeras, bacterial, archaeal and metazoan sequences, the remaining sequences were grouped into Operational Taxonomic Units (OTUs) based on 90, 95, 98, and 99% rRNA gene sequence similarity levels. This was achieved by first making all possible pairwise sequence alignments by using ClustalW (Thompson et al., 1994), calculating % sequence identities, followed by clustering the sequences by using the unweighted pair group method with arithmetic mean (UPGMA) as implemented in the OC clustering program (http://www.compbio.dundee.ac.uk/Software/OC/oc.html). All protistan sequences have been deposited in Genbank, see the accompanying paper in this issue, Edgcomb et al., for details (Chapter 1 in this dissertation).

Phylogenetic analyses

For our phylogenetic analyses we used “centroid” sequences, each representing Operational Taxonomic Units (OTUs) sharing 95% identity (for practical considerations). The centroid sequence was defined as the least dissimilar sequence relative to all sequences within a cluster at the 95% threshold. A centroid representative from each OTU was compared against the GenBank-nt database using BLASTn in search of their closest sequence relatives. The highest
scoring cultured and uncultured sequence relatives of each centroid were retrieved, and aligned using ARB (Ludwig et al., 2004).

**Statistical analyses of clone library data for richness predictions**

For each sample, we obtained "frequency count" data, i.e., the numbers of OTUs registered in the corresponding clone library only once, twice, etc., at the 90, 95, 98 and 99% levels of sequence identity. This data was used to estimate, at each % identity level, the total number of OTUs, representing the sum of seen and unseen. The clustering of the entire data set was used to test the fit of a wide range of parametric and non-parametric models for predicting total protistan diversity in the Cariaco Basin (see Edgcomb et al. this issue for a complete description [Chapter 1 in this dissertation]), a methodology that has been used in previous studies (Hong et al., 2006; Jeon et al., 2006) for the estimation of microbial richness. This method was implemented here using the software package CatchAll (Bunge, 2011). The entire data set was then separated by sites (A vs combined B and C) and the 4 depths from site A (40 m above the oxic/anoxic interface, interface, 40 m below, and 670-900 m) and clustered separately. These data were used to predict taxon richness of the different environments based on clustering data at 99% sequence similarity (for details see Supplementary Materials and Methods in Edgcomb et al. this issue [Chapter 1 in this dissertation]).

**Canonical correspondence analysis (CCA), Multi-Response Permutation Procedure (MRPP), and Nonmetric multidimensional scaling (NMS)**

CCA is an ordination method that allows for the exploration of community responses to environmental gradients by reducing the dimensionality of the analysis such that the axes reflect a linear combination of the environmental variables and the operational taxonomic unit data (McCune and Grace, 2002). CCA was used here to elucidate the relationships between protistan community structure and concentrations of dissolved O₂ and sulfide. MRPP is a tool used for testing the hypothesis of no effect of a variable on two different groups of entities and can help determine whether a variable exerts a statistically significant influence on sampled
communities (McCune and Grace, 2002). We used MRPP to assess the effect of sulfide, oxygen, season, site, and depth on the distribution of OTUs. In cases where species composition is influenced by external factors such as oxygenated water or terrestrial sediment intrusions from seasonal riverine inputs instead of simply internal factors or environmental gradients, NMS is an ordination method that may be more appropriate (McCune and Grace, 2002). We ran a NMS analysis using PC_ORD v4 (MJM Software Design, Gleneden Beach, OR) and selected the autopilot run with the slow and thorough option (McCune and Grace, 2002). Stress values were used as a measure of goodness of fit and values less than 15 indicate a low probability of drawing the wrong inferences from the results. Monte Carlo tests were used to identify dimensions with solutions that were significantly different from those arising strictly by chance. CCA, NMS, and MRPP analyses were applied to our sequence dataset clustered at the 95% sequence identity threshold.

*Urn-model hypothesis test*

As a further statistical test of equal distribution of any particular OTU across the sampled depths, we used an "urn model", which assumes that the occurrence of each sequence is equally likely in all urns (every depth in this case). The probability of 0, 1, 2 or 3 empty urns was determined by a well-established formula (Kolchin and Chistyakov, 1975). This allows computation of the probability that at least one urn will be empty (OTU does not appear at any one depth), thus obtaining the p-value for the test of the null hypothesis of random distribution.
Results

A total of 64 clone libraries were produced after applying the different primer sets to DNA extracted from each of the samples (16 samples x 4 primer sets). From each of the 64 libraries, we sequenced an average of 250 clones, resulting in over 16,000 rRNA gene sequences. Rigorous culling of confirmed or suspected metazoan, chimeric, and short (< 800 bases,) sequences left 6,498 high quality (average length 1100 bases) 18S rRNA protistan gene sequences (for details on the number of rRNA gene sequences analyzed in this study classified by season, biogeochemical habitat, site, and PCR primer pair from each sample see Supplementary Table 2 in Edgcomb et al., this issue [Chapter 1 in this dissertation]). Clustering our dataset at the 95% and 99% sequence identity thresholds produced a total of 822 and 2106 protistan centroids, respectively. The percentages of these centroids affiliated with different higher-level taxonomic groups of protists based on BLASTn searches are presented in Figure S1 for each depth sampled. Rhizaria, followed by Stramenopiles, dominate our clone libraries at all depths while the Ciliophora, Dinoflagellata, Syndiniales, and Euglenozoa represent progressively less abundant clones in the libraries.

Predicting OTU richness

For a complete discussion of the results of OTU clustering see Edgcomb et al., (this issue [Chapter 1 of this dissertation]). The overlap of empirically observed OTUs at the 99% identity level between the sampled depths and sites is presented in Figure S3. Between the communities above and within the interface only 9% of OTUs are shared, and less overlap (8%) exists between the communities above and below the interface. Additionally, over 90% of OTUs present at site A are not found at sites B and C and vice versa. Figure 2A presents the statistical predictions of total OTUs at 99% sequence identity for sampled depths from 40m above the oxic/anoxic interface, the interface, and 40m below, as well as pooled data for the first and second, and second and third depths. These three depths were selected for
presentation because they are in closest proximity (~40 m) to one another. The logic behind this analysis is that if one habitat contains X species, the other habitat contains Y species, and the two habitats combined contain Z species, then \[(X+Y) – Z\] is the maximum number of species shared by the two habitats. The results show a predicted overlap of 904 (+/- 728) OTUs between communities above the interface and within the interface (Figure 2A). Less overlap is predicted for the communities above and below the interface (277 +/- 1066 OTUs). Following the same logic, we obtained the richness estimated for site A and the combined estimation for sites B and C. These roughly sum to equal the lower bound of the standard error (SE) for the overall basin prediction of diversity suggesting that many OTUs are unique between these sites (Figure 3A).

**Canonical correspondence analysis (CCA), Multi-Response Permutation Procedure (MRPP) and nonmetric multidimensional scaling (NMS)**

CCA identified environmental parameters that may explain a large proportion of the observed protistan community distribution pattern. The analysis shows a separation of protistan communities within samples from oxygenated, microoxic, and sulfidic waters of the Cariaco Basin along \(O_2\) and \(H_2S\) gradients (Figure 2B). The species-environment correlations were 0.971 and 0.990 for axis 1 and 2, respectively, and the two-dimensional plot shows that axis 1 and 2 explained 34.6 and 28.4% of the variance in OTU distribution, respectively. A Monte Carlo test for significance of the Eigenvalues yielded a p-value \(\leq 0.02\).

NMS ordination indicates that there are distinct assemblages of protists recovered from site A samples vs. sites B/C and between the two seasons (Figure 3B). After 43 iterations the stress of the final 2-dimensional solution was 14.1 (below the cutoff of 15) and therefore the solution was considered stable. Axes 1 and 2 explain 28 and 20% of the variance, respectively. A Multi-Response Permutation Procedure (MRPP) was preformed for season, site, sulfide,
oxygen and depth, which were all found to have a significant effect (≤ 0.05) on the observed distribution of OTUs.

**Phylogenetic analyses**

Among rhizarians, Radiolaria were by far the most abundant phylum represented by 2,216 sequences, and 150 centroids. Over 50% of these fell into one of the five radiolarian sequence clades (RAD 1-5) described from the Sargasso Sea by Not et al. (2007), principally into the RAD-3 clade, detected from all depths, both seasons and sites (Table 1). In addition to these five “RAD” clades, several other radiolarian centroids from our dataset form 14 lineages (Figures 4 and 5) labeled RAD 6-19. We identified two new lineages (RAD-11 and 12) within the Acantharea and 2 new lineages within the Taxopodia (RAD-13, 14). Four RAD lineages (7, 8, 9, 10) were related to the Polycystinea and do not contain sequences from oxygenated samples. The RAD-9 lineage contained 77 sequences that derive from 7 of the 12 total samples from the anoxic water column.

Fourteen percent of our dataset was affiliated with the Stramenopiles, which formed 148 centroids. On our phylogenetic tree of all Stramenopile centroids (data not shown), 339 sequences clustered with the heterotrophic Stramenopiles. Most lineages detected were uncultured marine Stramenopiles (MAST) (Kolodziej and Stoeck, 2007; Massana et al., 2004). The number of sequences within each sample affiliated with MAST clades is organized by depth in Table 1. We discovered 6 lineages, MAST 14 - 19, that did not group with the 13 established MAST clades (Kolodziej and Stoeck, 2007; Massana et al., 2004) or the “mystery heterokont” groups (Richards and Bass, 2005) (Figure 6). Several of these new lineages were not detected in any oxygenated samples (Table 1). The MAST-14, 15, and 16 lineages were affiliated with the Bicosococida, although the latter two are supported weakly. In contrast, MAST 17 and 19 branch within the Thraustrochytrids and Labyrinthulids, respectively, with strong support from both methods.
Sequences affiliated with the Syndiniales (Alveolata) represent approximately 16% of our clone library sequences (Figure S1). We detected representatives from all five groups of Syndiniales (Guillou et al., 2008) with most of the sequences being related to groups I (247 sequences) and II (140 sequences). Within group I, we identified a new clade (100% support from both methods) labeled I.9 (Figure 7) that does not cluster with the established 8 group-I sub-clades of Syndiniales (Guillou et al., 2008). Within the NA1.4 clade, 7 centroids from our dataset form two supported subclades (1.0 posterior probability and 93% bootstrap). Both I.9 and I.4A were detected in 3 of our anoxic samples and were absent from our oxygenated water column samples (Table 1). Clade 1.4A was detected only from anoxic waters in the pooled samples from sites B/C and was absent from all site A samples.

We detected 90 sequences affiliated with the Euglenozoa that fell into 17 centroids at the 95% identity threshold, with the majority detected only in the 900m anoxic samples (Figure S1). Ten centroids, along with a sequence reported from an earlier survey of the Cariaco Basin (Stoeck et al., 2003) form a clade with 100% supported separation from all other major groups of Euglenozoa, and is presented here as a “putative novel euglenozoan class” (Figure 8). This clade was detected in 10 out of 13 anoxic samples (9 from our survey plus 1 from (Stoeck et al., 2003) and it was absent from all oxygenated samples (Table 1). The closest named relative to this group of sequences is the recently described *Calkinsia aureus* (Yubuki et al., 2009). Three sister clades (labeled Symbiontida clades A, B, C) to *Calkinsia aureus* were also identified all with strong support on our trees. Clades A and B contain sequences from anoxic samples of the Cariaco Basin only. Clade C consists of sequences from anoxic Framvaren and Mariager Fjords (Behnke et al., 2006; Zuendorf et al., 2006).

Having observed a number of protistan clades apparently restricted to specific geochemical regimes, we chose representative clades from several Kingdoms to statistically estimate whether their association with the respective regime could be due to chance using the
Urn Model Test (Supplementary Figure S2). In all cases, such probability proved infinitesimally low, down to e.g. $10^{-10}$ in case of the novel rhizarian clade RAD-9.

**Discussion**

*Spatiotemporal distribution patterns and responses to environmental variables*

Using phylogenetic and multivariate community analyses coupled with parametric richness estimations, we show that specific environmental variables impact the distribution of protistan taxa leading to their habitat preference. Three of the sampled habitats are particularly interesting for this type of analysis because they are sufficiently close to each other (40 m), and the water density gradient is sufficiently weak between these layers to allow, in principle, for frequent redistribution of the individual species by advective water mixing or active motility. The null hypothesis is that given no physical barrier to dispersal, mixing would ensure that each species is present throughout these closely-spaced layers of the water column. Based on results from several parametric statistical analyses, we reject the null hypothesis and show that protistan communities in closely spaced – but geochemically different - habitats are dissimilar and may share as little as 5% of species (Figure 2A). This conclusion is supported by a direct comparison of empirically observed OTUs at these depths showing that over 90% of detected OTUs are unique (Figure S3). Furthermore, canonical correspondence analysis of our dataset showed that many of the empirically observed OTUs were organized along oxygen and sulfide gradients (Figure 2B), indicating these environmental variables are responsible, in part, for the observed OTU distribution in our dataset. We note that distributions of protistan ecotypes most likely are shaped by integrated responses to the complex chemical milieu and the composition/activity of prokaryotic/eukaryotic prey communities and cannot be solely linked to a specific chemical species. In this study, we track distributions of protistan clades through the geochemically stratified water column, recognizing that oxygen and sulfide distributions correlate with numerous chemical variables that define biogeochemical habitats. We also
acknowledge that sampling limitations inherent to microbial oceanography studies may mask many spatiotemporal patterns. However, we argue that the relatively large sampling effort made here, across multiple seasons, sites, and geochemical habitats, reveals gross patterns in community structure of pelagic protists.

Our NMS ordination indicates that there are distinct assemblages of protists recovered from site A vs. sites B/C, and clear divisions in most (but not all) samples between the two seasons (Figure 3B). A MRPP p-value of 0.012 for the test of significance of geographic sites in the observed distribution of OTUs lends support to this conclusion. These results are corroborated by the minimal overlap in OTU numbers estimated and empirically observed for the locations compared (Figure 3A, Figure S3), supporting the notion that spatial separation and local biogeochemistry drives, in part, the distribution of these protistan assemblages. The separation of samples from site A vs the combined B and C samples may be due to the fact that sites B/C are closer to channels providing input of oxygenated surface water from the Caribbean Sea. Site A is closest to riverine inputs from coastal Venezuela, which may drive community divergence among sites. Lin et al. (Lin et al., 2008) found that among-site variation (site A vs. sites B/C) in bacterial communities within the redoxcline was greater than the vertical variation between depths, possibly due to differences in primary production and lateral intrusions of oxic water at the different sites. The western basin has lower surface primary production than the eastern basin (Muller-Karger et al., 2001a; Richards, 1975; Scranton et al., 1987) and such differences likely contribute to variations in microbial communities (Lin et al., 2008).

Water column productivity varies seasonally and spatially across the Basin due to nutrient upwelling from late January through June due to intensifying Trade Winds. (Muller-Karger et al., 2001b). Seasonal upwelling events would explain the fact that the site A samples from May in the eastern sub-basin are distinct from the samples taken at the same location in January (Figure 3B). It also explains why a 100% seasonal distinction is not observed for the
combined samples from sites B/C because site B is in the eastern sub-basin and site C is in the western sub-basin.

Significant temporal shifts in bacterial community composition along the redoxcline were observed over a 2-year period at the same three sites (Lin et al., 2008). Trophic interactions between aquatic protists and bacteria, better known as the 'microbial loop,' are thought to play a major role in controlling bacterial dynamics (Azam et al., 1983; Taylor, 1982). Phagotrophic protists can chemically sense and congregate at aggregations of prey (Fenchel and Blackburn, 1999). Their growth rates can keep pace with those of their bacterial prey (Fenchel, 1987; Finlay and Fenchel, 2001; Sherr and Sherr, 1994; Sherr and Sherr, 2002), and as bacterial population size and composition changes from site to site and season to season, it is logical to expect local protist populations to exhibit similar variations as the local habitat changes and potential prey species select for community dominance by different protists. Thus trophic responses could be contributing to the relatively high standard error in some of our estimates of richness (Figure 3A).

Many OTUs were observed to occur exclusively in a particular geochemical regime. Among the 19 radiolarian clades, 5 were detected in only one habitat (RAD 7, 8, 9, 10, 11), but most of these are under-represented in our data set, and are perhaps members of the “rare biosphere”. The RAD 9 clade was never observed in the Basin’s oxygenated waters, but appeared 77 times in over 50% of all samples from oxygen-depleted and sulfidic depths. The probability of such a distribution being due to chance is essentially zero (Figure S2). Thus, we conclude that RAD 9, and possibly several other radiolarian clades, are indeed restricted to anoxic regimes.

The distributions of some clades within the alveolate Order Syndiniales were confined to suboxic and/or anoxic habitats similar to the RAD clades described above, perhaps indicating interesting ecological relationships. The Syndiniales is a diverse group of picoeukaryotes (Guillou et al., 2008; Moon-van der Staay et al., 2001) exhibiting a parasitic lifestyle.
(Chambouvet et al., 2008; Skovgaard et al., 2005), representing 30% of all dinoflagellate sequences in public databases, and have only been retrieved from marine environments (Guillou et al., 2008). The distribution of Syndiniales may be constrained by the observed endemic distribution of their hosts (Guillou et al., 2008), among them the Radiolaria (Dolven et al., 2007). Indeed, the Syndiniales clade I.9 does seem to be absent from oxygenated waters of the Cariaco Basin, and present exclusively under sub and anoxic conditions, mirroring the habitat specialization of several RAD clades (Figures 4, 5; Table 1).

Further examples of apparent habitat specialization can be found among Stramenopiles. While several of the more frequently encountered MAST taxa seem to be distributed rather homogeneously throughout the water column (e.g., MAST 1, 3, 8), notable exceptions were evident: MAST 4, 14, 15, 16, and 19 were absent from oxygenated waters. In addition, some of these MAST clades appear to prefer anoxic waters immediately below oxic-anoxic interface, which coincidently exhibits peaks in prokaryotic biomass and production (Stoeck et al., 2003; Taylor et al., 2001b). This distribution is interesting because MAST species have been reported from anoxic environments in the past (Behnke et al., 2006; Stoeck et al., 2003; Zuendorf et al., 2006) and because at least some MAST species are known to be phagotrophic (Massana et al., 2006; Massana et al., 2009). Considering our observations, the possibility that some MAST species are anaerobic bacterivores and thus involved in regulating bacterial dynamics within and below the redoxcline is suggested.

One of the strongest cases of apparent habitat specialization is represented by the novel Euglenozoa clades we discovered. Some of the newly detected OTUs appear to form a novel class-level clade that affiliates strongly with the H52 clone from an earlier survey of the deep anoxic Cariaco Basin (Stoeck et al., 2003) (Figure 8). As importantly, this clade separates from all other Euglenozoa OTUs described to date from other anoxic marine environments (Behnke et al., 2006; Lara et al., 2009; Lopez-Garcia et al., 2007; Zuendorf et al., 2006). Additional euglenozoan signatures detected here form two sister clades to Symbiontida, a recently
proposed new class of Euglenozoa represented by only a single species, *Calkinsia aureus* (Bernhard *et al*., 2000; Yubuki *et al*., 2009). Unique α-taxonomic as well as genomic characteristics support the separation of Symbiontida from the three traditional euglenozoan classes Euglenids, Kinetoplastids, and Diplonemids (Yubuki *et al*., 2009). Many euglenozoan 18S rRNA gene sequences revealed here seem unique to suboxic, anoxic, or sulfidic habitats (Table 1), and the probability of such skewed distributions being exclusively due to chance is essentially zero (Figure S2).

This study expands the taxonomic groups recovered in smaller-scale previous studies of the Cariaco Basin protist community (Stoeck *et al*., 2003; Stoeck *et al*., 2006). Those studies also reported a differential recovery of particular taxonomic groups among depths from one site, but the depth of sequencing in those studies did not allow further testing of these observations.

*Relevance to questions concerning protistan distributions*

Clades unique to our study are highly suggestive of local speciation. The number of unique clades* is high (Figures 4-8, Table 1), and these are spread over multiple Kingdoms. A potentially important example is the new putative euglenozoan class (Figure 8), which is distinct from all other euglenozoan taxa described so far from all locations. It is abundant in our sequence collection (a total of 50 individual sequences forming 10 centroids), and is apparently distributed throughout the Basin’s anoxic waters (detected in over 80% of all the anoxic samples), and does not contain related sequences from the growing number of surveys of anoxic environments reported to date (Alexander *et al*., 2009; Behnke *et al*., 2006; Edgcomb *et al*., 2009; Edgcomb *et al*., 2002; Stoeck *et al*., 2010; Stoeck *et al*., 2009; Stoeck *et al*., 2003; Zuendorf *et al*., 2006). Numerous other examples of apparent geographic distribution patterns include 14 novel lineages of Rhizaria a new group-I uncultured novel Alveolate clade, and 6 new MAST Stramenopile clades. Also noteworthy is the absence of representatives from the Sargasso Sea in all of the novel clades reported here, as the Sargasso Sea has been sampled
heavily (Countway et al., 2007; Not et al., 2007; Piganeau et al., 2008; Piganeau and Moreau, 2007; Venter et al., 2004) and lies just northeast of the Caribbean Sea. As different sites within even the Basin itself harbor distinct species assemblages (Figure 3), it is not unexpected that the Cariaco Basin and Sargasso Sea likely contain widely divergent communities. The simplest interpretation of these observations is that at least some microbial eukaryotes speciate, diversify, and become restricted to a specific biogeochemical niche, and do not successfully disperse to either similar or dissimilar environments elsewhere on the globe.

____________________________

*We acknowledge that species of protists whose 18S rRNA genes have not yet been sequenced may bias our observations and interpretations regarding the potential for the global distribution of selected protist groups due to their underrepresentation in sequence databases.

At the same time, over 250 rhizarian, Dinoflagellate, Syndiniales, Stramenopile and various fungal OTUs (95% identity) appearing in our Cariaco Basin 18S rRNA libraries have previously been detected elsewhere (data not shown), under both similar and dissimilar biogeochemical conditions (Edgcomb et al., 2002; Lopez-Garcia et al., 2001; Massana et al., 2002; Moon-van der Staay et al., 2001; Not et al., 2007; Stoeck et al., 2006; Stoeck et al., 2003). This indicates that some protist taxa are dispersed widely, and are perhaps even cosmopolitan in their distribution.

**Size-related dispersal of protists**

The influence of microbial body size on dispersal has been debated, with arguments made for (Finlay, 2002) and against (Heino et al., 2010; Heino and Soininen, 2006; Soininen and Heino, 2007) the existence of a connection between the two. One hypothesis (Fenchel and Finlay, 2004) asserts that with an increase in organism size there is a decrease in dispersal rate, while microbes of smaller sizes are more likely to have a wider dispersal range, leading to
cosmopolitan distribution. We analyzed the apparent cosmopolitan and ‘endemic’ clades in our dataset to see if a particular body size range dominated either type of clades. We defined clades detected in the Cariaco Basin as likely cosmopolitan if their representatives had been registered in at least three locations worldwide (including our survey), and as potentially endemic if all the clade’s representatives were so far unique to the Basin. To make realistic size assignments, all clades must fulfill an additional criterion: to be closely related (>90% 18S rRNA gene identity) to a described taxonomic group, the latter with known and relatively narrow size range of its members. Screening our sequence collection lead to identification of 32 likely cosmopolitan and 36 possibly endemic clades that satisfied the above criteria, which we then placed into one of three size groups based on the length and/or diameter of their known relatives: 1-20 µm, 25-50 µm, and 80-200 µm. The results of this classification are presented in Supplementary Table 1. Clearly, our endemic group contains as many clades within the smallest and largest among our size categories. Likewise, the smallest clades are equally distributed between ‘endemic’ or ‘cosmopolitan’ groups. In spite of uncertainties about definitions of endemcity and size assignments, our analysis does not show a clear connection between the size of a cell and its dispersal potential. This in line with the findings of Heino and Soninen (2006) who, using freshwater diatoms as a model to study microbial dispersal capacity, did not observe strong evidence for a relationship between dispersal and body size.

Conclusions

An in-depth survey of the protistan community in the Cariaco Basin reveals an apparent restriction of some protists to different sites in the Basin, as well as oxygenated, suboxic, or anoxic/sulfidic environments. Phylogenetic and multivariate community analyses along with parametric estimations of richness imply that geographic location, seasonality, and geochemical gradients define the community structure of marine protists in the Cariaco Basin. We conclude that substantially different communities can exist in close proximity to one another, which
permits speciation to proceed differently at different sites and depths in the Basin. These analyses support the moderate endemcity model (Bass et al., 2007; Foissner, 2006), which views protists as an assemblage of species some of which are cosmopolitan and others endemic, and thus biogeographically similar to macroorganisms.

**Acknowledgements**

We thank the captain and crew of the B/O Hermano Gines and the staff of the Fundación La Salle de Ciencias Naturales, Margarita Island for their assistance during our field work in Venezuela, particularly Yrene Astor. We are grateful to Mary I. Scranton and the dedicated researchers of the CARIACO biogeochemical time series program without whom this work would not have been possible. We would like to thank Valentin Ilyin (Northeastern University), David Mark Welch (MBL), and Jed Goldstone (WHOI) for helpful discussions on bioinformatics approaches, Hilary Morrison and Rich Fox (MBL) for the use of their pipeline scripts for sequence data processing, and the Academic Research Computing User Group at Northeastern University for permitting us to use the Opportunity computer cluster. Thank you to Vladimir Bondarev and Anne-Christin Gimth for assistance with preliminary quality control of our Sanger dataset. Comments from two anonymous reviewers significantly improved the manuscript. This research was conducted using the resources of the Cornell University Center for Advanced Computing, which receives funding from Cornell University, New York State, the National Science Foundation, and other leading public agencies, foundations, and corporations. This research was supported by grants from NSF (MCB-0348341 and DEB-0816840 to SSE, DEB-0816638 to JB, MCB-0348407 to VE and OCE 03-26175 and MCB-03-47811 to GTT), and from Venezuela FONACIT (no. 2000001702 to RV). This is a contribution # 269 from the Marine Science Center, Northeastern University, Nahant MA, U.S.A.
Figure 1. Map of sample sites in the Cariaco Basin, Venezuela. Station A represents the site of the U.S.-Venezuelan CARIACO biogeochemical time series program in the eastern sub-basin. Station B is a shallower station south of the Tortuga channel, the source of incoming Caribbean waters. Station C is centered in the less productive western sub-basin to the east of Canal Centenela.
Figure 2. (a) Total protist richness predictions for selected environments (A: 40 m above interface, I: interface, B: 40 m below interface) of the chemically stratified water column in the Cariaco Basin at Station A. (b) A biplot generated from a canonical correspondence analysis (CCA) of our 18S rRNA dataset clustered at the 95% sequence identity level. Environmental variables sulfide and oxygen are represented by arrows. Sampled depths are represented by circles, the size and color of which indicate the detected amount of either oxygen or sulfide in that sample. Samples that cluster together on the biplot are highlighted in gray.
Figure 3. (a) Predicted richness of protistan assemblages in two different geographic locations within the Cariaco Basin compared to that predicted for the entire Basin. (b) Non-metric Multidimensional Scaling (NMS) ordination of our 18S rRNA dataset clustered at the 95% sequence identity level. Triangles represent the 16 samples taken, white and black triangles correspond to combined samples from sites B and C, and site A, respectively. A sample name ending in M or J (i.e. BCIM, AIJ, etc.) corresponds to May and January samples, respectively. Lines are drawn onto the biplot to separate samples that cluster together.
Figure 4. Phylogenetic relationships of selected Polycystinea (Radiolaria). The tree was constructed under Bayesian inference using an alignment of 925 unambiguous positions under the GTR+I+Gamma model of sequence evolution. Bootstrap (RAxML) and posterior probability values greater than 50% are shown at the nodes in the order PP/ML. Black circles at nodes represent full posterior probability and bootstrap support. Radiolarian clades new to this study are labeled in accordance with the naming scheme established by Not et al. (2007). Centroids from this survey are in bold font.
Figure 5. Phylogenetic relationships of selected Acantharea (Radiolaria). The tree was constructed under Bayesian inference using an alignment of 887 unambiguous positions under the GTR+I+Gamma model of sequence evolution. Clades new to this study are labeled in accordance with the naming scheme established by Not et al. (2007). For more information see Figure 4.
Figure 6. Phylogenetic relationships of selected Marine Stramenopiles (MAST) (Massana et al. 2004). The tree was constructed under maximum likelihood using an alignment of 927 unambiguous positions under the GTR+I+Gamma model of sequence evolution. A total of 94 centroid sequences were included in the alignment. Clades of MAST sequences new to this study are labeled in accordance with the naming scheme established by Massana et al. (2004). For more information see Figure 4.
Figure 7. Phylogenetic relationships of selected Syndiniales. The tree was constructed under Bayesian inference using an alignment of 869 unambiguous positions under the GTR+I+Gamma model of sequence evolution. The new NA1.9 clade is labeled in accordance with the naming scheme established by Guillou et al. (2008). For more information see Figure 4.
Figure 8. Phylogenetic relationships of selected Euglenozoa. The tree was constructed under Bayesian inference using an alignment of 766 unambiguous positions under the GTR+I+Gamma model of sequence evolution. For more information see Figure 4.
Figure S1. Pie charts showing the number of 95% identical OTUs affiliated with the major protistan taxonomic groups at each of the four depths sampled.
Figure S2. Urn model hypothesis test of non-randomness in distribution of 5 selected clades representing Radiolaria, Euglenozoa, Stramenopiles, and Fungi (A: 40 m above interface, I: interface, B: 40 m below interface, D: 900 m). The null hypothesis for the test was that the spatial distribution of respective organisms was random, and accumulation in one or another habitat was due to chance; p values for each clade tested indicate the likelihood of the null hypothesis being true.
Figure S3. A comparison of richness in observed OTUs at the 99% identity threshold across four geochemical layers and two sites in the Cariaco Basin. (a) Total OTUs present at depths above and interface, compared to the number of OTUs within each depth individually. The overlap shows the number of OTUs shared. This comparison was also done for depths above and below (b), above and 900 m (c), and sites A and B/C (d). above: 40 m above interface, interface: oxic/anoxic interface, below: 40 m below interface, 900 m: 900 m deep anoxic sample.
Table 1: The number of occurrences of various environmental clades detected in this study at each depth sampled. “seq” indicates the overall number of occurrences and “samp” is the number of samples in which the clade was detected. Clades marked with an asterisk are unique to this study. A: 40 m above interface, I: interface, B: 40 m below interface, D: 900 m.

<table>
<thead>
<tr>
<th>Environmental clade</th>
<th>A seq (samp)</th>
<th>I seq (samp)</th>
<th>B seq (samp)</th>
<th>D seq (samp)</th>
<th>Total seq (samp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAD-1</td>
<td>2 (1)</td>
<td>2 (1)</td>
<td>0</td>
<td>2 (1)</td>
<td>6 (3)</td>
</tr>
<tr>
<td>RAD-3</td>
<td>693 (4)</td>
<td>333 (4)</td>
<td>50 (4)</td>
<td>142 (4)</td>
<td>1361 (16)</td>
</tr>
<tr>
<td>RAD-4</td>
<td>2 (2)</td>
<td>3 (1)</td>
<td>1 (1)</td>
<td>0</td>
<td>6 (4)</td>
</tr>
<tr>
<td>RAD-6*</td>
<td>5 (2)</td>
<td>19 (3)</td>
<td>157 (4)</td>
<td>1 (1)</td>
<td>182 (10)</td>
</tr>
<tr>
<td>RAD-7*</td>
<td>0</td>
<td>2 (1)</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td>4 (3)</td>
</tr>
<tr>
<td>RAD-8*</td>
<td>0</td>
<td>2 (2)</td>
<td>2 (2)</td>
<td>0</td>
<td>4 (4)</td>
</tr>
<tr>
<td>RAD-9*</td>
<td>0</td>
<td>12 (1)</td>
<td>57 (3)</td>
<td>8 (3)</td>
<td>77 (7)</td>
</tr>
<tr>
<td>RAD-10*</td>
<td>0</td>
<td>0</td>
<td>3 (2)</td>
<td>0</td>
<td>3 (2)</td>
</tr>
<tr>
<td>RAD-11*</td>
<td>0</td>
<td>2 (2)</td>
<td>0</td>
<td>0</td>
<td>2 (2)</td>
</tr>
<tr>
<td>RAD-12*</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td>2 (1)</td>
<td>0</td>
<td>4 (3)</td>
</tr>
<tr>
<td>RAD-13*</td>
<td>0</td>
<td>2 (1)</td>
<td>2 (2)</td>
<td>1 (1)</td>
<td>5 (4)</td>
</tr>
<tr>
<td>RAD-14*</td>
<td>10 (2)</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td>0</td>
<td>12 (4)</td>
</tr>
<tr>
<td>RAD-15*</td>
<td>2 (1)</td>
<td>0</td>
<td>1 (1)</td>
<td>0</td>
<td>3 (2)</td>
</tr>
<tr>
<td>RAD-16*</td>
<td>3 (2)</td>
<td>5 (2)</td>
<td>2 (2)</td>
<td>18 (1)</td>
<td>28 (7)</td>
</tr>
<tr>
<td>RAD-17*</td>
<td>2 (2)</td>
<td>0</td>
<td>2 (2)</td>
<td>2 (2)</td>
<td>6 (6)</td>
</tr>
<tr>
<td>RAD-18*</td>
<td>38 (4)</td>
<td>11 (3)</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td>51 (9)</td>
</tr>
<tr>
<td>RAD-19*</td>
<td>125 (2)</td>
<td>62 (1)</td>
<td>2 (1)</td>
<td>82 (1)</td>
<td>271 (5)</td>
</tr>
<tr>
<td>MAST-1</td>
<td>9 (3)</td>
<td>10 (2)</td>
<td>5 (3)</td>
<td>5 (2)</td>
<td>29 (10)</td>
</tr>
<tr>
<td>MAST-3</td>
<td>7 (4)</td>
<td>17 (4)</td>
<td>19 (4)</td>
<td>20 (4)</td>
<td>83 (16)</td>
</tr>
<tr>
<td>MAST-4</td>
<td>0</td>
<td>0</td>
<td>2 (2)</td>
<td>6 (1)</td>
<td>8 (3)</td>
</tr>
<tr>
<td>MAST-5</td>
<td>2 (2)</td>
<td>3 (1)</td>
<td>2 (1)</td>
<td>0</td>
<td>7 (5)</td>
</tr>
<tr>
<td>MAST-7</td>
<td>5 (2)</td>
<td>3 (3)</td>
<td>3 (1)</td>
<td>1 (1)</td>
<td>12 (7)</td>
</tr>
<tr>
<td>MAST-8</td>
<td>12 (3)</td>
<td>8 (3)</td>
<td>23 (3)</td>
<td>28 (4)</td>
<td>71 (13)</td>
</tr>
<tr>
<td>MAST-9</td>
<td>6 (3)</td>
<td>14 (2)</td>
<td>28 (2)</td>
<td>4 (1)</td>
<td>52 (8)</td>
</tr>
<tr>
<td>MAST-12</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td>0</td>
<td>0</td>
<td>2 (2)</td>
</tr>
<tr>
<td>MAST-14*</td>
<td>0</td>
<td>5 (2)</td>
<td>0</td>
<td>0</td>
<td>5 (2)</td>
</tr>
<tr>
<td>MAST-15*</td>
<td>0</td>
<td>0</td>
<td>1 (1)</td>
<td>0</td>
<td>1 (1)</td>
</tr>
<tr>
<td>MAST-16*</td>
<td>0</td>
<td>2 (1)</td>
<td>2 (1)</td>
<td>0</td>
<td>4 (2)</td>
</tr>
<tr>
<td>MAST-17*</td>
<td>8 (2)</td>
<td>2 (2)</td>
<td>1 (1)</td>
<td>0</td>
<td>9 (5)</td>
</tr>
<tr>
<td>MAST-18*</td>
<td>1 (1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4 (2)</td>
</tr>
<tr>
<td>MAST-19*</td>
<td>0</td>
<td>7 (1)</td>
<td>2 (2)</td>
<td>4 (1)</td>
<td>13 (4)</td>
</tr>
<tr>
<td>NA1.1</td>
<td>0</td>
<td>7 (2)</td>
<td>1 (1)</td>
<td>2 (2)</td>
<td>10 (5)</td>
</tr>
<tr>
<td>NA1.2</td>
<td>10 (4)</td>
<td>30 (3)</td>
<td>13 (2)</td>
<td>2 (1)</td>
<td>45 (10)</td>
</tr>
<tr>
<td>NA1.3</td>
<td>8 (4)</td>
<td>6 (2)</td>
<td>10 (3)</td>
<td>0</td>
<td>22 (9)</td>
</tr>
<tr>
<td>NA1.4</td>
<td>21 (4)</td>
<td>40 (4)</td>
<td>47 (4)</td>
<td>25 (2)</td>
<td>134 (14)</td>
</tr>
<tr>
<td>NA1.5</td>
<td>0</td>
<td>0</td>
<td>1 (1)</td>
<td>1 (1)</td>
<td></td>
</tr>
<tr>
<td>NA1.6</td>
<td>0</td>
<td>0</td>
<td>4 (1)</td>
<td>2 (2)</td>
<td>6 (3)</td>
</tr>
<tr>
<td>NA1.7</td>
<td>5 (2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5 (2)</td>
</tr>
<tr>
<td>NA1.8</td>
<td>2 (1)</td>
<td>3 (3)</td>
<td>4 (2)</td>
<td>0</td>
<td>9 (6)</td>
</tr>
<tr>
<td>NA1.9*</td>
<td>0</td>
<td>3 (1)</td>
<td>1 (1)</td>
<td>3 (1)</td>
<td>7 (3)</td>
</tr>
<tr>
<td>NA1.4A*</td>
<td>1 (1)</td>
<td>0</td>
<td>2 (2)</td>
<td>0</td>
<td>3 (3)</td>
</tr>
<tr>
<td>NA1.4B*</td>
<td>0</td>
<td>1 (1)</td>
<td>4 (2)</td>
<td>0</td>
<td>5 (3)</td>
</tr>
<tr>
<td>Putative Novel Euglenozoa Class*</td>
<td>0</td>
<td>4 (1)</td>
<td>25 (4)</td>
<td>21 (5)</td>
<td>50 (10)</td>
</tr>
<tr>
<td>Symbiontida Sister Clade A*</td>
<td>0</td>
<td>1 (1)</td>
<td>0</td>
<td>9 (1)</td>
<td>10 (2)</td>
</tr>
<tr>
<td>Symbiontida Sister Clade B*</td>
<td>0</td>
<td>4 (2)</td>
<td>18 (6)</td>
<td>10 (2)</td>
<td>32 (10)</td>
</tr>
</tbody>
</table>
Table S1. The number of clades falling into three different sizes ranges, defined here as being either cosmopolitan or endemic based on established criteria (see discussion).

<table>
<thead>
<tr>
<th>Size Range</th>
<th>Cosmopolitan</th>
<th>25-50 μm **</th>
<th>80-200 μm ***</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-20 μm *</td>
<td>20</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Endemic</td>
<td>17</td>
<td>5</td>
<td>14</td>
</tr>
</tbody>
</table>

* Sydniales, *Pelagomonas, Pseudobodo, Cafeteria, Labyrinthuloides, Paraphysomonas, Skeletonema, Chaetoceros

** Symbiontida, *Cyclidium, Dictyocha*

*** Chaunacanthida, Arthracanthida, Taxopodida, *Tetrapyle, Collosphaera, Plagiopyla, Rhizosolenia, Entomoneis*
REFERENCES


Beijerinck MW (1913). *De infusies en de ontdekking der backteriën*. Jaarboek van de Koninklijke Akademie voor Wetenschappen. Müller: Amsterdam, the Netherlands.


Chapter 3: ‘candidatus Cariacotrichea’
‘Candidatus Cariacotrichea’, a novel ciliate taxon from the anoxic Cariaco Basin, Venezuela.

William Orsi¹, Tine Hohmann¹, Virginia Edgcomb², Craig Taylor², William H. Fowle¹, Gordon T. Taylor³, Jose Faria⁴, Paula Suarez⁴, Wilhelm Foissner⁷, Peter Vdacny⁸, and Slava S. Epstein¹,⁵

¹ Department of Biology, Northeastern University, 313 Mugar Building, Boston, MA 02115, USA.
² Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA.
³ Marine Sciences Research Center, State University of New York, Stony Brook, New York 11794, USA.
⁴ Departamento de Biología de Organismos, Universidad Simón Bolívar, Sartenejas, Baruta, Estado Miranda, Venezuela.
⁵ Marine Science Center, Northeastern University, 430 Nahant Rd., Nahant, MA 01908, USA.
⁶ Department of Organismal Biology, University of Salzburg, Hellbrunnerstrasse 34, Salzburg, A 5020, Austria.
⁷ Department of Zoology, Comenius University, Mlynská dolina B-1, Bratislava, SK 84215, Slovak Republic.

Running Title: Cariacotrichea, a novel candidate ciliate taxon.
Subject Category: Ciliophora

*Corresponding Author

Phone: 617-373-4048

Fax: 617-373-3724

E-mail: slava.epstein@gmail.com
Abstract

The majority of microorganisms identified with the ribosomal RNA (rRNA) approach from environmental samples have never been seen. Thus, their reliable classification and taxonomic assignment is often difficult or even impossible. In our preliminary 18S rRNA sequencing work in the world’s largest anoxic marine environment, the Cariaco Basin, we detected 264 18S rRNA gene sequences affiliated with a novel ciliate clade, designated previously as “CARH”. Here, we confirm the phylogenetic separation of the CARH sequences from all other ciliate classes and reveal an outstanding morphological feature of this group by combining the traditional rRNA detection method of Fluorescent In Situ Hybridization with Scanning Electron Microscopy (SEM). The oral apparatus of these ciliates forms a unique archway-shaped area delineated by the paroral membrane. Based on this specific feature and molecular phylogenies, we propose a novel class, ‘candidatus Cariacotrichea’. 
Introduction

Ciliates are the most highly differentiated and specialized microbial eukaryotes. The most widely recognized unifying trait of ciliates are dimorphic nuclei, namely a large macronucleus accompanied by a small micronucleus. A further characteristic feature is a large number of cilia that are present at least in some stage of their life cycle. Ciliates occur across almost every habitat on Earth, including internal organs of invertebrates and vertebrates, a variety of terrestrial and semi-terrestrial habitats, inside of ice, in freshwater and marine habitats including the deep sea, sediments, and anoxic regimes (Lynn, 2008). After over 200 years of research on their diversity, morphology, physiology, and evolution, the discovery of new ciliate taxa continues (Foissner et al. 2002; Song et al. 2009).

The application of the rRNA approach to environmental samples suggests the presence of an enormous “unseen” diversity of microorganisms including ciliates (Epstein & Lopez-Garcia, 2008). Anoxic environments in particular have been shown to house a diverse assemblage of microbial eukaryotes branching at high taxonomic levels (Stoeck, Taylor & Epstein 2003; Edgcomb, Kysela, Teske et al., 2002; Alexander, Stock, Breiner et al., 2009) but for the majority of these clades there exists no morphological data. The Cariaco basin is the largest marine anoxic basin in the world exhibiting numerous geochemical gradients within the redox transition zone (Taylor, Scranton, Labichella et al., 2001), below which many novel lineages of 18S rRNA genes have been detected, branching from the species to class levels (Orsi, Edgcomb, Jeon et al., 2011 [Chapter 2 of this dissertation]; Stoeck, Taylor & Epstein, 2003; Stoeck, Hayward, Taylor et al., 2006). One of these clades, designated previously as “CARH” (Stoeck, Taylor & Epstein, 2003), is affiliated with the ciliates and is phylogenetically separated from all of the described classes. Such a high level of phylogenetic novelty is remarkable considering the widespread belief that ciliates are the best studied protistan phylum, such that its species have been mostly discovered (Finlay et al. 1996; for a different view, see Foissner et al. 2008), and this prompted us to visualize the cells of the “CARH” clade. While more conventional methods such as silver
staining would have been optimal for a proper description, we were never able to visualize or enrich any living protists from anoxic samples. Additionally, the abundance of protists within these samples is so low as to render silver impregnations impractical. Despite these frustrations, we were able to obtain in situ preserved specimens using a deep-sea in situ sampler (Taylor & Doherty, 1990, Edgcomb, Orsi, Taylor et al., 2011 [Chapter 4 of this dissertation]), which allowed for FISH and SEM analyses of the cells of interest (Stoeck, Fowle & Epstein 2003). We visualized cells of CARH ciliates from in situ samples using a FISH probe specific to a group of rRNA sequences from the CARH clade. Analysis of SEM images of the cells hybridized with the FISH probe shows that they retain a unique feature not yet seen in any known ciliate, an archway-shaped paroral membrane. This feature supports the level of novelty implied by sequence data. We suggest that the ciliates from the CARH clade be considered as a novel taxon, ‘candidatus Cariacotrichea’, that corresponds to a new, class-level group of ciliates. The taxonomic status ‘candidatus’ is used by bacteriologists for uncultured taxa with novel molecular signatures and phenotypes (Murray & Stackebrandt, 1995), and has not yet been applied to microbial eukaryotes. However, it seems to be suitable for the ciliates studied here because they are currently uncultured and have a high level of both molecular and phenotypic novelty.
Materials and Methods

Sampling site. The sampling site was located in the eastern portion of the Cariaco Basin located at 10.50° N, 64.66° W. This is the location of the time series station of the cooperative U.S.-Venezuelan Carbon Retention in a Colored Ocean (CARIACO) program (Taylor, Scranton, Iabichella et al., 2001). Samples were taken aboard the R/V Hermano Ginés operated by Estación de Investigaciones Marinas (EDIMAR), Fundación la Salle de Ciencias Naturales, located on Margarita Island, Venezuela.

Sample collection and fixation. Samples for FISH-SEM were taken in January of 2009. The position of the oxic/anoxic interface in the water column was determined just prior to sample collection and was determined using a CTD scanner equipped with a YSI oxygen probe mounted on a Niskin rosette. The location of the interface was defined by a presence of a particle maximum at roughly the same depth in which the oxygen concentration dropped to zero. At the time of sampling this depth was 250 meters. Samples for FISH-SEM were taken from a depth of 900 m using a Large sample Volume Submersible Incubation Device (LV-SID) in situ water column sampler (Taylor & Doherty, 1990; Taylor, Howes & Doherty, 1993), which allows for sample fixation at the point of sampling. The LV-SID sample collection chambers were prefilled before deployment with a fixative mixture of Bouin’s and glutaraldehyde. The volume of the fixative was such that the final concentration of the fixed sample would contain 0.2% glutaraldehyde and 50% Bouin’s fixative. The LV-SID was programmed on deck to take a 4 L sample of water at 900 m, to be mixed in situ with the pre-loaded fixative. Fixed samples were transferred on deck to a 4 L carboy, stored at 4 degrees Celsius and were processed within 24 hours at the EDIMAR shore lab on Margarita Island.

Phylogenetic analysis. The sequence BCB5F14RJ2E06, from our earlier survey of the Cariaco Basin, was found to be a member of the CARH clade through a BLASTN search against the GenBank nt database (Orsi, Edgcomb, Jeon et al., 2011 [Chapter 2 of this dissertation]). Here, we improve this analysis by using a sequence alignment incorporating representatives from all ciliate
classes, ‘ca. Cariacotrichea’, and ciliates that appear similar to ‘ca. Cariacotrichea’. Only sites that could be reliably aligned were included in subsequent phylogenetic analyses. Alignments of the original sequences, along with GenBank reference sequences, were analyzed using Bayesian and Maximum Likelihood inference methods using RAxML (Stamatakis, Hoover & Rougemont, 2008) and MrBayes (Ronquist & Huelsenbeck, 2003). Both phylogenetic analyses were performed on the CIPRES portal (www.phylo.org) under the GTR+I+Gamma model. The reliability of the ML tree was assessed using 1000 bootstrap replicates. The Bayesian analysis consisted of two independent runs with 5x10^6 generations. Trees were sampled every 1000 generations with 25% discarded as burn-in. Topologies of ML and Bayesian trees were compared and the tree with the best log likelihood was chosen for presentation.

**Probe design and optimization.** The CARH group specific Cy3 probe; CARH658 (5’ – UACUGAUACCCCCGACUGUUC – 3’, 22 nucleotides, GC content 59%) used in this study was designed using the Probe Design tool available in the ARB Software Package (Ludwig, Strunk, Westram et al., 2004). The location of the hybridization region of CARH658 is within the best accessible portion of the 18S rRNA gene (position 658) based on *Saccharomyces cerevisiae* secondary structure (Behrens, Ruhland, Inacio et al., 2003). The probe was designed against 264 sequences sharing 95% identity, represented by the “centroid” sequence BCB5F14RJ2E06 from a previous rRNA survey of the Cariaco Basin (Edgcomb, Orsi, Bunge et al., 2011 [Chapter 1 of this dissertation]). The CARH658 probe was found to have no base pair mismatches with any sequences in this OTU. We also compared the probe against the GenBank-nt database using BLASTn, and the ARB-SiLVA database using the Probe Match tool, to confirm that it is specific exclusively to these sequences. The closest species match to the probe was the anaerobic ciliate *Epalxella antiquorum*, with one mismatch. However, this strain does not yet exist in culture and could not be used in a negative control experiment to test the specificity of the probe. The closest cultured species with a match to the probe was *Chlamydomonas monadina*, which exhibited five mismatches with the probe. A culture of *C. monadina* was obtained and fixed (50% Bouin’s
fixative, 0.2% glutaraldehyde) for a negative control experiment in order to optimize hybridization parameters for the probe. A range of 0-40% formamide concentrations in the hybridization buffer as well as incubations ranging from two to four hours were tested to determine the stringency required to eliminate visible non-specific binding of CARH658. The set of incubation parameters that produced the lowest amount of FISH signal with C. monadina was chosen for use with environmental samples to detect ciliates from the CARH clade.

**FISH-SEM and SEM preparation.** Samples fixed *in situ* for FISH-SEM were processed within 24 hours according to the protocol developed by Stoeck *et al.* (2003), with some minor modifications. In short, fixed samples were filtered onto 0.4 µm polycarbonate transwell membrane filters (Corning, USA) and washed with 1X PBS (pH 7.4). PBS was gradually replaced by three cycles of filtering the fluid over the filter down to 500 µl and then adding 3 ml of a hybridization buffer containing 40% formamide. With 3 ml of hybridization buffer remaining in the well, 500 µl of the Cy3 labeled probe CARH658 (30 ng/µl) was added. Transwells were incubated at 46 ºC for two hours in the dark, followed by incubation with a washing buffer preheated to 48 ºC for 10 minutes, and washing with distilled water. Transwells were then taken through a dehydration series in preparation for SEM and fixed with 100% hexamethyldisilizane (Electron Microscopy Sciences, Hatfield, Pennsylvania) before air-drying. As described by Stoeck *et al.* (2003), it was critical not to expose the transwell filters to air at any point during the protocol, until this final step, as such exposure would cause most fixed protists to collapse. The entire procedure was completed within 24 hr after sampling, and the air dried transwell filters were wrapped in aluminum foil and shipped back to the United States at 4 ºC. Upon arrival they were immediately placed in a -20 ºC freezer until further processing.

**Fluorescence Microscopy and SEM.** FISH-SEM prepared filters were cut out of the transwells using a scalpel and placed on a glass slide for visualization via epifluorescence. All filters were examined with dry objectives, because of the potential for SEM contamination with oil-immersed filters. Filters were first scanned at 200X under appropriate illumination using an AxioPlan 2
epifluorescence microscope equipped with a HBO 100-W mercury bulb; 10X Neofluar, 20X Neofluar, and 40X (dry) Neofluar objectives; 10x eye pieces; DAPI and Cy3 filter sets; and a Hamamatsu CCD camera (Hamamatsu, Hamamatsu City, Japan). Photographs of positively Cy3 labeled CARH ciliates were taken with at 400X. Positive Cy3-labeled cells had their location on the filter marked for downstream (SEM) observations by making puncture marks with tweezers in the filter next to the cell’s location. A minimum of three different punctures around the cell was needed to facilitate finding the labeled cell under SEM. The filter was then mounted onto a SEM specimen holder with a carbon adhesive tab and sputter coated with 10-15 nm of a mixture of platinum and palladium with a Tousimis Samsputter 2A. SEM was preformed on a Hitachi S-4800 scanning electron microscope. Approximate location of the marked cell was achieved under low-magnification mode by finding the puncture marks in the filter. The exact location of the cell was then determined in reference to the previously noted position of the cell (during epifluorescence microscopy) in relation to the puncture marks. High magnification mode was used to take detailed photographs of the Cy3 labeled cells for morphological assessments.
Results and Discussion

‘candidatus Cariacotrichea’

**Diagnosis.** Small anaerobic ciliates with archway-shaped paroral membrane and at least two adoral organelles, containing a molecular signature ‘GAAACAGUCGGGGGUAUCAGUA’ spanning nucleotide positions 658-680 in its ribosomal RNA gene.

**Type locality.** Anoxic, sulfidic water column from the Cariaco Basin, Venezuela, 10° 50’ N 64° 66’ W.

**Etymology.** Composite of the acronym CARIACO (for both the geographic name of the region and the multi-institutional ocean times series project Carbon Retention In A Colored Ocean), the stem of the Greek noun thrix (hair ~ cilium), and the class suffix –ea.

**Type material.** SEM prepared filters containing *in situ* samples from the deep anoxic portion of the Cariaco Basin. Filters with specimens are deposited in the Department of Biology at Northeastern University in Boston, Massachusetts, U.S.A. and are available upon request.

**Sequences affiliated with ‘ca. Cariacotrichea’ and evaluation of a group specific FISH probe.** The following sequences recovered from the anoxic portion of the Cariaco basin in previous studies (Stoeck, Taylor & Epstein, 2003; Stoeck, Hayward, Taylor *et al.*, 2006) are affiliated here with the ‘ca. Cariacotrichea’: CAR_H1b, T41D10, T41H4, T37D5, p15E10, and BCB5F14JR2E06 (Fig 1). Further, the sequence C1_E031, retrieved from an anoxic sediment in the Guaymas Basin (Edgcomb, Kysela, Teske *et al.*, 2002) is also assigned to the ‘ca. Cariacotrichea’. We discovered a conserved region of 18S rRNA in the sequence BCB5F14JR2E06 extending from bases 658-680 that is specific only to this sequence and its 264 sequence representatives. A BLAST search of this sequence within the GenBank-nt database, as well as a search with the ARB-SILVA database using the Probe Match tool, yielded 100% matches only to previously deposited sequences from the CARH clade. In a two hour incubation
using 40% formamide at 46 °C, the fluorescence from CARH658 was equal to that of the nonsense probe when used with the fixed culture of *Chlamydomonas monadina* (Fig 2). Thus, we used these hybridization parameters in FISH-SEM for the detection of ‘ca. Cariacotrichea’.

**Morphological description.** In total we located 13 different cells of ‘ca. Cariacotrichea’ by scanning filters under epifluorescence using the Cy3 filter set. No cells were visible on filters stained with a nonsense probe. The size of the FISH-labeled cells is 15-20 x 6-7 µm after processing samples for FISH-SEM. We assume that the ‘ca. Cariacotrichea’ might have an in vivo size of 30–40 x 15–25 µm. This calculation is based on a *Pleuronema* specimen treated with the same method that decreased in size by about 70% (data not shown). The body is bluntly fusiform and possibly flattened ventrally (Fig 4 a-c, Movie S1). There is a single, ellipsoidal macronucleus and a globular micronucleus consistently located in the posterior half of the body as visualized after DAPI staining (Fig 3 a-e). However, we cannot exclude a possibility that this unusual localization of the nuclear apparatus is an artifact caused by strong cell shrinkage. In SEM images, the cortex is almost completely covered by epibiotic bacteria approximately 1 µm long (Fig 4e, Movie S1). There is a special structure on the left side, spanning the entire body, and composed of narrowly spaced globular protrusions containing either unciliated basal bodies or cortical granules (Movie S1). The cilia are 10 to 20-µm long and 0.2-µm thick, as typical for most ciliates. They are loosely arranged and long in the anterior body region, while closely spaced in the posterior region where they are slightly elongated and partially appear as cirrus-like structures. At the posterior body end, there are two to several prominent caudal cilia 20 to 30-µm long. There are about five to six meridionally extending ciliary rows on each cell side indicating 10-12 kineties in total (Fig 5). The oral apparatus occupies about one third of the body length in the SEM and is located on the anterior ventral side. It is composed of a buccal cavity delineated by an archway-shaped paroral membrane composed of a single row of very narrowly spaced basal bodies with 10 to 30 µm long cilia (Fig 4a, c, d, Movie S1). In the specimen shown in Fig 4 a, c, the paroral is almost unciliated. We
believe that this is an artifact caused by the FISH-SEM preparation procedures. Quite similar pictures can be obtained in ciliates deciliated with a detergent (Fig. 4g): the paroral cilia disappear while some adoral cilia remain. On the right side of the buccal cavity, there are at least two adoral organelles with rather short cilia (~ 5µm) cilia.

**Autecology.** Cells of the ‘candidatus Cariacotrichea’ were found to have an *in situ* abundance of 0.2 cells/ml at a depth of 900 m in the Cariaco Basin. This region of the basin is permanently anoxic and exhibits a concentration of hydrogen sulfide up to 53 µM (Edgcomb, Orsi, Bunge et al., 2011 [Chapter 1 of this dissertation]). This depth has a typical temperature of 17 °C and a salinity of 36.2‰.

**Phylogeny and comparison with other ciliate classes.** In our phylogenetic analyses (Fig 1), the ‘candidatus Cariacotrichea’ forms a clade that is fully supported by both methods and is affiliated with the ciliate subphylum Intramacronucleata. However, ‘ca. Cariacotrichea’ does not group within any of the eleven described classes of ciliates (Lynn, 2008). This suggests that the ‘ca. Cariacotrichea’ represents a new class of ciliates, probably restricted to anoxic marine environments. The microeukaryotic community in the Cariaco Basin has likely diversified and speciated in response to the biogeochemical gradients of oxygen and sulfide (Orsi, Edgcomb, Jeon et al., 2011 [Chapter 2 of this dissertation]). Thus, the unique geochemistry of anoxic environments may explain, in part, the divergent phylogenetic position and outstanding morphology of the ‘ca. Cariacotrichea’. The closest class to the ‘ca. Cariacotrichea’ is the Spirotrichea, branching as a sister clade with bootstrap and posterior probability support of 64% and 0.89, respectively. The spirotricheans (Fig 1) are a species-rich, rather diverse taxon, containing the oligotrichids, eublotids, hypotrichs, and, possibly, the “ribo-subclasses” Protocruziida, Phacodiniida, and Licnophorida (Lynn, 2008). However, none of these closely resembles the ‘ca. Cariacotrichea’ morphologically, which instead superficially looks similar to some scuticociliates, namely *Cyclidium*, *Tetrahymena*, and *Wilbertia* Fan et al. (2009). However,
the paroral membrane of scuticociliates is typically J- or U-shaped, while the paroral membrane of the ‘ca. Cariacotrichea’ is archway-shaped delineating the anterior half of the buccal cavity on both right and left sides. There is also some similarity to Protocruzia and Palmarella (previously Palmarium, see Aescht 2001), whose classification and detailed ciliary pattern are still unknown (Fig. 5). As concerns Protocruzia, see Li et al. (2010).

‘Candidatus’ status for the Cariacotrichea. Due to significant collection challenges, particularly difficulties with handling and processing of samples caused by pressure differentials and exposure to oxygen, we were not able to observe any live cells from ‘candidatus Cariacotrichea’ (and indeed any live protist from this depth). Also, their low in situ abundances limit our ability to study the specimens using various silver staining methods. Both live observation and silver impregnation are considered to be prerequisites for description of new ciliate taxa (Foissner 1991). Despite the lack of such observations and preparations, we were able to obtain molecular and phenotypic evidence showing a high novelty of the CARH clade. Thus, we propose that this group be considered as ‘candidatus Cariacotrichea’, a new class-level group of ciliates. The taxonomic status candidatus has not yet been applied to microbial eukaryotes, but is used for novel bacterial taxa that are uncultured (Murray & Stackebrandt, 1995), and seems suitable in our case. Therefore, we will make a bid to the International Commissions on Zoological and Botanical Nomenclature to include the status candidatus into the International Codes. It should be noted that bacteriologists conventionally use an upper-case letter “C” in the Latin noun candidatus incorrectly. This may be in conflict with the Principle of Binomial Nomenclature (article 5.1 of the ICZN 1999 and article 23.1 of the ICBN 2006) because the statement ‘Candidatus’ appears as the generic name of a binomen. For the sake of clarity, we use a lower-case “c” in the spelling of candidatus throughout this manuscript.
ACKNOWLEDGEMENTS

We would like to thank the captain and crew of the B/O Hermano Ginés and the staff of the Fundación La Salle de Ciencias Naturales (FLASA) for their assistance during our fieldwork in Venezuela, without whom this work would not be possible. We are also deeply grateful to Yrene Astor and Ramon Varela who were instrumental in the transportation of the DEEP-SID to and from Isla Margarita, as well as Mary Scranton for her logistical support. We appreciate advice by Dr. Rudolf Amann from the Max Planck Institute for Marine Microbiology in Bremen, Germany, on optimization of FISH probes in the absence of proper positive control. This research was supported by grants from NSF (MCB-0348341 and DEB-0816840 to SE, MCB-0348407 to VE and OCE 03-26175) and the Austrian Science Foundation to WF (FWF, P-20360-B17). This is a contribution # 271 from the Marine Science Center, Northeastern University, Nahant MA, U.S.A.
Figure 1. Phylogenetic position of the proposed ciliate class ‘candidatus Cariacotrichaea’ relative to the other eleven classes. Bootstrap (BT) and posterior probability (PP) values of >50% are given at the nodes in the order PP/BT. A black circle denotes full bootstrap and posterior probability support. Dashed lines represent a bootstrap or posterior probability value of less than 50% or 0.50. The tree is based on an alignment of 46 sequences and 942 nucleotide positions.
Figure 2. Comparison of the probe CARH658 (a; lack of hybridization) to a universal eukaryotic probe Euk1209R (b; presence of hybridization), when used in FISH with a fixed culture of *Chlamydomonas monadina*. 
Figure 3. DAPI (a-e) and Cy3 (f-j) staining of cariacotrichean ciliates from the deep anoxic portion of the Cariaco Basin visualized with the CARH658 FISH probe. Bars, 30 µm. ma: macronucleus, mi: micronucleus.
Figure 4. (a-g) Scanning electron micrographs of cariacotrichean ciliates. (a) Ventral view showing the bluntly fusiform body with prominent caudal cilia and archway-shaped paroral membrane (arrowheads), an important diagnostic feature. (b) A second specimen where the anterior and posterior end cannot be unambiguously differentiated. Likely, the denser ciliated end is that with the oral apparatus. (c) Ventral view of the specimen shown in Fig. 4a, where the buccal cavity is delineated by the archway-shaped paroral membrane (arrowheads). Note the adoral organelles emerging from the buccal cavity. (d) Posterior region of the specimen shown in Fig. 4a. Note the narrowly spaced cilia (arrows) in the posterior region of the rows. (e) Surface view of the cortex covered with epibiotic bacteria. (f) Side view of the specimen shown in Fig. 4a, c, d. The oral apparatus (OA) is well recognizable and the ventral side slightly flattened. Arrowheads delineate the structure on the left side that spans the entire body. (g) A still undescribed oligohymenophorean ciliate from a tree-hole in Jamaica. When treated with a detergent, the oral apparatus looks rather similar to that of the candidatus Cariacotrichea because the paroral cilia were lost and some adoral cilia were maintained. AO, adoral organelles; CC, caudal cilia; PM, paroral membrane cilia; Bars, 1 µm (d, e), 2 µm (c), 5 µm (f), and 20 µm (a, b, g).
Figure 5. Line drawings of a cariacotrichean ciliate (a, b) and of *Palmarella* (c, from Kahl 1932 after Gaievskaia 1925). (a) Ventral view reconstructed from the specimens shown in Fig. 4a, b (sh. also movie). (b) Lateral outline showing flattened ventral side. (c) Ventral view of *Palmarella salina* from saline inland puddles in Russia, length about 30 µm. AO, adoral organelles; CC, caudal cilia; NC, narrowly spaced cilia in posterior region; NU, nuclear apparatus; PM, paroral membrane. Scale bar 10 µm.
REFERENCES


Chapter 4: *In situ* access to marine protists
Accessing Marine Protists from the Anoxic Cariaco Basin

V. Edgcomb¹, W. Orsi², G.T. Taylor³, P. Vdacny⁴, C. Taylor², P. Suarez⁵, and S. Epstein¹,⁶*

¹ Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA.

² Department of Biology, Northeastern University, 313 Mugar Building, Boston, MA 02115, USA.

³ School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York 11794, USA.

⁴ Department of Zoology, Comenius University, Mlynská dolina B-1, Bratislava, SK 84215, Slovak Republic.

⁵ Departamento de Biología de Organismos, Universidad Simón Bolívar, Sartenejas, Baruta, Estado Miranda, Venezuela.

⁶ Marine Science Center, Northeastern University, Nahant, MA 01908, USA

Keywords: protists, diversity, anoxic, 18S rRNA

Subject Categories: Microbial ecology and functional diversity of natural habitats

*Corresponding Author
The earliest microbial cells evolved in an anoxic ocean (Martin and Muller, 1998), and many argue that the earliest Eukarya arose in the absence of oxygen as well. The microbe-mediated biogeochemical processes taking place below oxic/anoxic interfaces are of undeniable importance and interest and the Cariaco Basin off the coast of Venezuela is a superb model ecosystem in this regard. It is the world’s largest body of truly marine anoxic waters and has been almost continuously anoxic for at least the last 2M years (Schubert 1982). Consequently, the Basin became the site of many investigations, including a long-term biogeochemical time series initiated in 1995 (Muller-Karger et al., 2001; Taylor et al., 2001), and an NSF-funded Microbial Observatory established in 2004. This Observatory was a multi-year, multi-investigator international effort. Its principal goals were:

- Combine molecular and cell-based approaches to survey microbial Eukarya among geochemically diverse habitats in the Basin.
- Analyze the community structure and its dynamics across time and space.
- Discover novel organisms, and to gain insights into what governs their distribution dispersal, and biogeography.

Overall, this observatory became arguably the largest exploration of protistan life in a single geographic location to date. We summarize its key results in two papers published in this issue (Edgcomb et al., 2011 and Orsi et al., 2011 [Chapters 1 and 2 of this dissertation, respectively]). Although the data on protists in the Basin presented in the graphs, phylogenetic trees, and statistical plots of these papers tell some interesting stories, other intriguing and sometimes unexpected observations also deserve mention. Although such observations may not support firm conclusions, they can provide important insights into protistan ecology and help guide future research, as discussed in this Commentary.
From the beginning, a major goal was to step beyond sequence-based discovery, and, having identified interesting 18S rRNA gene signatures, to get hold of the novel species, preferably in the form of their cultures. However, all the samples brought to the surface from below the oxic/anoxic interface, typically positioned at a depth of 250 meters, remained lifeless. Despite the utmost precautions to avoid exposure to oxygen, or to minimize shifts in pressure, no protists ever grew.

This challenge turned into an opportunity to develop a sampling device to capture and fix the cells in situ, something that apparently had not been done before, at least for microbial eukaryotes. A modified deep-sea fixation device (Taylor and Doherty, 1990) enabled preservation of cells in situ with glutaraldehyde for scanning electron microscopy (SEM) and with Bouin’s solution for combined fluorescent in situ hybridization (FISH) - SEM analysis (Stoeck et al., 2003). Although the resulting cells were dead, and thus still not cultivable, it was now possible to see them in all their glory (Figure 1 a-i, and also the cover of this issue).

This approach revealed an abundance of apparently novel protists from anoxic waters of the Cariaco Basin, whose novel morphologies reflected the divergent 18S rRNA gene sequences found below the oxic/anoxic interface (Edgcomb et al., 2011 and Orsi et al., 2011 [Chapters 1 and 2 of this dissertation, respectively]). A large number of small cells had spherical shapes that did not leave much room for interpretation. However, cells exhibiting more complex morphology showed unique features. Some examples include nanoflagellates that possessed conspicuously long flagella almost five times longer than their body (Figure 1c) and what appeared to be cryptomonad flagellates exhibited a furrow extending from the vestibulum to the posterior body end (Figure 1a). We also observed numerous ciliates, sharing with the known scuticociliates their prominent paroral membrane and one to several caudal cilia. As common for anaerobic
ciliates, these species were covered with morphologically diverse epibiotic bacteria (Figure 1e-i and the cover of this issue).

The observed flagellates and ciliates confirmed that at least some protists detected from anoxic depths by the rRNA gene approach were not dead: their cells were not collapsed as would be expected of naked cells upon death, and the presence of putative symbionts suggested activity. A cDNA approach brought another line of evidence for the presence of live protists in this habitat, whereby reverse transcription lead to detection of short lived 18S rRNA. This revealed an abundance of likely live and active cells in the deep anoxic layers of the Cariaco Basin, such as ciliates (Figure 2), fungi, and stramenopiles. Our observation of many diverse and novel ciliate phylotypes is in line with earlier molecular surveys of the Cariaco Basin (Stoeck et al., 2003b).

Despite these advances, the questions remained, “Was there anything particularly special about the anoxic depths of the Cariaco? Were these habitats somehow enriched with novel protists? And if so, to what extent?”. To address these questions is a challenge, but should be possible using a combination of SEM and FISH (Kolodziej and Stoeck, 2007; Stoeck et al., 2003). Counting intact cells on SEM preparations, which is rarely done because it is so labor-intensive, would provide arguably the best possible total cell count. FISH with a universal eukaryotic probe such as euk1209R (Giovannoni et al., 1988) performed on the same preparations, would provide an approximation of the number of cells belonging to known species. Given the fact, that such probes are designed on the basis of known sequences, mismatches in target binding sites would be more likely to occur in previously undiscovered organisms. Thus, the ratio between the FISH and SEM counts provides insight into the relative fraction of known to total protists in the sample.
We compared such fractions for samples collected at four depths, from the fully oxygenated upper to highly sulfidic 900 m deep layers (Figure 1j). Expectedly, SEM counts exceeded those obtained by FISH in all samples, since no single universal FISH probe is perfect. However, such a probe could be expected to visualize the majority of cells in better-studied habitats, such as the ocean surface, towards which such probes are biased. Predictably, 91% of cells from the oxygenated layer were detected in our FISH/SEM assay (Figure 1j). But, we were surprised by a dramatic drop in the ratio of FISH-to-SEM detected cells with depth, down to mere 6% in the most sulfidic samples (Figure 1j). The likeliest explanation is that, the more rRNA novelty was contained in the progressively deeper samples, the more organisms the conventional rRNA-based technique missed. This is indirectly supported by the following observation. In silico, over 40% of OTUs that we detected in anoxic samples, and clustered at 99% sequence identity level, have at least one mismatch with the 1209R probe binding site: 15 out of 39 OTUs in the cDNA dataset, and 313 out of 765 in the rRNA gene sequence dataset (the latter is published in the two companion papers in this issue by Edgcomb et al., 2011 and Orsi et al., 2011 [Chapters 1 and 2 of this dissertation, respectively]). Indeed, the shortcomings of rRNA probes targeting diverse microbial groups are well known (Lucker et al., 2007), and some evidence suggests that a conventional application of the rRNA approach likely accesses only half of microbial eukaryotic diversity (Jeon et al., 2008). On the other hand, because rRNA copy number is an indication of microbial growth and activity (Klappenbach et al., 2000), it is also possible that some of the cells seen by SEM but missed by FISH had low ribosome content, growth rates, or were dead. We consider this rather unlikely because, for the last possibility, such cells must be dead yet look alive and remain intact, which is almost impossible for naked protists (surrounded only by a cell membrane). Finally, it is possible that our probe did not penetrate the cell membrane of the majority of cells from the deepest samples. However, how likely is it that probe exclusion explains low hybridization efficiencies considering that the majority of cells at this depth are comprised of naked cells? We conclude that it might indeed be possible that the
majority of protists from this deep anoxic layer are divergent at this ‘universal’ eukaryotic FISH probe binding site. It follows that it is equally possible that most of these protists would not be recovered using available eukaryotic PCR primers either. Thus, many of our SEM observations may represent truly novel organisms that had escaped detection to date by standard molecular approaches. Techniques free of PCR biases, such as metagenomics and metatranscriptomics, might provide access to this unknown source of diversity but these have yet to be applied to communities of microbial eukaryotes. These results also highlight a need to renew efforts in direct observation and novel cultivation methods. Traditional microscopy approaches seem to have fallen out of fashion, but clearly remain invaluable, especially as a synergy between as transmitted light, fluorescence, and electron microscopy observations, particularly for exploring taxa rich in morphological characters.

This brings us back to our original point. While admittedly preliminary, the most critical data discussed here were obtained exclusively because we were able to fix cells in situ. We note that even “simple” fixation in situ was not a trivial exercise, and actual in situ experimentation will be all the more difficult. However, if results and surprises to date are any indication, this might be well worth the effort.

Acknowledgements

We thank the captain and crew of the B/O Hermano Gines and the staff of the Fundación La Salle de Ciencias Naturales, Margarita Island for their assistance during our fieldwork in Venezuela. We thank particularly Yrene Astor and Ramon Varela for their strong logistical support during our sampling. We are grateful to Mary I. Scranton and the dedicated researchers of the CARIACO biogeochemical time series program, without whom this work would not have been possible. This research was supported by grants from NSF (MCB-0348341 and DEB-
0816840 to SSE, MCB-0348407 to VE and OCE 03-26175 and MCB-03-47811 to GTT). This is a contribution # 270 from the Marine Science Center, Northeastern University, Nahant MA, U.S.A. The cDNA sequences from this study have been deposited in GenBank with the accession numbers HM443081 - HM443437.
Figure 1: Scanning electron micrographs of protists from the Cariaco Basin’s anoxic, sulfidic, deepwaters (900 m). (a-d), Flagellates; (g-i) possible scuticociliates with epibiotic bacteria (e, f). FU, furrow; V, vestibulum; F, flagellum; PS, pseudopodia; PM, paroral membrane; SK, somatic cilia; CC, caudal cilia; B, epibiotic bacteria. Scale bars: 2 µm (c, e, f), 5 µm (b, d, h, i), 10 µm (g), and 20 µm (a). (j) The number of protistan cells visualized by FISH with a ‘universal’ eukaryotic probe (solid line) versus their total number determined by SEM (dotted line). A: 40 m above oxic/anoxic interface, I: interface (approximately 250 m depth), B: 40 m below interface, D: 900 m.
Figure 2: Phylogenetic relationships of ciliate 18S rRNA gene sequences obtained by the cDNA approach and direct amplification of the gene from environmental samples. Bootstrap (RAxML) and posterior probability values greater than 50% are shown at the nodes in the order PP/ML. Black circles at nodes represent full posterior probability and bootstrap support. Centroid sequences from our 18S rDNA survey (see Chapters 1 and 2) are highlighted in brown and sequences from GenBank are black. All cDNA sequences fell into the ‘cDNA clade’. The number of cDNA sequences and those from our rDNA survey are shown in the triangle in the order cDNA, rDNA.
<table>
<thead>
<tr>
<th>OTUs (total number of sequences)</th>
<th>Closest named species</th>
<th>Genbank accession #</th>
<th>Identity</th>
<th>Closest environmental clone</th>
<th>Genbank accession #</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (15)</td>
<td>Cyclidium porcatum</td>
<td>Z29517</td>
<td>88-89%</td>
<td>Uncultured eukaryote E2</td>
<td>AY256242</td>
<td>93-99%</td>
</tr>
<tr>
<td></td>
<td>Urocentrum sp.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (15)</td>
<td>Altsee 2</td>
<td>EF114302</td>
<td>88%</td>
<td>Uncultured eukaryote E2</td>
<td>AY256242</td>
<td>95%</td>
</tr>
<tr>
<td></td>
<td>Blastocladielae sp.</td>
<td></td>
<td></td>
<td></td>
<td>AB468672</td>
<td></td>
</tr>
<tr>
<td>1 (16)</td>
<td></td>
<td>EF565163</td>
<td>93%</td>
<td>Uncultured fungus S08B38</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Uncultured eukaryote</td>
<td>AB252777</td>
<td>98%</td>
</tr>
<tr>
<td>1 (4)</td>
<td>Rhizophyllum sp.</td>
<td>AY635821</td>
<td>90%</td>
<td>Uncultured eukaryote</td>
<td>AB275064</td>
<td>99%</td>
</tr>
<tr>
<td></td>
<td>Blastocladielae sp.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (2)</td>
<td>Physoderma dulichii</td>
<td>EF565163</td>
<td>92%</td>
<td>Uncultured eukaryote</td>
<td>AB252777</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (4)</td>
<td>Thalassiosira punctigera</td>
<td>DQ536472</td>
<td>88%</td>
<td>NAMAKO-37</td>
<td>AB252777</td>
<td>98%</td>
</tr>
<tr>
<td>9 (221)</td>
<td>Thalassiosira</td>
<td>DQ514893</td>
<td>96-99%</td>
<td>Uncultured eukaryote</td>
<td>DQ093369</td>
<td>98-99%</td>
</tr>
<tr>
<td></td>
<td>aestivalis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thalassiosira ritscheri</td>
<td>DQ0933690</td>
<td>99%</td>
<td>Uncultured marine eukaryote</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 (16)</td>
<td>Thalassiosira antarctica</td>
<td>DQ514891</td>
<td>96-99%</td>
<td>NPK57_37</td>
<td>DQ310225</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Uncultured marine eukaryote</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 (45)</td>
<td>Stephanodiscus hantzschii</td>
<td>EF140621</td>
<td>98.94</td>
<td>Uncultured marine eukaryote</td>
<td>DQ093369</td>
<td>98%</td>
</tr>
<tr>
<td>1 (1)</td>
<td></td>
<td>GQ844874</td>
<td>96%</td>
<td>NPK57_7</td>
<td></td>
<td>99%</td>
</tr>
</tbody>
</table>

Table 1: Taxonomic affiliation of cDNA sequences obtained from a deepwater anoxic (900 m) sample.
REFERENCES


Summary

With the evidence presented in this dissertation, one of the oldest questions concerning the biogeography of protists, “Is everything everywhere?,” has finally been laid to rest. Of course, the full quote is “Everything is everywhere, but the environment selects,” and most would agree that environmental selection for dominance of different microbial groups does occur. However, the first portion of the quote has been debated by microbial ecologists for nearly a century, with evidence presented for and against this statement. Another way of phrasing this question is “What is the extent of the species pool that the environment is exerting selection upon?” Are all species or just a subset present during environmental selection pressure? The data presented here suggest that the latter applies. The sequence dataset described and analyzed here is the largest and least biased of its kind and thus it enabled the first reliable estimation of protistan species richness in a large and varied environment using state-of-the-art statistical tools developed for this specific purpose. The number of protistan species estimated to exist in the Cariaco Basin falls far short of the total number of described protists, and thus it is concluded that not all species of protists are present in the Cariaco Basin (see Chapter 1). Additional supportive evidence comes from the finding that distinct geochemistry at different depths in the water column select for many protistan taxa that speciate and diversify in response to a biogeochemical niche and do not successfully disperse to other locations (see Chapter 2).

Many of the taxa that appear to be confined to a specific geochemical niche in the water column were found to be new to science, and divergent from species to class levels. One example of this is the discovery and description of the ‘candidatus Cariacotrichea’, a new class-level lineage of ciliates found within the anoxic waters of the Cariaco Basin (see Chapter 3). An additional example of the protistan novelty held within the anoxic waters of the Basin is the discovery that the majority of protists in these waters are divergent in one of the most conservative regions of the 18S rRNA gene (a universal FISH probe binding site) (see Chapter
4). This is exciting because if the majority of protists have mismatches to the binding site for a universal 18S rRNA FISH probe, they may also have mismatches to “universal” PCR primer binding sites as well. In that case the majority of protistan cells below the oxic/anoxic interface may not be detectable by PCR-based methods utilizing “universal” primers.

The discovery and description of a new, class-level clade of ciliates in the Cariaco Basin has many far-reaching implications. First, it is an excellent demonstration of the morphological novelty that can lie behind divergent ribosomal RNA sequences. Furthermore, a discovery of a new group at such high taxonomic level within one of the best-studied protistan phyla suggests that less studied groups likely contain many undiscovered high-level taxonomic lineages as well. As these, yet to be discovered groups likely have important ecological functions, their discovery, description, and study is necessary in order to obtain a comprehensive understanding of microbial ecology in aquatic ecosystems. Lastly, the discovery of the ‘candidatus Cariacotrichea’ lends strong support to the high estimate of species richness within the Cariaco Basin (roughly 36,000 species), as the phylogenetic diversity within this class as well as the morphological novelty of the cells corresponding to the novel sequences, suggest that the Cariacotrichea likely contains a multitude of species that await description.

What does the future hold for the field of protistan ecology? Molecular based surveys of 18S rRNA gene diversity over the last decade have indicated protists are an extremely diverse group that likely houses a large degree of taxonomic novelty. Indeed, it appears we have only “scratched the surface” of protists within aquatic ecosystems, with anoxic marine environments (such as the Cariaco Basin) proving to be exceptionally promising “hotspots” for microbial discovery. While diversity-based studies of protists have yielded important and interesting findings, there are many logical questions that arise from such discoveries. What are the in situ abundances of protistan groups, both novel rRNA sequence clades and described species, in aquatic ecosystems? How does the diversity of protists correlate to bacterial diversity? What is the in situ function of protists in aquatic ecosystems? What are the different groups of protists
that impact the biogeochemical cycles of nitrogen, sulfur, carbon, and phosphorus in the world’s oceans, and how is this impact exerted? What influence does the diversity of symbiotic modes between bacteria and protists have on the environment? To answer these questions is a challenge, and will require researchers to supplement ribosomal RNA sequence-based discovery of protists with additional approaches.

One of the most important of such approaches is the culturing of representative species that lie behind novel ribosomal RNA sequence clades, as well as those clades that have been described based on sequence data alone but for which a culture does not exist. Such protists may constitute a significant fraction of protistan communities in aquatic ecosystems. Culturing of these strains will enable detailed study and experimentation that will provide insights into their physiology and functional role in the environment. Cultures of novel eukaryotes could also be used for in situ-based experiments to garner additional information about their environmental role and importance.

The study of protistan communities in situ should provide new and interesting insights. For example, utilization of an in situ filtration device would permit a comparison of in situ diversity to that recovered with traditional Niskin sampling. Such a comparison would be useful as the biases introduced by Niskin sampling in ribosomal RNA based studies are not well known. Additional in situ experiments to examine protistan grazing, utilizing fluorescent and or isotope labeled bacterial prey will help to identify how much bacterial biomass is removed by protistan grazing in different environments as well as what types of bacterial prey are selected for by different groups of protists. Different ‘omics’ based approaches, such as metagenomics and metatranscriptomic analyses should also provide insights into the metabolic diversity and function of protistan communities in aquatic environments. It is my prediction that a combination of the aforementioned approaches will substantially supplement our existing knowledge of protistan diversity and provide a richer understanding of the role of aquatic protists and their impact on the environment.
Oxygen minimum zones (OMZs) represent one environment in particular that is of significant importance to global biogeochemical cycles and society in general. Marine oxygen minimum zones are naturally occurring and widespread regions of low dissolved oxygen concentration found throughout the global oceans. Commonly referred to as “dead zones”, OMZs are anathema to macrobiota yet house diverse assemblages of microbes that impact numerous biogeochemical cycles such as the cycling of nitrogen, phosphorus, sulfur, and carbon. Microbial mediated biogeochemical transformations within OMZs remove biological nitrogen in the form of dinitrogen gas (N\textsubscript{2}) and mediate the production of the greenhouse gases nitrous oxide (N\textsubscript{2}O) and methane (CH\textsubscript{4}). Changes in ocean temperature and circulation patterns have been compounded by human activities and are leading to OMZ expansion. This anthropogenic influence has in turn has caused concomitant alteration in nutrient and climate active trace gas cycling. Therefore, studies designed towards understanding microbial community structure and systems metabolism within OMZs are necessary for developing probabilistic models of ocean ecosystem resilience and responses to global climate change. It is likely that as OMZs continue to expand, global shifts of microbial populations will occur in these regions. This shift will result in an increase in the bacterial mediation of anaerobic processes such as the anaerobic oxidation of ammonia, sulfate reduction, and denitrification. Protistan populations will also shift in response to expanding OMZs. In these regions, protists typical of oxygenated waters will gradually be replaced by those with anaerobic capabilities, and selection will occur for diverse symbiotic modes providing microbial fitness advantages.

Within marine OMZs, protists (along with viruses) likely play a significant role in regulating the abundance of bacteria responsible for the production of toxic hydrogen sulfide (H\textsubscript{2}S), the climate active trace gases nitrous oxide and methane (N\textsubscript{2}O, CH\textsubscript{4}), and biological nitrogen loss (N\textsubscript{2}). Thus, as the formation and expansion of these regions continues, a comprehensive understanding of the protistan community and its role within OMZs is needed. Multidisciplinary approaches should bring us closer to that end.
Protistan community patterns within the brine and halocline of deep hypersaline anoxic basins in the eastern Mediterranean Sea

Virginia Edgcomb · William Orsi · Chesley Leslin · Slava S. Epstein · John Bunge · Sunok Jeon · Michail M. Yakimov · Anke Behnke · Thorsten Stoeck

Received: 8 September 2008 / Accepted: 27 October 2008 / Published online: 5 December 2008 © Springer 2008

Abstract Environmental factors restrict the distribution of microbial eukaryotes but the exact boundaries for eukaryotic life are not known. Here, we examine protistan communities at the extremes of salinity and osmotic pressure, and report rich assemblages inhabiting Bannock and Discovery, two deep-sea superhaline anoxic basins in the Mediterranean. Using a rRNA-based approach, we detected 1,538 protistan rRNA gene sequences from water samples with total salinity ranging from 39 to 280 g/Kg, and obtained evidence that this DNA was endogenous to the extreme habitat sampled. Statistical analyses indicate that the discovered phylotypes represent only a fraction of species actually inhabiting both the brine and the brine-seawater interface, with as much as 82% of the actual richness missed by our survey. Jaccard indices (e.g., for a comparison of community membership) suggest that the brine/interface protistan communities are unique to Bannock and Discovery basins, and share little (0.8–2.8%) in species composition with overlying waters with typical marine salinity and oxygen tension. The protistan communities from the basins’ brine and brine/seawater interface appear to be particularly enriched with dinoflagellates, ciliates and other alveolates, as well as fungi, and are conspicuously poor in stramenopiles. The uniqueness and diversity of brine and brine-interface protistan communities make them promising targets for protistan discovery.

Keywords Anoxic · Brine · Community structure · Deep-sea · DHAB · Hypersaline · Molecular diversity · Protists

Non-standard abbreviations

DHAB Deep hypersaline anoxic basin
UMA Uncultured marine alveolate clade

Introduction

Explorations of biological diversity under some of the most extreme conditions present in Earth’s biosphere will likely lead to the discovery of microorganisms with novel structures and metabolic/physiological capabilities, and will increase our understanding of the extent and characteristics of life on Earth. Deep hypersaline anoxic basins (DHABs) in the Mediterranean and Red Sea have provided exciting new insights into novel microbial diversity and have already extended our knowledge of the environmental factors that define the limits of life (Eder et al. 2001; 1999,
2002; Sass et al. 2001; van der Wielen et al. 2005; van der Wielen and Heijs 2007; Yakimov et al. 2007b). Several DHABs (Tyro, Urania, Bannock, l’Atalante and Discovery), which originated from the dissolution of ancient subterranean Miocene salt deposits that became exposed to seawater in response to tectonic activity, have been discovered only relatively recently in the Eastern Mediterranean Sea (Camerlenghi 1990; de Lange et al. 1990). These basins are typically more than 3,000 m below sea level, and the high density of the brines within these basins results in limited mixing with overlying seawater. The combination of nearly saturated salt concentration and corresponding high density, high hydrostatic pressure, absence of light, anoxia, and a sharp chemocline make these basins some of the most extreme habitats on Earth.

Each of the Eastern Mediterranean DHABs is geochemically distinct (van der Wielen et al. 2005). For example, a striking difference between the Discovery basin and the other basins is the extremely high concentration of Mg$^{2+}$ (up to 5,000 mM compared to 300–650 mM in the other basins, and ca. 60 mM in regular seawater) and low concentration of Na$^+$ (ca. 70 mM compared to 3,500–4,700 mM in the other basins and ca. 500 mM in regular seawater). Such concentrations of divalent cations are the highest recorded to date in a natural environment (Wallmann et al. 1997, 2002) and are considered anathema to life and biogeochemical dead ends (Coleman 1993; Horowitz et al. 1972; Oren 1999; Siegel et al. 1979; Coleman 1993; Horowitz et al. 1972; Oren 1999; Siegel et al. 1979).

Yet, studies of prokaryotes in the Discovery, L’Atalante, Urania, and Bannock basins incorporating DAPI counts, analyses of 16S ribosomal RNA (16S rRNA) gene sequences, and measurements of sulfate reduction, methanogenesis, and heterotrophic activity have revealed metabolically active bacterial and archaeal communities in the Eastern Mediterranean DHABs (Daffonchio et al. 2006; Hallsworth et al. 2007; van der Wielen et al. 2005; van der Wielen and Heijs 2007; Yakimov et al. 2007a, b). The unique nature of these microbial communities in the Eastern Mediterranean deep-sea brines became evident by the observation that 16S rRNA gene sequences found in the brines are related to phylogenetic groups that are not found in normal seawater (van der Wielen et al. 2005). Furthermore, in a cluster analysis of archaeal and bacterial 16S rRNA gene data, van der Wielen et al. (2005) revealed distinctive structural differences between the microbial communities in the Urania, Bannock, L’Atalante and Discovery basin, with the latter being the most dissimilar. These data indicated that microbial community structures in the DHABs and geochemical conditions are directly linked. The authors assumed that the distinct geochemical conditions in each of the basins together with their physical separation from each other, as well as their existence for thousands of years resulted in the evolution of specific microbial communities in each of the basins. This assumption finds support in the discovery of novel microbial enzymes with unusual structures and biochemical characteristics that are adapted to function in these basins (Ferrer et al. 2005).

Thus far, the above hypothesis has gone untested for the eukaryote representatives of microbial communities (protoists). In chemosynthetic (anoxic) deep-sea systems a high abundance of bacteria usually supports a secondary food web (Taylor et al. 2001, 2006) that consists of a diverse assemblage of unicellular eukaryotes (Stoeck et al. 2003b, 2006). However, protoists in deep-sea brines, specifically in MgCl$_2$-rich environments, are entirely unexplored. Despite the fact that protoists play an integral role in the functioning of any ecosystem (Azam et al. 1983), only a limited number of studies, most of them microscopy-based, have focused on protistan diversity in hypersaline habitats in general. The emerging picture, as summarized in a recent review (Hauer and Rogerson 2005), is that moderately hypersaline systems sustain a rich and diverse community of mostly halotolerant microbial eukaryotes; however, once the salinity exceeds 30%, protoists are either missing or very rare (Elloumi et al. 2006; Oren 2000; Pedrós-Alió et al. 2000; Por 1980; Ramos-Cormenzana 1991) [for criticism see (Finlay 1990)]. In this study, we used statistical and phylogenetic analyses of eukaryote small subunit ribosomal RNA (SSU rRNA) genes to address the following questions: (1) Are there signs of microeukaryotic life in the super-extreme environments of the Bannock and Discovery basins? (2) If so, is this life represented by a few potentially highly specialized (endemic) taxa or is it as diverse and abundant as in other, non-hypersaline, marine environments? And (3), do the distinct biogeochemical characteristics of the two basins under study select for different protistan communities?

**Materials and methods**

Sample collection and preparation

Samples were collected in fall 2003 aboard the R/V *Urania* from (1) the Bannock brine/seawater interface (cast AB27SCI, 34°17.488’N, 20°00.692’E) with total salinity up to 246 g/kg (3.07 mol/kg of Na$^+$ and 470 mmol/kg of Mg$^{2+}$); (2) Bannock brine (cast AB29SCI, 34°17.397’N, 20°00.709’E) with total salinity of 280 g/kg in Bannock basin (3.50 mol/kg of Na$^+$ and 540 mmol/kg of Mg$^{2+}$); and (3) from the Discovery brine/seawater interface (cast DISC3R1, 35°16.611’N, 21°41.384’E) with total salinity up to 320 g/kg (45 mmol/kg of Na$^+$ and 3.30 mol/kg of Mg$^{2+}$). Two reference samples were collected from overlying seawater in the DHAB area (34°13.850’N, 21°28.400’E) with 39 g/kg of total salinity (420 mmol/kg Na$^+$.
of Na\(^+\) and 55 mmol/kg Mg\(^{2+}\)). For a detailed physico-chemical description of the sampling sites we refer to the available literature (Daffonchio et al. 2006; Hallsworth et al. 2007; van der Wielen et al. 2005). To prevent mixing and perturbation of the samples, we applied a specific high-precision sampling technique that was developed for this particular purpose (Daffonchio et al. 2006). This strategy employed the Modus-Scipack system as described in detail previously (Daffonchio et al. 2006) equipped with a CTD and a series of 10-l Niskin bottles as well as a pressure sensor for recording the pressure at which the Niskin bottles were closed. As the halocline samples in the Niskin bottles exhibited a steep salinity gradient, the contents of the Niskin bottles were carefully fractionated on board ship by slowly recovering only the fraction that matched the salinity of the upper or lower halocline (fractionated sampling). For details see Daffonchio et al. (2006). Water samples were filtered through 47 mm diameter, 0.22-µm pore-size Nuclepore filters (Millipore, Billerica, MA, USA). Collected material was re-suspended in 100 µL of TE buffer (pH 8.0) containing lysozyme (5 mg mL\(^{-1}\)), lysed by addition of 300 µL of lysis buffer QRL1 (Qiagen, Milan, Italy) and stored at \(-20^\circ\)C until processing. Total genomic DNA was extracted from filters using a Qiagen RNA/DNA Mini Kit (Qiagen, Milan, Italy). The extraction was carried out according to the manufacturer’s instructions. DNA was stored in isopropanol at \(-20^\circ\)C before precipitation. The quality of the DNA samples was examined by agarose electrophoresis and concentrations were determined using a NanoDrop\(^\text{TM}\) ND-1000 Spectrophotometer (Wilmington, DE, USA).

18S rRNA gene analysis

PCR amplification of the 18S rRNA gene was performed using a nested PCR. The first reaction used the universal eukaryotic primer set EukA 1–21 and EukB 1795–1772 (Medlin et al. 1988) and HotStar Taq DNA polymerase (Qiagen, Valencia, CA, USA). The PCR protocol consisted of an initial hot-start incubation of 15 min at 95°C followed by 30 identical amplification cycles (94°C for 1 min, 55°C for 30 s, 72°C for 2 min 30 s; and finally 72°C for 10 min). This was followed by amplification using the primer combinations Euk360FE and the universal primer 1492R (Medlin et al. 1988). Negative control reactions included bacterial (Escherichia coli) and archaeal (Sulfolobus solfataricus) DNA. For the second PCR amplification, 1 µL of the first PCR mixture was used to amplify the target fragments using the same PCR protocol as described above. Amplified DNA was checked for quality by agarose gel electrophoresis, and cloned into the vector TopoXL (Invitrogen). Separate clone libraries were generated for each of the original DNA samples. Plasmid DNA from four 96-well plates of clones from each library (ten 96 well plates in case of the Bannock library) was prepared using a MWG Biotech RoboPrep2500, and inserts were sequenced bi-directionally using an Applied Biosystems 3730XL capillary sequencer at the Josephine Bay Paul Center at the Marine Biological Laboratory (MBL), Woods Hole, MA. Processing of the data used PHRED, PHRAP (Ewing and Green 1998; Ewing et al. 1998) and a pipeline script to call bases from chromatograms, perform quality control procedures including checks for data consistency, data integrity, and data quality, to trim vector and low quality data and to assemble the sequences into contigs. The sequences were checked for chimeras using the Bellerophon Chimera Check program and the Check,Chimera utility (Ribosomal Database Project (Cole et al. 2003). After removal of short sequences (<800 bp) and putative chimeras, the remaining sequences were clustered using an all-to-all sequence comparison at percent similarity cutoffs ranging from 99% to 50% as described previously (Stoeck et al. 2007). We selected the OTUs at 98% for further phylogenetic analysis, and for each such OTU designated one, randomly chosen sequence as a representative. These representative sequences were subsequently aligned to 18S rRNA sequences available in the ARB package ((Ludwig et al. 2004), http://www.arb-home.de). The rRNA alignment was corrected manually according to secondary structure information and alignment uncertainties were omitted. Only unambiguously aligned positions were used to construct phylogenetic trees. These sequences have been deposited in Genbank under the accession numbers FJ000071-FJ000279.

Phylogenetic analyses

Partial alignments comprising sequences of defined taxonomic groups were exported from ARB for phylogenetic analyses to construct detailed subtrees. Evolutionary distance analyses under maximum likelihood criteria were carried out in PAUP × v4.0b8 (Swofford 2002), with all characters equally weighted and unordered. The TBR heuristic option was used to search tree space, running ten random additions with the MulTree option on. The evolutionary model that best fit each of our aligned data sets was chosen among 56 possible models using Modeltest (Posada and Crandall 2001). Support for the evolutionary distance analyses under maximum likelihood came from 1,000 bootstrap replicates using heuristic searches. Bayesian analyses were carried out using MrBayes v3.2.1 (Huelsenbeck and Ronquist 2001) with posterior probability support values calculated using four chains/two runs and running 10 million generations for each alignment. Trees were sampled every 1,000th generation. The first 25% of sampled trees were considered ‘burn-in’ trees and were
discarded. A 50% majority rule consensus of the remaining trees was used to calculate posterior probability values. The GTR + I+G evolutionary model was estimated using hLTr in MrModeltest v2.

Phylotype richness estimates

We estimated the total number of phylotypes (operational taxonomic units) in each sampled community using statistical procedures described previously (Hong et al. 2006; Jeon et al. 2006; Zuendorf et al. 2006). In brief, we fit seven candidate parametric abundance models to the observed phylotype frequency counts, selecting a preferred model based on the Pearson chi-square and Akaike Information Criterion (AIC) statistics, to obtain a final parametric richness estimate and associated standard error (SE). The candidate abundance distributions included the equal-species-sizes model, and the gamma, lognormal, Pareto, inverse Gaussian, and mixtures of two and of three exponential distributions. The corresponding mixed-Poisson distributions (e.g., the gamma-mixed Poisson) were fitted to the frequency count data (derived from the clone libraries) via maximum likelihood, using custom software on the Velocity Cluster in Cornell’s Center for Advanced Computing. The selected parametric model in each case is given in Table 1. We also computed various nonparametric statistics using the software SPADE (Chao and Shen 2003–2005), particularly the Abundance-based coverage estimators (of total richness) ACE and ACE1; either ACE or ACE1 was selected as the preferred nonparametric analysis based on an empirical cutoff value for the coefficient of variation of the frequency count data, as given in the literature (Chao and Shen 2003–2005). The selected nonparametric estimator in each case is given in Table 1.

Community similarity

The program package SPADE (Chao and Shen 2003–2005) was used to calculate the Jaccard index as a measure of similarity between two communities based on incidence ($J_{\text{incidence}}$), abundance ($J_{\text{abundance}}$), and abundance with adjustment for the effect of unseen shared phylotypes, in order to reduce bias due to undersampling ($J_{\text{adjusted}}$) (Chao et al. 2006).

DNA recovery experiments

We considered the possibility that brine DNA may be exogenous, settling from the overlying water with, e.g., dead cells and/or detrital material. This possibility was tested in a laboratory experiment designed to replicate as closely as possible the interface between the Bannock brine and the overlying seawater. We prepared artificial seawater with chemistry matching that of the Bannock brine (Daffonchio et al. 2006; van der Wielen et al. 2005). Fifteen liters of this brine solution was put into each of four 25 L carboys. Two duplicated experimental treatments included brine solution that was carefully over-layered with either 5 L of water containing live protists (5,000 cells/L of Paramecium), or 5 L of artificial seawater containing DNA from dead cells (5,000 cells/L of heat-killed Paramecium heated to 60°C for 30 min). In the absence of actual counts, we chose this cell abundance as representative of in situ conditions because heterotrophic protists are typically found in 1:1,000 ratio to their principal (prokaryotic) prey (Zubkov et al. 2007); the latter is about $5 \times 10^8$ mL$^{-1}$ in waters adjacent to Bannock’s brine (Daffonchio et al. 2006; van der Wielen et al. 2005). Continuous separation of the two layers was confirmed by regular measurements of salinity and density of the upper (3% salinity) and lower (the brine) layers. One carboy from each treatment was sacrificed immediately ($T_s$), and the two remaining carboys were sacrificed after 6 weeks of incubation ($T_i$), by collecting two 5 L samples of the brine layer by carefully dripping the brine solution from the bottom of the carboy into sterile flasks. Additionally, two 500 mL halocline samples at $T_i$. The brine samples

Table 1

<table>
<thead>
<tr>
<th>Site</th>
<th>BI (756)$^a$</th>
<th>DI (283)$^a$</th>
<th>BB(165)$^a$</th>
<th>SW(334)$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>38</td>
<td>39</td>
<td>38</td>
<td>31</td>
</tr>
<tr>
<td>Total richness, parametric</td>
<td>103.3 ± 48</td>
<td>141.8 ± 83.5</td>
<td>119 ± 64.1</td>
<td>55.5 ± 15.2</td>
</tr>
<tr>
<td>model</td>
<td>Two mixed exponential</td>
<td>Inverse Gaussian</td>
<td>Inverse Gaussian</td>
<td>Two mixed exponential</td>
</tr>
<tr>
<td>Total richness, nonparametric</td>
<td>104.8 ± 46.7</td>
<td>103.5 ± 42.9</td>
<td>89 ± 31.1</td>
<td>42.7 ± 9.1</td>
</tr>
<tr>
<td>Estimator</td>
<td>ACE-1</td>
<td>ACE-1</td>
<td>ACE-1</td>
<td>ACE-1</td>
</tr>
</tbody>
</table>

$^a$ Library size (number of clones sequenced)

---

* BI Bannock interface, DI Discovery interface, BB bannock brine, SW seawater reference community, $S$ estimate, $SE$ standard error

Estimates are presented at the 95% sequence similarity cutoff

---

* Springer
from each carboy were individually filtered on 0.22 μm Nucleopore filters, and DNA was extracted using the same protocol as was used during the original sample collection in the field. This was followed by PCR amplification of 18S rRNA genes using the same protocol as above.

Results and discussion

The principal purpose of this study is to answer the following three questions: (1) Does microeukaryotic life exist at the extremes of salinity and osmotic pressure found in the thalassic pure halite Bannock basin and the athalassic pure bischofite Discovery basin with chaotrophic conditions? (2) If yes, is this life diverse and abundant? (3) Do the geochemically distinct basins exhibit similar protistan communities or do they select for different protistan lineages? To address these questions, we explored the protistan communities in these two extreme environments using statistical and phylogenetic analyses of eukaryote rRNA genes retrieved from (1) the interface and (2) the brine body of Bannock basin, (3) the chaotrophic interface of the Discovery basin, and (4) a seawater reference sample collected immediately above the haloclines.

Does microeukaryotic life exist at the extremes of salinity and osmotic pressure in the Bannock and Discovery basins?

We successfully constructed four SSU rRNA gene clone libraries from the environmental genomic DNA extracts, which supports the notion that eukaryotes are present in all sampled environments, including the extremes of the Bannock brine. We considered a possibility that this DNA was not indigenous to the respective environments. Indeed, the high density of the haloclines of the DHABs may act as a barrier that traps organic material (including cell debris) originating from the upper water column (de Lange et al. 1990), and the high-salt concentration of the brine may help preserve DNA (Danovaro et al. 2005). To experimentally check the likelihood of an allochthonous origin for brine and interface DNA, we conducted a mesocosm experiment in which we simulated the deep-sea water, the halocline, and the brine of the Bannock Basin. The experiment was designed to investigate the likelihood of the (1) accumulation of PCR-amplifiable DNA from the overlying seawater in the halocline, and (2) permeation of such DNA into the brine and its persistence there. Live or heat-killed Paramecium cells were layered above the simulated halocline and incubated for 6 weeks. Figure 1a shows that at the start of the experiment positive amplification of eukaryotic genomic DNA was only achieved for the seawater overlayers. After the incubation period, no PCR-amplifiable (same protocol as used for the clone library construction and at the beginning of the mesocosm experiment) DNA was detected in the overlying seawater, the interface or the brine (Fig. 1b).

Even though our experimental setup only approximated the physico-chemical conditions at the Bannock interface, and tested a single organism, the data strongly suggest a low probability of PCR contamination from the accumulation of amplifiable extracellular DNA bound to detritus or from the accumulation and preservation of intact dead allochthonous planktonic organisms. This is likely due to the specific density of the brine (1.23 at 280 g/Kg salinity, the highest sampled in this study in the Bannock brine), as detrital particles as a rule have a higher density (Epstein 1995) and should simply sink to the bottom. Allochthonous DNA in the brine should therefore be limited to the transient amount, further diminished by its apparent degradation as indicated by the failure of our PCR reactions (see faint smear of non-amplifiable DNA in Fig. 1b).
assume that the relatively high temperatures in the deep waters of the basins under study (ca. 14°C) fuel a faster degradation of transient high-molecular weight free DNA compared to ‘normal’ cold deep-sea environments. Live cells on the other hand would not be able to penetrate even the uppermost layers of the interface as their typical specific density is 1.025 (Beaver and Crisman 1982), or less (Gates et al. 1982), and even hard-bodied benthic metazoa are lighter than brine (Epstein 1995). It therefore stands to reason that our samples from at least the denser brines were largely free of allochtonous cells and their DNA, and that at least the vast majority of detected 18S rRNA gene sequences represent the indigenous protistan communities. Despite this reasoning, we note that we cannot exclude the detection of non-indigenous or inactive organisms in our DNA-derived clone libraries. Evidence for the indigenous nature of these organisms can only come from direct isolation and culturing or RNA-targeted direct hybridization techniques (Massana et al. 2002; Stoeck et al. 2003a). Therefore, we consider the indigenous nature of the analyzed protistan communities merely as a working hypothesis.

How abundant and diverse is protistan life in Bannock and Discovery Basins?

We monitored phylotype accumulation (library saturation) during sequencing. As three of the libraries (Discovery interface, Bannock brine, seawater) were approaching saturation after analyzing ca. 150–200 protistan clones (Fig. 2) we did not continue sequencing these libraries. This was different for the Bannock interface library. Even after the analysis of more than 1,500 clones, the phylotype accumulation curve is still increasing steeply. Richness estimates revealed that the protistan community is quite diverse within each of the four environments sampled, with predicted numbers of OTUs that are higher in the interface and brine environments than in the overlying seawater. At 95% sequence similarity parametric estimators predict 103.3 ± 48 OTUs in the Bannock interface, 141.8 ± 83.5 OTUs in the Discovery interface, 119 ± 64.1 OTUs in the Bannock brine, and 55.5 ± 15.2 OTUs in the overlying seawater control. Nonparametric estimators were consistent with these predictions (Table 1). As estimates for selected cases are too inaccurate (high SE), we here refrain from discussing phylotype richness estimates at higher sequence similarities (e.g., 98 and 99%).

A general overview of all sequences (Fig. 3) shows a high diversity: with the exception of excavates and amoebzoa we found representatives of all major eukaryote taxonomic groups (opisthokonts, plants, cercozoa, alveolates, stramenopiles and discicristates). All domain-specific libraries are dominated by sequences that have
affiliations to single cell organisms and metazoa play only a negligible role (data not shown). The vast majority of all phylotypes retrieved in this study falls within the alveolates (74%), 12% of which are constituted by ciliates and 62% by dinoflagellates (including syndiniales) and uncultured marine alveolates. Also fungi are very abundant (17%)
while other taxonomic groups like stramenopile (2%) and euglenozoa (4%) flagellates as well as Polycystinea (2%) are relatively rare. This community composition is distinct from the typical picture of open ocean communities in the photic zone that are usually dominated by stramenopiles and pigmented picoplankton taxa (Courtway et al. 2007; Massana et al. 2004b; Moon-van der Staay et al. 2001; Not et al. 2007), but largely agrees with observed diversity and total diversity estimates of other molecular protistan diversity surveys in (anoxic) deep-sea environments (Edgcomb et al. 2002; Stoeck and Epstein 2003; Stoeck et al. 2006; Takishita et al. 2007). This further supports the notion that our interface and brine libraries are not primarily reflections of detritus from the water column but at least to a large extent reflect indigenous protistan assemblages.

Are protistan communities from the Bannock and Discovery basins unique?

The phylotypes and diverse taxonomic groups we recovered are not equally distributed in the four samples that we studied, and the number of phylotypes shared between the different samples is only marginal (Table 2). We calculated Jaccard indices as a measure of community similarity (Table 3). Regarding community membership (Jincidence), community structure (Jabundance) and community structure after adjustment for unseen phylotypes (Jadjusted), a very high dissimilarity (Jaccard values range between 1 and 0 with 1 indicating identical communities) was found between the Bannock and Discovery interfaces communities (0.015–0.02) (Table 3). A higher similarity was detected in all cases when comparing the seawater reference sample with the two DHABs interface communities. The brine community proved to be most dissimilar from all other communities in all cases. These data show increasing changes in community membership and structure from the deep-water brine through the halocline and into the brine. This is in agreement with previous studies on bacterial and archaeal communities in Eastern Mediterranean DHABs (Daffonchio et al. 2006; van der Wielen et al. 2005; Yakimov et al. 2007b).

Van der Wielen et al. (2005) observed most pronounced differences in bacterial community composition between the Discovery basin and the other DHABs in the Eastern Mediterranean (Urania, Bannock, L’Atalante). The authors attribute this finding to the high concentrations of MgCl$_2$ found in Discovery. Most known and described hypersaline organisms have difficulties coping with high MgCl$_2$ concentrations. MgCl$_2$-rich environments like bittern ponds have even been described as sterile (Javor 1989). The molecular mechanisms that enable protists to cope with extremely high concentrations of divalent cations in their natural habitat are literally unknown (Oren 2002). Different from sodium salts, magnesium salts behave as salting-in salts, increasing protein solubility and reducing protein stability. Clearly, high MgCl$_2$ concentrations seem to impose more of a restriction on microbial life in some of the DHABs due to its chaotropicity (Hallsworth et al. 2007). Thus, it is highly likely that life in the Discovery interface requires specific adaptations, resulting in microbial communities that are distinctly different from the ones in NaCl-rich environments like Bannock.

A striking difference in the microeukaryote community composition between the NaCl-rich Bannock and the MgCl$_2$-rich Discovery environments is the distribution of fungal phylotypes (Fig. 4). No fungi were detected in the Bannock brine and the Discovery interface, while they are numerous in the Bannock interface library (24 basidiomycetes and ten ascomycetes, one of which is shared with the reference seawater library, clone 572A06). Regarding their adaptation to high-salt environments, fungi, specifically ascomycete yeasts and basidiomycetes, reaffirmed themselves as one of the ecologically most successful eukaryotic lineages (Gunde-Cimerman et al. 2000, 2005; Kis-Papo et al. 2001). However, this situation is not quite as obvious regarding the adaptation of fungi to anoxia. Apart from a few exceptions (e.g., some chytridiomycetes) no strictly anaerobic fungi have been described and they have been thought to play only a minor role in ecosystem processes in anaerobic systems (Dighton 2003; Mansfield and Barlocher 1993). This contrasts with molecular diversity surveys in anaerobic aquatic environments, where fungal sequences sometimes account for a significant fraction of the clones (Dawson and Pace 2002; Edgcomb et al. 2002; Luo et al. 2005; Stoeck and Epstein 2003; Stoeck et al. 2006, 2007). Support for the survival and growth of fungi in anaerobic environments comes from

<table>
<thead>
<tr>
<th>Libraries</th>
<th>n shared phylotypes</th>
<th>Sequences in trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bannock interface-seawater</td>
<td>7</td>
<td>1041H09 (UMAII), 521B09 (UMAII), 521B08 (DF), 571G08 (DF), 521G06 (DF), 572A06 (fungi), 1041E04 (ciliates)</td>
</tr>
<tr>
<td>Discovery interface-Bannock brine-seawater</td>
<td>2</td>
<td>BB1D12 (ciliates), 212C04 (ciliates)</td>
</tr>
<tr>
<td>Discovery interface-Bannock interface-seawater</td>
<td>1</td>
<td>571A05 (DF)</td>
</tr>
</tbody>
</table>

DF dinoflagellates, UMAII uncultured marine alveolates groups II

 Springer

143


Table 2 Phylotypes shared by two or more libraries
Fig. 4 Minimum evolution phylogenetic tree of eukaryotic small-subunit rRNA gene sequences showing the position of fungal phylotypes in the Bannock interface (green), Discovery interface (brown), brine (red) and overlaying oxygenated deep seawater (blue). The tree was constructed using maximum likelihood criteria using a GTR + I + G DNA substitution model with the variable-site gamma distribution shape parameter \( \Gamma \) at 0.2179 and base frequencies and a rate matrix for the substitution model as suggested by Modeltest (Posada and Crandall 2001) based on 809 unambiguously aligned positions. The first numbers are distance bootstrap values over 50% from an analysis of 1,000 pseudoreplicates. The second numbers are posterior probabilities from 1,502 collected trees (2 runs, \( 10^6 \) generations, 1,000 sampled trees per run, burnin = \( 250 \)). Black circles at nodes indicate full support from both analyses. Numbers behind a DHAB sequence indicate a phylotype that was observed in both the upper and lower face of the halocline with color coding as described above.
Table 3  Jaccard indices (J) for a comparison of the eukaryote community membership (J_{incidence}) and the community structure under consideration of the unseen phylotypes (J_{abundance}) for different phylotype definitions (sequence similarity). Values for Jaccard indices range between 0 and 1, with the latter indicating identical communities. The analyses are based on a phylotype definition of 98% sequence similarity.

<table>
<thead>
<tr>
<th></th>
<th>Bannock interface</th>
<th>Discovery interface</th>
<th>Bannock brine</th>
<th>Seawater</th>
</tr>
</thead>
<tbody>
<tr>
<td>J_{incidence}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bannock interface</td>
<td>0.0159</td>
<td>0.0000</td>
<td>0.1000</td>
<td></td>
</tr>
<tr>
<td>Discovery interface</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bannock brine</td>
<td>0.0000</td>
<td>0.1910</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seawater</td>
<td>0.5892</td>
<td>0.2519</td>
<td>0.0904</td>
<td></td>
</tr>
<tr>
<td>J_{abundance}</td>
<td></td>
<td></td>
<td>Bannock brine</td>
<td>Seawater</td>
</tr>
<tr>
<td>Bannock interface</td>
<td>0.0203</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discovery interface</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bannock brine</td>
<td>0.0000</td>
<td>0.1910</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seawater</td>
<td>0.6843</td>
<td>0.2519</td>
<td>0.0904</td>
<td></td>
</tr>
</tbody>
</table>

reports of some ascomycetes (Dumitru et al. 2004; Sonderegger et al. 2004) and some basidiomycetes (Fell et al. 2001) that are capable of fermentation and anaerobic growth, some of which have even been isolated from anaerobic deep-sea environments (Nagahama et al. 2003). Thus, we assume that fungi indeed may play an ecologically important role in anoxic hypersaline environments. This assumption needs to be verified in future studies.

In our study, large alveolate sequence clusters are populated almost exclusively by sequences recovered from the Bannock interface and by only a few sequences from the Discovery interface and none from the Bannock brine or the reference seawater (Fig. 5). These clusters fall within the uncultured marine alveolate group II (UMAII). This lineage was originally discovered in the Antarctic deep-sea (López-García et al. 2001) and since then, UMAII sequences have been reported from a variety of different marine environments including diverse anaoxic deep-sea environments (Edgcomb et al. 2002; López-García et al. 2003; Stoeck et al. 2006; Takishita et al. 2007) but also from oxygenated open ocean or coastal waters (Massana et al. 2004a; Not et al. 2007; Romari and Vaulot 2004; Worden 2006). The closest BLASTn match of all DHAB UMAII sequences exhibiting only 90% sequence similarity was to two previously deposited sequences of the flagellated parasite Amoebophrya, namely Amoebophrya sp. ex. Gymnodinium instriatum (dinoflagellate) and Amoebophrya sp. ex. Alexandrium affine (dinoflagellate). However, due to the large phylogenetic distance of the environmental DHAB UMAII sequences to Amoebophrya, it is not possible to infer the life styles of these DHAB organisms.

One large clade (“uncultured marine alveolate Bannock clade”, Fig. 6) consisting of exclusively Bannock sequences sits as a possible sister group to UMAII, deeply branching with the “true” dinoflagellates and the UMAI. This novel group is also populated by sequences exclusively from the interface in Bannock basin. There is no support for its inclusion in UMA I, in fact different analyses place this clade within the dinoflagellates (with very weak support, unpublished data), rendering the branching position of this clade, and thus, the cellular identity and ecological role of the respective organisms, uncertain. Another environmental clade (“uncultured dinoflagellate clade”) is located “between” the “true” dinoflagellates and the UMAI, and includes sequences from uncultured organisms recovered from the nanoplanктон size fraction of the deep chlorophyll maximum in the Sargasso Sea (Armbrust et al., unpublished), and a supersulfidic anoxic fjord in Norway (Behnke et al. unpublished). The closest BLASTn hits of these environmental sequences to a described species are to the dinoflagellates Gyrodinium. Like numerous other dinoflagellates, Gyrodinium is characterized as a mixotroph feeding type, that is capable of phagotrophy, and thus, is able to thrive in light-independent environments (Stoecker 1999). Neither Bayesian nor Evolutionary Distance offers strong support to hold this group of sequences together or to hold them outside of either the “core” dinoflagellates or UMAI. Despite weak support within the dinoflagellate analysis for most groupings, Figure 6 illustrates a tremendous diversity of dinoflagellate-like taxa present in the Bannock interface environment, with some obviously specific and possibly endemic Bannock clades.

Analysis of ciliate sequences (Fig. 7) reveals that representatives of this alveolate phylum comprise part of all DHAB communities under study. However, while ciliates were relatively rare in the reference seawater library (4 phylotypes at 98% sequence similarity), the ciliated protists constitute a significant proportion of the brine library (10 phylotypes at 98% sequence similarity). We found representatives from six out of eleven described (ribo) classes, with some of them occurring exclusively in a single environment. For example, the strictly anaerobic Armophorea are unique to the chaotrophic Discovery sample, while heterotrichs occur exclusively in the Bannock brine. The vast majority of all ciliate phylotypes discovered in this study branch as closest neighbors of either environmental clones from other marine anoxic sampling sites such as, the Great Sippewisset Saltmarsh (CCW111 and CCA70, Stoeck and Epstein 2003), the Guaymas deep-sea basin (C1_E007, (Edgcomb et al. 2002), the Mid-Atlantic Ridge (AT6-4, (López-García et al. 2003), and a Danish fjord (M1_18C08, (Zaendorf et al. 2006). The closest related
Fig. 5 a and b Minimum evolution phylogenetic tree of eukaryotic small-subunit rRNA gene sequences showing the position of uncul-
tured marine alveolate group II (UMAII) phylotypes in the Bannock
interface (green), Discovery interface (brown), Bannock brine (red)
and overlaying oxygenated deep seawater (blue). The tree was
constructed using a Tamura-Nei substitution model with the variable-
site gamma distribution shape parameter ($G$) at 0.5958, the propor-
tion of invariable sites at 0 and base frequencies and a rate matrix for the
substitution model as suggested by Modeltest (Posada and Crandall
2001) based on 1,007 unambiguously aligned positions. The first
numbers are distance bootstrap values over 50% from an analysis of
1,000 pseudoreplicates. The second numbers are posterior probabil-
ities from 1,502 collected trees (2 runs, $10^6$ generations, 1,000
sampled trees per run, burnin = 250). Black circles at nodes indicate
full support from both analyses.

Uncultured marine alveolate group I

Uncultured marine alveolate group II


Springer
named species are ciliates with a facultative or strict anaerobic or microaerophilic life style, such as, *Trimyema*, *Strombidium*, *Metopus*, and *Peritromus*. A highly supported clade consisting exclusively of possibly endemic sequences recovered from the thalassic Bannock interface is highly divergent to *Cryptocaryon irritans*, an enigmatic parasite with a weak assignment to the class Prostomatea (Wright and Colorni 2002). This Bannock clade shares a common ancestor with another environmental sequence IN2411 retrieved from an anoxic deep-sea sediment sample (López-García et al. 2003) and cannot be assigned to one of the 11 described classes, suggesting a possible novel ciliate candidate-class. This finding once more (compare (Behnke et al. 2006; Stoeck et al. 2003b, 2006; Zuendorf et al. 2006)) emphasizes that we are still far from the discovery of all higher taxonomic groups in the domain Eukarya, even in groups of organisms that have been studied for more than two centuries (Ehrenberg 1838; Foissner 2008; Kahl 1930–
Minimum evolution phylogenetic tree of eukaryotic small-subunit rRNA gene sequences showing the position of uncultured marine alveolates, dinoflagellates and Syndiniales phylotypes in the Bannock interface (green), Discovery interface (brown), Bannock brine (red) and overlaying oxygenated deep seawater (blue). The tree was constructed under Tamura-Nei substitution model with the variable-site gamma distribution shape parameter ($G$) at 0.6935, the proportion of invariable sites at 0 and base frequencies and a rate matrix for the substitution model as suggested by Modeltest (Posada and Crandall 2001) based on 862 unambiguously aligned positions. The first numbers are distance bootstrap values over 50% from an analysis of 1,000 pseudoreplicates. The second numbers are posterior probabilities from 1,502 collected trees (2 runs, $10^6$ generations, 1,000 sampled trees per run, burnin = 250). Black circles at nodes indicate full support from both analyses. Numbers behind a DHAB sequence indicate a phylotype that was observed in both the upper and lower face of the halocline with color coding as described above.

Fig. 6 Minimum evolution phylogenetic tree of eukaryotic small-subunit rRNA gene sequences showing the position of uncultured marine alveolates, dinoflagellates and Syndiniales phylotypes in the Bannock interface (green), Discovery interface (brown), Bannock brine (red) and overlaying oxygenated deep seawater (blue). The tree was constructed under Tamura-Nei substitution model with the variable-site gamma distribution shape parameter ($G$) at 0.6935, the proportion of invariable sites at 0 and base frequencies and a rate matrix for the substitution model as suggested by Modeltest (Posada and Crandall 2001) based on 862 unambiguously aligned positions. The first numbers are distance bootstrap values over 50% from an analysis of 1,000 pseudoreplicates. The second numbers are posterior probabilities from 1,502 collected trees (2 runs, $10^6$ generations, 1,000 sampled trees per run, burnin = 250). Black circles at nodes indicate full support from both analyses. Numbers behind a DHAB sequence indicate a phylotype that was observed in both the upper and lower face of the halocline with color coding as described above.
Fig. 7 Minimum evolution phylogenetic tree of eukaryotic small-subunit rRNA gene sequences showing the position of ciliate phylotypes in the Bannock interface (green), Discovery interface (brown), Bannock brine (red) and overlaying oxygenated deep-sea water (blue). The tree was constructed under maximum likelihood criteria by using a GTR + I + G DNA substitution model with the variable-site gamma distribution shape parameter (G) at 0.4344, the proportion of invariable sites at 0.0886 and base frequencies and a rate matrix for the substitution model as suggested by Modeltest (Posada and Crandall 2001) based on 986 unambiguously aligned positions. The first numbers are distance bootstrap values over 50% from an analysis of 1,000 pseudoreplicates. The second numbers are posterior probabilities from 1,502 collected trees (2 runs, 10⁶ generations, 1,000 sampled trees per run, burnin = 250). Black circles at nodes indicate full support from both analyses. Numbers behind a DHAB sequence indicate a phylotype that was observed in both the upper and lower face of the halocline with color coding as described above.
1935). Specifically, extreme environments that have been difficult to access with traditional cultivation- and microcopy-based techniques repeatedly emerge as a potential excellent source of novel taxa at high taxonomic levels.

Conclusions

The environmental conditions that define the limits of eukaryote life are still far from being established. Our studied revealed that even in one of the most extreme marine environments that has been considered anathema to life for a long time, a diverse assemblage of protists can be found. This assemblage seems to be supported by a high abundance of bacteria as inferred from the trophic mode of most of the organisms that are related to the 18S rRNA gene sequences retrieved from the Discovery and Bannock DHABs. Considering the high proportion of sequence clades that are unique to the environments under study and their high divergence to previously described 18S rRNA gene sequences, these basins are an excellent source for the discovery of novel organisms with physiological capabilities and survival strategies that are very likely still unknown to science. Our data suggest that the protistan communities in the brine and the brine/seawater interfaces of Bannock and Discovery basins are unique to these habitats and share little similarity in species composition with the overlying seawater. Thus, it seems likely, that as reported for bacteria, the distinct biogeochemical conditions prevailing in the different basins exert a specific selection pressure on protistan communities resulting in the evolution of specialized, possibly geographically restricted, protistan assemblages.

Acknowledgments

This study was supported by grant STO414/2-4 of the Deutsche Forschungsgemeinschaft, the EuroDEEP program of the European Science Foundation under 06-EuroDEEP-FF-004 MIDDLE project and NSF-grant MCB-0348341 VE wishes to acknowledge Dr. Hilary Morrison and Rich Fox of the Marine Biological Laboratory, Woods Hole for the development and adaptation of their pipeline scripts for processing the sequence data for this study. We thank Linda Woodard for overseeing the richness calculations. This research was conducted using the resources of the Cornell University Center for Advanced Computing, which receives funding from Cornell University, New York State, the National Science Foundation, and other leading public agencies, foundations, and corporations. We thank the captain and the crew of RV Uranus for their expert handling of our casts and equipment and for highly productive oceanographic cruises.

References


Springer
Molecular and morphological evidence for a sister group relationship of the classes Armophorea and Litostomatea (Ciliophora, Intramacronucleata, Lamellicorticata infraphyl. nov.), with an account on basal litostomeans

Peter Vďačnýa,b,*, William Orsic, Wilhelm Foissnerd

a Department of Organismal Biology, Faculty of Natural Sciences, University of Salzburg, Hellbrunnerstrasse 34, A-5020 Salzburg, Austria
b Department of Zoology, Faculty of Natural Sciences, Comenius University, Mlynská dolina B-1, SK-84215 Bratislava, Slovak Republic
c Department of Biology, 305 Magar Life Sciences, Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA

Received 22 April 2010; received in revised form 7 July 2010; accepted 7 July 2010

Abstract

Based solely on the localization of the cytostome, Cavalier-Smith (2004) divided the ciliate subphylum Intramacronucleata into three infraphyla: the Spirotrichia, including Armophorea and Spirotrichea; the Rhabdophora, containing exclusively Litostomatea; and the Ventrata, comprising the remaining six intramacronucleate classes. This scheme is supported largely by 18S rRNA phylogenetic analyses presented here, except for the placement of the Armophorea. We argue that this group does not belong to the Spirotrichia but forms a lineage together with the Litostomatea because the molecular sister relationship of the Armophorea and Litostomatea is supported by two morphological and morphogenetic synapomorphies: (i) plate-like arranged postciliary microtubule ribbons, forming a layer right of and between the ciliary rows and (ii) a telokinetal stomatogenesis. Thus, we unite them into a new infraphylum, Lamellicorticata, which replaces Cavalier-Smith’s Rhabdophora. Further, our phylogenetic analyses consistently classify the most complex haptorian genus Dileptus basal to all other litostomeans, though morphological investigations suggest dileptids to be highly derived and possibly originating from a spathidiid ancestor. These discrepancies between molecular and morphological classifications have not as yet been investigated in detail. Thus, we propose an evolutionary scenario, explaining both the sister relationship of the Armophorea and Litostomatea, as well as the basal position of the morphologically complex dileptids.

© 2010 Elsevier GmbH. All rights reserved.

Keywords: Apicalization; Dileptids; Phylogeny; Postciliary microtubule ribbons; Stomatogenesis; Toxicysts

Introduction

The phylum Ciliophora contains two subphyla, the Postciliodesmatophora and the Intramacronucleata. The relationships among the nine intramacronucleate classes are largely unresolved and only two, Prostomatea and Oligohymenophorea, are consistently placed together (for review, see Lynn 2008). Based solely on the localization of the cytostome, Cavalier-Smith (2004) proposed the grouping of the nine intramacronucleate classes into three infraphyla (Fig. 3): the Spirotrichia, including Armophorea and Spirotrichea; the Rhabdophora, containing exclusively Litostomatea; and the Ventrata, comprising the remaining six classes (Phyllopharyngea, Nassophorea, Colpodea, Plagiopylea, Oligohymenophorea, and Prostomatea). However,
Classification of Armophorea and Litostomatea based on light microscopical features (Bütschli 1889), on ultrastructural data (Small and Lynn 1985), on localization of the cytostome (Cavalier-Smith 2004), and on a combination of morphological and molecular data (Lynn 2008; present study).

<table>
<thead>
<tr>
<th>Category</th>
<th>Presence</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bütschlian and Kahlian system</td>
<td>1</td>
<td>Morphology and arrangement of somatic fibrillar system</td>
</tr>
<tr>
<td>Small and Lynn (1985)</td>
<td>2</td>
<td>Morphology and arrangement of somatic fibrillar system</td>
</tr>
<tr>
<td>Cavalier-Smith (2004)</td>
<td>3</td>
<td>Morphology and arrangement of somatic fibrillar system</td>
</tr>
<tr>
<td>Lynn (2008)</td>
<td>4</td>
<td>Morphology and arrangement of somatic fibrillar system</td>
</tr>
<tr>
<td>Present study</td>
<td>5</td>
<td>Morphology and arrangement of somatic fibrillar system</td>
</tr>
</tbody>
</table>

Figs 1-5. Classification of Armophorea and Litostomatea based on light microscopical features (Bütschli 1889), on ultrastructural data (Small and Lynn 1985), on localization of the cytostome (Cavalier-Smith 2004), and on a combination of morphological and molecular data (Lynn 2008; present study).

This classification was widely ignored by most ciliatologists, for instance, it was not mentioned in the recent monograph of Lynn (2008). Based on 18S rRNA gene phylogeny and comparative analyses of ontogenesis and arrangement of the somatic fibrillar system, we slightly modify Cavalier-Smith’s framework and argue that the classes Armophorea and Litostomatea should be united as a novel infraphylum.

Morphologists never hypothesized a close relationship of the Armophorea (e.g. Metopus) and Litostomatea (e.g. Dileptus or Balantidium) because they appear very different. The former are anaerobic bacterivores with somatic dikinetids and a complex oral ciliature (paroral membrane and adoral membranelles), while the latter are aerobic predators (haptorid litostomeans) or anaerobic endosymbionts (trichostomatid litostomeans) with somatic monokinetids and a comparatively simple oral ciliature (oral dikinetids and/or oralized somatic monokinetids). Thus, it was a great surprise when early molecular phylogenetic studies suggested a sister relationship of the armophoreans and litostomeans, though with low bootstrap support ranging from 53% to 72% (Embley and Finlay 1994; Hammerschmidt et al. 1996; Hirt et al. 1995).

The monophyly of these two groups was not rejected when further species from all main ciliate lineages were added in the phylogenetic analyses, but the support usually remained poor (e.g. Gong et al. 2009; Strüder-Kypke et al. 2006). Nobody commented in detail on these remarkable results.
Only Foissner and Agatha (1997) found some common ontogenetical features for these two groups, but they did not argue for a sister relationship due to the prevailing morphological and ecological dissimilarities.

In the past, the Armophorea were classified as a subgroup of the Heterotricha, which were assigned to the Spirotricha due to the prominent adoral zone of membranelles (Fig. 1; Bütschli 1889; Small and Lynn 1985). However, based on morphological characters (somatic kineties with postciliary microtubules and macronucleus divided by extramacronuclear microtubules) and gene sequences, the heterotrichs were separated from the Spirotricha and were found as a sister group of the Karyorelictea (Figs 3–5; Hammerschmidt et al. 1996; Hirt et al. 1995; Lynn 2008). Further, these analyses showed that the armophores do not belong to the heterotrichs, but could form a monophyletic group with the litostomateans.

The Haptoria were assigned, together with the Prostomatea, to the Holotricha because of the completely ciliated body and simple oral apparatus (Fig. 1). Later, both were integrated into the Rhabdophora due to the transverse microtubule ribbons longitudinally lining the wall of the oral basket (Fig. 2; Small and Lynn 1985). However, the inclusion of the Prostomatea was caused by a misinterpretation of the fibrillar associates of the oral dikinetids (Lynn 2008). Further, neither molecular phylogenies nor ontogenetical data support a close relationship of the Haptoria and Prostomatea (Figs 3–5; Bardele 1999; Foissner 1996; Lynn 2008; Strüder-Kypke et al. 2006).

Small and Lynn (1981) recognized that the Haptoria and the Trichostomatia share a unique ultrastructural pattern of the somatic kineties which are single basal bodies bearing a convergent postciliary microtubule ribbon, a short kinetodesmal fibre, and two transverse microtubule ribbons. Consequently, Small and Lynn (1981) united the Haptoria and Trichostomatia into the class Litostomatea whose monophyly is strongly supported by the molecular phylogenies of the 18S rDNA gene (e.g. Lynn 2008; Strüder-Kypke et al. 2006) and by three additional synapomorphies: (i) the cytopyharynx is of a rhobdos type, i.e., it is lined by transverse ribbons (see Foissner and Foissner 1985, 1988 for haptorians and Grain 1966a,b for trichostomatians), whereas by postciliary ribbons in all other ciliate classes, including Prostomatea; (ii) the ciliate of at least one somatic kinety is differentiated to clavate cilia, forming a dorsal brush in the haptorians and a "clavate field" in the trichostomatians (Foissner 1996); (iii) the micronucleus conspicuously increases in size during the first maturation division and the conjugation mode is heteropolar, except for the pleurostomatid haptorians in which it is homopolar (Raikov 1972; Vďačný and Foissner 2008; Xu and Foissner 2004).

The molecular phylogenies of the haptorian ciliates (Gao et al. 2008; Strüder-Kypke et al. 2006) are conflicting with morphology-based evolutionary scenarios and classifications (Foissner 1984; Foissner and Foissner 1988; Lipscomb and Riordan 1990; Vďačný and Foissner 2008, 2009; Xu and Foissner 2005). Specifically, the morphologically most complex genus, Dileptus, branches off at the base of the molecular trees (Gao et al. 2008; Strüder-Kypke et al. 2006), while morphological traits (complex oral ciliature, hybrid circumoral kinety, transiently formed spatulid and polar ciliary patterns during ontogenesis and conjugation) indicate dileptids as highly derived and possibly originating from a spatulid ancestor (Foissner 1984; Vďačný and Foissner 2008, 2009; Xu and Foissner 2005). All these discrepancies between molecular and morphological classifications have as yet not been investigated in detail. Thus, we propose an evolutionary scenario, explaining both the sister relationship of the Armophorea and Litostomatea, as well as the basal position of the morphologically complex dileptids.

Material and Methods

Metopex es (Müller, 1776) Lauterborn, 1916 was found in a puddle from the surroundings of the city of Salzburg, Austria. *Enchelys polynucleata* (Foissner, 1984) Foissner et al., 2002 was collected from the upper soil layer (0–2 cm) of a meadow near Salzburg (Schaming near Eugendorf). Species were identified using live observation and protargol impregnation technique (Foissner 1991). *Enchelys polynucleata* was processed for transmission electron microscopy following the procedure of Foissner (1991). For explanation of morphological and ontogenetical terms, see Foissner (1996) and Lynn (2008).
Table 1. List of ciliate species with GenBank accession numbers of corresponding 18S rRNA gene sequences included in the phylogenetic analyses.

<table>
<thead>
<tr>
<th>Species name</th>
<th>GB number</th>
<th>Species name</th>
<th>GB number</th>
<th>Species name</th>
<th>GB number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amphileptus australis</td>
<td>EU242510</td>
<td>Dytrya proceria</td>
<td>DQ057347</td>
<td>Nyctotheroides parvus</td>
<td>AF145382</td>
</tr>
<tr>
<td>Anoplophrya marylandica</td>
<td>AY547546</td>
<td>Entodinium caudatum</td>
<td>U57765</td>
<td>Nyctotherus ovalis</td>
<td>AY007454</td>
</tr>
<tr>
<td>Arcuwastigmatida muscorum</td>
<td>DQ411859</td>
<td>Epeholota gennipara</td>
<td>DQ834370</td>
<td>Obertrumia georgiana</td>
<td>X65149</td>
</tr>
<tr>
<td>Balantium ciliolatum</td>
<td>AF019763</td>
<td>Epispathidium papilliferum</td>
<td>DQ411857</td>
<td>Ophyrocladula catenula</td>
<td>U17355</td>
</tr>
<tr>
<td>Blepharisma americanum</td>
<td>M97969</td>
<td>Eudiplodinum magnitisc</td>
<td>U57766</td>
<td>Ophyroclados parkynjei</td>
<td>U57768</td>
</tr>
<tr>
<td>Bryomatopus pseudochilodon</td>
<td>EU098873</td>
<td>Ejufoliculina ubligi</td>
<td>U57760</td>
<td>Orthodentia apophamatus</td>
<td>DQ232761</td>
</tr>
<tr>
<td>Bursaria truncatella</td>
<td>EU2126</td>
<td>Frontonia vernalis</td>
<td>U97110</td>
<td>Oxysorica granulifera</td>
<td>AF164122</td>
</tr>
<tr>
<td>Childodella uncinata</td>
<td>AF030281</td>
<td>Furgasia blochmanni</td>
<td>X65150</td>
<td>Paramecium calkinsi</td>
<td>AF003011</td>
</tr>
<tr>
<td>Chlamydodon exocellatus</td>
<td>AY331790</td>
<td>Helioxia erhardi</td>
<td>AY007445</td>
<td>Philia salinarum</td>
<td>EU242510</td>
</tr>
<tr>
<td>Chlamydodon triguerus</td>
<td>AY331794</td>
<td>Homaloxon vermiculare</td>
<td>L26447</td>
<td>Plagiopyla nasuta</td>
<td>Z29442</td>
</tr>
<tr>
<td>Coleps hirtus</td>
<td>AM82311</td>
<td>Isoschona sp</td>
<td>L264119</td>
<td>Pseudostramonius</td>
<td>X71140</td>
</tr>
<tr>
<td>Colpoda magnas</td>
<td>EU039896</td>
<td>Isotricha intestinalis</td>
<td>U57770</td>
<td>Pseudomicrothorax dubius</td>
<td>X65151</td>
</tr>
<tr>
<td>Copododium caudatum</td>
<td>EU264560</td>
<td>Lacymaria marina</td>
<td>DQ77746</td>
<td>Spathidium stanmerii</td>
<td>DQ411862</td>
</tr>
<tr>
<td>Cyrtophysis micocola</td>
<td>EU039899</td>
<td>Leptopharynx costatus</td>
<td>EU286811</td>
<td>Strombidium purpureum</td>
<td>U97112</td>
</tr>
<tr>
<td>Daystyricha ruminantium</td>
<td>AM18463</td>
<td>Litonotus peracynicus</td>
<td>EU242509</td>
<td>Tetrahymena cortissi</td>
<td>U17356</td>
</tr>
<tr>
<td>Didinium nasatum</td>
<td>U57771</td>
<td>Luxodasie striatus</td>
<td>U242458</td>
<td>Trachelothrix sp</td>
<td>L31520</td>
</tr>
<tr>
<td>Dilophus sp.</td>
<td>AF029764</td>
<td>Leptopharynx ostatum</td>
<td>DQ1900465</td>
<td>Trimemya compressum</td>
<td>Z39438</td>
</tr>
<tr>
<td>Diophysopus bistrix</td>
<td>EF486861</td>
<td>Metopas palaeformis</td>
<td>AY007450</td>
<td>Tritygystoma steini</td>
<td>X71134</td>
</tr>
<tr>
<td>Diplodinium dentatum</td>
<td>U57764</td>
<td>Nassula sp</td>
<td>EU286810</td>
<td>Zosterodasys transversus</td>
<td>EU286812</td>
</tr>
<tr>
<td>Discophysa collini</td>
<td>L26446</td>
<td>Notosoma parabryophryides</td>
<td>EU039903</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results and Discussion

Phylogenetic analyses

All phylogenetic analyses (BI, ML, MP) consistently placed the class Armophorea as sister to the class Litostomatea (Fig. 6). This clade was strongly supported by Bayesian interference with a posterior probability (PP) of 0.94, and moderately supported by the 75% ML and 67% MP bootstraps. Thus, based on two morphologic synapomorphies (unique arrangement of the somatic fibrillar system and ontogenetic mode), we unite the armophorans and litostomateans into a new infraphylum, Lamellicorticata. In all trees, the Lamellicorticata and the Spirotrichidae formed a super-clade that was, however, only weakly supported (0.65 PP, 56% ML, 52% MP). Within the large subphylum Infraciliata, another super-clade, comprising six classes (Colpodea, Oligohymenophorea, Prostomatea, Plagiopylae, Phyllopharyngea, and Nassophorea), was depicted with very strong (1.00 PP) to moderate (75% ML, 82% MP) support.

The monophyly Armophorea and Litostomatea were fully supported in all analyses (1.00 PP, 100% ML, 100% MP). However, the internal relationships of the class Litostomatea were rather poorly resolved (Fig. 6). The subclass Haptoria was paraphyletic in all analyses, consistent with Gao et al. (2008) and Strüder-Kypke et al. (2006). The genus Dileptus branched basal to all other litostomeans, justifying at least the ordinal rank suggested by Jankowski (1980). This node was strongly supported only in the BI analysis (0.94). The subclass Trichostomatia was classified as a monophyletic group with full posterior probability (1.00 PP) and very strong bootstrap support (99% ML, 99% MP). The trichostomatians branched rather deep within the subclass Haptoria where they clustered together with the free-living, aerobic haptorian Epispathidium papilliferum in the BI and ML trees, while they formed a polytomy pattern along with Arcuwastigmatida muscorum, Epispathidium papilliferum, and Spathidium stanmeri in the MP analysis. The trichostomatian Balantium ciliolatum was placed basal to the cluster formed by other vestibuliferans and entodiniomorphids. The support values for this node were very high in the BI (posterior probability 1.00) and ML (bootstrap 99%) analyses, while moderate in the MP tree (bootstrap 84%).

Morphological evidence and evolutionary scenario for a sister relationship of Armophorea and Litostomatea

Arrangement of somatic fibrillar system

The classic set of fibrillar associates of ciliate somatic kinetids includes a kinetodesmal fibre, a postciliary fibrisom,
Fig. 6. Phylogenetic tree of 59 18S rRNA gene sequences from the phylum Ciliophora, showing the sister relationship of the classes Armophorea and Litostomata. Three methods (Bayesian inference, maximum likelihood, and maximum parsimony) were used to construct trees, all resulting in the same topology. Posterior probabilities (PP) and bootstrap values for the maximum-likelihood (ML) and maximum-parsimony (MP) analyses are shown at nodes.
Figs 7–11. Comparison of light microscopic (Fig. 7, protargol impregnation) and electron microscopic (Figs 8–11) appearance of the somatic fibrillar system in armophoreans (Fig. 7, *Metopus es*, original; Fig. 9, *M. contortus*, micrograph kindly supplied by Finlay and Esteban) and haptorid litostomateans (Figs 8, 10, 11, *Enchelys polynucleata*, originals). The well-developed postciliary microtubule ribbons are arranged in a unique pattern: they form a single, plate-like layer right of and between the somatic kineties (Figs 7–11, arrows). The cortical granules, which impregnate strongly, follow the slightly oblique course of the postciliary microtubule ribbons (Fig. 7). ER – rough endoplasmic reticulum, G – cortical granules, KD – kinetodesmal fibre, M – mitochondria, PMT – postciliary microtubule ribbons, SK – somatic kineties (kinetosomes), TM – transverse microtubule ribbons.
the armophoreans (relative of the Lamellicorticata; (ii) during binary fission of
tians: (i) the body is oblong in the Spirotrichia, the closest
evolved convergently in the armophoreans and trichostoma-
ancestral body shape was oblong and the twisted body
feature of the Haptoria. However, the aerobic lifestyle is an
of dikinetids whose anterior basal body is, however, often
omokinetal except for the dorsal brush which is dikineti-
dal. This is certainly an apomorphic state for which the
perizonal stripe is homologous with the
ite which it is monotelokinetal
and the trichostomats in which it is cryptotelokinetal
nard, excluding dikinetids with ciliated basal
ent in the ventral kinetics in all
other ciliates. Secondly, the paroral membrane and
the circumoral kinety originate from kinetofragments that
detach from the somatic kinetics and unite into a single
hetorichricans and karyorelicteans are consistently placed basal to
litostomateans and armophoreans. In accordance with the
structural conservatism hypothesis (Lynn 2008), we con-
sider this unique pattern as a highly important phylogenetic
marker.

Stomatogenic mode
The Armophorea and Litostomea have a telokinetal and
purely somatic stomatogenesis. This is pleurotelokinetal in the
former and typically holotelokinetal in the latter, except for
the pleurostomatid haptorians in which it is monotelokinetal
and the trichostomats in which it is cryptotelokinetal
(Cameron and O’Donoghue 2001; Foissner 1996). There-
fore, Foissner and Agatha (1999) did not particularly argue
for a close relationship between armophoreans and litosto-
mateans. However, our detailed comparison revealed further
ontogenetic similarities, all considered here as apomorphies.
Firstly, the proliferation of basal bodies commences in the
dorsal or dorsolateral kinetics in the armophoreans and
litostomateans. This mode is unique to these two groups,
as basal body proliferation begins in the ventral kinetics in all
other ciliates. Secondly, the paroral membrane and
the circumoral kinety originate from kinetofragments that
detach from the somatic kinetics and unite into a single
kinety. The paroral membrane of all other ciliates develops
from an oral primordium or anarchic field. Thirdly, the
adoral membranelles and the preoral kinetics of the dileptids
are migrating kinetofragments, indicating the latter as
highly reduced membranelles (Foissner 1996; Foissner and
Agatha 1999; Vďačný and Foissner 2009; Xu and Foissner
2005). In all other ciliates with a prominent adoral zone
of membranelles, the new membranelles differentiate from
a long oral primordium, and thus they are not migrating
kinetofragments.

Body shape
Most armophoreans and some trichostomatids are
twisted, and thus spiralization seems to be plesiomorphic for
the Lamellicorticata. However, there are three indications that
the ancestral body shape was oblong and the twisted body
evolved convergently in the armophoreans and trichostoma-
tians: (i) the body is oblong in the Spirotrichia, the closest
relative of the Lamellicorticata; (ii) during binary fission of
the armophoreans (Metopus and Caenomorpha), the compli-
cated cell shape becomes oblong and all ciliary rows arrange
meridionally (Foissner and Agatha 1999; Martín-González
et al. 1987), indicating the ancestral body organization; and
(iii) in the molecular phylogenies, the oblong trichostoma-
tians (e.g. Balantidium) are placed basally, while twisted ones
(e.g. Ophryoscolex) appear to be derived (Strüder-Kypke
et al. 2006, 2007).

Structure of somatic ciliature
The complex ciliate cortex undoubtedly evolved from a
flagellate dikinetid (Lynn 2008; Orias 1976). Thus, somatic
dikinetids are considered as plesiomorphic feature for cili-
ates. The somatic ciliature of armophoreans is still composed
of dikinetids whose anterior basal body is, however, often
not ciliated. In litostomateans, the somatic ciliature is
monokinetal except for the dorsal brush which is dikineti-
dal. This is certainly an apomorphic state for which the
most parsimonious explanation is that the somatic dikinetids,
except for those of the dorsal brush, lost the anterior basal
body and became monokinets. The somatic ciliature of the
last common ancestor of armophoreans and litosto-
mateans was very likely condensed in the anterior body
portion. Our rationale is based on the assumption that
the armophorean perizonal stripe is homologous with the
litostomean dorsal brush. Both are a specialized field
in the anterior portion of the ciliary rows and are com-
posed from narrowly spaced dikinetids with ciliated basal
bodies.

Localization and structure of oral apparatus
The oral apparatus in most ciliates, including
the armophoreans and the basal litostomateans (dileptids), occurs
on the ventral side. Thus, the more or less apically or even
dorsally located oral opening of all other haptorians is consid-
ered as an apomorphy. The paroral membrane is very likely
homologous with the circumoral kinety and the adoral mem-
branelles are very likely homologous with the preoral kinetics
(see “Stomatogenic mode”). Based on the homology of the
armophorean and dileptic oral ciliature, we conclude that
the oral apparatus of their last common ancestor was com-
plex and included a paroral membrane and an adoral zone
of membranelles.

Lifestyle
The armophoreans are typically free-living anaerobes, the
haptorid litostomateans are free-living aerobes, and the tri-
chostomatid litostomateans are anaerobic endosymbionts in
a variety of vertebrates. Thus, more parsimonious would be
to infer that anaerobiosis is ancestral within the Lamellicor-
ticata and a transition to an aerobic way of life is a derived
feature of the Haptoria. However, the aerobic lifestyle is an
old plesiomorphy in ciliates, occurring also in the outgroup,
i.e., Spirotrichia. Further, the 18S rRNA gene phylogenies
indicate that anaerobic way of life evolved convergently in
the armophoreans and trichostomatid litostomateans because
trichostomateans branch rather deep within the Haptoria and
are grouped with the free-living, aerobic haptorian Epis-
Fig. 12. An evolutionary scenario for a sister relationship of Armophorea and Litostomatea. The Armophorea maintained the ancestral oral apparatus and the somatic dikinetids, but their body became spiralized and the anterior cilia condensation was transformed into a perizonal stripe. In the Litostomatea, the anterior condensation developed into a dorsal brush, the paroral membrane became a circumoral kinety, and the multi-rowed membranelles were reduced to single-rowed preoral kineties. Asterisks mark convergently evolved features of armophoreans and trichostomatians (from Xu and Foissner 2005).
pathidium papilliferum (Fig. 6; Strüder-Kypke et al. 2006, 2007).

Last common ancestor of Lamellicorticata

Based on the morphology and ontogeny of the extant armphoreans and litostomateans, we hypothesize that their last common ancestor was a bacterivorous ciliate living in aerobic aquatic environments and having the following apomorphies: (i) plate-like arranged postciliary microtubule ribbons, forming a layer right of and between the ciliary rows; (ii) telokinetal stomatogenesis commencing in the dorsal or dorsolateral kinetics, and with migrating oral kinetofragments; (iii) oblong and not twisted body; (iv) ventrally located oral apparatus composed of a dikeniditid paroral membrane and several multi-rowed adoral membranelles; and (v) dikeniditid somatic ciliature condensed in the anterior body portion (Fig. 12).

Last common ancestor of Armophorea

The armphoreans maintained the complex ancestral oral apparatus and the somatic dikeniditids, but their body became more or less spiralized and the anterior cilia condensational was transformed into a perizonal stripe. Finally, the armphoreans exploited oxygen-depleted habitats, developed hydrogenosomes from mitochondria, and gained endosymbiotic methanogenic bacteria (Fig. 12).

Last common ancestor of Litostomata

In the haptorid litostomates, (i) the anterior dikeniditid cilia condensation was partially reduced and developed into a dorsal brush, while all other somatic dikeniditids became monokinetids, loosing the anterior basal body which is often barren in the armphoreans; (ii) the paroral membrane elongated to a U-shaped pattern, forming the circumoral kinety which is still open anteriorly in the dileptids; and (iii) the multi-rowed adoral membranelles were transformed into single-rowed preoral kinetics as indicated by the localization (left body margin) and the origin (migrating kinetofragments) of these structures in metopids and dileptids (Foissner and Agatha 1999; Vd’ačny and Foissner 2009). Finally, the haptorians evolved toxicysts and became predators (Fig. 12). Based on the molecular phylogenies, we propose that the trichostomatid litostomates evolved from a microaerophilic haptorian having oralized somatic monokinetids. Finally, the trichostomatians became anaerobic endosymbions of vertebrates, simplifying the oral structures, loosing toxicysts, and transforming the mitochondria to hydrogenosomes (convergence to the Armophorea). For the review of trichostomatian morphological evolution, see Cameron and O’Donoghue (2004).

Are dileptids basal litostomateans?

Data on ontogenesis and conjugation (e.g., transiently formed spathidiiid and polar ciliary patterns) as well as some morphological traits (complex oral ciliature, hybrid circumoral kinety) indicate that dileptids are highly derived, possibly originating from a spathidiiid ancestor by developing a proboscis with a complex ciliature (Vd’ačny and Foissner 2008, 2009; Xu and Foissner 2005). In contrast, the molecular phylogenies of the small subunit rRNA gene (Fig. 6; Gao et al. 2008; Strüder-Kypke et al. 2006) and the present scenario (Fig. 12) suggest the dileptids as ancestral litostomateans because their oral apparatus still displays important plesiomorphic features, such as a ventral oral opening and preoral kinetics possibly homologous to adoral membranelles. There is a further morphological trait sustaining the basal position of the dileptids, viz., the occurrence of small kinetofragments (adesmakinetids) in the myriokaryonid and some spathidiiid haptorians (Xu and Foissner 2005). These kinetofragments resemble dileptid preoral kinetics in location (left of the oral bulge) and structure (short, oblique rows), and thus may be their vestiges (Fig. 12). Accordingly, all other haptorians originated from a Dileptus-like ancestor by reduction of a proboscis-like anterior body portion, causing apicalization (polarization) of the body and loss of the preoral kinetics. As concerns trichostomatians, their oral ciliature is considered to be secondarily simplified, comprising only oralized somatic kinetids (Lipscomb and Riordan 1990). Further, there are obvious trends towards simplification of trichostomatid oral structures: the vestibuliferids have a vestibulum with extension of densely packed somatic kinetics lining it, while the more derived entodiniomorphids have only short oral kinetofragments, the so-called polybrachykineties (Cameron and O’Donoghue 2004).

A similar kind of evolution occurred in the bacterivorous Oligohymenophorea, from which the rapacious Prostomatea evolved. In the prostomates, the paroral membrane became an apical circumoral kinety and the subapical adoral membranelles (polykineties) were strongly reduced becoming the brosse, as first recognized by Wilbert and Schmall (1976) and Foissner (1984), and later confirmed by the detailed studies of Bardele (1999). Based on the oral apparatus (paroral membrane and three to several adoral organelles), Pytorac et al. (1993) united prostomates with oligohymenophoreans and nassophoreans in the superclass Membranellophora. However, molecular trees (e.g., Gao et al. 2008; Kim et al. 2007; Strüder-Kypke et al. 2006) and somatic kinetid ultrastructure (slightly divergent postciliary ribbon, anteriorly directed ketododesmal fibre, and radially oriented transverse ribbon) recover only a close relationship of the prostomates and oligohymenophoreans (see Lynn 2008 for a review). Accordingly, we confine the Membranellophora to these two taxa (Fig. 5). On the other hand, Lynn (2008) synonymized pro parte the Membranellophora with the Heterotrichia, Spirotrichia, and Oligohymenophorea.

A new infraphylum Lamellicorticata and macrosystem of the phylum Ciliophora

In the mid-1990s, the phylum Ciliophora was divided into two subphyla – Postciliodesmatophora and Intramacronu-
cleata (for review see Lynn 2008). The former subphylum comprises two classes (Karyorelictea and Heterotrichacea), whereas the latter includes nine classes which Cavalier-Smith (2004) grouped according the localization of the cytosome into three infraphyla: the Spirotrichia, including the class Spirotrichea and Armophorea; the Rhabdophora, containing the single class Litostomatea; and the Ventrata, comprising all other intramacronucleate classes (Phylloparyngea, Nassophorea, Colpodea, Oligohymenophorea, Plagiopylea, and Prostomatea). We were surprised to find that our 18S rRNA gene phylogeny matches Cavalier-Smith’s (2004) higher classification of intramacronucleate ciliates rather well. Both differ only in the placement of the Armophorea. According to our molecular and morphological data, the class Armophorea does not belong to the infraphylum Spirotrichia, but forms a separate lineage together with the class Litostomatea, i.e., Rhabdophora sensu Cavalier-Smith (2004). Unfortunately, Rhabdophora is not eligible as a name for an infraphylum containing only Litostomatea, as it was originally proposed to unite Litostomatea and Prostomatea (Small and Lynn 1985). Therefore, we replace Cavalier-Smith’s Rhabdophora with a new infraphylum Lamellicorticata which includes two classes Litostomatea and Armophorea. Thus, the classification of the subphylum Intramacronucleata has been slightly changed as follows: Spirotrichia (excluding Armophorea), Lamellicorticata infraphyl. nov. (including Armophorea and Litostomatea), and Ventrata (as proposed by Cavalier-Smith 2004).

Why did some of the previous phylogenetic analyses fail to reveal the Lamellicorticata?

We analyzed several datasets with different taxon selections and alignments based on primary as well as secondary structure of the 18S rRNA molecule. When a high number of heterotrichian or spirotrichian sequences was included in our analyses, we observed that some armophoreans (i.e., Metopus spp.) cluster within the Spirotrichia, while others (i.e., Caenomorpha uniserialis) cluster within the Litostomatea (data not shown), an observation reported also by Miao et al. (2009a,b). Thus, the stability of the armophorean clade is somewhat dependent on the number of sequences from other groups included in the alignment. This explains why the sister relationship between Armophorea and Litostomatea may have been missed in some studies (e.g. Miao et al. 2009a,b), while indicated in others (Embley and Finlay 1994; Gong et al. 2009; Hammerschmidt et al. 1996; Hirt et al. 1995; Strüder-Kypke et al. 2006). The consistency of the three phylogenetic analyses used here with our comparative analyses of the ontogenesis and somatic fibrillar system of these two groups strongly suggests that they form a monophylum.

Taxonomic Summary

Lamellicorticata infraphyl. nov.

Diagnosis: Intramacronucleate ciliates with postciliary microtubules arranged in a single layer right of and between the ciliary rows. Somatic dikinetids typically very narrowly spaced in anterior body portion, forming a perizonal stripe (in armophoreans), dorsal brush (in haptorid litostomateans) or clavate field (in trichostomatid litostomateans). Stomato- genesis telokinetial, commencing in dorsal or dorsolateral somatic kineties, and with migrating oral kinetofragments. 

Etymology: Composite of the Latin noun lamella, the thematic vowel i, and the Latin noun cortex, referring to the lamellar arrangement of the postciliary microtubules in the cortex.


Acknowledgments

Financial support was provided by the Austrian Science Foundation (FWF projects P-19699-B17 and P-20360-B17). The technical assistance of Robert Schörghofer, Andreas Zankl, and Mag. Barbara Harl is greatly appreciated.

References


Foissner, W., 1984. Infraciliatur, Silberliniensystem und Biome trie einiger neuer und wenig bekannter terrestrischer, limnischer und mariner Ciliaten (Protozoa: Ciliophora) aus den Klassen
alignment editor and analysis program for Windows 95/98/NT.
Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Hig-
Vďačný, P., Foissner, W., 2008. Morphology, conjugation, and postconjugational reorganization of Dileptus tirijakowae n. sp.
Vďačný, P., Foissner, W., 2009. Ontogenesis of Dileptus terrenus and Pseudomonilicaryon brachyproboscis (Ciliophora, Hapto-
Xu, K., Foissner, W., 2004. Body, nuclear, and ciliary changes during conjugation of Protospathidium serpens (Ciliophora, Hapto-
Xu, K., Foissner, W., 2005. Morphology, ontogenesis and encyst-
ment of a soil ciliate (Ciliophora, Haptoidea), Arcuospathidium cultriforme (Penard, 1922), with models for the formation of
the oral bulge, the ciliary patterns, and the evolution of the
Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities

Thorsten Stoeck1, Anke Behnke1, Richard Christen2, Linda Amaral-Zettler3, Maria J Rodriguez-Mora4, Andrei Chistoserdov4, William Orsi5 and Virginia P Edgcomb*6

Address: 1Department of Ecology, University of Kaiserslautern, Kaiserslautern, Germany, 2Université de Nice et CNRS UMR 6543, Laboratoire de Biologie Virtuelle, Center de Biochimie, Parc Valrose, F06108 Nice, France, 3Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, USA, 4University of Louisiana at Lafayette, Lafayette, LA, USA, 5Northeastern University, Boston, MA, USA and 6Woods Hole Oceanographic Institution, Woods Hole, MA, USA

Email: Thorsten Stoeck - stoeck@rhrk.uni-kl.de; Anke Behnke - behnke@rhrk.uni-kl.de; Richard Christen - Richard.CHISTEN@unice.fr; Linda Amaral-Zettler - amaral@mbl.edu; Maria J Rodriguez-Mora - mjr9766@louisiana.edu; Andrei Chistoserdov - ayc6160@louisiana.edu; William Orsi - William.orsi@gmail.com; Virginia P Edgcomb* - vedgcomb@whoi.edu

* Corresponding author

Abstract

Background: Recent advances in sequencing strategies make possible unprecedented depth and scale of sampling for molecular detection of microbial diversity. Two major paradigm-shifting discoveries include the detection of bacterial diversity that is one to two orders of magnitude greater than previous estimates, and the discovery of an exciting ‘rare biosphere’ of molecular signatures ('species') of poorly understood ecological significance. We applied a high-throughput parallel tag sequencing (454 sequencing) protocol adopted for eukaryotes to investigate protistan community complexity in two contrasting anoxic marine ecosystems (Framvaren Fjord, Norway; Cariaco deep-sea basin, Venezuela). Both sampling sites have previously been scrutinized for protistan diversity by traditional clone library construction and Sanger sequencing. By comparing these clone library data with 454 amplicon library data, we assess the efficiency of high-throughput tag sequencing strategies. We here present a novel, highly conservative bioinformatic analysis pipeline for the processing of large tag sequence data sets.

Results: The analyses of ca. 250,000 sequence reads revealed that the number of detected Operational Taxonomic Units (OTUs) far exceeded previous richness estimates from the same sites based on clone libraries and Sanger sequencing. More than 90% of this diversity was represented by OTUs with less than 10 sequence tags. We detected a substantial number of taxonomic groups like Apusozoa, Chrysomonadophytes, Centrhopiloza, Eustigmatophytes, haptophytes, diatoms, chrysophytes, ichthyophytes, Oikomonads, Phaeothamniosphytes, and rhodophytes which remained undetected by previous clone library-based diversity surveys of the sampling sites. The most important innovations in our newly developed bioinformatics pipeline employ (i) BLASTN with query parameters adjusted for highly variable domains and a complete database of public ribosomal RNA (rRNA) gene sequences for taxonomic assignments of tags; (ii) a clustering of tags at k differences (Levenshtein distance) with a newly developed algorithm enabling very fast OTU
this paper we applied the 454 sequencing technique to Macaque gut microbiota. In the project described in has subsequently unveiled the richness and complexity of (see [16] for review, but see also [17]). Pyrosequencing separation lead to non-random distributions of microbes indications that environmental factors and geographic that these distinct population structures reflect the differ-

distinct hydrothermal vents. These data sets demonstrated V6 tag sequences obtained from two biogeochemically lysed nearly 700,000 bacterial and ca. 200,000 archaeal sphere diversity detected in this study (the so called dant populations accounting for most of the phylogenetic

nities are one to two orders of magnitude more complex the NE Pacific. The study revealed that bacterial commu-
nies in an ecological and biogeographical context [10-12]. Massively parallel tag sequencing (454 sequencing, pyro-
sequencing) is a promising remedy and offers a means to more extensively sample molecular diversity in microbial communities [13]. For example Sogin et al. [14] analyzed up to 23,000 tags per sample of the V6 hypervariable region of the bacterial SSU rDNA genes from deepwater masses of the North Atlantic and hydrothermal vents in the NE Pacific. The study revealed that bacterial commu-
nities are one to two orders of magnitude more complex than previously reported, with thousands of low abund-
ant populations accounting for most of the phylogenetic diversity detected in this study (the so called rare bi-
osphere). This was confirmed by Huber et al. [15] who ana-
alyzed nearly 700,000 bacterial and ca. 200,000 archaeal V6 tag sequences obtained from two biogeochemically distinct hydrothermal vents. These data sets demonstrated that these distinct population structures reflect the differ-
ent local biogeochemical regimes, corroborating previous indications that environmental factors and geographic separation lead to non-random distributions of microbes (see [16] for review, but see also [17]). Pyrosequencing has subsequently unveiled the richness and complexity of soil bacterial communities [18], human [19] and Macaque [20] gut microbiota. In the project described in this paper we applied the 454 sequencing technique to eukaryotes to analyze the complexity of microbial eukary-

otic communities in two environmentally contrasting anoxic basins (Cariaco and Framvoren).

The Cariaco Basin is the world’s largest truly marine anoxic body of water located on the northern continental shelf of Venezuela [21,22]. Primary production in Cari-
aco, microbial biomass, and midwater dark CO₂ fixation vary strongly with factors such as seasonal riverine inputs, seasonal upwelling intensity, lateral intrusions of water from the Caribbean Sea, and trade-wind intensity [22-24]. The basin exhibits pronounced vertical chemical gradients controlled by physical transport of oxygen downwards and reduced compounds upwards countered by biologi-
cal demands. Typically, oxygen concentrations decrease from saturation at the surface to 0 μM between 250 and 300 m. Deeper waters have remained anoxic and sulfidic down to the basin’s bottom at ca. 1,400 m over timescales of centuries to millennia [25]. Significant enrichments in abundance of bacteria, bacterial activity and protists are routinely observed in the redoxcline and in the sulfidic waters underlying the redoxcline [23,26,27]. The Fram-

varen Fjord located in southwest Norway shares the fea-
ture of a defined oxic/anoxic interface with the Cariaco Basin. Yet, this fjord varies in many physico-chemical parameters (see Table 1) from the latter. For example, while the Cariaco Basin is truly marine with a redoxcline below the photic zone and relatively low sulfide concen-
brations below the redoxcline, the oxic-anoxic boundary layer of the fjord is located at shallow depth (ca. 18 m) with high sulfide concentrations below the redoxcline and steep biogeochemical gradients down to the bottom waters (180 m). Sulfide levels in bottom waters are 25 times greater than those in the Black Sea [28]. Initial stud-
ies of these two sites ([10,29,30]; Edgcomb et al. unpub-
lished) based on clone-library construction and traditional Sanger sequencing indicate evidence for adap-
tation of protistan communities to differing environmen-
tal conditions along, O₂/H₂S gradients. In spite of tremendous efforts in these previous studies, the sequenc-
ing depth was still significantly less than predicted total diversity and one might argue that additional sequencing would reveal homogeneous communities along these gra-
dients. Massively parallel tag sequencing (in total, we ana-
lyzed 251,648 tag sequences obtained from the hypervariable V9 region of the SSU rRNA gene offers the opportunity to evaluate if the structuring of microbial communities observed in these two contrasting basins still holds true at significantly increased sequencing efforts, whether richness predictions based on clone library analyses are supported and how well severely undersampled clone libraries reflect the “true” protistan diversity at a specific locale.

Results

The number of high-quality eukaryotic reads we obtained from each sample ranged from 16,256 (FV3) to 38,280 (FV1). After dereplication (consolidating all sequences that are identical in primary structure into one OTU), the numbers of unique eukaryotic tags ranged from 3,765 (FV3) to 5,983 (CAR1). After exclusion of metazoan tags, we were left with numbers of unique tags ranging from 2,070 (CAR4) to 5,597 (CAR1), most of which could be assigned to protists and fungi (Table 1) for further analyses. The number of tags from non-eukaryotic domains was only marginal (0-0.02% of total tag reads, see Table 1) indicating the high domain-specificity of the primers used.

Sampling saturation

Despite substantial sequencing effort, the communities under study did not show saturation (Figure 1) in unique OTU richness. When clustering OTUs at one nucleotide difference, the number of OTUs detected decreased sharply, but still did not saturate. Only when clustering the tags at two, three, five and ten nucleotide differences (OTUs_xnt, where x is the number of nucleotide (nt) differences), did the sampling saturation profiles show a tendency of leveling off. The collapse of detected OTUs when comparing unique tags with OTUs based on two nucleotide differences (roughly 1.5% difference in primary structure), is remarkable: in the same sample (FV1) up to 6.3 times more unique OTUs were detected compared to OTUs_2 nt. In contrast, the number of detected OTUs varied noticeably less when comparing OTUs over a clustering range of three to ten nucleotides, indicating that most of the tag variation was within two nucleotide differences between tags. Interestingly, regardless of the initial number of unique tags that varied greatly among the eight samples, all samples showed similar numbers of OTUs when tags were clustered at two, three, five and ten nucleotide differences.
Sampling saturation of V9 tag libraries

Figure 1

Sampling saturation profiles of tag libraries generated for samples collected from anoxic waters of the Norwegian Framvaren Fjord (FV1-4) and the Caribbean Cariaco Basin (CAR1-4) at different levels of nucleotide differences for operational taxonomic units (OTUs). Only protistan and fungal tags were taken into account. Tags are clustered at k differences from k = 0 to 10 differences as described in pipeline 2 of the sequence data processing paragraph in the methods section. A difference can be an insertion or a mutation necessary to align the two sequences. At k differences, two tags having k or fewer differences are placed in the same cluster; if they have more than k differences, they are in two different clusters. Unique tags are tags clustered at 0 differences.
Rank abundance
In all eight samples, the frequency distribution of protistan tags within unique OTUs was very uneven (Figure 2): Only few populations were dominating the individual data sets, while the majority of OTUs contained less than ten sequences. The combined frequencies of these low-abundance unique phylotypes in the individual amplicon libraries accounted for 0.14%-0.03% of total protistan tags analyzed in each sample and thus, were considered as rare. Regardless of the sampling effort, this proportion of rare taxa remained similar for all samples (for example 96% rare populations in sample CAR4 and 95% in sample CAR1).

Community comparisons
An UPGMA linkage distance analysis of unique OTUs based on Incidence (Figure 3) identified two distinct clusters one of which consisted of all FV samples, another of samples CAR4, CAR3 and CAR2, all from below the interface. The deep-sea sample from the Cariaco interface (CAR1) was the most distinct of all CAR samples regarding protistan community membership with higher affinity to the other CAR samples rather than to the FV samples. In the Framvaren fjord, the two samples that were taken at different seasons from below the interface of the central basin were most similar to each other (FV2 and FV4), while the below-interface sample from the upper basin (FV3) - 3 km apart from the central basin station - was less similar to both FV2 and FV4. Neither samples CAR2 and CAR3, which were sampled from below the interface in the same season but at different locations, nor samples CAR2 and CAR4, which were sampled from below the interface at the same site but in different seasons clustered together. Instead, samples CAR3 and CAR4, were most similar in terms of community membership. These two samples were collected at two different seasons from below the interface at two different locations (Station B and Station A, respectively).

Protistan community structures
The vast majority of all unique tags could be confidently assigned to a defined taxonomic rank, at least at class-level (Figures 4, 5, 6, 7, 8 and 9). Between 3.5% (FV4) and 21% (CAR3) of unique tags could not be reliably assigned a taxonomic rank because sequence similarity to their best BLAST match was too low (<80%, see methods section). We attribute this to mainly two reasons. First, numerous sequences of described species that are deposited in GenBank lack the nucleotide positions that correspond to the V9 region of the SSU rRNA gene (ca. 1,620-1,790) in part or completely; second these unassignable tags correspond to as yet unsequenced taxonomic groups. Unfortunately it is currently not possible to discriminate between these two categories, rendering any interpretation of the proportion of unassignable tags speculative. We do not consider chimeras as a major contributor to unassignable tags because, as our protocol amplifies short DNA sequences with a negligible likelihood of chimera formation [31]. The proportion of unique tags that had only environmental sequences as the nearest match, without a sequence of a named species falling into the minimum 80% sequence-similarity boundary was large (up to 21% for sample FV4), reflecting the paucity of cultured representatives and the taxonomic annotation of environmental sequence data in public databases. In future studies, the implementation of specifically curated and annotated databases like KeyDNATools ([32] and http://www.pc-informatique.fr/php-fusion/news.php) will be beneficial for the taxonomic assignment of tags that have a good BLASTN match to environmental sequences but lack a species-match within a defined sequence similarity threshold. A tremendous number of higher taxonomic groups represented by tags that accounted for at least 1% of the overall number of protistan tags were discovered in each sample. For example, in sample FV3 we detected 17 such groups. When tag sequences that account for <1% of all protistan tags were taken into account (category ‘others’ in Figure 4), this number was even larger. Such groups included: Euglenozoa, Rhodophyta, Jakobida, Ichthyosporea, Telonema, Cryothecomonas and Apusozoa. In sum, all major eukaryotic lineages have been detected in each individual sample. However, the proportion of the different taxonomic groups in the individual samples varied considerably. Generally, all samples were dominated by alveolate OTUs, accounting for up to 64% of all unique protistan tags in an individual sample (FV1). In all CAR samples, Dinozoa contributed to the largest proportion of alveolate OTUs, followed by Ciliophora. The latter were noticeably less abundant in the CAR1 and CAR3 samples. In the Framvaren samples, Ciliophora comprised a decidedly larger proportion of the Alveolata, in FV4 and FV2 reaching or even exceeding the percentage of Dinozoa, respectively (Figure 4).

Interestingly, in nearly all dominant phyla occurring at both locales, the taxon composition in the non-sulfidic anoxic water depth was distinctly different from the sulfidic waters below the redoxcline (Figures 4a-f). For example, in the Cariaco Basin, unique tags assigned to Polycystinea accounted for 31% of all protistan tags in the interface (CAR1), while below the interface this number ranged between 3% (CAR3) and 7% (CAR4). However, generally, the genotype diversity in the sulfidic waters was decisively higher in a variety of taxon groups compared to the anoxic, non-sulfidic waters. Ciliophora tag proportion varied more than five-fold between CAR1 and CAR4, Cercozoa 18-fold between CAR1 and CAR2, Bacillariophyta seven-fold. Ascomycota nearly 10-fold, Heterokontophyta and Chlorophyta ca. seven-fold between these samples (Figure 4). This, even though to a lesser extent, was
Figure 2
Rank abundance distribution of unique protistan OTUs. Protistan (including fungi) rank abundance distribution of unique operational taxonomic units (OTUs) obtained from four samples of the Caribbean Cariaco Basin (CAR1-4) and from four samples of the Norwegian Framvaren Fjord (FV1-4). For sampling sites information see Table 1. Curves were obtained when clustering tags at zero differences as described in pipeline 2 of the sequence data processing paragraph in the methods section. Subsequently tags were ordered according to decreasing rank (number of replicates present for each tag).
also true for the Framvaren fjord. Just to mention a few examples, the proportion of Dinozoa-assigned tags decreased from 42% in the interface (FV1) to 14% below the interface (FV2), ascomycota increased nearly three-fold, Cercozoa increased ca. four-fold and Haptophyta, Chlorophyta and Heterokontophyta nearly five-fold (Figure 5) were exclusively found in all Cariaco samples. For examples Blastodiniales and Noctilucales (Dinozoa, Figure 5) were exclusively found in all Cariaco samples. Rhizosoleniophycid Bacillariophyta (Figure 7) and Pseudoscourfieldiales (Chlorophyta, Figure 8) were noticeably higher in all Cariaco samples and Choricystis (Chlorophyta, Figure 8) genotypes were much more diverse in Framvaren.

Discussion

The application of the 454 sequencing technique to the investigation of protistan communities in two anoxic marine basins revealed three significant findings. First, even a sampling effort that was one to two orders of magnitude larger than that achieved by environmental clone library construction and Sanger sequencing, was not successful in retrieving all unique SSU rRNA gene sequences present in a single sample (Figure 1). Up to 5,600 unique tags could be identified in a 7-L water sample from the Cariaco basin without reaching saturation (sample CAR1). However, this is unlikely to reflect the true species richness, because (i) not all SSU rRNA gene copies within a species are necessarily identical [34,35], (ii) some of the observed tag variability may be due to extreme variability of the V9 region in specific taxonomic groups, and [36] even when minimizing the effect of sequencing and PCR errors using a systematic trimming procedure (see Methods section and [14]) the accuracy of the 454 pyrosequencing strategy (GS-technology) is 99.75% - 99.5% for small subunit rRNA genes [37]. Indeed, in sample CAR1 the number of OTUs drops from 5,600 to ca. 2,600 when phylotypes are clustered based on one nucleotide difference (accounting for ca. 0.8% sequence similarity). Thus, about half of the unique protistan tags retrieved from this sample are potentially afflicted with an error and/or represent the same taxon. The detected number of unique tags would likely represent an overestimation of taxon richness. On the other hand, clustering OTUs at ten nucleotide differences (OTUs_{10 nt} reflecting ca. 8% sequence similarity) resulted most likely in an underestimation because different taxa may be lumped together into the same OTU. Consequently, it is reasonable to assume that the true taxon richness is reflected in the range between 2,600 and 5,600 species.
OTUs<sub>1 nt</sub> (ca. 1,700 in sample CAR1) and OTUs<sub>5 nt</sub> (ca. 1,200 in CAR1).

Interestingly, even the number of detected OTUs<sub>10 nt</sub> exceeded previous parametric and non-parametric richness estimates from the same sites, based on clone-library derived OTUs called at 99% or 98% sequence similarity, respectively [10,38,39]. Explanations for this may be several fold: (i) even though the sample sizes obtained from previous Cariaco and Framvaren clone libraries were relatively large, the sample size may still have been too small to obtain adequate resolution of the complex communities. If so, this makes previous clone library-based richness estimates severe underestimations; (ii) the statistical error of previous richness estimates may be too large, which cannot be assessed due to a lack of good confidence intervals; [36] abundance-based richness estimates may not reflect the true community richness or relative species abundance in a sample but rather the PCR-amplicon richness. The reasoning for the latter is that in contrast to bacteria, the copy number of SSU rRNA genes varies widely among protists [8,40,41]. Thus, the relative amplicon copy number after PCR does not necessarily reflect the relative abundance of a specific taxon in a sample, rendering abundance-based species richness estimates highly erroneous. It is likely that these factors and probably other factors that we cannot account for at present resulted in severe richness underestimations. We hypothesize that the protistan richness in marine anoxic waters by far exceeds previous estimates, and that anaerobic protistan
communities are substantially more complex than previously reported. It will be interesting to further investigate how sequence divergence of a hypervariable SSU rRNA gene region translates into taxonomic entities. This will help interpreting the vast diversity of tags generated by massively parallel tag sequencing.

Most of the observed complexity was found in the low-abundance populations. Even when calling OTUs at five nucleotide differences, the proportion of rare OTUs (represented by less than 10 tags) ranges between 71% and 81% in FV samples and between 78% and 83% in CAR samples (data not shown), indicating that the high number of rare taxa is not an artifact based on high intra-species heterogeneity in the V9 region. This corroborates, to a somewhat lesser extent, the previous findings in the bacterial world [14,15,18]. The origin and meaning of this complexity is still unclear [42,43]. Actually, to date there is no evidence that this high frequency of low-abundant genotypes describes a true diversity. It could result from the amplification of detrital or extracellular DNA. On the other hand, it is reasonable to assume that a liter of water is only inhabited by a few individuals of a protist species that never meet in this volume and are therefore subjected to allopatric speciation. The result would be tremendous microheterogeneity that is reflected in these rare genotypes. One hypothesis suggests that these rare genotypes (if real) may represent a large genomic pool, which helps the protistan community to react to any biotic or abiotic changes [43]. In this seed-bank scenario, the species that are best adapted to prevailing environmental conditions would always be abundant in a community.

The second significant finding is the phylum-richness of protistan communities that is missed by the clone library/Sanger sequencing approach. Previous environmental protistan diversity surveys in the same sites of the Framvaren Fjord ([10] and Behnke et al. unpublished, accession numbers [DQ310187 to DQ310369 and EF526713 to EF527205]) did not retrieve any sequences assigned to Apusozoa, Chrysomerophytes, Centroheliozoa, Eustigmatophytes, Ichthyosporea, Oikomonads, Phaeothamniophytes, and rhodophytes, all of which have been recovered with the massively parallel tag sequencing approach. Similarly, a vast array of higher taxon ranks detected in this tag-sequencing project could not be detected with an extensive clone library sampling in Cariaco ([26,30] Edgcomb et al. in preparation). Inter-

![Diagram](https://example.com/diagram.png)

**Figure 5** Taxonomic distribution of V9 tags assigned to Dinozoa. The data that served as a basis for the taxonomic bar chart are available as supplemental material (Table S4 in Additional file 5).
Interestingly, the tags that could be assigned to taxonomic groups not detected via clone libraries all account for <1% of the unique protistan tags, explaining why they have been missed with the clone library approach [26,30]. Regarding taxonomic groups that were represented by large relative abundances of tags (e.g. alveolates and stramenopiles), the 454 data sets corroborate well with clone library-obtained data. Evidence of and tentative explanations for the dominance of these taxonomic groups in anoxic marine systems have already been intensively discussed elsewhere (e.g. [30,44,45]).

The broad taxonomic representation of 454 tags nicely demonstrates the efficiency of the primers used to target the hypervariable V9 region of eukaryote SSU rRNA genes. However, up to 50% of unique 454 tag sequences in our data sets were metazoa. This is a general problem also observed in SSU clone libraries (even though probably to a lesser extent) and not specific to 454 technology [46-48]. The consequence is that this large proportion of potential non-target tags has to be taken into account when designing protistan diversity studies using 454 technology. Either sequencing effort needs to be increased 1.5-fold to get the desired number of protistan tags, or group-specific 454 primers need to be applied subsequently to focus on selected protistan groups.

Our findings also reveal that higher sampling efforts can be obtained in a cost- and time-efficient way by the application of pyrosequencing, which therefore paints a substantially more comprehensive picture of protistan communities. The degree of undersampling inherent in most published clone library-based studies may be so high that it is possible that they cannot be compared in a meaningful manner to other equivalent surveys of diversity. Getting a comprehensive picture of a microbial community is critical to addressing fundamental questions in protistan ecology on the basis of molecular diversity surveys. Such questions include for example, determining the true richness and evenness of microbial communities, which is important in defining microbial ecosystem dynamics [15], and determining the biogeographic distribution of specific taxonomic groups, the stability of protistan communities over time, as well as local patchiness of protists. All of these community attributes are cornerstones for understanding microbial diversity, ecology, and evolution [16,49,50].

Figure 6
Taxonomic distribution of V9 tags assigned to Ciliophora. The data that served as a basis for the taxonomic bar chart are available as supplemental material (Table S5 in Additional file 5).
Some of these subjects frame the third important finding of this study. The eight sites sampled differed markedly in community composition. Based on community membership, it appears that protistan communities from the supersulfidic Framvaren Fjord with an interface located in the photic zone are distinct from the ones of a less sulfidic anoxic deep-sea site. Similarly, anaerobic protistan communities exposed to hydrogen sulfide are distinct from those that thrive in sulfide-free oxygen-depleted habitats. Even though we cannot unequivocally identify H$_2$S as the single most important driving force shaping these protistan communities using this dataset, this observation is not unexpected: H$_2$S-detoxification requires specific adaptation that is not necessarily present in all facultative or strictly anaerobic protists [51,52]. For example, Atkins et al. [53] found a significant difference in the hydrogen sulfide tolerance of different hydrothermal vent species they isolated, including the closely related sister taxa Cafeteria and Caecitellus. Cafeteria strains isolated by these authors could tolerate up to 30 mM sulfide under anoxic conditions over the 24 hr course of their experiment, Rhynchomonas nasuta could tolerate up to 5 mM sulfide, and Caecitellus could only tolerate up to 2 mM sulfide.

Symbioses between protists and sulfide-oxidizing bacteria are another adaptive strategy observed in micro-oxic environments with high hydrogen sulfide concentrations. For example, the peritrich ciliate Zoothamnium niveum found in mangrove channels of the Caribbean Sea depends on its sulfur oxidizing ectobionts for detoxification of its immediate environment [54]. Scanning electron microscopy has revealed a visible diversity of ectobiotic prokaryotic associations with ciliates in the anoxic water column of Cariaco, and these associations are likely to be dependent on the distinct chemical nature of the basin’s water column (see Additional file 1). The environmental selection pressure that acts on the phylogenetic composition of protistan communities can be of interest for the design of environment-specific phylo-chips (for example of application see Sunagawa et al. [55] that may help to monitor the global distribution of specific protistan communities.

The temporal and spatial resolution of our sampling strategy is insufficient to deduce temporal and spatial patterns in protistan communities under study. Yet, possible explanations for the observation that in the Cariaco deep-sea basin, samples collected from the same depth at two dif-
Different points in time are distinctively less similar to each other (samples CAR2 and CAR4 in Figure 3, UPGMA), compared to the shallow Framvaren Fjord (samples FV2 and FV4) are obvious: Surface waters of the Cariaco Basin are subject to strong seasonal upwelling, driving as much as 13-fold excursions in net primary production (NPP) between upwelling and non-upwelling seasons [22]. This causes significant seasonal variations in vertical carbon fluxes, which seems to be not only very important for the dynamics of viral [27] and bacterial communities [56] in such systems, but also for protistan communities, even though the exact mechanisms for how vertical carbon flux variations may act on protistan communities are largely unknown. One possibility could be that due to selective interactions of protist with specific bacteria [57-59], changes in vertical carbon flux that have a direct influence on bacteria can act indirectly on protistan communities.

At first glance it seems disturbing that metazoa accounted for up to ca. 50% of all eukaryote tags (Figure 10). Because most metazoa are very sensitive to anoxia and hydrogen sulfide, this raises the question about the nature of these tags, whether they represent organisms that could plausibly live in the geochemical environments under study or rather represent contamination. Such high proportions of unique metazoan tags are indeed not unexpected after careful consideration: body parts, eggs or planktonic larvae of an individual taxon that may have been present in 5 to 10 liter water samples used for DNA extraction would contribute tremendous amounts of genomic DNA compared to the few individuals of a protistan taxon. Therefore, the SSU rRNA gene copies of this individual metazoan taxon would outnumber any protistan SSU rRNA gene copy numbers by far, resulting in high proportions of metazoan tags. For example, one individual copepod contributes almost 9,000 nearly identical amplicons to the FV1 amplicon library (Additional file 2). In order to account for intrinsic error rates of the pyrosequencing technique (see above) and for intraspecies SSU rDNA polymorphisms as described above for protistan data, we also clustered all metazoan tags at one to five nucleotides differences in a separate analysis. Indeed, it turned out that the proportion of unique metazoan tags decreased decisively (Additional file 3), accounting for only 3.9% to

Figure 8
Taxonomic distribution of V9 tags assigned to Chlorophyta. The data that served as a basis for the taxonomic bar chart are available as supplemental material (Table S7 in Additional file 5).
11.4% (Additional file 4) of total eukaryote tags when clustered at five nt differences (ca. 2% sequence divergence). Data serving as the basis for the relative distribution of taxonomic groups presented in Figures 4-9 can be found in Additional file 5.

Only a few taxa accounted for most of these metazoan tags, which belonged predominantly to copepods, cnidaria, ctenophores, molluscs and polychaetes (Additional file 2). Copepods can survive anoxia and high hydrogen sulfide concentrations for long periods of time [60]. Also several molluscs [61], cnidarians, ctenophores [62] and polychaetes are tolerant of anoxia [63]. Even Bryozoa that were detected in three of the samples (Additional file 2) are capable of thriving under anoxic conditions [64]. Thus, the detection of metazoan sequences in anoxic environments retrieved by domain (Eukarya)-specific PCR primers is not surprising. Yet, with the exception of copepods, which we can observe frequently at least in the oxic-anoxic interfaces of our sampling sites we did not confirm the presence of these metazoan taxa in the water samples under study by visual inspection. This is mainly due to the fact that we only screened 20-μl aliquots microscopically (for protistan target taxa). Because of this, small forms (life stages) of larger metazoans or small metazoans like bryozoa represented in our amplicon libraries may have been easily overlooked. It is reasonable to assume that the metazoan amplicons may represent a mixture of allochthonous material (see the detection of a hymenoptera phylotype in FV4 that is represented by nearly 5,000 amplicons) and autochthonous organisms. However, taking into account the low proportion of unique metazoan taxa when clustered at 5 nt differences and the high likelihood of the indigenous nature of most of the metazoans represented by the non-protistan tags it is reasonable to consider contamination in general as an insignificant issue.

This study shows that when 454 pyrosequencing of the V9 region is paired with rigorous downstream data processing, this method is more time- and cost-efficient, and produces a much more comprehensive picture of the protist community than Sanger sequencing of clone libraries, allowing for better estimates of community complexity. While direct comparison of the Framvaren and Cariaco communities is complicated by multiple physico-chemical differences between these two sampling locations, it is possible to distinguish protistan communities on the

![Taxonomic distribution of V9 tags assigned to Heterokontophyta. The data that served as a basis for the taxonomic bar chart are available as supplemental material (Table S8 in Additional file 5).](image)
basis of community composition in the supersulfidic Framvaren Fjord with an interface located in the photic zone from those in the deep-sea anoxic and less sulfidic environment. Furthermore, protistan populations in the sulfide-free oxic/anoxic interface in both Framvaren and Cariaco are distinct from those that are exposed to hydrogen sulfide. However, the specific environmental factors structuring protistan communities remain unknown.

Conclusion
This study combined high-throughput parallel tag sequencing with a highly conservative bioinformatic analysis pipeline to investigate protist community complexity in two contrasting anoxic marine ecosystems (Framvaren Fjord, Norway and Cariaco Basin, Venezuela). Our data suggest that both ecosystems under study are highly variable regarding the dynamics of protistan communities on a spatial and temporal scale. However, high-resolution sampling will be necessary to reliably assess the true extent of this variability. Furthermore, this study illustrates that our current understanding of the ecological complexity of protist communities, and of the global species richness and genome diversity of protists, is severely limited. A deep sequencing of ca. 250,000 V9 SSU rRNA gene tags in total allowed us to recover eukaryotic taxonomic groups that were not detected by previous clone library-based diversity surveys of both sampling sites. Furthermore, the number of detected OTUs far exceeded previous richness estimates from the same sites based on clone libraries and Sanger sequencing, even when tag sequences were clustered at up to ten nucleotide differences (reflecting ca. 8% sequence similarity). Our data highlight the magnitude of the under-sampled protistan gap in the eukaryotic tree of life, and support the notion that not only in bacteria but also in protists over 90% of the observed diversity was represented by rare OTUs that had less than 10 sequence tags assigned to them. Even though 454 pyrosequencing is not a panacea, it allows for more comprehensive insights into the diversity of protistan communities, and combined with appropriate statistical tools, enables improved ecological interpretations of the data and projections of global diversity.

Methods
Sampling sites and collection procedure
Samples were collected from two locales, the Cariaco Basin, located on the northern continental shelf of Venezuela, and the Framvaren Fjord located in Southwest Norway.
way. Sampling protocols for both sites were as described elsewhere, as well as the protocols for measurement of physico-chemical and biological parameters [10, 26]. Depths of samples, volumes of water collected, and physicochemical characteristics at the time of sampling are presented in Table 1. The sampling design accounted for three features: (i) temporal effects (same site sampled at different seasons), (ii) local patchiness (same depth sampled at two distant sites at each locale), [36] environmental factors (vertical water column gradient at each site and distinct locale-characteristics). Cariaco samples were collected at the Cariaco Time Series Station A (10.30° N, 64.40° W) and at Station C (10.40° N, 65.35° W). These two stations are located in the eastern and western subbasins of the Cariaco system, respectively. Cariaco samples were collected from Station A at the depth corresponding to the oxic/anoxic interface (250 m, oxygen and sulfide not detectable) in January 2005 (CAR1) and from below the interface (300 m) in January 2005 (CAR2) and May 2005 (CAR4). Sample CAR3 was collected at Station C from below the interface (320 m) ca. 100 km distant from Station A, in January 2005. Framvaren Fjord samples were collected in the central basin (58.11° N, 06.45° E) from the oxic/anoxic interface at 20 m sampled in September 2005 (FV1) and at Station A, in January 2005. The sample FV3 was taken from below the interface (36 m) in September 2005 (FV2) and May of 2004 (FV4). The sample FV3 was taken from the sampling location in the central basin) from below the interface at 36 m in September 2005.

### DNA isolation, PCR amplification, and 454 pyrosequencing

DNA was isolated from environmental samples and quality-checked as described previously [26]. In short, samples were taken with Niskin bottles and drawn onto 0.45 μm Durapore membranes (Millipore, Billerica, MA, USA) under anoxic conditions with no prefiltration step. Samples were frozen immediately in liquid nitrogen until further processing in the laboratory. The nucleic acid extraction protocol employed a high-salt extraction buffer (100 mM Tris HCl (pH 8), 100 mM sodium phosphate buffer (pH 8), 1.5 M NaCl, 100 mM EDTA (pH 8)) with 1% cetrimonium bromide. Approximately 3 ml of this buffer was added to one filter and the total genomic DNA was extracted using chloroform-phenol extraction and isopropanol precipitation. In order to minimize bias caused by sampling the extracts from three filters per sample site were combined prior to polymerase chain reaction-amplification. Our strategy targeted the V9 hypervariable region of the SSU rRNA genes [65]. This region was chosen because it is (i) among the most variable of eukaryote SSU rRNA hypervariable regions [66], represents a good marker for the taxonomic complexity of protistan communities, (ii) allowed for the use of conserved PCR-primers that target most described major eukaryote lineages, [36] has only marginal length variability among different taxonomic groups (127-150 bp) and (iv) could be fully sequenced using the Roche GS FLX system (up to 250 bp-reads) developed by 454 Life Sciences ([65], Stoeck T., Richards T, and Bass D., unpublished). PCR amplification and pyrosequencing followed the protocol of Amaral-Zettler et al. [65]. The PCR primers we used flanked the V9 region of eukaryote SSU rRNA genes. These primers were 1,380F (forward 1), 1,389F (forward 2), and 1,510R (reverse). Separate 1380F/1510R and 1389F/1510R reactions were run for each sample to recover the broadest eukaryotic diversity possible. The 454 Life Science's A or B sequencing adapters were fused to the 5' end of the primers. For each individual environmental DNA extract we ran three independent 30-μl PCR reactions with reaction mix consisting of 5 μl of Pfu Turbo polymerase (Stratagene, La Jolla, CA, USA), 1 μl Pfu reaction buffer, 200 μM dNTPs (Pierce Nucelic Acid Technologies, Milwaukee, WI, USA), a 0.2 μM concentration of each primer in a volume of 100 μl, and 3-10 ng genomic DNA as template. The PCR protocol employed an initial denaturation at 94°C for 3 min; 30 cycles of 94°C 30 s, 57°C for 45 s, and 72°C for 1 min; and a final 2 min extension at 72°C. PCR products from the same DNA sample were pooled and cleaned by using the MinElute PCR purification kit (Qiagen, Valencia, CA, USA). The quality of the products was assessed on a Bioanalyzer 2100 (Agilent, Palo Alto, CA, USA) using a DNA1000 LabChip (Agilent). Only sharp, distinct amplification products with a total yield of >200 ng were used for 454 sequencing. The fragments in the amplicon libraries were bound to beads under conditions that favor one fragment per bead. The emulsion PCR (empPCR, [67]) was performed by emulsifying the beads in a PCR mixture in oil, with PCR amplification occurring in each droplet, generating >10 million copies of a unique DNA template. After breaking the emulsion, the DNA strands were denatured, and beads carrying single-stranded DNA clones were deposited into wells on a PicoTiter-Plate (454 Life Sciences) for pyrosequencing on a Genome Sequencer FLX system (Roche, Basel, Switzerland) at the Marine Biological Laboratory (Woods Hole, MA, USA). In total, we recovered 251,648 sequence reads for the eight samples that were subjected to quality control. Removal of low quality sequences [14] left us with 222,593 high-quality reads for further consideration. Tag sequences have been deposited in the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) under the accession number SRP001212.

### Sequence data processing

We developed three automated analysis pipelines to analyze quality-checked 454 reads: The first pipeline for taxonomic assignment of V9 tags, the second pipeline for
clustering V9 tags at different sequence similarity levels and dereplication, and the third pipeline to construct a global tag matrix for sample comparison.

1.) Taxonomic assignment of V9 tags

The first pipeline was aimed at assigning taxonomy to our full-length eukaryote SSU rRNA gene sequences with a described taxonomy. A BLASTN analysis of the V9 fragments against the GenBank nr database revealed that the short V9 fragments could reliably be assigned to order-level when the closest BLAST hit (the original respective full-length sequence excluded) was at least 80% (see also [33,34]). Taxonomy of protists is according to Adl et al. [69] and for fungi according to Hibbett et al. [70]. We note that because Synurophyceae and Chrysophyceae are hardly distinguishable even when full-length 18S rRNA gene sequences are available, we united tags that were putatively assigned to Synurophyceae with Chrysophyceae to Heterokontophyta.

2.) Similarity clustering of 454 tags and dereplication

The second pipeline was dedicated to the clustering of tags at a given level of similarity. This is usually done by first using a multiple sequence alignment (MSA) program (usually MAFFT [71] or MUSCLE [72] to align the tags, followed by the calculation of a distance matrix (using QuickDist [14] for example) and finally statistical analyses. Our experience with the V9 domain indicated that none of the MSA programs was able to output alignments of high enough quality. We therefore implemented a completely new approach (Shahbazkia & Christen, in preparation). Our key hypothesis was that the frequency at which a given sequence occurs, the more likely it represents a real sequence. Conversely, there is a probability that a sequence found only once is the result of a PCR or sequencing error, or due to the presence of variations in some operons within a single genome [73]. First a python program allowed for a strict dereplication, i.e. clustering strictly identical sequences. This led to a 5 to 10 fold reduction in the number of sequences. Strict dereplication allowed for the second step, but also allowed for the construction of rank abundance curves. The resulting file (of strictly dereplicated tags) was sorted by decreasing abundance of tags in each cluster. Then, instead of computing a percentage of similarity between sequences (which is difficult because we don’t know how to implement a good substitution matrix for hypervariable regions of rRNA sequences) we implemented a Levenshtein distance calculation for clustering sequences. Levenshtein distance [74] is a measure of the similarity between two strings, which we will refer to as the source string (s) and the target string (t). The distance is the number of deletions, insertions, or substitutions required to transform s into t. Taking successively each dereplicated tag, the following tags were clustered with this representative if they had a Levenshtein of k or less (k ranging from 1 to 10). A number of checks were performed to analyze such clusters. A comparison of these clusters to the taxonomic assignments performed by the first pipeline showed an almost perfect agreement when taxonomic assignments had been possible by BLAST (k = 1,2,3). Above these k values many non-
assigned tags could be assigned to clusters containing assigned tags.

Comparisons of operational taxonomic units (OTUs) based on V9 domains and (almost) complete SSU rRNA sequences are almost impossible on large data sets of sequences because none of the multiple sequence alignment software is able to properly align SSU rRNA sequences within their divergent domains, and this problem is exacerbated for short divergent tag sequences (Guillou & Christen unpublished). For this reason, published 454 studies have relied heavily on BLAST alignment to public sequences to cluster tags. We used a completely new algorithm (Shahbazkia & Christen, unpublished) that directly clusters tag sequences having less than k differences (k = 0, 1,...,10) and does not rely on a multiple sequence alignment. We validated this approach in a separate analysis (Guillou and Christen unpublished) by demonstrating that our tag clustering method based on word counting instead of percent sequence similarity identified correctly the almost full-length sequences of a separate large, well-curated SSU rRNA alignment from which tags were extracted, and that using the clustering approach here, the same cluster ID was attributed to sequences that were phylogenetically close to the original tag sequence (Guillou and Christen, unpublished).

3. Tag matrix for sample comparison
A final pipeline was designed for the global statistical analysis of all eight samples. The entire data set consisting of all eight samples was this time considered and globally treated as described above in pipelines 2 and 3. This led to the construction of an abundance matrix at various clustering values as explained above where each column was a given sample and each line a cluster, values being either the number of occurrences of the tag in the sample, or simply 1 or 0 to indicate presence or absence of sequences belonging to that cluster. However, we here refrained from further analysis of the abundance-matrix, because due to different genome sizes and rRNA gene copy numbers among protists [75] and PCR primer selectivity [30] the abundance of PCR-amplicons from a sample does not necessarily reflect the relative abundance of the respective organisms in this sample.

The script for data analyses (Linux, Windows and Macintosh operating systems) is provided online [http://bioinfo.unice.fr/biodiv/prootst_data/]

Community comparisons
We calculated the Jaccard index, based on incidence (I_{incidence}) of unique OTUs as obtained from the third data processing pipeline described above, as a measure of community similarity between the eight samples under study using the program package SPADE [76]. Analyses were performed as recommended by the authors. Similarity values were transformed into a distance matrix and used for an Unweighted Pair Group Method with Arithmetic Mean analysis (UPGMA) of the eight unique libraries [77].

Data from the authors cited as unpublished are available from the authors upon request.

Abbreviations
OTU: operational taxonomic unit; PCR: polymerase chain reaction; MSA: multiple sequence alignment; DNA: deoxyribonucleic acid; RNA: ribonucleic acid; UPGMA: Unweighted Pair Group Method with Arithmetic Mean; NPP: net primary production; NE: northeast; CAR: Cariaco Basin; FV: Framvaren Fjord; BLAST: Basic Local Alignment Search Tool; SSU rRNA: small subunit ribosomal RNA.

Authors’ contributions
TS, VE, RC and AB conceived and designed the experiments. TS, VE, RC, AB and MR-M performed the experiments. TS, VE, RC, AB, WO and LAZ analyzed the data. RC contributed analysis tools. VE, TS, AB and LAZ wrote the paper.

Additional material

Additional file 1
Scanning electron micrograph of an unidentified ciliate isolated from anoxic, sulfidic waters of the Cariaco Basin. Figure S1. The ciliate in the picture, isolated from anoxic waters of the Cariaco basin, is covered with bacterial ecyotymbions. Protists with bacterial ecyotymbions are frequently recovered from sulfidic waters of both, the Cariaco basin as well as the Framvaren Fjord. It is not unlikely that these as yet unidentified bacteria may play a role as an adaptive mechanism for some protists to thrive in anoxic sulfidic environments. This picture is courtesy of Orsi W., Edgcomb V., Hohmann T. and Uiptar S.S. as part of a study on bacterial ecyotymbions on protists from the Cariaco Basin (Orsi et al., in preparation for publication).

Click here for file [http://www.biomedcentral.com/content/supplementary/1741-7007-7-72-S1.pdf]

Additional file 2
Taxonomy and proportion of abundant metazoan operational taxonomic units. Table S1. Taxonomy and proportion of abundant metazoan operational taxonomic units (OTUs) accounting for at least 1% of all metazoan OTUs of a specific amplicon library from four anoxic water samples from the Caribbean Cariaco deep-sea basin (CARI-1) and four anoxic water samples of the Norwegian Framvaren Fjord (FY1-4). OTUs were established based on identical best GenBank hit. For each OTU the best GenBank match is given (accession no., organism description, and taxonomy), as well as the number of total and unique tags. Unique tags are tags clustered at 0 differences.

Click here for file [http://www.biomedcentral.com/content/supplementary/1741-7007-7-72-S2.doc]
Additional file 3
Numbers of unique metazoan operational taxonomic units. Figure S2. Number of unique metazoan operational taxonomic units (OTUs) obtained from four samples of the Caribbean Cario Sea basin (CAR1-4, FV1-4) and four samples V9 tag rDNA sequences from four anoxic water samples of the Caribbean Cario Sea basin (CAR1-4) and from four anoxic water samples of the Norwegian Framvaren Fjord (FV1-4). Figure S2-B) at different levels of nucleotide differences. Tags were clustered at nt differences zero to five differences as described in pipeline 2 of the sequence data processing paragraph in the methods section. A difference caps in at least one of the eight libraries used for 454 sequencing. At k differences, two tags having k or fewer differences are placed in the same cluster; if they have more than k differences, they are in two different clusters. Click here for file [http://www.biomedcentral.com/content/supplementary/1741-7007-7-2-S3.pdf]

Additional file 4
Relative contribution of metazoan operational taxonomic units to total eukaryote operational taxonomic units. Table S2. Relative contribution of metazoan operational taxonomic units (OTUs) to total eukaryote OTUs when clustered at 5 nt differences (OTUs5) in the Framvaren Fjord (FV1-FV4) and the Caribbean basin (CAR1-CAR4) amplicon libraries. Click here for file [http://www.biomedcentral.com/content/supplementary/1741-7007-7-2-S4.doc]

Additional file 5
Relative taxonomic distribution of unique protistan and fungal V9 tags. Table S3. Accompanying data to Figure 4. Relative taxonomic distribution of unique protistan and fungal V9 tags generated from four anoxic water samples of the Caribbean Cario Sea basin (CAR1-4) and from four anoxic water samples of the Norwegian Framvaren Fjord (FV1-4) and from four anoxic water samples of the Norwegian Framvaren Fjord (FV1-4) within the Heterokontophyta. Table S7. Accompanying data to Figure 8. Relative taxonomic distribution of unique protistan and fungal V9 tags generated from four anoxic water samples of the Caribbean Cario Sea basin (CAR1-4) and from four anoxic water samples of the Norwegian Framvaren Fjord (FV1-4) within the Chlorophyta. Table S8. Accompanying data to Figure 9. Relative taxonomic distribution of unique protistan and fungal V9 tags generated from four anoxic water samples of the Caribbean Cario Sea basin (CAR1-4) and from four anoxic water samples of the Norwegian Framvaren Fjord (FV1-4) within the Heterokontophyta. Click here for file [http://www.biomedcentral.com/content/supplementary/1741-7007-7-2-S5.doc]

Acknowledgements
The International Census of Marine Microbes and the W.M. Keck Foundation award to the Marine Biological Laboratory at Woods Hole (MA) supported the pyrosequencing part of this study. We would like to thank L. Graham Murphy and E. Andreicheva for assistance with pyrosequencing. Further financial support came from a grant from the Deutsche Forschungsgemeinschaft to TS (STO414-3-1). Support for the unpublished work on Cario Sea basin protists came from NSF MCB-0348407 to VE (collaborative project with S. Epstein at Northeastern University, Boston, MA, USA). Financial support to AC was provided by NSF MCB 0348405. Financial support to RC was provided by the ANR-Biodiversité project Aquaparadox. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We would like to thank H. R. Shahbashzadeh (DEEL-FCU Universidad del Algarve Campus de Gambelas, Faro, Portugal) for programming parts of the clustering algorithm, the captain and crew of the B/O Hermann Gies and the staff of the Fundacion La Salle de Ciencias Naturales for their assistance during our fieldwork in Venzuela and HH Bruner for technical assistance. We are grateful to M. Scranton and G. Taylor for some of our Cario Sea metadata. We would like to thank four anonymous reviewers who substantially helped to improve the manuscript.

References


