ORIGINAL ARTICLE

Microbial community structure across the tree of life in the extreme Río Tinto

Linda A Amaral-Zettler[1,2,3], Erik R Zettler[4,5], Susanna M Theroux[1,2,3], Carmen Palacios[1,2,7], Angeles Aguilera[6] and Ricardo Amils[5,6]

[1]Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, USA; [2]NASA Astrobiology Institute, Marine Biological Laboratory, Woods Hole, MA, USA; [3]Department of Geological Sciences, Brown University, Providence, RI, USA; [4]Sea Education Association, Woods Hole, MA, USA; [5]Centro de Biología Molecular Severo Ochoa, Universidad Autónoma de Madrid, Cantoblanco, Spain and [6]Centro de Astrobiología, INTA-CSIC, Torrejón de Ardoz, Spain

Understanding biotic versus abiotic forces that shape community structure is a fundamental aim of microbial ecology. The acidic and heavy metal extreme Río Tinto (RT) in southwestern Spain provides a rare opportunity to conduct an ecosystem-wide biodiversity inventory at the level of all three domains of life, because diversity there is low and almost exclusively microbial. Despite improvements in high-throughput DNA sequencing, environmental biodiversity studies that use molecular metrics and consider entire ecosystems are rare. These studies can be prohibitively expensive if domains are considered separately, and differences in copy number of eukaryotic ribosomal RNA genes can bias estimates of relative abundances of phylotypes recovered. In this study we have overcome these barriers (1) by targeting all three domains in a single polymerase chain reaction amplification and (2) by using a replicated sampling design that allows for incidence-based methods to extract measures of richness and carry out downstream analyses that address community structuring effects. Our work showed that combined bacterial and archaeal richness is an order of magnitude higher than eukaryotic richness. We also found that eukaryotic richness was highest at the most extreme sites, whereas combined bacterial and archaeal richness was highest at less extreme sites. Quantitative community phylogenetics showed abiotic forces to be primarily responsible for shaping the RT community structure. Canonical correspondence analysis revealed co-occurrence of obligate symbionts and their putative hosts that may contribute to biotic forces shaping community structure and may further provide a possible mechanism for persistence of certain low-abundance bacteria encountered in the RT.

Introduction

Few scientists would deny that the search for life on planets beyond our own should be microbial in nature. Such expectations have fueled numerous studies on extreme environments as analogs for extraterrestrial habitable environments that can help predict whether life existed or still exists beyond our planet. Many of these extreme analog environments are dominated by chemolithotrophy as the principal form of metabolism and include examples such as Iron Mountain in California (Edwards *et al.*, 1998; Schrenk *et al.*, 1998) and the Río Tinto (RT) (Amils *et al.*, 2007; González-Toril *et al.*, 2003a, b) in southwestern Spain. The RT flows 100 km through the world's largest pyritic belt and is distinct from other extremely acidic, iron-rich sites in that both chemolithotrophy and phototrophy drive the metabolic machinery of this environment. As a consequence, bacteria, archaea and eukaryotes often occur in abundance (López-Archilla and Amils, 1999; Amaral-Zettler *et al.*, 2002; González-Toril *et al.*, 2003b; Aguilera *et al.*, 2007). Furthermore, it is an ancient ecosystem and geological evidence suggests that the microbial communities that exist there today are similar to those that existed millions of years ago (Fernández-Remolar *et al.*, 2005). Given its antiquity and size, one might predict an absence of a 'rare biosphere' (Sogin *et al.*, 2006) characterizing microbial diversity there, as the most successful organisms might be expected to outcompete and dominate in this extreme habitat. Before this

investigation and an earlier study (Palacios et al., 2008) that focused strictly on bacteria, we understood bacterial diversity in the RT to be limited to a small number of taxa that dominate the river (González-Toril et al., 2003b; Amils et al., 2004). These include the main iron-oxidizing and -reducing bacteria *Leptospirillum ferrooxidans*, *Acidithiobacillus ferrooxidans* and *Acidiphilium* spp. that help to shape the extreme conditions found in the RT (Amils et al., 2002). More recently, García-Moyano et al. (2007) reported on the bacterial and archaeal composition of floating macroscopic filaments. These investigators found a total of seven operational taxonomic units (OTUs) with an upper estimate of 36 for the total sequence diversity present. However, modern molecular high-throughput sequenced-based approaches to microbial diversity have transformed our ability to measure ecological diversity in the microbial realm and have revealed many diverse low-abundance taxa (Palacios et al., 2008) even in extreme environments such as the RT.

In this study, we explored the diversity of the RT tree of life (encompassing bacteria, archaea and eukarya) using high-throughput sequencing of an 800–1000 bp fragment of the small-subunit ribosomal RNA (*rRNA*) gene (covering the V4–V8 hypervariable region of the small-subunit *rRNA* coding region) and report on the results of a survey of 8002 sequences from three geochemically distinct stations in the RT. Important factors that set our study apart from previous work include a replicated study design and a simultaneous look at all three domains of life in the RT interpreted in the context of the underlying physicochemical environment. Before this investigation, diversity studies in the RT were either focused on the bacteria and archaea ('prokaryotes') or the eukaryotes as independent entities. And yet we know that this ecosystem, as most ecosystems, is composed of consortia including bacteria, archaea and eukaryotes that interact with each other and their environment; in some areas of the RT, eukaryotic microbes even dominate the biomass (López-Archilla et al., 2001).

It has been hypothesized that habitat filtering is largely responsible for shaping the community structure in extreme environments (Kraft et al., 2007). We tested this hypothesis directly by using phylogenetic approaches to compare and contrast biotic versus abiotic forces that drive community assembly in the RT using the Net Relatedness Index (NRI) and the Nearest Taxa Index (NTI) (Webb et al., 2002). NRI is a mean pairwise distance metric sensitive to phylogeny-wide patterns and NTI is a nearest-taxon-based measure sensitive to patterns at the 'tips' of a phylogenetic tree. We calculated these metrics for each of our sites for each domain separately to weigh the relative contributions of biotic versus abiotic forces shaping RT community structure. To explore the environmental factors that best explain the underlying patterns in species distributions, we applied canonical correspondence analysis (CCA) to our data set. In addition, we also directly compared bacterial and archaeal richness against eukaryotic richness. These methods have allowed us to generate new hypotheses with regard to the relationship between geochemistry and species distributions in the RT, as well as challenge the current perception of the relative contributions of different domains of life to the overall diversity in this extreme environment.

Materials and methods

Sample collection
In October 2002, surface water was sampled in triplicate from three different sites at each of three stations in the RT (Supplementary Figure S1): Origin (OR, located at N 37° 43.32′ × W 6° 33.06′), Anabel's Garden (AG, located at N 37° 43.49′ × W 6° 33.62′) and Berrocal (BE, located at N 37°35.58′ × W 6° 33.04′). The RT samples have the following naming convention: station abbreviation, site number and sample replicate number; for example, AG1.2 is the second replicate sample from site 1 at AG station. Sample collection, *in situ* and *ex situ* geochemical measurements and DNA extraction are described in Palacios et al. (2008).

Polymerase chain reaction (PCR) amplification, cloning and sequencing
Genomic DNA from each of the 27 samples was PCR amplified in triplicate using universal primers designed to target all three domains of life (bacteria, archaea and eukarya; Sinigalliano et al., 2007). This region spans *E. coli* positions 517–1391 corresponding to hypervariable regions V4–V8. We confirmed the efficacy of these primers using probeCheck (http://www.131.130.66.200/cgi-bin/probecheck/probecheck.pl) and the Greengenes database (bacteria, archaea and eukaryotes). Our queries indicated a 94.2% zero-mismatch level for 517F, whereas the 1391R primer yielded a 95.2% zero-mismatch level when we constrained the search to *Escherichia coli* positions 1200–1500 for bacteria and archaea. A separate eukaryotic search using a custom eukaryotic database created in ARB (Munich, Germany; Ludwig et al., 2004) and using the 1391R primer yielded a zero-mismatch level of 88.3% for this primer and 90.7% for 517F. These are all conservative estimates, as they reflect the total number of sequences in these respective databases, some of which may be missing parts of the queried region. PCR amplifications contained 1 μl of DNA template (approximately 2–5 ng), 5 μl of both the forward and reverse primers at 10 ng μl^{-1}, 2 μl Taq polymerase (Promega, Madison, WI, USA), 5 μl dNTPs at 2 ng μl^{-1} each (Promega), 2.5 μl of PrimeSafe (Smiths Detection, Boston, MA, USA) to suppress mispriming and improve overall efficiency of the reactions, 5 μl 10 × Taq buffer

solution (Promega) and 28.5 μl of PCR grade water. The PCR conditions were as follows: 2 min at 94 °C, followed by 30 cycles at 94 °C for 1 min, 45 °C for 1 min, 72 °C for 2 min; and 72 °C for 10 min, carried out using an Eppendorf Mastercycler. PCR products were confirmed using gel electrophoresis.

Two or three separate PCR amplifications were pooled before cloning. A Purelink purification kit (Invitrogen, Carlsbad, CA, USA) was used to purify PCR products for cloning and to exclude any fragments <300 bp. An A-tailing reaction and an additional purification with the Purelink kit preceded cloning using the Invitrogen TOPO TA kit.

From each sample, 384 clones were selected and grown in 1.2 ml of SuperBroth (Qbiogene Inc., Carlsbad, CA, USA) with 0.01% kanamycin in a 96-well block. Plasmids were purified using an alkaline lysis method on a Beckman Coulter BiomekFX machine (Brea, CA, USA). Sequencing of plasmid DNA was carried out using an ABI 3730 XL (Applied Biosystems, Foster City, CA, USA) capillary sequencing machine at the Marine Biological Laboratory Keck Facility. A total of 384 clones were sequenced from each library in the forward and reverse directions using M13F (5′-GTAAAACGAC GGCCAGT-3′) and M13R (5′-AACAGCTATGACC ATG-3′) primers. Sequences were assembled and quality checked using an automated pipeline (STRAW) developed at the Josephine Bay Paul Center (Woods Hole, MA, USA).

We ran BLAST searches against the nucleotide and environmental sequence databases to identify those relatives of the sequences observed that were closest to those in GenBank. We imported all sequences into the ARB (Ludwig et al., 2004) program and aligned them using the fast aligner option in the Integrated Aligners menu and further adjusted the alignment manually. We then used the 'quick add sequences to tree' feature, to add all the sequences to the reference tree found in the SILVA 93 REF database (Pruesse et al., 2007). This allowed us to assign domains more accurately to a given sequence and to distinguish mitochondrial and chloroplast sequences from bacterial sequences. Only eukaryotic chloroplast and nuclear genes were used for downstream analyses. Base-calls were verified and sequences were manually edited with the program Consed for chromatogram viewing (Gordon et al., 1998). Potential chimeras called by Bellerophon (Huber et al., 2004) and MALLARD (Ashelford et al., 2005, 2006) were visually inspected for confirmation. All sequences have been deposited in EMBL under accession numbers FN860149-FN868148.

Community phylogenetics
The phylogenetic tree shown in Figure 1 was constructed using RAxML (Stamatakis, 2006) and 857 positions for all 8002 sequences from the sequenced data set that remained after chimera checking and quality control screening. Only double-stranded sequences were used in downstream analyses and resulted in a reduction in the original total number of sequences. This consisted of 5753 bacterial sequences, 626 archaeal sequences, 439 eukaryotic sequences, 923 chloroplast sequences and 261 mitochondrial sequences. The RAxML tree was generated using the following settings: GTRGAMMA (General Time Reversible model with gamma) using a starting tree created by the quick add species to tree option generated in ARB and an initial rearrangement setting of 10, and the number of rate categories set to 50. Individual 'domain' trees for downstream Phylocom 4.0.1 (Webb et al., 2008) analyses were constructed by running separate analyses in RAxML with a subset of the 8002 sequences (857 positions), by (1) removing all mitochondria from the analyses and (2) removing any suspected duplicates for nuclear and chloroplast genes. Only unique sequences were used to construct these trees run under the GTRGAMMA model. To generate NTI and NRI values for all sites, we implemented the comstruct command in Phylocom with random draws of species from the phylogeny pool and 9999 randomizations.

Multivariate analyses
OTUs were defined by 1% dissimilarity for bacterial and archaeal sequences after Palacios et al. (2008) and using phylogenetically derived clusters for eukaryotic sequences. The 1% clustering level for bacteria and archaea is conservative, as this level of dissimilarity is considered over the entire V4–V8 region that includes both conserved and hypervariable regions. There is no universally agreed upon clustering level to define eukaryotic OTUs; therefore, we chose to define our eukaryotic OTUs on the basis of phylogenetically derived clusters. Such an approach is largely missing from the literature because most studies do not carry out phylogenetic analyses on a given data set for lack of sufficient sequence data or computational cost. Eukaryotic OTUs were extracted from clades found after adding the sequences to the reference tree in ARB. Sequences that branched with distinct and different known sequences or environmental clones were considered as distinct OTUs. We converted all abundance data to incidence data (presence or absence) before proceeding with multivariate analyses. Our data consisting of a 732 OTU by 27 sample matrix and 15 geochemical parameters measured simultaneously from the same 27 samples were used for further multivariate analyses.

CANOCO 4.5 (Microcomputer Power, Ithaca, NY, USA) (ter Braak and Šmilauer, 2002) was used to perform the CCA. Of the 25 elements analyzed using Total Reflection X-ray Fluorescence, we retained the following 12 parameters that had no missing values: As, Ca, Co, Cr, Cu, Fe, Mn, Ni, S, Si, Sr and Zn. We also used the pH, redox and conductivity values
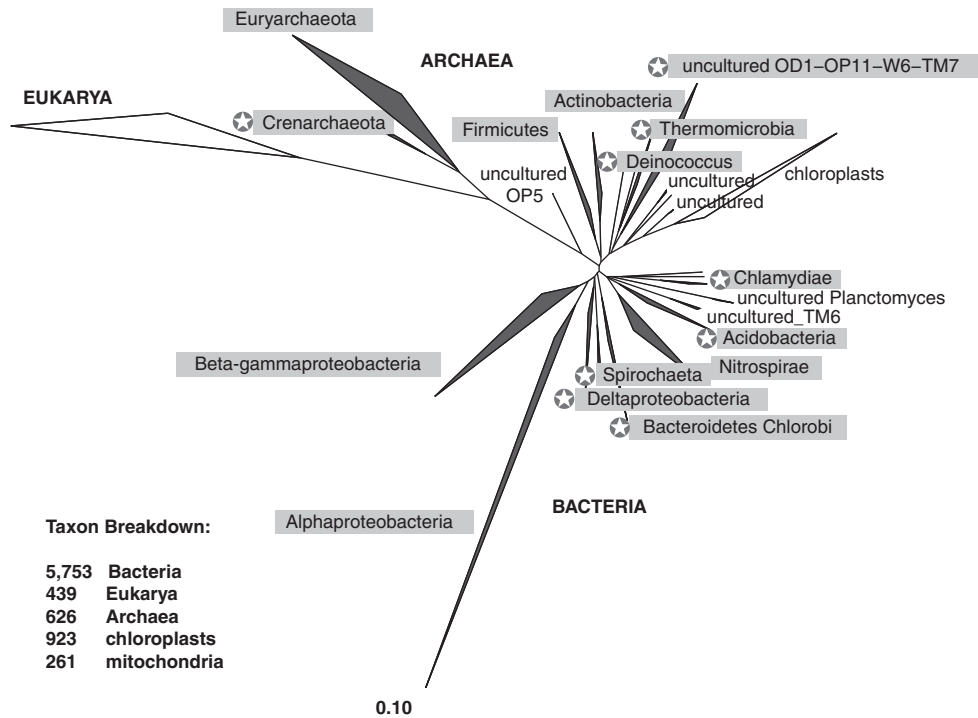
Figure 1 The RT tree of life based on the V4–V8 portion of the small-subunit rRNA coding region and inferred using maximum likelihood as implemented in RAxML. Stars designate phyla not reported before in clone libraries from river water samples. The bar represents the number of substitutions per site.

measured in the field with probes (see Palacios *et al.* (2008) for methods on environmental parameters). Environmental data were transformed using $\ln(x + 0.1)$ and normalized to zero mean and unit variance. CCA was used to investigate how the microbial community varied with the geochemical conditions. CCA is a robust multivariate analysis method for directly relating species to environmental gradients. It facilitates both pattern recognition in noisy data and statistical testing of the relationships between organisms and environment. A preliminary detrended correspondence analysis yielded gradients of 3.9 and 4.6 for axes 1 and 2; therefore, we chose to continue our analyses using a unimodal approach. As many of the environmental parameters were strongly correlated, we reduced the number of environmental variables further. This simplifies interpretation of the results. CANOCO provides an automatic selection feature to ensure that parameters that have low cross-correlation and explain the most variance are included. Because of the importance of iron, both as a metabolic substrate and as a pH buffer for the river, we manually selected Fe to be included in the analysis. After that, we used automatic selection of environmental parameters until additional variables no longer explained a significant amount of the variance. These criteria yielded four variables: Fe, Cu, Ni and redox. The significance of the first CCA axis and all CCA axes combined was tested using Monte Carlo permutation tests (ter Braak and Šmilauer, 2002).

Incidence-based richness estimates
We used presence-absence OTU matrices from replicate samples (three per site) as input for the SPADE program (Chao and Shen, 2010). We calculated the incidence-based estimator using the SPADE 'Presence/Absence Data for Multiple Samples/Quadrats' option with default settings for our bacteria, archaea and eukaryotes from our nine sites. The estimated number of OTUs per site is shown with Bonferroni-corrected 95% confidence intervals.

Results and Discussion

New phylogenetic diversity across domains
This study targeted a region of the *small-subunit rRNA* gene bounded by priming sites that are among the most conserved regions across the domains of life. Although not as informative as full-length sequencing, our strategy has the advantage of allowing for the simultaneous recovery of bacterial, archaeal and eukaryotic genes, as well as associated mitochondrial and chloroplast genes, and their subsequent OTU assignment by a combination of clustering based on sequence similarity and phylogenetic approaches. Supplementary Figure S1 depicts our study sites overlain with the relative recovery of genes represented from bacteria, archaea or eukarya. Bacterial genes dominated all our sites, except Anabel's Garden site 2 (AG2), in which 65% of the clones recovered were eukaryotic in origin.

Archaeal genes were highest in AG1 and AG3. In many cases, the recovery of chloroplast genes was favored over nuclear genes for a given taxon, but this bias was reproducible between samples and only in rare cases were we unable to identify the taxonomic source of the organellar gene. From our data, we constructed a global phylogeny for all 8002 sequences recovered from our survey. This also served to differentiate between nuclear and organellar genes and helped to place each sequence in its proper domain.

In this survey we have been able to capture a broader phylogenetic breadth of taxa than in any previous study of the RT, including sequences of several taxa not detected before in the RT. In Figure 1, we depict our phylogenetic tree with the entire eukaryotic domain and all but the major bacterial and archaeal lineages compressed. Sequences of major phyla not reported before from clone library surveys of RT river water are shown with a star next to the group name. Among them were members of the archaeal phylum Crenarchaeota, including clones from Origin sites 1 and 2 (OR1, OR2) related to sequences derived from an acid-mine drainage study from the Chinese Tong Shan-Kou mine (96% similar to DQ901270). These were phylogenetically distinct from the crenarchaeal clones found in an RT rhizosphere study (Mirete *et al.*, 2007) that branched among the South African gold mine crenarchaeal group-1. We further detected a singleton crenarchaeal clone from Berrocal site 2 (BE2) related to an archaeal sequence from a subsurface acid-mine drainage system (99% similar to AY082452). We also recovered several euryarchaeal sequences from AG and OR stations that clustered with nanoarchaea (89% similarity to AY652726) reported from Iron Mountain, CA, USA (Baker *et al.*, 2006). Baker *et al.* argued that the discovery of their novel nanoarchaea was due to their metagenomic approach and cited the use of traditional primers covering a larger portion of the gene as the cause for failed amplification attempts in the past. Here we suggest that the use of universal primers can also uncover a larger fraction of the microbial diversity present, including putative nanoarchaea that have escaped past amplification attempts.

The bacterial domain also yielded sequences from several phyla novel to the RT and some first recovered from our SARST-V6 approach (Palacios *et al.*, 2008). An advantage that this study has over our bacterial-targeted SARST-V6 study is that the sequencing recovered much longer reads of 800–1000 bp using the V4–V8 hypervariable region, versus 60–100 bp using the V6 region in our earlier study. Hence, we have been able to better characterize many of these new sequences in a phylogenetic context. These taxa included members of the Bacteroidetes Chlorobi, Deltaproteobacteria, Spirochaeata, Acidobacteria, Chlamydiae, *Deinococcus*, Thermomicrobia and members of the uncultured OD1-PO11-W6-TM7 clade. The largest numbers of

these sequences newly reported from RT were Deltaproteobacteria and were related to sequences derived from acid-mine drainage sites in California and China. At least some of the members of the Bacteroidetes-related sequences branched with sequences identified as symbionts of *Acanthamoeba* spp. In addition to new acidobacterial sequences, some of the Acidobacteria that we obtained from OR2 were related to sequences from a metagenomic survey from the RT rhizosphere, whereas others were related to those recovered from acid-mine tailings, including acidobacterial sequences from a nearby copper mine in the Odiel River basin (Rowe *et al.*, 2007). The presence of Acidobacteria in the RT was detected previously by an oligonucleotide array experiment (Garrido *et al.*, 2008), but here we report the actual sequences from acidobacteria new to the RT. The *Deinococcus* clones were related to sequences derived from a rejected coal pile (Brofft *et al.*, 2002).

Among the novel eukaryotic sequences recovered during this study were amoebae which are opportunistic human pathogens that were not detected in our earlier work (Amaral-Zettler *et al.*, 2002). These included sequences related to *Acanthamoeba* spp. and also one sequence recovered from our most extreme site (OR3) related to *Balamuthia mandrillaris* (92% similarity). We also detected sequences that branched among the amoeboid genus *Naegleria*, but additional sequence information is required to determine whether these sequences are related to the opportunistic human pathogenic species of this genus, *Naegleria fowleri*. At OR1, we detected sequences related to an RT-isolated vannellid (a lobose amoeba) that we have in culture but that has never been detected before by *rRNA* gene surveys. Another group of novel sequences came from microsporidia-related taxa. We hypothesize that these sequences may be from allochthonous sources or from insects that we have seen in the river, including aquatic beetles (Coleoptera), water striders (Hemiptera) and midge larvae (Diptera).

Abiotic forces largely drive RT community structure
A common tenet popularized by Darwin himself is that closely related species often occur in similar habitats and share ecological characteristics. Modern-day ecologists have adopted phylogenetic methods to explore this question (Martin, 2002). Most of these studies have concentrated on multicellular life, with the exception of a few studies that are microbial in focus (Horner-Devine and Bohannan, 2006; Bryant *et al.*, 2008). Methods developed by Webb *et al.* (2002) provide a method to use phylogenetic structuring in microbial communities to quantify the drivers of community assembly. Two possible patterns emerge: phylogenetic overdispersion, which is predicted for communities shaped primarily by biotic factors, with competition and facilitation being the two most widely cited causes (Cavender-Bares *et al.*, 2004; Kembel and Hubbell,

2006); and phylogenetic clustering, which is most often attributed to habitat filtering such as environmental drivers shaping community structure (Cornwell et al., 2006; Kembel and Hubbell, 2006).

The majority of the communities we sampled showed significant phylogenetic clustering, consistent with taxa being more closely related to each other than expected by chance. This is shown in Figure 2, which shows the values we obtained for NRI and NTI. Values >0 indicate clustering and values <0 indicate overdispersion in the trees. Solid symbols represent communities that are significantly structured by phylogenetic clustering or phylogenetic overdispersion at the 95% confidence level. On closer examination, this structure was most prevalent in bacterial communities for both NRI and NTI and to a lesser extent, in eukaryotic and archaeal communities. This suggests that environmental factors are more important than biological factors (competition, predation, parasitism) in constraining phylogenetic diversity in the RT. The one notable exception in which we saw evidence for phylogenetic overdispersion occurred with the NRI analysis for the archaeal community at site OR2 ($P<0.05$). The NRI for bacteria at this site was also <0, suggesting overdispersion albeit at the $P<0.10$ level; this result is striking because at all other sites the bacterial communities had NRI values >0, meaning that community structuring was consistent with phylogenetic clustering. Although the NTI for OR2 did not indicate significant overdispersion, it was <0 for bacterial communities.

Overdispersion was not echoed in the archaeal community for the NTI analysis nor did eukaryotes show a significant trend for this site. A closer look at the environmental properties and geochemistry of site OR2 reveals that it has a higher pH (2.72) than any of the other sites (Palacios et al., 2008). However, pH is not the only factor that sets OR2 apart. In examining a 6-year record of geochemistry at this site, we found that OR2 stands out for its geochemical stability. We speculate that perhaps environmental stability has led to different forces structuring OR2 bacterial and archaeal communities. Eukaryotic communities at this and other sites seemed phylogenetically clustered for both NRI and NTI at sites AG2 and OR3, two sites with the lowest pH values (pH 1.97 and 1.71, respectively) during our sampling period. NRI values at these sites were also much higher than at the others.

Combined bacterial and archaeal richness exceeds eukaryotic richness in the RT

A long-standing paradigm in RT literature is that eukaryotic diversity surpasses bacterial and archaeal (prokaryotic) diversity (Amils et al., 2002, 2007) in this extreme ecosystem. In this study, for the first time, we directly compared incidence-based (presence or absence) richness estimates (Chao and Shen, 2010) of bacterial and archaeal communities (i.e., 'prokaryotic communities') compared with eukaryotic communities. We favored incidence-based methods over abundance-based ones because differences in eukaryotic gene copy number in different taxa bias comparisons of relative abundance of eukaryotes in a sampled environment (e.g., as many as 12 000 copies of rRNA genes occur in some species, whereas as few as one in others; Zhu et al., 2005). We found that combined bacterial and archaeal richness estimates were an order of magnitude higher than eukaryotic richness estimates (Figure 3).
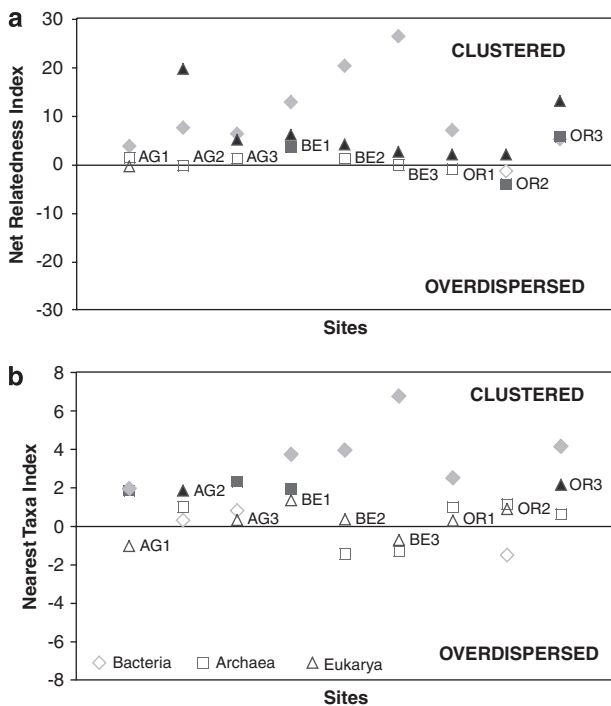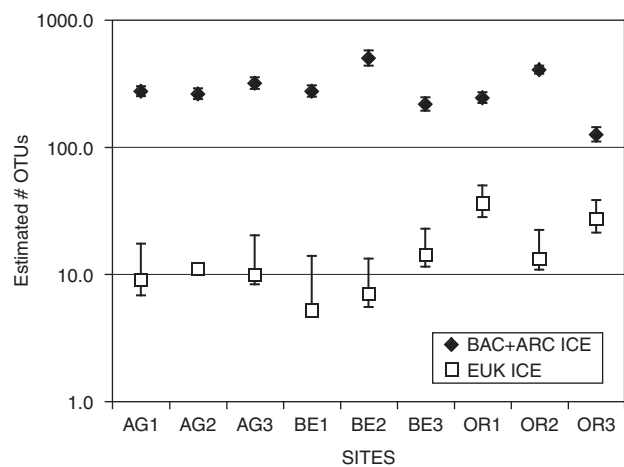

Figure 2 NRI and NTI for the nine sites from our study. Solid symbols represent communities that are significantly structured by phylogenetic clustering (values >0) or by phylogenetic overdispersion (values <0) at the $P=0.05$ level.

Figure 3 Incidence-based (ICE) diversity estimates for bacteria and archaea (BAC + ARC) versus eukaryotes (EUK) for our nine sites. Error bars represent Bonferroni-corrected 95% confidence intervals. The y-axis is represented as a log scale.

48

This result contradicts the long-standing perception that eukaryotic diversity is much greater than bacterial and archaeal diversity in the RT. Eukaryotic richness was highest at our most extreme sites (OR1 and OR3), being significantly higher for OR1, whereas bacterial and archaeal diversity was significantly higher at our less extreme sites (BE2 and OR2) and significantly lower at the most extreme site (OR3) (Figure 3). The observation that eukaryotic diversity was higher at our most extreme sites is a remarkable finding and deserves deeper investigation. Other high-throughput methods such as 454 pyrosequencing will likely further unveil the complexity of this 'simple' ecosystem.

RT environmental structuring and co-occurrence patterns

Our phylogenetic community comparisons suggest that community structure in the RT is largely shaped by abiotic factors, but provide little information about the environmental drivers that might be responsible. To address this question, we explored the relationship between samples, OTUs and environmental factors using CCA. The CCA triplot in Figure 4 summarizes our results showing the relationship between environmental variables, samples and OTUs. Samples plot in different areas of the diagram depending on their environmental characteristics. Replicate samples from a given site (represented by shape and color) generally clustered together. CCA axis 1 was most closely correlated with copper, separating all three stations (OR, AG and BE). CCA axis 2 was most closely correlated with iron and redox, and negatively correlated with the supplementary variable pH. OR sites separated well along CCA axis 2 and showed a strong gradient from the most acidic iron-rich site OR3 to the least acidic site OR2. OR2 samples plot more closely to AG samples located 1 km away than to the other OR samples located just meters away, confirming that differences are not simply due to spatial separation. The diagram also plots each bacterial, archaeal and eukaryotic OTU recovered from our analyses at the centroid of its occurrence in samples; therefore, each OTU plots at the center of its environmental 'niche'. OTUs that are close to each other co-occur more frequently than those that are far apart.
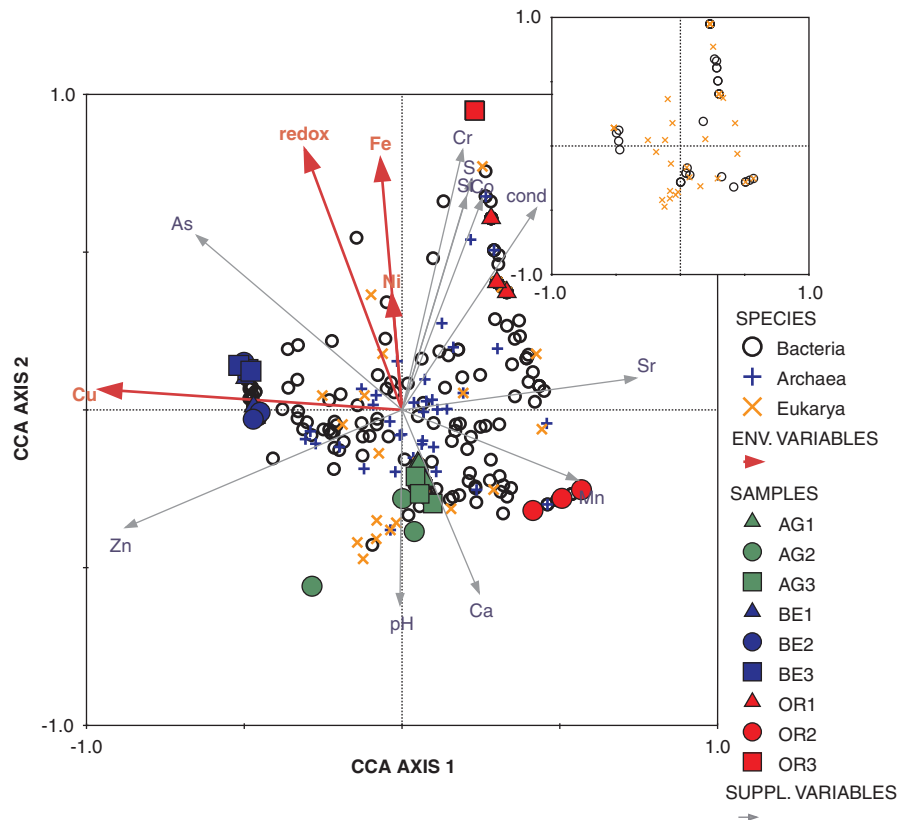

Figure 4 Triplot of CCA axes 1 and 2 accounting for 54% of the variance of species (OTUs) with respect to environmental variables. The eigenvalue for axis 1 (horizontal) is 0.61 and for axis 2 (vertical) is 0.55, indicating that both are strong gradients. Presence or absence data were used for bacterial, archaeal and eukaryotic OTUs. Replicate samples are shown. Separate sites are indicated by different symbols, separate stations by different colors. Explanatory variables used in the analysis are shown in red arrows, whereas supplementary variables are shown in gray. The inset highlights eukaryotic OTUs and their relationship to putative symbiotic bacterial OTUs. See main text for explanation and interpretation and Supplementary Figure S2 for further details on the inset figure.

Although CCA provides a window into the environmental factors influencing RT community structure, it also provides a rare glimpse at potential biotic interactions that occur in this environment by the examination of co-occurrence patterns of different domain constituents. Such biotic interactions include symbioses that might contribute to the overall bacterial richness in the RT by enabling the persistence of certain low-abundance bacterial OTUs within eukaryotic hosts. The inset in Figure 4 and the corresponding detailed blowup and table in Supplementary Figure S2 revealed co-occurrence patterns for select bacterial and eukaryotic OTUs. Most noteworthy is the prevalence of putative symbiotic OTUs affiliated with our most extreme site—OR3. Here, we see the co-occurrences of several amoeboid protists alongside putative symbionts. These included the protistan *Acanthamoeba*, heterolobosean amoebae and *Balamuthia*-related OTUs co-occurring with bacterial OTUs related to *Rickettsiales* and Captivus-related clades, as well as OTUs related to Chlamydiae and Rhabdochlamydia—all obligate intracellular symbionts or parasites. Captivus-related OTUs were specifically related to *Candidatus* Captivus acidiprotistae from Iron Mountain (Baker *et al.*, 2003). These are all members of Holosporaceae and include the named species *Holospora obtusa*, which is a macronuclear symbiont of the ciliate *Paramecium caudatum* (Fujishima and Fujita, 1985). We also discovered several OTUs related to the bacterial genus *Legionella*, a potential human pathogen. In addition to co-occurrence with amoebae, we also detected at least one ciliate OTU associating with these symbiotic bacterial OTUs; therefore, it is possible that our symbiotic OTUs are also associated with ciliates in addition to amoebae. Although endosymbionts have been detected before in acidic environments (Baker *et al.*, 2003), this study points to the possible breadth of symbiont diversity associated with their acidophilic hosts. If these symbionts are neutrophiles, it follows that these associations are also very ancient ones. The role of eukaryotic hosts facilitating bacterial diversity in extreme systems remains to be explored.

In summary, this is the first three-domain comparison of microbial richness in an extreme ecosystem and to the best of our knowledge the first study to quantify phylogenetic structure across all three domains of life in any ecosystem. In contrast to earlier studies in this extreme system, our work showed that combined bacterial and archaeal richness is an order of magnitude higher than eukaryotic richness. We also found that eukaryotic richness was highest at the most extreme sites, whereas combined bacterial and archaeal richness was highest at less extreme sites. A phylogenetic community structure analysis in the RT supports our conclusion that environmental forces largely determine the underlying patterns of diversity in the communities inhabiting this ecosystem. However, our analysis revealed that poorly studied interactions such as symbiosis in the RT may contribute to biotic forces shaping community structure (by facilitation) and may further provide a possible mechanism for persistence of certain low-abundance bacterial OTUs encountered in the RT.

Acknowledgements

References

Aguilera A, Zettler E, Gomez F, Amaral-Zettler L, Rodriguez N, Amils R. (2007). Distribution and seasonal variability in the benthic eukaryotic community of Rio Tinto (SW, Spain), an acidic, high metal extreme environment. *Syst Appl Microbiol* **30**: 531–546.

Amaral-Zettler LA, Gómez F, Zettler E, Keenan BG, Amils R, Sogin ML. (2002). Eukaryotic diversity in Spain's River of Fire. *Nature* **417**: 137.

Amils R, Gonzalez-Toril E, Fernandez-Remolar D, Gomez F, Aguilera A, Rodriguez N *et al.* (2007). Extreme environments as Mars terrestrial analogs: the Rio Tinto case. *Planet and Space Sci* **55**: 370.

Amils R, Gonzalez-Toril E, Fernandez-Remolar D, Gomez F, Rodríguez N, Duran C. (2002). Interaction of the sulfur and iron cycles in the Tinto River ecosystem. *Rev Environ Sci Biotechnol* **1**: 299–309.

Amils R, Gonzalez-Toril E, Gomez F, Fernandez-Remolar D, Rodríguez N, Malki M *et al.* (2004). Importance of chemolithotrophy for early life on earth: the Tinto River (Iberian Pyritic Belt) case. In: Seckbach J (ed). *Origins*. Kluwer Academic Publishers: Netherlands, pp 463–480.

Asshelford KE, Chuzanova NA, Fry JC, Jones AJ, Weightman AJ. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl Environ Microbiol* **71**: 7724–7736.

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

Baker BJ, Hugenholtz P, Dawson SC, Banfield JF. (2003). Extremely acidophilic protists from acid mine drainage host Rickettsiales-lineage endosymbionts that have intervening sequences in their 16S rRNA genes. *Appl Environ Microbiol* **69**: 5512–5518.

Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE *et al.* (2006). Lineages of acidophilic Archaea revealed by community genomic analysis. *Science* **314**: 1933–1935.

50

Brofft JE, McArthur JV, Shimkets LJ. (2002). Recovery of novel bacterial diversity from a forested wetland impacted by reject coal. *Environ Microbiol* **4**: 764–769.

Bryant JA, Lamanna C, Morlon H, Kerkhoff AJ, Enquist BJ, Green JL. (2008). Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proc Nat Acad Sci* **105**: 11505–11511.

Cavender-Bares J, Ackerly DD, Baum DA, Bazzaz FA. (2004). Phylogenetic overdispersion in Floridian oak communities. *Am Nat* **163**: 823–843.

Chao A, Shen T-J. (2010). Program SPADE (Species Prediction And Diversity Estimation). Program and User's Guide published at http://chao.stat.nthu.edu.tw.

Cornwell WK, Schwilk DW, Ackerly DD. (2006). A Trait-based test for habitat filtering: convex hull volume. *Ecology* **87**: 1465–1471.

Edwards KJ, Schrenk MO, Hamers R, Banfield JF. (1998). Microbial oxidation of pyrite; experiments using microorganisms from an extreme acidic environment. *Am Mineral* **83**: 1444–1453.

Fernández-Remolar DC, Morris RV, Gruener JE, Amils R, Knoll AH. (2005). The Rio Tinto Basin, Spain: mineralogy, sedimentary geobiology, and implications for interpretation of outcrop rocks at Meridiani Planum, Mars. *Earth and Planet Sci Lett* **240**: 149–167.

Fujishima M, Fujita M. (1985). Infection and maintenance of *Holospora obtusa*, a macronucleus-specific bacterium of the ciliate *Paramecium caudatum*. *J Cell Sci* **76**: 179–187.

García-Moyano A, González-Toril E, Aguilera A, Amils R. (2007). Prokaryotic community composition and ecology of floating macroscopic filaments from an extreme acidic environment, Río Tinto (SW, Spain). *Syst Appl Microbiol* **30**: 601–614.

Garrido P, González-Toril E, García-Moyano A, Moreno-Paz M, Amils R, Parro V. (2008). An oligonucleotide prokaryotic acidophile microarray: its validation and its use to monitor seasonal variations in extreme acidic environments with total environmental RNA. *Environ Microbiol* **10**: 836–850.

González-Toril E, Gomez F, Rodriguez N, Fernandez-Remolar DC, Zuluaga J, Marin I et al. (2003a). Geomicrobiology of the Tinto River, a model of interest for biohydrometallurgy. *Hydrometallurgy* **71**: 301–309.

González-Toril E, Llobet-Brossa E, Casamayor EO, Amann R, Amils R. (2003b). Microbial ecology of an extreme acidic environment, the Tinto River. *Appl Environ Microbiol* **69**: 4853–4865.

Gordon D, Abajian C, Green P. (1998). Consed: a graphical tool for sequence finishing. *Gen Res* **8**: 195–202.

Horner-Devine MC, Bohannan BJM. (2006). Phylogenetic clustering and overdispersion in bacterial communities. *Ecology* **87**: S100–S108.

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

Kembel SW, Hubbell SP. (2006). The phylogenetic structure of a neotropical forest tree community. *Ecology* **87**: 86–99.

Kraft NJB, Cornwell WK, Webb CO, Ackerley DD. (2007). Trait evolution, community assembly, and the phylogenetic structure of ecological communities. *Am Nat* **170**: 271–283.

López-Archilla AI, Amils R. (1999). A comparative ecological study of two acidic rivers in southwestern Spain. *Microb Ecol* **38**: 146–156.

López-Archilla AI, Marín I, Amils R. (2001). Microbial community composition and ecology of an acidic aquatic environment: the Tinto River, Spain. *Microb Ecol* **41**: 20–35.

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

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

Mirete S, de Figueras CG, Gonzalez-Pastor JE. (2007). Novel nickel resistance genes from the rhizosphere metagenome of plants adapted to acid mine drainage. *Appl Environ Microbiol* **73**: 6001–6011.

Palacios C, Zettler E, Amils R, Amaral-Zettler L. (2008). Contrasting microbial community assembly hypotheses: a reconciling tale from the Rio Tinto. *PLoS ONE* **3**: e3853.

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

Rowe OF, Sánchez-España J, Hallberg KB, Johnson DB. (2007). Microbial communities and geochemical dynamics in an extremely acidic, metal-rich stream at an abandoned sulfide mine (Huelva, Spain) underpinned by two functional primary production systems. *Environ Microbiol* **9**: 1761–1771.

Schrenk MO, Edwards KJ, Goodman RM, Hamers RJ, Banfield JF. (1998). Distribution of *Thiobacillus ferrooxidans* and *Leptospirillum ferrooxidans*: implications for generation of acid mine drainage. *Science* **279**: 1519–1522.

Sinigalliano CD, Gidley ML, Shibata T, Whitman D, Dixon TH, Laws E et al. (2007). Impacts of hurricanes Katrina and Rita on the microbial landscape of the New Orleans area. *Proc Nat Acad Sci* **104**: 9029–9034.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR et al. (2006). Microbial diversity in the deep sea and the underexplored, 'rare biosphere'. *Proc Nat Acad Sci* **103**: 12115–12120.

Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

ter Braak CJF, Šmilauer P. (2002). CANOCO reference manual and CanoDraw for Window's user's guide: software for canonical community ordination (version 4.5). Microcomputer Power: Ithaca, pp 500.

Webb CO, Ackerley DD, McPeek MA, Donoghue MJ. (2002). Phylogenies and community ecology. *Ann Rev Ecol Syst* **33**: 475–505.

Webb CO, Ackerley DD, Kembel SW. (2008). Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, **24**: 2098–2100.

Zhu F, Massana R, Not F, Dominique M, Vaulot D. (2005). Mapping of picoeukaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52**: 79–92.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)