

## ORIGINAL ARTICLE

# Global ecological patterns in uncultured Archaea

Jean-Christophe Auguet, Albert Barberan and Emilio O Casamayor

Group of Limnology-Department of Continental Ecology, Centre d'Estudis Avançats de Blanes, CEAB-CSIC, Accés Cala Sant Francesc, Girona, Spain

**We have applied a global analytical approach to uncultured Archaea that for the first time reveals well-defined community patterns along broad environmental gradients and habitat types. Phylogenetic patterns and the environmental factors governing the creation and maintenance of these patterns were analyzed for c. 2000 archaeal 16S rRNA gene sequences from 67 globally distributed studies. The sequences were dereplicated at 97% identity, grouped into seven habitat types, and analyzed with both Unifrac (to explore shared phylogenetic history) and multivariate regression tree (that considers the relative abundance of the lineages or taxa) approaches. Both phylogenetic and taxon-based approaches showed salinity and not temperature as one of the principal driving forces at the global scale. Hydrothermal vents and planktonic freshwater habitats emerged as the largest reservoirs of archaeal diversity and consequently are promising environments for the discovery of new archaeal lineages. Conversely, soils were more phylogenetically clustered and archaeal diversity was the result of a high number of closely related phylotypes rather than different lineages. Applying the ecological concept of 'indicator species', we detected up to 13 indicator archaeal lineages for the seven habitats prospected. Some of these lineages (that is, hypersaline MSBL1, marine sediment FCG1 and freshwater pISA1), for which ecological importance has remained unseen to date, deserve further attention as they represent potential key archaeal groups in terms of distribution and ecological processes. Hydrothermal vents held the highest number of indicator lineages, suggesting it would be the earliest habitat colonized by Archaea. Overall, our approach provided ecological support for the often arbitrary nomenclature within uncultured Archaea, as well as phylogeographical clues on key ecological and evolutionary aspects of archaeal biology.**

*The ISME Journal* (2010) 4, 182–190; doi:10.1038/ismej.2009.109; published online 22 October 2009

**Subject Category:** microbial population and community ecology

**Keywords:** Archaea; community; ecology; diversity; environmental factors; global patterns

## Introduction

The study of the biology and ecology of Archaea is currently among the most exciting and dynamic research topics in microbial ecology. In less than two decades the status of these enigmatic microorganisms has changed completely. The popularization of environmental ribosomal gene analysis has revolutionized the biased perception on their biology and ecology. The new tools have expanded archaeal ecological distribution and metabolic diversity far beyond expected, unveiling a widespread distribution and an unexpected diversity (Schleper *et al.*, 2005; Chaban *et al.*, 2006; Auguet and Casamayor, 2008; Llíros *et al.*, 2008; Casamayor and Borrego, 2009).

The earliest archaeal phylogenetic tree derived from laboratory cultures (hyperthermophiles, halophiles and methanogens) was composed of the two main phyla, Crenarchaeota and Euryarchaeota, and contained a few branches. However, environmental PCR-based 16S rRNA gene surveys quickly expanded the archaeal tree with the discovery of new uncultured lineages. One of the most noticeable advances during the nineties was the discovery of mesophilic Crenarchaeota inhabitants of marine plankton and soils that formed a deeply divergent clade distantly related to hyperthermophiles. Two main crenarchaeal lineages were observed within this new clade: the *1.1a* (DeLong, 1992; Fuhrman *et al.*, 1992) and the *1.1b* (Bintrim *et al.*, 1997; Ochsenreiter *et al.*, 2003). In the last years, the 16S rRNA gene sequences from uncultured Archaea in databases have increased several orders of magnitude above those available from the cultured counterparts. A precise taxonomic placement of the new sequences will remain, however, uncertain until microbiologists succeed bringing into culture more

Correspondence: J-C Auguet, Group of Limnology-Department of Continental Ecology, Centre d'Estudis Avançats de Blanes, CEAB-CSIC, Accés Cala Sant Francesc, 14, Blanes, Girona, 17300, Spain. E-mail: jcauguet@ceab.csic.es

Received 5 June 2009; revised 27 August 2009; accepted 29 August 2009; published online 22 October 2009

archaeal representatives from a larger range of phyla. In addition, almost half of the 16S rRNA gene sequences archived in GenBank database lack clear taxonomic information (DeSantis *et al.*, 2006). As a consequence, different authors use different names for uncultured clusters that lead to conflicting nomenclatures, and ecological or physiological information becomes often veiled behind confusing clusters naming.

At present, public databases hold a large number of archaeal 16S rRNA environmental sequences (c. 40 000) from a large set of environments. This data set contains information to extract general macroecological patterns and to bring some light on how archaeal communities are structured along global environmental gradients. The aims of this study are to use the information present in databases to (i) describe the global distribution of archaeal communities and understand the forcing environmental factors that shape archaeal diversity and (ii) detect the main taxa that can be considered as 'indicator species' for a given habitat. We also provided a framework to identify environments that contain the highest archaeal diversity and represent promising habitats for the discovery of new archaeal lineages.

## Methods

### *Construction of the archaeal 16S rRNA gene database*

We surveyed published literature and GenBank database for archaeal 16S rRNA clone libraries (that is, a collection of identified PCR products obtained from the same source) that matched each one of the following criteria: (i) communities obtained from natural environments (artificial and semi-artificial environments with human-induced dynamics, such as rice soils and chemical reactors, were excluded for detailed analyses. In fact when sorted into an ordination plot according to phylogenetic community similarity, rice soils significantly separated from typical natural soil environments and were closer to freshwater sediments (data not shown)); (ii) high-quality data (no nucleotide ambiguities present and sequences > 300 bp); and (iii) use of universal primers covering the same 16S rRNA gene region. We homogenized different methodologies and sampling efforts by clustering sequences at 97% identity threshold (Shaw, 2008). We ended with an archaeal database of ~2000 archaeal 16S rRNA sequences from 67 clone libraries globally distributed (see Supplementary Table 1). The sequences were treated by two methods (see below), that is, by using an explicitly phylogenetic approach, and by a taxon-based approach (where taxa were picked at a defined level and then treated as equally divergent).

The different clone libraries were grouped into seven distinct habitats (understood as a group of environments sharing a close geochemistry) as

follows: freshwater plankton (Fwc), freshwater sediment (Fsed), soil (S), marine plankton (Mwc), marine sediment (Msed), hypersaline planktonic environments (Hsal) and hydrothermal vents (Hdv). Next, we constructed a semiquantitative environmental matrix according to the range of environmental gradients present in these habitats: temperature (hydrothermal vents to polar waters), salinity (hyperhaline brines to freshwater), life environment (plankton, soil and sediment), trophic state (eutrophic to ultraoligotrophic) and oxygen concentrations (anoxic to full oxic).

The 16S rRNA gene sequences were automatically aligned with the NAST aligner (DeSantis *et al.*, 2006) and imported into the Greengenes database (<http://greengenes.lbl.gov/>) based on the ARB package (Ludwig *et al.*, 2004) (<http://www.arb-home.de>). A base frequency filter was applied to exclude highly variable positions before sequences were added using the ARB parsimony insertion tool to the original Greengenes tree calculated by maximum parsimony method and provided by default.

### *Phylogenetic approach*

Distance matrices were constructed using the UniFrac metric (<http://bmf2.colorado.edu/unifrac/index.psp>). UniFrac is a beta diversity metric that quantifies community similarity based on the phylogenetic relatedness (Lozupone *et al.*, 2006). To assess the sources of variation in the UniFrac matrix, we used permutational manova based on 1000 permutations (McArdle, 2001) with function *adonis* in *vegan* package (Oksanen *et al.*, 2008).

Phylogenetic diversity (PD) for each of the seven habitats was calculated as the sum of the branch length associated with the 16S rRNA gene sequences within this habitat (Faith, 1992). To correct for unequal number of sequences, we calculated the mean PD of 1000 randomized subsamples of each habitat (Barberan A and Casamayor EO, unpublished).

The phylogenetic structure for each habitat was calculated with the phylogenetic species variability (PSV) index (Helmus *et al.*, 2007). PSV quantifies how phylogenetic relatedness decreases the variance of a hypothetical neutral trait. The value is 1 when all species are phylogenetically unrelated (that is, a star phylogeny) and approaches 0 as species become more related. To statistically test whether habitats were composed of species that are more or less related to each other than expected, we compared the mean observed PSV with distributions of mean null values (1000 iterations) using two different randomization procedures. Null model 1 maintains species occurrence, whereas null model 2 maintains habitat species richness (Helmus *et al.*, 2007). Analyses were run with the R package *picante* (Kembel *et al.*, 2008).

A genetic distance matrix of the sequences from each habitat was constructed with a subset of studies that amplified the same 16S rDNA region.

This matrix was imported to DOTUR (Schloss and Handelsman, 2005) and used to determine OTUs and to calculate rarefaction curves.

#### Taxon-based approach

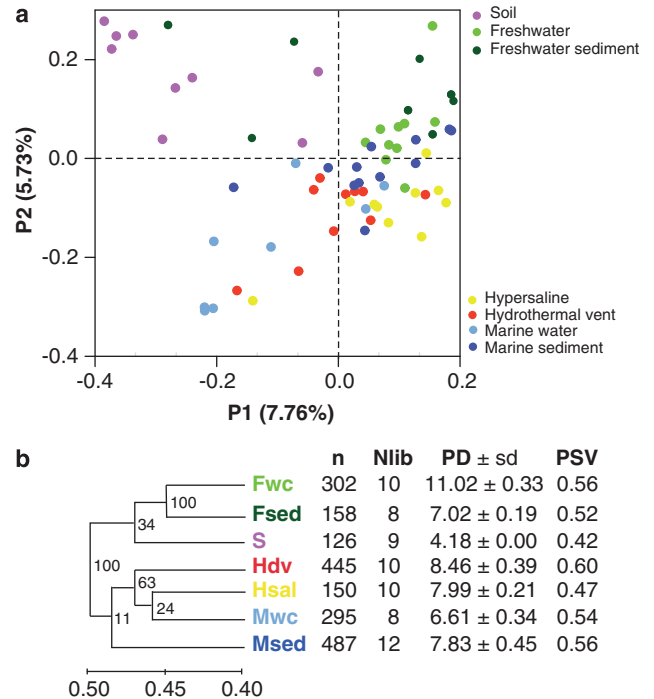
Archaeal lineages were named following the clusters or divisions naming immediately subordinate to the Crenarchaeota or Euryarchaeota phyla and provided by default in the Greengenes tree. However, several sequences seemed not related to any labeled cluster and would have remained unaffiliated at the lineage level. Accordingly, we named four new crenarchaeotal lineages *de novo* as follows: 1.1*d*, 1.1*e*, 1.1*f* and 1.1*g*, and one euryarchaeotal lineage as HV-Fresh (see Figure 4). The HV-Fresh lineage not only contained the already described DHVE3 (Deep Hydrothermal Vent group 3) and HV1 (Hydrothermal Vent group 1), but also a large number of single freshwater sequences. Grouping at a lower phylogenetic level was ruled out because of the high number of archaeal sequences not properly affiliated yet and the poor taxonomic agreement due to the lack of cultured representatives.

Microbes have a great capacity of dispersion and one sequence of any lineage can be retrieved in any habitat by chance. Furthermore, cross-contamination is very possible when sampling at the interface of two habitats (for example, sediment-water column). Hence, to identify archaeal lineages as analogous to the concept of 'indicator species' for each habitat with enough statistical support, we constructed a table of abundances and used the indicator value (IndVal) index, which combines relative abundance and relative frequency of occurrence (Dufrene and Legendre, 1997). A multivariate regression tree was computed with the R package *mvpart* (De'Ath, 2002) in order to represent the relationship between the table of lineage abundances and the environmental matrix.

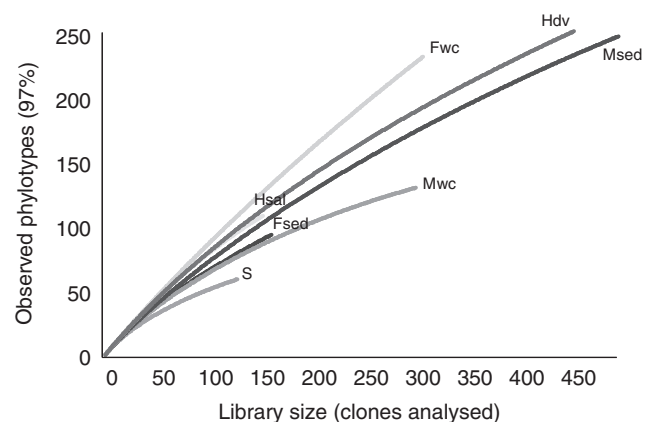
## Results

#### Environmental forces shaping the phylogenetic structure of archaeal community

Natural samples from 67 globally distributed archaeal clone libraries were sorted into an ordination plot according to phylogenetic community similarity (Figure 1a). Habitat classification was a strong structuring factor of the archaeal assemblages ( $R^2 = 0.20$ ,  $P < 0.001$ ) and communities grouped according to their habitat of origin (Figure 1a). Nonsaline environments clearly separated from saline environments (Figures 1a and b), and salinity was the strongest and the only significant environmental factor ( $R^2 = 0.03$ ,  $P = 0.024$ ). The remaining environmental factors explored (that is, temperature, life environment, oxygen concentration and trophic status) were not significant and explained only 8.4% of the total variance from the UniFrac matrix.



**Figure 1** (a) Principal coordinate analysis (PCoA) obtained with the UniFrac distance matrix comparing the 67 libraries summarized in Supplementary Table 1. Principal coordinate 1 (P1) vs principal coordinate 2 (P2) are represented. (b) Hierarchical clustering analysis (UPGMA algorithm with Jackknife supporting values, 126 subsampled sequences, 100 replicates) carried out on the libraries belonging to the seven habitats type previously defined by the PCoA analysis. Distances between clusters are expressed in UniFrac units: a distance of 0 means that two environments are identical, and a distance of 1 means that two environments contain mutually exclusive lineages. The number of sequence (*n*), number of libraries (Nlib), phylogenetic diversity with s.d. (PD ± s.d.) and phylogenetic species variability (PSV) in each habitat is given. S.d. for PSV index was  $< 0.001$  for all habitats.



**Figure 2** Rarefaction curves for archaeal diversity in the seven habitats prospected. OTUs were calculated at a 97% cutoff.

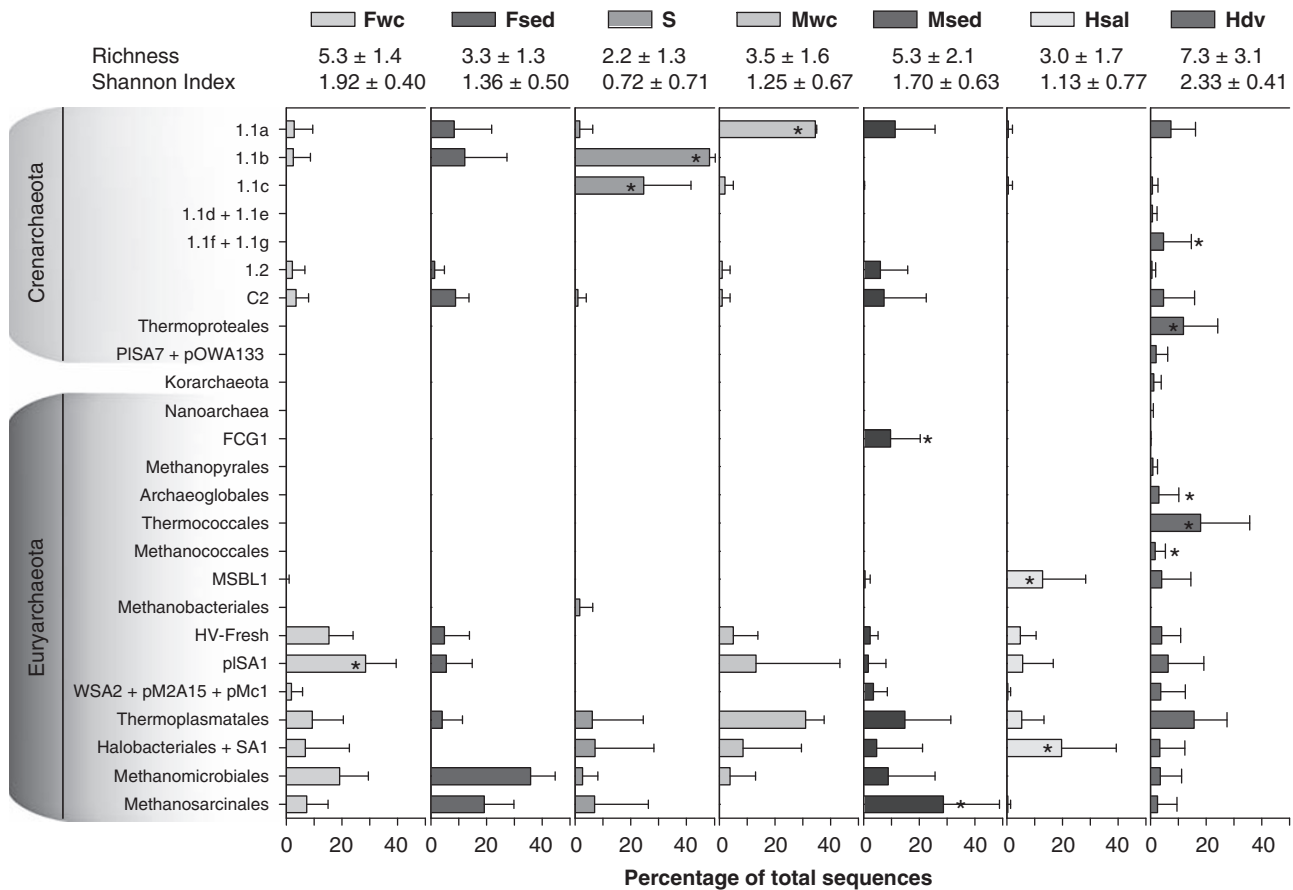
Rarefaction curves (Figure 2) and diversity indices (Figure 1b) were determined for the seven types of habitats. The linear rarefaction curves provided evidences that the archaeal diversity is far from exhaustively sampled, particularly in freshwater,

hydrothermal vent and hypersaline habitats. PD was higher in freshwater plankton (Fwc) and hydrothermal vents (Hdv), whereas soil (S) hold the lowest PD value (Figure 1b). Habitats showed a nonrandom sampling of phylotypes from the phylogeny pool, thereby indicating a significant phylogenetic structure. The mean observed PSV value (0.53) was significantly lower than the null distribution for model 1 (0.75,  $P < 0.05$ ) and for model 2 (0.60,  $P < 0.05$ ). Null model 1 test suggested nonrandom associations between phylotypes among communities, with habitats containing more closely related phylotypes than expected by chance (that is, phylogenetic clustering). The null model 2 suggested that phylotype composition represented nonrandom samples from the phylotypes pool (that is, significant pattern in phylotypes prevalence). Particularly, Hdv, Fwc and Msed habitats showed the highest PSV values (that is, more overdispersed), whereas S and Hsal the lowest (that is, more phylogenetically clustered).

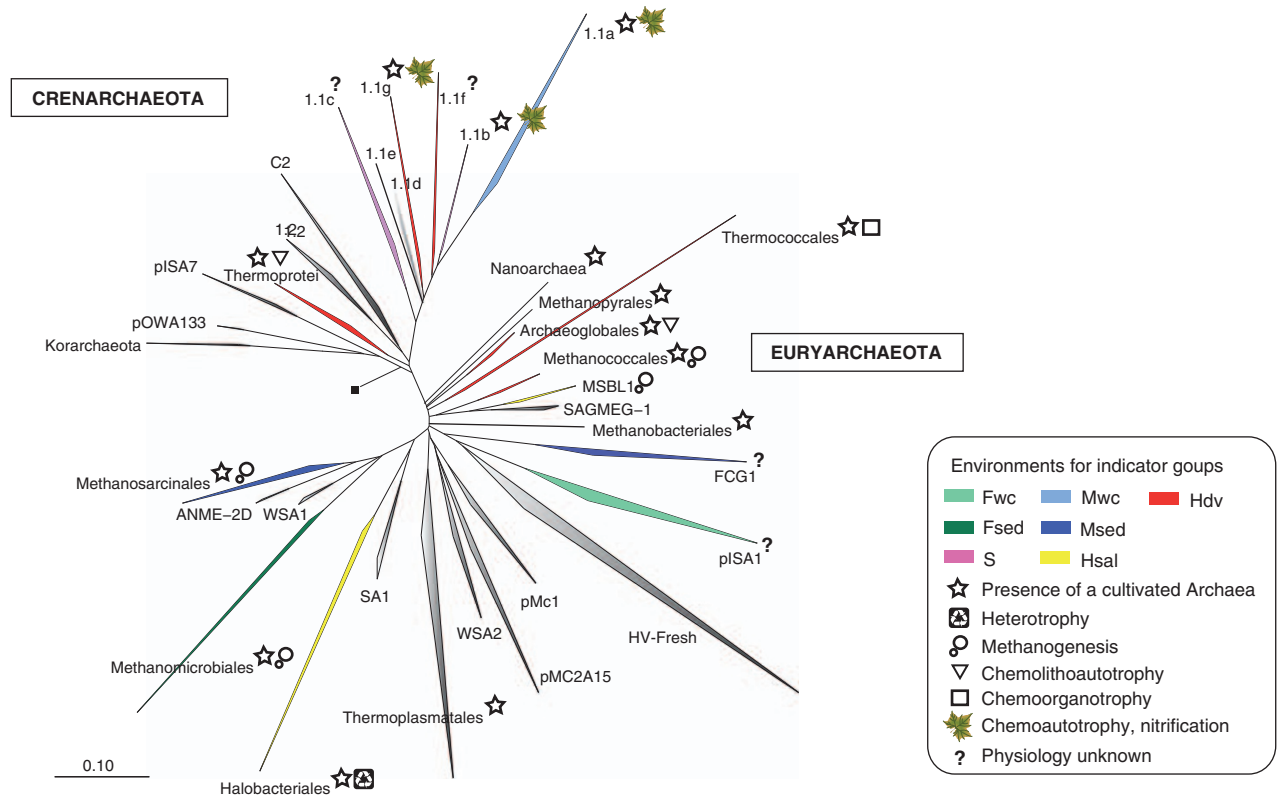
*Identifying indicator lineages and their distribution along gradients*

Both the Shannon index and the richness values (Figure 3 upper part) showed again that hydrothermal

vents (Hdv) and freshwater plankton (Fwc) were the most diverse habitats, whereas soil (S) was the lowest. Overall, 13 out of 25 archaeal lineages showed a significant IndVal ( $P < 0.01$ ) for one single habitat (labeled with asterisk in Figure 3). The 1.1b, FCG1, Thermococcales, and Thermoproteales had high IndVal values (range: 63–90), whereas the remaining lineages showed moderate values (range: 30–49). *Methanomicrobiales* and *Thermoplasmatales* were predominant in freshwater and marine habitats, respectively, but the analysis was not significant for any of them ( $P > 0.01$ ). Freshwater archaeal communities were dominated by the indicator archaeal group pISA1. Soil samples were dominated by the crenarchaeal lineages 1.1b and 1.1c (abundance  $76 \pm 33\%$ ). Conversely, Crenarchaeota were essentially absent from hypersaline samples, where Euryarchaeota from the Halobacteriales + SA1 lineages dominated. Remarkably, almost all the phylogenetic groups were present in hydrothermal vents with some specific groups exclusively found there. Hydrothermal vents also showed by far the largest number of indicator lineages (five lineages), most of them located close to the root of the tree (Figure 4). Curiously, for freshwater sediments none of the lineages were detected as indicator at 0.01 significance



**Figure 3** Relative proportion of archaeal lineages (based on sequence abundance) within each of the seven habitats identified. The number of libraries and sequences used for each habitat is given in Figure 1b. Error bars represent s.d. Asterisks show indicator archaeal lineages at a significance threshold of  $P = 0.01$ . The richness and Shannon diversity index are given in the upper part of the figure for each habitat.



**Figure 4** Phylogenetic archaeal tree based on 16S rRNA gene sequences present in the Greengenes database for the ARB software in January 2008. Sequences were inserted into the original Greengenes tree calculated by maximum parsimony method and provided by default by using parsimony criteria with the Archaea filter excluding highly variable positions. The nomenclature follows the labeled clusters or divisions provided by default in the Greengenes general tree and immediately subordinate to the Crenarchaeota or Euryarchaeota phyla. The black square in the centre indicates rooting to species in the domain Bacteria. Only the physiologies of indicator lineages were represented.

level, though *Methanomicrobiales* became significant at  $P < 0.05$ .

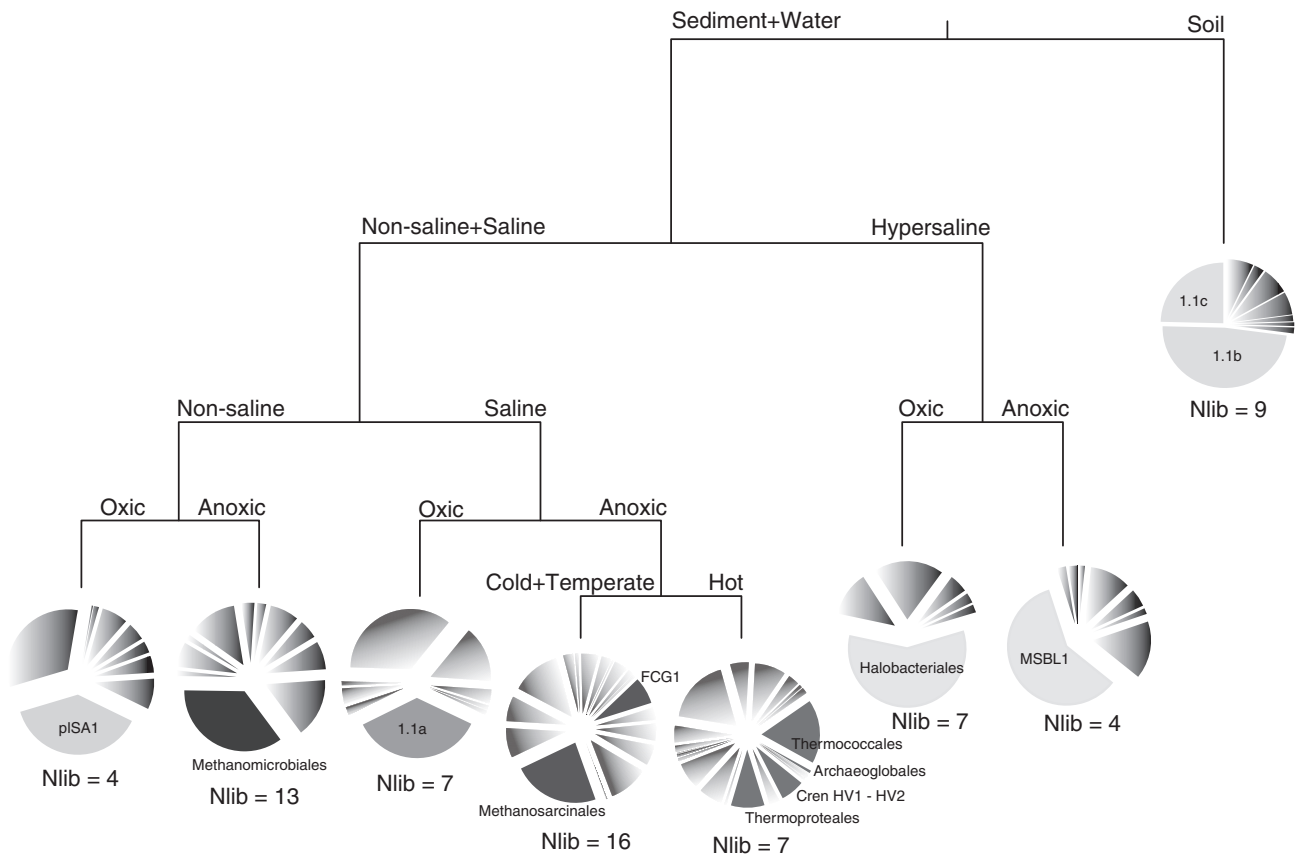
A multivariate regression tree analysis was carried out in order to link the abundance of the lineages and environmental data. The analysis showed an eight-leaf tree ordination (Figure 5) primarily based on life environment (soils vs sediment and plankton), and followed by salinity (hypersaline vs marine and freshwater), oxygen level and temperature. The ordination explained 38.5% of the phylogenetic lineage variance. As previously observed for UniFrac analyses, samples clustered in the leaves of the tree merely in function of their habitat of origin. Nonetheless, some samples from related habitats grouped together forced by other environmental parameters. Thus, anoxia tended to pool together Hdv and Msed (hot- and cold-temperate anoxic marine sites), as well as Fsed and Fwc (sediments and water column from anoxic freshwaters), whereas Hsal environments were separated between oxic and anoxic (Figure 5).

Pie charts in Figure 5 show in detail how the relative abundance of each phylogenetic group contributed to the separation and composition of the leaves. Indicator lineages previously identified by the IndVal index were mainly responsible for the

regression tree topology observed. For example, the crenarchaeotal lineages *1.1b* and *1.1c* largely determined the initial separation soil vs aquatic environments in agreement with the previous analyses that classified these lineages as typical soil inhabitants.

## Discussion

Archaeal ecology derived from cultured representatives (54 cultured species reported so far, and spread in 18 lineages, Schleper *et al.*, 2005) provided for many years a strongly biased view on the diversity and distribution of the third Domain of life in the biosphere. Basically, they were considered extremophiles thriving under severe extremes of environmental gradients. The emergence of culture independent molecular techniques unveiled the ubiquitous distribution of Archaea (see Casamayor and Borrego, 2009; Chaban *et al.*, 2006; Schleper, 2007) and also revealed a hidden PD in the Domain with, to date, up to 49 mostly uncultured lineages (Schleper, 2007; Schleper *et al.*, 2005). Obviously, to gain knowledge on the true ecology of the Domain, all the components should be analyzed as a whole. However, detailed comparative ecological studies



**Figure 5** Multivariate regression tree (MRT) analysis of the interaction between archaeal lineage abundance (in term of sequence number) and environmental parameters. The model explained 38.5% of the variance in the whole data set. Pies under each leaf represent the mean of normalized archaeal lineage abundance for each lineage significantly correlated with environmental parameters.

to fully appreciate the distribution, community patterns and environmental drivers of uncultured Archaea are missing. To fill this gap, our analytical approach revealed for the first time well-defined community patterns along global environmental gradients and habitat types for uncultured Archaea.

Archaeal communities were more similar within habitats than among habitats. This clustered phylogenetic structure (that is, more closely related phylotypes than expected from a random distribution within habitats) is consistent with the concept of habitat filtering (Helmus *et al.*, 2007). Curiously, salinity rather than temperature explained a significant part of these distribution patterns. Salinity was also recently recognized as a key environmental factor globally structuring bacterial communities (Lozupone and Knight, 2007). Other environmental factors such as oxic–anoxic conditions probably also had a significant role structuring the observed patterns. Overall, the two phylogenetically independent domains of life (that is, Archaea and Bacteria) shared similar broad trends, suggesting a commonality in the types of factors that are important for prokaryotes distribution.

The lack of environmental information associated with database sequences was, however, a strong

limitation of meta-analysis and may have hampered a better explanation for the global archaeal patterns observed here. As stated by other authors, it becomes crucial to gain a consensus rationale for measuring and reporting some basic environmental variables in microbial surveys (Robertson *et al.*, 2005, Field *et al.*, 2008). In addition, our approach contained a stochastic component inherent to any environmental study (Sloan *et al.*, 2006). And the 16S rDNA molecule may not be the most suitable marker to target fine biogeographical patterns because of its highly conserved nature. However, several studies agree that this approach remains still valuable as the first step for exploring general ecological patterns in uncultured microorganisms (Reche *et al.*, 2005, 2007, Ramette and Tiedje, 2007 and references therein).

Nonetheless, these limitations were not strong enough to blur the powerful effect of local environmental selection on archaeal diversity. Interestingly, both phylogenetic and taxon-based approaches revealed similar diversity patterns, suggesting that all the phylotypes of a lineage roughly shared the same distribution and probably the same physiology as in the case of methanogens, halophiles, thermophiles and ammonia oxidizers, all of them clustered

in specific functional groups. These observations agree with a recent study (von Mering *et al.*, 2007) that showed a significant correlation between habitat and evolutionary relatedness for microorganisms, even for taxa related at the order level, suggesting that truly adapted specialists acquired their abilities long time ago. To ascertain who were the true archaeal specialists, we use the concept of indicator species borrowed from plant and animal ecology.

We defined as specialists those archaeal lineages that were more frequently represented in most of the sites within a specific habitat, a definition closely related to the concept of indicator species used in ecology (Dufrene and Legendre, 1997). We applied this concept to the whole set of archaeal lineages, unveiling up to 13 lineages with significant IndVal values. We found at least one indicator lineage for each habitat at a significance threshold of  $P=0.01$ , except for freshwater sediments. Thus, this original approach provided a novel ecological support for the sometimes arbitrary nomenclature found in uncultured archaeal clusters, and heavily supported the attributes previously given for crenarchaeotal *1.1a* as the marine planktonic group ('marine plankton group 1', DeLong, 1998), or to the *1.1b* as the soil crenarchaeotal group.

In addition, this approach provided new phylogeographical clues on ecological and evolutionary aspects of the archaeal biology. As acquisition of the essential functions to be permanently adapted to a habitat requires an extended period of time (von Mering *et al.*, 2007), the lack of indicator archaeal lineages in freshwater sediments may indicate a late archaeal colonization. Conversely, hydrothermal vents exhibited by far the highest number of indicator lineages and this may indicate that we were dealing with the earliest habitat colonized by Archaea. Although the thermophilic origin of planktonic Archaea is still a matter of debate (DeLong, 1998; Brochier-Armanet *et al.*, 2008), the presence in hydrothermal vents of representatives from almost all archaeal lineages (especially those at the root of the tree in Figure 4) offers another piece of the puzzle supporting Hdv as the cradle of planktonic Archaea and probably of the origin for the common archaeal ancestor.

The indicator archaeal lineages identified by the IndVal index produced the clustering topology observed in the multivariate regression tree analysis, confirming that these groups were the best-adapted assemblages to the prevailing environmental conditions for each habitat. If environmental forcing selects microorganisms on the basis of their functional capacities, indicator lineages should be, consequently, among the main players in the pivotal ecological functions within the habitat. Good examples are the *Methanomicrobiales* and *Methanosarcinales* (indicator groups for freshwater and marine sediments, respectively), well known as central components for anaerobic organic matter degradation

coupled to methanogenesis in aquatic environments. Haloarchaeales also constitute the most active population for organic matter degradation in hyperhaline environments (for example, Gasol (2004)), whereas in hydrothermal vents chemolithoautotrophs members of the *Archaeoglobales* and the *Thermoprotei* are recognized as key primary producers under anaerobic conditions by coupling oxidation of hydrogen gas with sulfate reduction (Seegerer *et al.*, 1993 and references therein). Furthermore, from the recent cultivation of the autotrophic ammonia oxidizer Crenarchaeota *Nitrosocaldus yellowstonii* (de la Torre, 2008), we can hypothesize a significant role in the nitrogen cycle of Hdv by members of the *1.1g* group. Similarly, Crenarchaeota from the *1.1a* and *1.1b* groups are thought to be important nitrifiers in planktonic marine systems and soils (Francis *et al.*, 2007 and references therein). Finally, 3 out of 13 indicator archaeal lineages (that is, MSBL1, FCG1 and pISA1) did not contain cultivated counterparts or genomic fragments good enough to extract functional information (Figure 4). For the MSBL1 lineage, however, a putative methanogenic metabolism was inferred according to its phylogenetic allocation (van der Wielen, 2005). No physiological information is available so far for the FCG1 lineage (marine sediments) and, particularly, for the pISA1 lineage (characteristic of freshwater habitats) where peculiar fast evolving 16S rRNA gene sequences (long branches) were observed (Figure 4). These three lineages, for which ecological importance has remained unseen to date, deserve further and detailed attention as they represent potential key archaeal groups in term of distribution and ecological processes in their respective habitats.

In the coming future, new genomic tools will offer a wider picture of the archaeal diversity that will probably lead to substantial changes in current archaeal phylogeny (Brochier-Armanet *et al.*, 2008; Robertson *et al.*, 2005; Schleper *et al.*, 2005). A correct positioning of the lineages within the phylogenetic tree topology is a fundamental issue to extract insights into the evolution and metabolic capacities of uncultured Archaea. In essence, both success in bringing into culture more archaeal representatives from a larger range of phyla and a higher sequencing effort are still needed to get a more realistic picture of archaeal diversity and phylogeny. In this context, the diversity analyses reported here offered unexpected views unveiling hydrothermal vents and planktonic freshwater ecosystems as the largest reservoirs of archaeal diversity and, therefore, promising environments for the discovery of new archaeal lineages. This would encourage a new focus on sequencing efforts, as these two habitats are by far less thoroughly sampled than soil or marine habitats (for example, planktonic freshwater sequences only represent ~2.5 % of total archaeal sequences in GenBank). Conversely, archaeal diversity in soils was unexpectedly low even though soil microbial diversity is assumed to be one

of the highest on Earth (Torsvik and Øvreås 2002). As previously observed for soil bacteria (Lozupone and Knight, 2007), the archaeal soil diversity was the result of a high number of closely related phylogenetic lineages rather than different lineages. This is confirmed here by the lowest PSV value indicating a high degree of phylogenetic clustering in soil archaeal assemblages. This peculiar characteristic of soils as compared with aquatic habitats certainly deserves further attention. A first element to consider would be the lowest evolutionary rates observed in soils that could be related to the faculty of microorganisms to enter in dormancy during long stressing periods (for example, winter, desiccation) (von Mering *et al.*, 2007).

Overall, our approach revealed for the first time well-defined global patterns in the distribution of uncultured Archaea with a strong environmental filtering component. Archaeal indicator lineages were identified for specific habitats leading the classification of uncultured Archaea into a more comprehensive and ecological framework. Such lineages appear as good targets in future research for finely depicting the links between ecological drivers and archaeal biology. Emerging patterns will help to guide future research on archaeal biology and ecology.

## Acknowledgements

We are thankful to all the authors who provided valuable data for this work. We also acknowledge anonymous reviewers for valuable feedbacks and constructive comments. This research was supported by grant CRENYC CGL2006-12058 to EOC from the Spanish Ministerio de Educación y Ciencia (MEC), and CONSOLIDER-INGENIO 2010 project GRACCIE CSD2007-00004. JCA benefits from a Juan de la Cierva-MEC postdoctoral fellow, and AB is supported by an FPU-MEC postdoctoral scholarship.

## References

- Auguet JC, Casamayor EO. (2008). A hotspot for cold Crenarchaeota in the neuston of high mountain lakes. *Environ Microbiol* **10**: 1080–1086.
- Bintrim SB, Donohue TJ, Handelsman J, Roberts GP, Goodman RM. (1997). Molecular phylogeny of archaea from soil. *Proc Natl Acad Sci USA* **94**: 277–282.
- Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. (2008). Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Rev Microbiol* **6**: 245–252.
- Casamayor EO, Borrego CM. (2009). Archaea in inland waters. In: Likens G (ed). *Encyclopedia of Inland Waters*, Vol. 3. Academic Press, Elsevier: Oxford, UK, pp 167–181.
- Chaban B, Ng SYM, Jarrell KF. (2006). Archaeal habitats—from the extreme to the ordinary. *J Can Microbiol* **52**: 73–116.
- De'Ath G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology* **83**: 1105–1117.
- de la Torre J, Walker CB, Ingalls AE, Konneke M, Stahl DA. (2008). Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. *Environ Microbiol* **10**: 810–818.
- DeLong E. (1992). Archaea in coastal marine environments. *Proc Natl Acad Sci USA* **89**: 5685–5689.
- DeLong EF. (1998). Everything in moderation: Archaea as 'non-extremophiles'. *Curr Opin Gen Dev* **8**: 649–654.
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.
- Dufrene M, Legendre P. (1997). Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol Monogr* **67**: 345–366.
- Faith DP. (1992). Conservation evaluation and phylogenetic diversity. *Biol Conserv* **61**: 1–10.
- Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P *et al.* (2008). The minimum information about a genome sequence (MIGS) specification. *Nat Biotech* **26**: 541–547.
- Francis CA, Beman JM, Kuypers MMM. (2007). New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *ISME J* **1**: 19–27.
- Fuhrman JA, McCallum K, Davis AA. (1992). Novel major archaeobacterial group from marine plankton. *Nature* **356**: 148–149.
- Gasol JM, Casamayor EO, Join I, Garde K, Gustavson K, Benlloch S *et al.* (2004). Control of heterotrophic prokaryotic abundance and growth rate in hypersaline planktonic environments. *Aquat Microb Ecol* **34**: 193–206.
- Helmus MR, Bland TJ, Williams CK, Ives AR. (2007). Phylogenetic measures of biodiversity. *Am Nat* **169**: E68–E83.
- Kemmel S, Ackerly D, Blomberg S, Cowan P, Helmus M, Webb C. (2008). Picante: tools for integrating phylogenies and ecology. Version 0 4 0, <http://picante.r-forge.r-project.org/>.
- Llirós M, Casamayor EO, Borrego CM. (2008). High archaeal richness in the water column of a freshwater sulphurous karstic lake along an inter-annual study. *FEMS Microbiol Ecology* **66**: 331–342.
- Lozupone C, Hamady M, Knight R. (2006). UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7**.
- Lozupone CA, Knight R. (2007). Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104**: 11436–11440.
- Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucl Acids Res* **32**: 1363–1371.
- McArdle B. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**: 290–297.
- Ochsenreiter T, Selezi D, Quaiser A, Bonch-Osmolovskaya L, Schleper C. (2003). Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR. *Environ Microbiol* **5**: 787–797.
- Oksanen J, Kindt R, Legendre P, O'hara B, Simpson GL, Stevens MHH. (2008). Vegan: community ecology package. version 1 11 14, <http://vegan.r-forge.r-project.org>.
- Ramette A, Tiedje JM. (2007). Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microb Ecol* **53**: 197–207.



- Reche I, Pulido-Villena E, Morales-Baquero R, Casamayor EO. (2005). Does ecosystem size determine aquatic bacterial richness? *Ecology* **86**: 1715–1722.
- Reche I, Pulido-Villena E, Morales-Baquero R, Casamayor EO. (2007). Does ecosystem size determine aquatic bacteria richness? Reply. *Ecology* **88**: 253–255.
- Robertson CE, Harris JK, Spear JR, Pace NR. (2005). Phylogenetic diversity and ecology of environmental Archaea. *Curr Opin Microbiol* **8**: 638–642.
- Schleper C. (2007). Diversity of uncultivated Archaea: perspectives from microbial ecology and metagenomics. In: Garret RA and Klenk HP (eds). *Archaea: Evolution, Physiology and Molecular Biology*. Blackwell Publishing: Oxford, UK, pp 39–53.
- Schleper C, Jurgens G, Jonuscheit M. (2005). Genomic studies of uncultivated archaea. *Nature Rev Microbiol* **3**: 479–488.
- Schloss PD, Handelsman J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71**: 1501–1506.
- Seegerer AH, Burggraf S, Fiala G, Huber G, Huber R, Pley U *et al*. (1993). Life in hot springs and hydrothermal vents. *Orig Life Evol Biosph*, pp 77–90.
- Shaw A. (2008). It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10**: 2200–2210.
- Sloan WT, Lunn M, Woodcock S, Head IM, Nee S, Curtis TP. (2006). Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ Microbiol* **8**: 732–740.
- Torsvik V, Øvreås L. (2002). Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol* **5**: 240–245.
- van der Wielen P. (2005). The enigma of prokaryotic life in deep hypersaline anoxic basins. *Science* **307**: 121–123.
- von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ *et al*. (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**: 1126–1130.

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