The ISME Journal (2008) 2, 1052–1065
© 2008 International Society for Microbial Ecology All rights reserved 1751-7362/08 $32.00
www.nature.com/ismej

ORIGINAL ARTICLE

Genomic patterns of recombination, clonal divergence and environment in marine microbial populations

Konstantinos T Konstantinidis[1,2,3] and Edward F DeLong[1,2]

[1]*Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA and* [2]*Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*

Microorganisms represent the largest reservoir of biodiversity on Earth, both in numbers and total genetic diversity, but it remains unclear whether this biodiversity is organized in discrete units that correspond to ecologically coherent species. To further explore this question, we examined patterns of genomic diversity in sympatric microbial populations. Analyses of a total of ~200 Mb of microbial community genomic DNA sequence recovered from 4000 m depth in the Pacific Ocean revealed discrete sequence-defined populations of Bacteria and Archaea, with intrapopulation genomic sequence divergence ranging from ~1% to ~6%. The populations appeared to be maintained, at least in part, by intrapopulation genetic exchange (homologous recombination), although the frequency of recombination was estimated to be about three times lower than that observed previously in thermoacidophilic archaeal biofilm populations. Furthermore, the genotypes of a given population were clearly distinguishable from their closest co-occurring relatives based on their relative abundance *in situ*. The genetic distinctiveness and the matching sympatric abundances imply that these genotypes share similar ecophysiological properties, and therefore may represent fundamental units of microbial diversity in the deep sea. Comparisons to surface-dwelling relatives of the Sargasso Sea revealed that distinct sequence-based clusters were not always detectable, presumably due to environmental variations, further underscoring the important relationship between environmental contexts and genetic mechanisms, which together shape and sustain microbial population structure.

The ISME Journal (2008) **2,** 1052–1065; doi:10.1038/ismej.2008.62; published online 26 June 2008
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: metagenomics; marine; population; recombination; speciation

Introduction

Phylogenetic surveys of small-subunit ribosomal RNA genes (SSU rRNA) (Giovannoni *et al.*, 1990; Rappe and Giovannoni, 2003; Acinas *et al.*, 2004) and, to a lesser extent, single protein-coding genes (Palys *et al.*, 2000) have frequently recovered distinct sequence-based clades from the environment revealing one level of genetic structural organization of microbial diversity. Several explanations based on recombination frequency (Feil *et al.*, 2001; Spratt *et al.*, 2001) or population sweeps caused by periodic natural selection (Cohan, 2002; Acinas *et al.*, 2004) have been advanced to explain the maintenance of these patterns of ribotype diversity. Alternative explanations such as population bottlenecks and random birth/extinction are less favorable and probably applicable to more restricted microbial groups and habitats, such as the vertically transmitted microbial pathogens (Moran, 2007), compared to the groups surveyed in these previous studies. In any case, none of these theories has yet gained universal support for environmental microorganisms, in part due to lack of experimental data and the limited sub-species level resolution afforded by SSU rRNA sequence data.

In addition, emerging findings from higher resolution environmental genomic studies such as whole-genome shotgun (WGS) sequence analyses have recently revealed a considerable amount of genomic complexity within naturally occurring microbial communities. This genetic complexity has largely prevented the assembly of whole genomes from

Correspondence: EF DeLong, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 48-427 MIT, 15 Vassar Street, Cambridge, MA, USA.
E-mail: delong@mit.edu
[3]Current address: School of Civil and Environmental Engineering and School of Biology, Georgia Institute of Technology, Atlanta, GA, USA

Received 3 March 2008; revised 26 May 2008; accepted 26 May 2008; published online 26 June 2008

most microbial populations (Venter *et al.*, 2004; Tringe *et al.*, 2005; Rusch *et al.*, 2007), with a few exceptions from very simple communities (Tyson *et al.*, 2004; Garcia Martin *et al.*, 2006; Hallam *et al.*, 2006a). These data have not fully resolved whether a continuum of sequence variants (which would account for the incomplete assemblies) or discrete sequence-based clades most accurately typify natural microbial population diversity. Further, comparisons of isolates, either at the whole-genome level (Konstantinidis and Tiedje, 2005) or at multiple loci in the genome (for example, Multi Locus Sequence Analysis) (Hanage *et al.*, 2005), have frequently uncovered indiscrete ('fuzzy') microbial clades and species. Finally, genetic isolation between otherwise discrete microbial populations can be potentially reduced by the well-documented pervasiveness of horizontal gene transfer (HGT) (Lawrence, 2002; Sheppard *et al.*, 2008).

However, most known rRNA-based clades and clades supported by Multi Locus Sequence Analysis data are typically comprised of genotypes recovered from different populations and habitats (Feil *et al.*, 2001; Hanage *et al.*, 2005; Thompson *et al.*, 2005; Coleman *et al.*, 2006; Rusch *et al.*, 2007). Thus, the extent of ecological distinctiveness and genomic adaptation to the environment of isolation of the organisms under study remain speculative, severely confounding population genetic analyses and comparisons (Konstantinidis *et al.*, 2006). These limitations also apply to a recent WGS study (Rusch *et al.*, 2007), which reported ribotype genetic diversity in microbial populations from randomly sampled planktonic microbial communities. To further understand the drivers of genetic variation in microbial populations, within-community genomic variation should be analyzed, focusing ideally on abundant populations that are less likely to represent transient and/or allochthonous members of the community. To this end, we analyzed genomic variation in a 4000-m deep sample from the Pacific Ocean by examining WGS coverage patterns on large genomic scaffolds of predominant archaeal and bacterial populations.

Materials and methods

Fosmid clone sequencing and assembling

Fosmid (36 kb inserts, on average) and small-insert (1–2 kb long) plasmid libraries were constructed and sequenced from a 4000 m deep planktonic microbial assemblage as described previously, based on Sanger sequencing technology (DeLong *et al.*, 2006; Hallam *et al.*, 2006a). The assembly of fosmid sequences into contigs was performed, essentially as described previously (Hallam *et al.*, 2006a), requiring at least 5-kb long overlap of 95% nucleotide (nt) identity or higher to merge two fosmids or contigs (high stringency). Only two of the total 50 fully sequenced crenarchaeal fosmids originating from 4000 m depth in the Pacific Ocean did not form contigs with other crenarchaeal fosmids (that is, they were singletons). The six largest and more robust contigs, each consisting of at least three fosmid clones and ranging in length between 60 and 120 kb, were arbitrarily linked to provide a 450 kb scaffold presenting a snapshot of the mosaic genome of the deep-sea Crenarchaea. The average nt identity, approximate start and end positions, and the length of the total overlap of the 25 fosmids constituting the scaffold relatively to the scaffold are provided in Supplementary Table S2; all crenarchaeal fosmids have been deposited in GenBank. No attempt was made to close the crenarchaeal genome because the coverage by our fosmid or small-insert libraries was not adequate to allow genome closing or even accurate genome size estimation. Nonetheless, the high frequency of overlapping fosmids among the 50 randomly selected crenarchaeal fosmids available, the total unique sequence space contained in these 50 fosmids, that is, ~1 Mb, which is about half of their total sequence space, and the coverage of the *Cenarchaeum symbiosum* genome by fosmid and small-insert libraries (Supplementary Figure S1) clearly indicated that the genome size of the planktonic crenarchaea is substantially smaller than that of *C. symbiosum* (~2.1 Mb).

Sequenced reads from the small-insert library were mapped on a reference sequence such as a fosmid using the blastn (nt level) algorithm version 2.2.12 (Altschul *et al.*, 1997) with the following settings: $X = 150$ (drop-off value for gapped alignment), $q = -1$ (penalty for nt mismatch) and $F = \mathrm{F}$ (filter for repeated sequences); the rest of the parameters were at default settings with a minimum cutoff for a match of at least 500 aligned bases (of ~870 in total, on average). These settings can more robustly detect distantly related sequences compared to the default settings. The nt identities were obtained directly from the blastn output for the aligned region.

Identifying crenarchaeal sequences in small-insert or fosmid end-sequence data

(I) Previously published Hawaii Ocean Time fosmid end sequences: all end sequences in each of the seven available fosmid libraries (~10 000 sequences per library) (DeLong *et al.*, 2006) were searched against the crenarchaeal genomic scaffold from 4000 m depth. Crenarchaeal sequences and sequences representing close relatives were defined as those sequences that had a match of at least 50% nt identity over at least 500 bp against the scaffold and their sister sequence also matched the scaffold at the same cutoff. A number of unrelated non-crenarchaeal sequences might have also passed the cutoff used; however, this number is presumably relatively small because Crenarchaea were typically very genetically distinct from any other population in the same microbial community (see Figure 2a).

(II) Sargasso Sea metagenome: crenarchaeal sequences were identified as the previously assembled contigs (Venter *et al.*, 2004) that were highly homologous and syntenic to the deep-sea crenarchaeal genomic scaffold, similar to how crenarchaeal fosmid sequences were identified using the *C. symbiosum* (Hallam *et al.*, 2006a) as reference genome (see also Results). Visual inspection of the contigs was necessary to remove (rare) cases of assembly errors such as chimeric contigs, that is, contigs that were most likely composed of sequences representing another population in addition to the natural planktonic Crenarchaea (data not shown).

Visual inspection of recombination events

To assess recombination, it was first necessary to normalize for the variation in the sequencing error rate among the three small-insert WGS data sets used in this study, that is, the deep-sea Pacific Ocean (this study), the surface Sargasso Sea (Venter *et al.*, 2004) and the Acid Mine Drainage (AMD) biofilm community (Tyson *et al.*, 2004). For this, the raw WGS data were cleaned and trimmed using the same Q15 quality cutoff for base calling with the Phrap-Phred package (Green, 2006). Trimmed WGS sequences were subsequently aligned against a genomic sequence or a scaffold using the Sequencer v4.5 (Gene Codes, Ann Arbor, MI, USA). Visual inspection of the single-nucleotide polymorphism pattern among overlapping sequences for potential homologous recombination events was performed as reported previously (Tyson *et al.*, 2004). The RDP version 3 β05 program package (Martin *et al.*, 2005) was also used to identify potential recombination areas, which were subsequently visually verified.

Recombination detection using GARD

Genetic algorithm for recombination detection (GARD) (Kosakovsky Pond *et al.*, 2006b) was run with default settings (first running the GARD.bf module followed by the RecombinationProcessor_.bf one), using the HKY model for nt substitution for every alignment evaluated. The alignments were built using the following procedure: a reference scaffold (for example, deep-sea Crenarchaea) or a genomic sequence (for example, *Ferroplasma*) was cut into 800-bp long consecutive fragments, and these fragments were queried against the trimmed small-insert WGS or fosmid sequences for fully overlapping sequences of at least 90% nt identity to the reference fragment. Sequences that were at maximum 10 bp shorter at either end (but fully overlapping in their remaining length) were also included, after filling the missing bases with Ns (ambiguous sequence positions). Sequences longer than the 800-bp long fragments were trimmed to the exact 800-bp long overlap. All sequences for each fragment were finally aligned with ClustalW (Thompson *et al.*, 1994), and the alignment was analyzed by the GARD algorithm. Only multiple alignments of at least four sequences were evaluated. Evaluating shorter alignments, which would have increased the number of alignments available for evaluation, was not attempted as the sensitivity of GARD is decreasing with shorter sequences (SL Kosakovsky Pond, personal communication). For the deep-sea small-insert library, which has an average insert size of 1.5 kb (that is, the sister reads typically overlapped by about 400 bp; average read length was ∼870 bp after trimming), the trimmed sister reads were assembled in one long sequence prior to the search with the reference fragments. The insert size for the Sargasso Sea and AMD libraries is larger, about 4 kb long (that is, the two sister reads are, typically, 2–2.5 kb apart). Given also that the trimmed sequencing reads were, on average, ∼850 bp long, these data sets did not produce enough alignments to evaluate with the above strategy; for example, there were not many fully overlapping reads at the standards used. For these data sets, the reference genomes were cut into 400-bp long consecutive fragments, which were then searched against the full WGS data set as described above for 800-bp long fragments. The final sequences evaluated with GARD were composed of the artificial fusion of the sister reads that matched two 400 bp fragments separated by 2–2.5 kb in the genomic sequence (that is, final alignment was 800 bp long). Fragments, either 800 bp long or 400 bp long, typically contained one protein-coding gene, whose annotation was obtained from the original annotation of the genomic sequence or after manual inspection of the results of tblastx (protein level) searches with the fragment sequence against GenBank.

LDhat analysis

The LDhat software (McVean *et al.*, 2002) was run on 400-bp long gene alignments composed of fully overlapping WGS sequences representing the *Ferroplasma* or the deep-sea Crenarchaea natural population. Alignments were produced as described above for the GARD analysis. The genes evaluated were selected at random provided that they were single-copy, not hypothetical or mobile in function and had preferably about 20 × coverage in the WGS data set. Genes with substantially higher coverage than 25 ×, which met the requirements above and, particularly, the prerequisite for fully overlapping 400-bp long sequences, were not found, whereas genes with lower coverage than 10 × were excluded from the analysis as the sensitivity of LDhat is decreased with a low number of sequences in the alignment (G McVean, personal communication). LDhat was run with default settings, including all polymorphic nt positions and running the grid with maximum values allowed, that is, max. value of 4Ner to estimate = 100; the number of points to estimate for 4Ner = 201. θ values were calculated using the Watterson estimate for θ per site as calculate by LDhat. The LDhat estimates of ρ are per fragment (not per site), thus, the ρ/θ ratio per site for each gene was derived from the equation $[(\rho/400)/\theta]/2$. Division by 2 in the latter equation was necessary as the LDhat program assumes a eukaryotic model of bidirectional recombination, and the prokaryotic system studied here is more likely a unidirectional transfer similar to a gene conversion.

Phylogenetic analysis

Phylogenetic analysis of overlapping small-insert WGS and fosmid sequences was performed as follows: sequences were aligned using the ClustalW software (Thompson *et al.*, 1994) and phylogenetic trees were built from the alignment using the Maximum Likelihood or the Neighbor-Joining algorithms as implemented in the Phylip package (Felsenstein, 2004).

Results

Crenarchaeal sequence identification and sequence randomness

A combination of intermediate-sized DNA fragments (fosmid clones, ~36 kb on average) and WGS data (insert size ~1.5 kb) recovered from the same DNA sample from 4000 m depth in the Pacific Ocean was analyzed, focusing initially on an abundant autotrophic nitrifying archaeal group, the planktonic Crenarchaea (DeLong *et al.*, 1994; Karner *et al.*, 2001; Hallam *et al.*, 2006b; Lam *et al.*, 2007). The fosmid clones analyzed represent a subset of the fosmid library that was end-sequenced and reported previously as part of the Hawaii Ocean Time Series metagenome survey (DeLong *et al.*, 2006). To identify the crenarchaeal fosmid clones in the latter library, all available end sequences of the library (DeLong *et al.*, 2006) were searched against all published microbial-sequenced genomes (as of end of 2005) using the blastx (protein level) algorithm (Altschul *et al.*, 1997). A total of 50 fosmids, whose end sequences had a best match against the genome of *C. symbiosum* (Hallam *et al.*, 2006a), a marine crenarchaeal symbiont of animal marine sponges, were selected for complete sequencing.

Comparisons of the complete fosmid sequences against the *C. symbiosum* genome suggested that these 50 fosmids were indeed representative of pelagic Crenarchaea. A typical crenarchaeal fosmid had the majority and at least 20 out of an average of ~40 of its genes shared with *C. symbiosum*, and the amino-acid identities of the shared genes averaged ~60, ±10%. In contrast, a collection of 40 non-crenarcheal fosmids (selected independently and at random) had very few (<10) genes shared, of substantially lower amino-acid identity (<40%), with *C. symbiosum* (see Supplementary Table S1 for an example). Almost all of the putative crenarchaeal fosmids overlapped at high (>95%) nt identity over at least 5 kb of their sequences (Supplementary Table S2) and assembled into several large and robust contigs of, typically, 60–100 kb in size. These results further indicated that the fosmids were derived from very closely related organisms and were not representative of HGT events in unrelated genetic backgrounds (or such HGT events must had been very recent (to show >95% nt identity) and involved fragments much larger than 40 kb in size to be missed by our analysis, which is unlikely). Moreover, the Crenarchaea at 4000 m were genetically distinct from any other member of the indigenous microbial community (see below and Figure 2a), which greatly facilitated the accurate and robust identification of crenarchaeal sequences in our random community libraries. Finally, the randomness of the fosmid clones selected for sequencing is evident by a nearly complete and uniform coverage of the *C. symbiosum* genome by the fosmid sequences. For example, the parts of genome shared by *C. symbiosum* and planktonic crenarchaea had on average 2 × coverage by fosmids, with an s.d. of 1.17 (Supplementary Figure S1).

In summary, the 50 sequenced fosmid clones selected for further analyses were representative of the indigenous crenarchaeal population at 4000 m depth, and provided sampling at different parts of the genome in 50 (presumably) different crenarchaeal cells.

Discrete sequence-based clusters characterize many marine microbial populations

Nucleotide diversity analysis of the overlapping sections between these 50 fosmid sequences showed that the crenarchaeal population was dominated by genotypes sharing an average nt identity (ANI; calculated as described in the caption of Supplementary Table S2) between 94.5% and 100%. A smaller number of more divergent genotypes was syntenic, with a very small, if any, number of indels representing typically <5% of the total sequence of the fosmid clone (see also Figure 6), with those predominant genotypes, but shared only <~90% ANI with them (Figure 1a). The latter fosmids most likely represented divergent genotypes rather than horizontally transferred genomic islands, as they shared a large overlap with the scaffold, and the nt identity of the overlap was uniform.

To evaluate the relative *in situ* abundance of the crenarchaeal genotypes, the 50 fosmid sequences were queried against a total of ~197 Mb of WGS data (101 176 clones in total) from the same DNA sample, which provided for a much more comprehensive and unbiased data set due to its size and (presumably) fewer cloning biases (see also the Discussion section). The number of highly related (⩾98% nt identity) WGS reads assembling to each fosmid was used to estimate the relative abundance

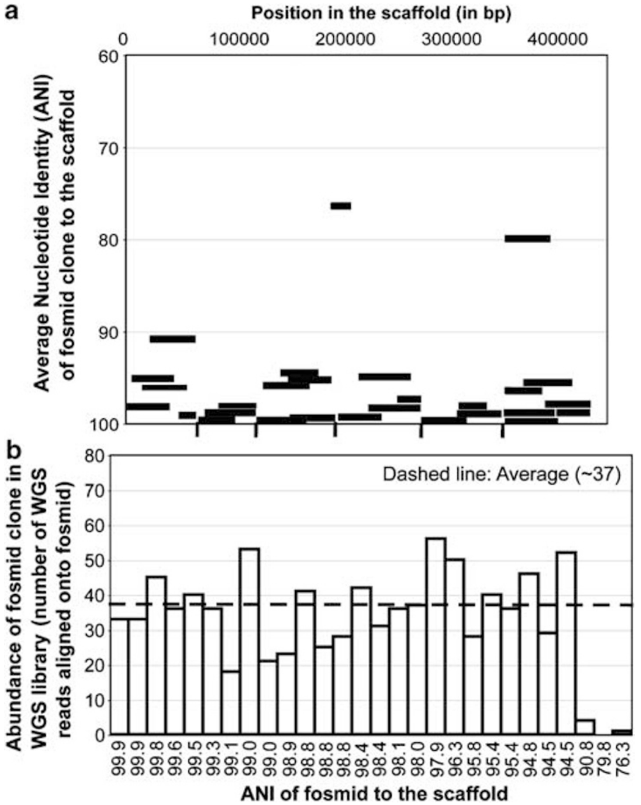

Figure 1 Population structure of planktonic Crenarchaea from 4000 m depth in the Pacific Ocean based on fosmid clones. (**a**) Each line represents an individual fosmid clone sequence mapped against the crenarchaeal consensus scaffold of the most-abundant genotype variants (that is, those represented by fosmid clones showing >94% ANI among themselves). The scaffold is 450-kb long and consists of six arbitrarily linked robust contigs assembled from overlapping fosmid sequences, with the break points between the six sequences denoted by the vertical lines along the *x* axis. (**b**) Each bar represents an individual fosmid clone and shows the number of WGS reads that assemble onto the fosmid clone sequence (that is, the coverage) at 98% nt identity cutoff (*y* axis) plotted against the average nt identity of the fosmid clone against the consensus scaffold (*x* axis). Fosmids are ordered according to decreasing identity to the scaffold on *x* axis. The dashed line represents the average coverage of the fosmids showing only >94% ANI to the scaffold. The analysis shows that the latter fosmids show comparable coverage by WGS reads in contrast to the more-divergent fosmids (<91% ANI to the scaffold), which show significantly lower coverage. Note that some (presumably small) variation in the coverage is expected as an effect of technical stochasticity in constructing and sequencing these random libraries. See Materials and methods for details; all underlying data are provided in Supplementary Table S2. ANI, average nucleotide identity; Nt, nucleotide; WGS, whole-genome shotgun.

of the corresponding genotype and its closest (showing >99% ANI) congeners. The 98% nt identity cutoff was also used to accommodate for the sequencing errors in WGS reads, which was calculated to be 2% at maximum (these were single-pass sequencing reads, trimmed at low stringency to provide for longer sequences, which was more important than sequence accuracy for the previous analysis. In contrast, fosmid sequences had substantially lower sequencing errors because they were sequenced at ~10× coverage instead). The most divergent genotypes represented very minor populations as those fosmids were nearly devoid of matching WGS reads compared to the predominant genotypes that averaged ~37 WGS reads per fosmid (that is, ~1× coverage at 98% nt identity cutoff; Figure 1b). The latter genotypes therefore appear ecologically differentiated from the rare divergent genotypes, given that under identical environmental conditions, they were significantly more abundant. The comparable abundances among the most-related genotypes indicate that the genotypes may share similar ecophysiological properties, which is reflected in their cohesiveness as a population.

To further assess the population structure of the deep-dwelling Crenarchaea, a 450-kb long scaffold, assembled from fosmids representing the consensus genome of the most abundant crenarchaeal variants, was queried against the WGS data at relaxed stringency ($\geqslant$50% nt identity). These comparisons indicated that the predominant deep-sea crenarchaeal population was genetically distinct from any other population sampled within the community as only ~200 WGS reads were found that shared a nt identity between 50% to ~85% with the scaffold. In contrast, ~2000 WGS reads (providing for ~4× coverage of the scaffold), representing the predominant genotypes that were identified in the fosmid-based analysis, assembled evenly across the scaffold in the ~85–100% nt identity range (Figure 2a, black histogram). Further, analysis of the genes contained in the few matching WGS reads in the 50–85% nt identity range indicated that the majority of these WGS were attributable to highly conserved—at the sequence level—housekeeping genes such as ribosomal rRNA genes, ribosomal proteins and polymerases, and—relatively old—HGT events from/to unrelated organisms rather than divergent crenarchaeal genotypes (see also Analyses of whole-genome sequences of isolates and Figure 3a). Even if some of the latter reads represented divergent genotypes (similar to those identified in the fosmid-based analysis), these would be at least 10-fold less abundant than the predominant genotypes according to the WGS-based analysis (<200 vs ~2000 matching WGS reads). Thus, the discreetness of the predominant genotypes would be evident at the quantitative abundance level in the latter case.

The intrapopulation diversity based on WGS data (Figure 2a) appeared higher than the diversity based on the fosmid sequences (Figure 1a) (~10% vs ~5% nt sequence diversity, respectively). This difference was not attributable to the fact that the scaffold was composed of fosmid sequences because the ANI values among the fosmids were only 0.8% lower, on average, compared to the ANI values of the fosmids against the scaffold. Instead, the difference was largely explained by the higher sequencing errors in our unassembled WGS data and the fact that shorter sequences show a larger spread around

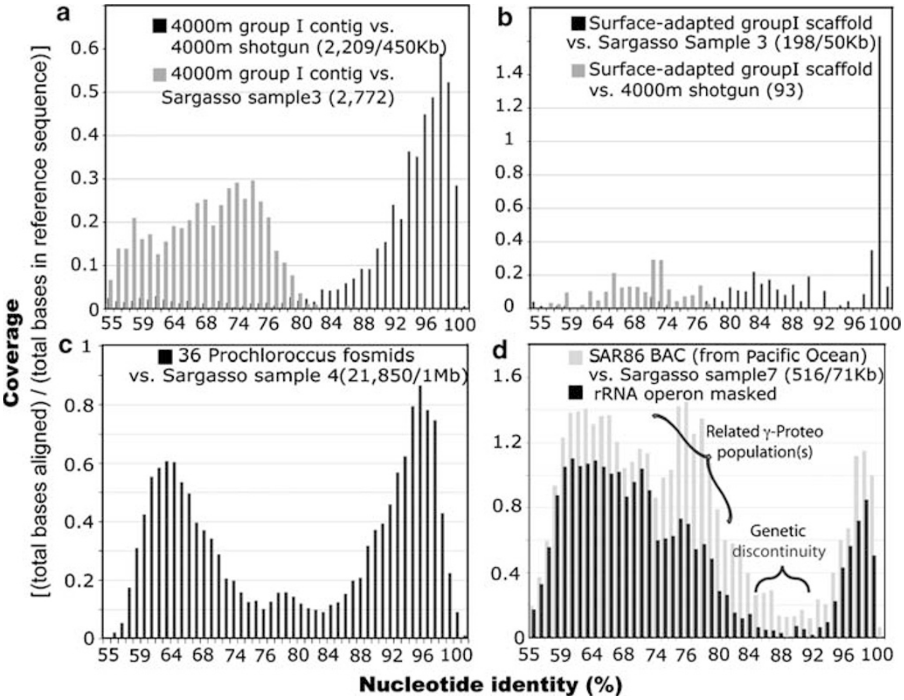

Figure 2 Populations of marine microorganisms form discrete sequence-based populations. The histogram indicates the coverage of a reference sequence by small-insert shotgun reads (*y* axes) per unit of nt identity (*x* axes). For example, there are ~330 reads in the 4000-m WGS library that assembled onto the 450-kb long crenarchaeal consensus scaffold with exact 97–97.99% nt identity over ~800 (on average) aligned nt bases each, which gives a (330 × 800)/450 000 = ~0.59 × coverage at 97% nt identity level (**a**, highest black bar). The total number of reads yielding the observed coverage and the total length of each reference sequence are shown in parentheses. The coverage is normalized to the length of the reference sequence used to make coverage directly comparable between the four panels and representative of the relative *in situ* abundances of the corresponding populations. The reference sequences were (**a**) 450-kb long crenarchaeal consensus scaffold from 4000 m, (**b**) first 40 kb of scaffold no. 44893849 from Sargasso Sea assembled metagenome (Venter *et al*., 2004), (**c**) concatenated sequence of 36 fosmid sequences (Coleman *et al*., 2006), (**d**) BAC clone, GenBank accession AY619685.1 (Sabehi *et al*., 2004), before (gray histogram) and after (black histogram) masking its rRNA operon.

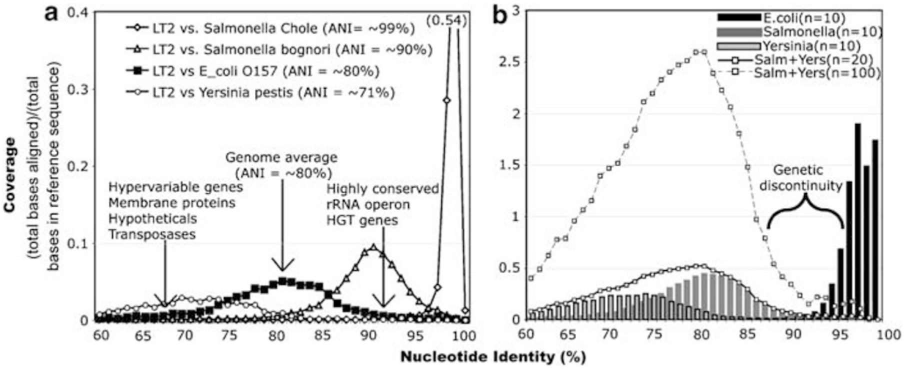

Figure 3 Simulated analysis of genomes of isolates identifies discrete sequence-based clusters. (**a**) Pairwise comparisons: a query genome was cut in 1-kb long consecutive fragments (to simulate WGS reads), which were subsequently queried against the reference whole-genome sequence of *Salmonella typhimurium* strain LT2. Figure shows the coverage (*y* axis) of the reference genome by the fragments of the query genome (graph legend) plotted against the nt identity of the matching fragments (*x* axis), similar to Figure 2. Note that the dispersion of the identities of shared fragments around the genome average (ANI, graph legend) follows a normal distribution and is tighter with smaller genetic distance between the two genomes compared. Thus, even very highly related genomes such as strain bognori vs strain Chole yielded discrete coverage plots. The classes of genes deviating the most from the genome average in terms of their degree of sequence conservation (outliers) are denoted in the *Salmonella* vs *Escherichia coli* comparison (solid squares). (**b**) Multiple genome comparison: all 1-kb long fragments of 10 available genomes of *Yersinia* (70–75% ANI to the reference genome, light-gray histogram), 10 *Salmonella* (80–83% ANI, dark-gray histogram) and 10 *E. coli* genomes (94–100% ANI, black histogram) were queried against the whole-genome sequence of *E. coli* strain Sakai. The *E.coli* genomes formed a discernible cluster in all cases, even when the *Salmonella* and *Yersinia* genomes were combined (solid line) and when each of the combined genomes was first amplified 10 times (dashed lines). All genomes used were obtained from GenBank. Similar results were obtained for other important bacterial groups with several sequenced representatives (data not shown). ANI, average nucleotide identity; WGS, whole-genome shotgun.

the genome-average ANI value compared to longer sequences or the whole genome (see also Figure 3a). Therefore, the actual intrapopulation diversity was presumably better-reflected by the fosmid rather than the WGS data. Further analyses indicated that the deep-dwelling population was related to, but genetically distinct from, relatives dwelling in other habitats, including shallow-water relatives found in the Sargasso Sea (Venter *et al.*, 2004) (Figure 2a, gray histogram) and a sponge-associated crenarchaeal population (Hallam *et al.*, 2006a) (data not shown).

A similar approach was used to assess the representation of other bacterioplankton genotypes represented by large genome fragments (BAC or fosmid clones) or whole-genome sequences from cultivated isolates. Other abundant marine microbial groups were also organized as discrete populations similar to those observed in deep-sea Crenarchaea. For example, SAR86-like γ-proteobacteria in Sargasso Sea populations (Venter *et al.*, 2004) shared 95–100% ANI with an 'SAR86' reference sequence (Sabehi *et al.*, 2004). Unlike deep-sea crenarchaeal populations, however, at least one other relatively abundant but distinct co-occurring SAR86-like population was present in Sargasso Sea surface waters, sharing ∼60–70% ANI with the same SAR86 reference sequence (Figure 2d). Comparable results were also observed for whole-genome reference sequences from *Prochlorococcus* (Coleman *et al.*, 2006) (Figure 2c) and *Pelagibacter* (Giovannoni *et al.*, 1990) (Supplementary Figure S2) in the Sargasso Sea data (Venter *et al.*, 2004) as well as for fosmid clones of other abundant bacterial members of the δ-proteobacteria, Chloroflexi and Planctomycetes in the 4000 m data. These data are not shown because the exact taxonomic affiliations of these as-yet 'unculturable' groups remain elusive as their fosmids typically lacked a 16S rRNA gene; the corresponding fosmid clones have been deposited to GenBank and are also available from the authors upon request. The diversity encompassed within each identified population varied, but never exceeded ∼10% ANI (WGS-based results), depending on the specific taxon examined and reflecting the composition of the corresponding population. For example, the *Prochlorococcus* population appeared more genetically diverse compared to the SAR86-like population in the Sargasso Sea (90–100% vs 95–100% nt diversity among the WGS reads representing each population, respectively; compare panel c with d in Figure 2). Although the number of clones representing the latter bacterial groups was too small to permit the full-scale analyses performed for Crenarchaea, the similarity in the coverage plots observed indicated that the patterns of diversity observed within Crenarchaea may be more broadly applicable to other planktonic microbial groups, including surface-dwelling ones.

Consistent with the findings based on the coverage of reference sequences by WGS data, phylogenetic analysis of a small subset of randomly selected fully overlapping WGS sequences also indicated discrete sequence-based populations for the deep-sea Crenarchaea and the surface-dwelling *Prochlorococcus* (Supplementary Figures S3 and S4). It is important to note, however, that a phylogenetic approach to address the same issues at the whole-genome level, and to the same extent addressed by our approach with the reference sequences, is not currently possible. This is due to the low coverage of the natural community by the WGS library, which does not allow for robust multiple alignments of sufficient length to be built. For example, Crenarchaea were the most abundant members in 4000-m deep community based on the coverage of fosmids by WGS reads; yet the coverage of their genome was, on average, 4 ×. This translates to (an average of) four sequences per 800-bp long alignment at maximum. That is, when the WGS reads were fully overlapping for their entire length, which rarely was the case due to the randomness of our plasmid libraries. The same limitation applies to the recently published large Global Ocean Survey (GOS) data set (Rusch *et al.*, 2007), as the coverage of each natural community sampled (that is, sequencing effort per sample) is comparable or smaller to the coverage achieved in our small-insert library.

Support from the analyses of whole-genome sequences of isolates

To better interpret environmental sequence comparisons, we compared ∼100 fully sequenced closely related (that is, >60% ANI) genomic sequences from cultivated isolates representing several important bacterial genera (*Escherichia*, *Shewanella* and *Burkholderia*). Pair-wise whole-genome comparisons revealed that even genomes sharing 90–100% ANI would be distinguished as discrete sequence-based clusters by our approach, reflecting the fact that nt identities of the majority of individual shared genes were normally and tightly distributed around the average ANI between entire genomes. Indeed, genes much more (or less) conserved than the aggregate genome ANI comprised too small a fraction of the genome (typically <5%) to obscure the clusters observed (Figure 3a). Even when a reference genome was queried against a mixture of 110 genomes, 100 of which showed 70–80% ANI to the reference genome and 10 which shared >94% ANI to the reference genome, the closely related genomes formed a clear sequence-based cluster (Figure 3b).

The outlier genes in the previous pair-wise comparisons were not randomly distributed among the genes in the genome. Genes much more conserved than the genome average encompassed primarily housekeeping functions, with the ribosomal rRNA operon being the strongest outlier with this respect. On the other hand, genes much less conserved than the average included genes that are known to be hyper-variable, such as membrane proteins and antigens, and/or genes frequently transferred horizontally such as phage and transposase genes (see also annotations on graph of Figure 3a). These results suggested that omitting highly conserved and mobile genes in the reference sequence would identify genetic discontinuities with higher sensitivity in environmental sequence comparisons. For instance, removing just the ribosomal rRNA operon from the SAR86 BAC clone revealed a clearer sequence-based cluster for the SAR-86 population in the Sargasso Sea (Figure 2d). The alternative scenario for interpreting the differences in the coverage plots of SAR-86 BAC clone sequence prior and after the masking of the ribosomal rRNA operon, namely that rRNA genes have been transferred frequently among unrelated organisms, is highly unlikely. These results further supported the assumption that the less-abundant sequence intermediates in the coverage plots (for example, *Prochlorococcus* in Figure 2c, 75–85% nt range; Crenarchaea in Figure 2a, 50–85% nt range) are attributable more to gene-specific patterns rather than to intermediate organisms. Hence, the sequence-based clusters described previously (for example, Figure 2) may, in fact, be more transparent when gene-specific patterns are taken into account.

Discrete populations may arise by several different mechanisms

Distinct sequence-based populations were observed for more than 20 different reference sequences evaluated. A few exceptions to this pattern, however, were also encountered. For instance, a discrete sequence-based population was not evident when an assembled crenarchaeal contig from the Sargasso Sea project (Venter *et al.*, 2004) was queried against the WGS data from the same sample. Instead, the crenarchaeal population in this environment appeared to consist of clonal or highly related genotypes showing >97–98% ANI to the reference sequence together with a collection of significantly less-abundant (0.2 × vs 1.5 × coverage, see Figure 2b) genotypes sharing ~76% to ~92% identity to the reference sequence. These latter genotypes likely represented differentiated populations on the basis of their lower abundance and genetic distinctiveness compared to the former genotypes. Thus, the total crenarchaeal population in the Sargasso Sea at the time of sampling appeared heterogeneous. This may be related to winter deep-water mixing of Atlantic Ocean waters, which might combine allopatric crenarchaeal populations originating from different depths. The lack of crenarchaeal-like shotgun reads in the stratified summer samples of Sargasso (Venter *et al.*, 2004; Hallam *et al.*, 2006a) is consistent with this interpretation and suggests that frequent environmental fluctuation may have strong effects on the relative abundance of various discrete populations.

Additional evidence for the influence of winter deep-water mixing was observed in comparisons of crenarchaeal populations obtained from seven different depths in the Pacific Ocean (DeLong *et al.*, 2006) against the Sargasso Sea crenarchaeal WGS reads. Consistent with previous environmental surveys that found larger populations of Crenarchaea in the sub-photic zone (Karner *et al.*, 2001; DeLong *et al.*, 2006), crenarchaeal sequences in the Pacific Ocean increased in numbers with depth down to 770 m, and then decreased slightly at 4000 m depth. About half of the Pacific Ocean crenarchaeal sequences from 130 m depth and, to a lesser extend, from 200 m had highly related matches (>95% nt identity) to crenarchaeal sequences from the Sargasso Sea surface waters. Crenarchaeal sequences from deeper or shallower waters were genetically more divergent from their Sargasso Sea counterparts (Figure 4; see also Supplementary Table S3), further corroborating the idea that populations are genetically distinct at different depths. These results also imply a crenarchaeal population with little geographic structure (global population) between the world's two largest oceans at similar depths (for stratified waters) and confirm that the crenarchaeal population in the Sargasso Sea metagenome may consist of several distinct populations that are seasonally bloomed during winter deep-water mixing events. Notably, a global population between the two oceans was also implied for the only other microbial group that there was enough data available to evaluate our approaches, the *Prochlorococcus* group (Supplementary Figure S4).

Recombination may mediate population cohesiveness

Finer scale genomic comparisons of deep-sea crenarchaeal genotypes provided some clues about mechanisms that may influence population cohesiveness or divergence. Substantial evidence for intrapopulation genetic exchange (homologous recombination) was detected for about one-fourth of the crenarchaeal genes evaluated using the GARD algorithm (Kosakovsky Pond *et al.*, 2006b), which employs an advanced likelihood-based phylogenetic approach for recombination detection (Kosakovsky Pond *et al.*, 2006a) (Figure 5). Visual inspection, performed essentially as described previously (Tyson *et al.*, 2004), of the single-nucleotide polymorphism patterns of the corresponding gene alignments as well as a total of ~2 Mb of aligned WGS and ~250 kb of fosmid sequences also indicated several genetic exchange events among the genotypes of the population (Supplementary Figure S5). Homologous recombination therefore does not appear to be negligible in these deep-sea crenarchaeal genotypes and could potentially constitute an

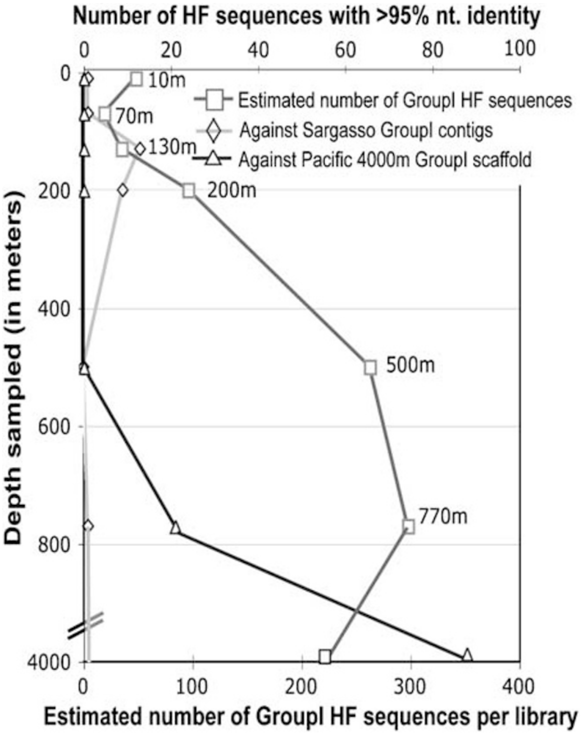

Figure 4 Vertical distribution of Crenarchaea in the Pacific Ocean and their genetic relatedness to their counterparts in the Sargasso Sea. The number of crenarchaeal sequences in published end sequences of seven fosmid libraries from the Pacific Ocean (DeLong *et al*., 2006) (squares, primary *x* axis) is plotted against the depth that each library was originated from (*y* axis). The number of crenarchaeal sequences in each library that was also highly related, that is >95% nt identity, to the crenarchaeal genomic scaffold from 4000 m Pacific Ocean (triangles, primary *x* axis) as well as to crenarchaeal sequences from the Sargasso Sea (diamonds, secondary *x* axis) is shown. Note the different scales between the two *x* axes. All underlying data are provided in Supplementary Table S3. Nt, nucleotide.

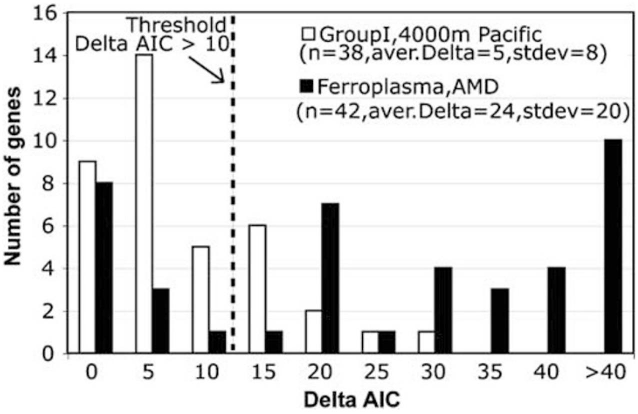

Figure 5 Lower recombination levels detected among deep-sea planktonic Crenarchaea compared to Euryarchaea from the acid mine drainage (AMD) biofilm community. Gene alignments were analyzed with the GARD algorithm (Kosakovsky Pond *et al*., 2006b) to detect instances of intrapopulation recombination. GARD employs a likelihood-based approach to calculate the Delta Akaike Information Criterion (Delta AIC) value between a candidate tree with one or multiple recombination break points and the default tree with no recombination. In general, Delta AIC values higher than ~10 represent significant evidence that the evaluated gene has undergone recombination; the higher the Delta AIC value, the more dramatic the effect of recombination it likely has. Graph shows the number of genes (*y* axis) plotted against the Delta AIC value of the gene (*x* axis). The vertical dashed bar represents the critical Delta AIC value = 10. The number of genes analyzed for each population, their average Delta AIC value and s.d. are shown in parentheses. Note that significantly more (χ^2 test, $P<0.01$) *Ferroplasma* (AMD) genes showed evidence of recombination and that the average Delta AIC value was four times higher for these genes compared to those for Crenarchaea genes. Genes were selected at random, provided that they included a similar number of sequences and comparable sequence diversity in their alignments and they were single-copy and not mobile or hypothetical, to make results directly comparable between the two populations studied. See Materials and methods for details and Supplementary Table S4 for the underlying data. GARD, genetic algorithm for recombination detection.

important force of population cohesion (see also below). Interestingly, comparable levels of recombination were detected in the *Prochlorococcus* population in the Sargasso Sea (Supplementary Table S4), suggesting that similar trends may be prevalent among many other pelagic microbial taxa. However, the recombination frequency in these marine populations was about three to four times lower than our parallel estimates for the *Ferroplasma* (euryarchaeal) population in an AMD biofilm community (Tyson *et al*., 2004), using an identical approach on a very comparable subset of the AMD metagenome. For example, we estimate that more than 70% of the genes in the *Ferroplasma* population (vs ~25% in Crenarchaea) have possibly undergone recombination, consistent with the findings of the original AMD study (Tyson *et al*., 2004), and that the effect of recombination was typically more pronounced than in the deep-sea crenarchaeal population with multiple recombination events occurring on the same gene (Figure 5 and Supplementary Table S4).

An important parameter for determining whether a microbial population is predominantly sexually reproducing or clonal in nature is the ratio of recombination rate (ρ) to mutation rate (θ), with ρ/θ ratios greater than one being indicative of a population with sexually reproducing characteristics. Recombination rates can be estimated by several methods, although the accuracy of the estimates for natural microbial populations remains uncertain. The uncertainty emerges from the facts that several important population parameters such as the effective population size (N_e) remain elusive, and that the methods have not been optimized for WGS data yet (see also Supplementary Table S5 and related discussion in the table caption). For the latter reason, Eppley *et al*. (2007) developed an independent method for estimating the ρ/θ ratio in WGS data, based on manual identification of recombination events among overlapping WGS reads, and calculated the ratio to be between 2:1 and 4:1 for the *Ferroplasma* population in the AMD community. Given that recombination frequency is

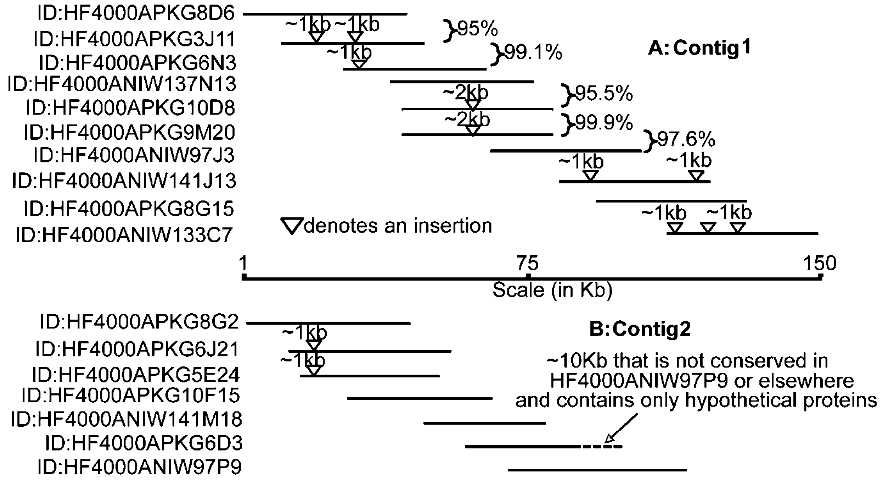

Figure 6 Intrapopulation gene-content diversity for the deep-sea planktonic Crenarchaea based on overlapping fosmid clones. Graphs show the composition of the two largest crenarchaeal contigs assembled (contigs 1 and 2) by individual overlapping fosmids. Gene-content differences as well as the average nt identities between the overlapping fosmids are also denoted on the graph. Note that fosmids were absolutely syntenic in gene content, with no rearrangements, unless interrupted by small, typically 1–2-kb long, exact gene insertions or deletions (depending on which fosmid was considered as reference sequence; insertions were used here for consistency purposes only). The patterns observed in the two contigs shown were representative of the patterns in the remaining contigs assembled from crenarchaeal fosmid sequences as well.

three to four times lower in the deep-sea crenarchaeal population, the ρ/θ ratio is expected to be smaller and hence, this population appears predominantly clonal or, at most, marginally sexual based on the latter approach. Consistent with these interpretations, our estimation of the ρ/θ ratio for 20 different genes of the deep-dwelling Crenarchaea, based on an alternative method that employs coalescent theory and a maximum likelihood approximation as implemented in the LDhat software (McVean *et al.*, 2002), ranged between ~0.01 and 0.3, with an average of ~0.1, and was significantly lower (about threefold) than our parallel estimate for *Ferroplasma* genes (Supplementary Table S5). It is important to note, however, that a ρ/θ ratio of 0.3 might be high enough to account for sexually reproducing populations according to a recent computer simulation study (Fraser *et al.*, 2007). Thus, although it cannot be more firmly established whether homologous recombination is a strong force of population cohesion for these marine populations, these results do indicate a lower recombination levels in the planktonic vs the biofilm-associated (*Ferroplasma*) microbial populations based on three independent approaches (GARD, visual inspection and LDhat).

The role of horizontal gene transfer

It is important to point out that the frequency of illegitimate recombination (HGT), evidenced by foreign-sequence insertions or highly diverged alleles in the population consensus, appeared low but not undetectable. For example, about half of the fosmids evaluated contained 1- to 2-kb long insertions (for example, Figure 6). Similar levels of HGT were noted when examining the frequency with which the sister read of a read that matched the scaffold did not match the scaffold or a crenarchaeal fosmid, indicating that the corresponding clone may be representative of an HGT event to/from non-crenarchaeal genetic background (data not shown). Therefore, HGT may constitute another important evolutionary process for these deep-sea populations despite the remarkable stability of the environment at 4000 m depth. HGT appeared, however, quantitatively less frequent than homologous recombination. That is, it affected a maximum of 5% vs one-fourth of genes in the genome for recombination (for example, Figure 5), based on this small data set examined. Functional annotation of the genes subjected to horizontal transfer did not offer any definitive information as to whether or not the transferred genes were of any ecological importance for the planktonic crenarchaeal populations, mostly because the genes were primarily of hypothetical or conserved hypothetical function (data not shown).

Discussion

Do sequence-based clusters encompass more than one discrete population?

The discrete sequence-based populations identified here (for example, Figure 2) appear to represent an important organizational level in planktonic microbial populations. The matching sympatric abundances of similar but non-identical genotypes comprising these populations (for example, Figure 1b) indicate a high level of ecological coherence. These data, however, do not fully resolve whether there may be even finer scale levels of ecological differentiation. One concern common to all such studies is whether the coverage of the microbial community by sequencing is great enough to expose the finer levels of differentiation. If the sequenced libraries provide a relatively unbiased sampling of the natural community, however, then the results based on relatively low coverage (for example, this study) should be representative and reproducible in high-coverage libraries as well. We believe that the metagenomic libraries used in our study, particularly the small-insert one due to its very small insert size (~1.5 kb), which makes the cloning of intact toxic genes for the laboratory *Escherichia coli* vector more unlikely (Sorek *et al.*, 2007), provide such unbiased sampling. Moreover, statistical comparisons of the diversity of the 16S rRNA genes recovered in the 4000-m deep fosmid library relative to the small-insert library from the same DNA sample revealed that although fosmids sampled significantly fewer α-proteobacteria 16S rRNA genes, the differences were overall rather than minor (Pham *et al.*, 2008). Finally, the 50 sequenced crenarchaeal fosmids used in this study provided a rather extensive and uniform coverage of the *C. symbiosum* genome (Supplementary Figure S1). Although no clone library should be expected to be free of any biases, our results collectively indicate that the biases associated with our metagenomic libraries are not strong enough to prevent meaningful comparisons and conclusions to be made.

In addition, three independent lines of evidence suggest that finer levels of ecological differentiation may be absent. First, if the identified populations consisted of distinct sub-populations, and these sub-populations differed (genetically or in relative abundance), this would have been evidenced by uneven coverage plots or uneven abundances of the corresponding genotypes in the WGS libraries, respectively. Such uneven coverage was not observed except in very specific cases such as the crenarchaeal population in the Sargasso Sea metagenome (see Figure 2b and related discussion above). Second, the analyses of whole genomes of isolates confirmed that even genotypes sharing 90–100% ANI would have been distinguishable as discrete sequence-based clusters by our approach. These analyses also revealed that organismal discontinuities might appear more pronounced when gene-specific patterns of sequence conservation and horizontal transfer are taken into account (Figure 3). Third, the substantial levels of homologous recombination detected (Figure 5) indicate that the genotypes of a population might cohere together via means of genetic exchange. Further, the documented logarithmic drop in recombination frequency with increasing evolutionary divergence of the recombining sequences, particularly in the range of 80–90% sequence identity (Zawadzki *et al.*, 1995; Vulic *et al.*, 1997; Eppley *et al.*, 2007), which roughly corresponded to the areas of genetic discontinuities in our coverage plots (for example, Figure 2), makes homologous recombination an intriguing candidate mechanism responsible for the discreetness of the populations observed here.

Consistent with the idea of a discrete ecological unit, extrapolation from overlapping fosmid clones indicated that gene-content differences between genotypes of the crenarchaeal population comprised less than 5% of all the genes in their genomes. For instance, there is a maximum of ~2 kb insertions (5% of the total) between fully overlapping 36-kb long (on average) fosmid sequences (Figure 6 shows a representative example of the fosmids evaluated). It is possible that some of these gene content differences are due to transposition events to other parts of the genome (not cloned in our libraries), and hence the actual gene content differences may be significantly smaller than 5%. Only a single exception to this rule was encountered for fosmid HF4000ANIW97P9, which contains a ~10-kb long fragment comprised almost exclusively of hypothetical genes. This sequence appears to represent a genomic island or acquisition of a prophage genome specific to the corresponding genotype (Figure 6). The extremely low frequency of such cases identified, that is one fosmid clone in ~50 overlapping fosmids evaluated, indicates that large genomic differences among the genotypes comprising the population may be rare, in general. Comparative analyses have shown that strains of the same bacterial species frequently differ by up to 30–35% of their total gene content (Konstantinidis and Tiedje, 2005), suggesting that these discrete populations are phenotypically much more homogeneous than are many commonly defined cultivated bacterial species.

It cannot be completely ruled out that distinct sub-populations have emerged only recently, and so have not differentiated enough (or differences occur in only a small number of genes) to be detectable by our analyses. It is also possible that different microniches, each containing discrete sub-populations, might exist in the 670 l of filtered seawater that made up our deep-sea sample. The relative homogeneity of the nutrient-limited planktonic environment, however, renders this explanation less likely. Furthermore, our results collectively indicate that the deep-sea crenarchaeal populations identified here represented a consistent, evolutionarily and ecologically discrete unit. If this proves correct, then our data suggest that the most plausible scenario for the emergence of such units would have been selection (or niche invasion) followed by neutral diversification fostered by the remarkably stable physicochemical environment found at 4000 m depth (DeLong *et al.*, 2006). Alternative scenarios such as rapid diversification, after severe population bottlenecks, are less favorable, given the stability of this environment.

Implications for the species definition

Interestingly, the genetic diversity within several of the populations we observed (for example, Figures 1a, and 2 and related discussion above) was roughly comparable to the most frequently used standards for demarcating microbial species, that is, ~5–6% genomic sequence divergence (Konstantinidis and Tiedje, 2005; Goris *et al.*, 2007). Yet, this current operational species definition has been criticized for being overly broad, encompassing phenotypically heterogeneous groups of microorganisms under the same 'named' species (Cohan, 2002; Konstantinidis *et al.*, 2006; Staley, 2006; Ward, 2006). This debate may be largely attributable to the fact that many species collections represent heterologous assemblages of genotypes, recovered from different populations and habitats. Our data indicate that when genotypes share the same ecological trajectory, they not only conform to a discrete genotypic boundary (for example, Figure 2), but also appear to represent much more uniform phenotypes (for example, Figure 6). When the environment of isolation and the ecological success of the organism within this environment remain unknown, then more stringent standards such as the 99% ANI proposed earlier (Konstantinidis and Tiedje, 2005) may present more robust and dependable standards for identifying homogeneous collections of organisms. Finally, the substantial intrapopulation genomic diversity observed seems to imply that selective sweeps, thought to cause microbial speciation, might be more rare or propagate more slowly than previously anticipated (Cohan, 2002), at least in some microbial habitats such as the deep sea.

Concluding summary

Determining how accurately rRNA-based microbial clades (Giovannoni *et al.*, 1990; Rappe and Giovannoni, 2003; Acinas *et al.*, 2004) (sometimes equated with species (Staley, 2006; Ward, 2006)) represent functionally uniform phenotypes is relevant to interpreting their ecological significance. Intraclade diversity is frequently defined as ranging between 1% and 5% 16S rRNA (Rappe and Giovannoni, 2003) sequence difference. In the 4000-m crenarchaeal population, we observed a maximum of ~5–6% genomic sequence divergence that corresponded to $\ll$1% rRNA gene sequence divergence (DeLong *et al.*, 2006). This relationship between genome sequence divergence and rRNA sequence variation is similar to that observed in whole-genome sequence comparisons of cultivars (Konstantinidis and Tiedje, 2005, 2007). These results suggest that only a small proportion of related genotypes typically reported within the known rRNA clades may be active or ecologically significant at a given time and place. The remaining genotypic variants (typically, <90% identical at the genome level to the dominant type) apparently occupy different niches or habitats, perhaps reproducing at diminished rates or remaining latent until genotype-specific environmental conditions allow them to flourish. Our results therefore tend not to support 'neutral' or 'functionally redundant' diversity patterns between (but not within) the different populations constituting the known rRNA-based clades. They also indicate that sequence divergence (as indicated by clear genetic discontinuities, Figure 2) and niche overlap (reflected by comparable sympatric abundances, Figures 1b and 2a) may determine cohesiveness or divergence of populations, more than recombination does. Thus, our findings contrast with recent results from other microbial habitats, in particular biofilms that exhibit much higher levels of intrapopulation recombination (Tyson *et al.*, 2004; Nesbo *et al.*, 2006). These habitat-related differences underscore the important relationships between ecological setting, biotic interactions and genetic mechanisms that together shape and sustain microbial population structure and function.

Acknowledgements

We are grateful to Virginia Rich and Gene Tyson for helpful discussions regarding the manuscript and Sergei Kosakovsky, Gilean McVean and Thierry Wirth for their help with GARD and LDhat. We also thank Chris Preston, Lynne Christianson and Professor David Karl, and the Hawaii Ocean Time series staff and crew for assistance in sample collection at station ALOHA. We thank the JGI production sequencing staff for whole fosmid and shotgun sequencing. This work was supported by the Gordon and Betty Moore Foundation, Department of Energy Microbial Genome Sequencing program, US Department of Energy's Office of Science, Biological and Environmental Research Program and the University of California, Lawrence Livermore National Laboratory, under contract no. W-7405-ENG-48, Lawrence Berkeley National Laboratory under contract no. DE-AC03-765F00098 and Los Almos National Laboratory contract no. W-7405-ENG-36.

Data release

The genomic scaffold used in Figure 1 is provided as a supplementary fasta-formated file. All fully sequenced fosmids and the WGS data from 4000-m depth sample are now available in GenBank under the accession numbers: EU016559–EU016674 and ABEF00000000, respectively.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL *et al.* (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**: 551–554.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* **25**: 3389–3402.

Cohan FM. (2002). What are bacterial species? *Annu Rev Microbiol* **56**: 457–487.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF *et al.* (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**: 1768–1770.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.

DeLong EF, Wu KY, Prezelin BB, Jovine RV. (1994). High abundance of Archaea in Antarctic marine picoplankton. *Nature* **371**: 695–697.

Eppley JM, Tyson GW, Getz WM, Banfield JF. (2007). Genetic exchange across a species boundary in the archaeal genus *Ferroplasma*. *Genetics* **177**: 407–416.

Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC *et al.* (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci USA* **98**: 182–187.

Felsenstein J. (2004). *PHYLIP (Phylogeny Inference Package)*, version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington: Seattle.

Fraser C, Hanage WP, Spratt BG. (2007). Recombination and the nature of bacterial speciation. *Science* **315**: 476–480.

Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC *et al.* (2006). Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24**: 1263–1269.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**: 60–63.

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**: 81–91.

Green P. (2006). *Phrap-Phred Assembly Software*. Distributed by the author. Genome Sciences Department, University of Washington: Seattle.

Hallam SJ, Konstantinidis KT, Putnam N, Schleper C, Watanabe Y, Sugahara J *et al.* (2006a). Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc Natl Acad Sci USA* **103**: 18296–18301.

Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, Richardson PM *et al.* (2006b). Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine *Crenarchaeota*. *PLoS Biol* **4**: e95.

Hanage WP, Fraser C, Spratt BG. (2005). Fuzzy species among recombinogenic bacteria. *BMC Biol* **3**: 6.

Karner MB, DeLong EF, Karl DM. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**: 507–510.

Konstantinidis KT, Ramette A, Tiedje JM. (2006). The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* **361**: 1929–1940.

Konstantinidis KT, Tiedje JM. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102**: 2567–2572.

Konstantinidis KT, Tiedje JM. (2007). Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* **10**: 504–509.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. (2006a). Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* **23**: 1891–1901.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. (2006b). GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22**: 3096–3098.

Lam P, Jensen MM, Lavik G, McGinnis DF, Muller B, Schubert CJ *et al.* (2007). Linking crenarchaeal and bacterial nitrification to anammox in the Black Sea. *Proc Natl Acad Sci USA* **104**: 7104–7109.

Lawrence JG. (2002). Gene transfer in bacteria: speciation without species? *Theor Popul Biol* **61**: 449–460.

Martin DP, Williamson C, Posada D. (2005). RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **21**: 260–262.

McVean G, Awadalla P, Fearnhead P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**: 1231–1241.

Moran NA. (2007). Symbiosis as an adaptive process and source of phenotypic complexity. *Proc Natl Acad Sci USA* **104**(Suppl 1): 8627–8633.

Nesbo CL, Dlutek M, Doolittle WF. (2006). Recombination in Thermotoga: implications for species concepts and biogeography. *Genetics* **172**: 759–769.

Palys T, Berger E, Mitrica I, Nakamura LK, Cohan FM. (2000). Protein-coding genes as molecular markers for ecologically distinct populations: the case of two *Bacillus* species. *Int J Syst Evol Microbiol* **50**(Part 3): 1021–1028.

Pham VD, Konstantinidis KT, Palden T, DeLong EF. (2008). Phylogenetic analyses of ribosomal DNA-containing bacterioplankton genome fragments from a 4000 m vertical profile in the North Pacific Subtropical Gyre. *Environ Microbiol* (doi:10.1111/j.1462-2920.2008.01657.X).

Rappe MS, Giovannoni SJ. (2003). The uncultured microbial majority. *Annu Rev Microbiol* **57**: 369–394.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

Sabehi G, Beja O, Suzuki MT, Preston CM, DeLong EF. (2004). Different SAR86 subgroups harbour divergent proteorhodopsins. *Environ Microbiol* **6**: 903–910.

Sheppard SK, McCarthy ND, Falush D, Maiden MC. (2008). Convergence of *Campylobacter* species: implications for bacterial evolution. *Science* **320**: 237–239.

Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**: 1449–1452.

Spratt BG, Hanage WP, Feil EJ. (2001). The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Curr Opin Microbiol* **4**: 602–606.

Staley JT. (2006). The bacterial species dilemma and the genomic-phylogenetic species concept. *Philos Trans R Soc Lond B Biol Sci* **361**: 1899–1909.

Thompson JD, Higgins DG, Gibson TJ. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.

Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J *et al.* (2005). Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307**: 1311–1313.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW *et al.* (2005). Comparative metagenomics of microbial communities. *Science* **308**: 554–557.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Vulic M, Dionisio F, Taddei F, Radman M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci USA* **94**: 9763–9767.

Ward DA. (2006). A macrobiological perspective on microbial species. *Microbe Mag* **1**: 269–278.

Zawadzki P, Roberts MS, Cohan FM. (1995). The log-linear relationship between sexual isolation and sequence divergence in Bacillus transformation is robust. *Genetics* **140**: 917–932.

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