Bacterial diversity of metagenomic and PCR libraries from the Delaware River

Introduction

Metagenomic clone libraries of environmental DNA enable the exploration of the phylogenetic and metabolic diversity of microbes in the environment without cultivation. New aspects of microbial metabolism in uncultured microbes have been uncovered using this metagenomic approach, as first shown in a study of marine *Bacteria* and *Archaea* (Stein et al., 1996), and then later in soils (Rondon et al., 2000). Intriguing examples include studies revealing different types of phototrophy in the ocean (Beja et al., 2000a, 2001; Sabehi et al., 2003). Other studies have examined heterotrophic metabolism, including one study that identified chitinase genes of uncultured marine microbes (Cottrell et al., 1999). In soils, different types of metabolism have been targeted such as those involving alcohol oxidoreductases (Knietsch et al., 2003), lipases, amylases and nucleases (Rondon et al., 2000). Soil clones have also been obtained that produce broad-spectrum antibiotics (Gillespie et al., 2002) while other clones have hemolytic properties (Rondon et al., 2000).

The data on the metabolic diversity revealed by metagenomic libraries are most valuable when viewed in the context of community structure. Phylogenetic information is needed together with estimates of metabolic potential in order to link specific members of the community to biogeochemical processes. However, the phylogenetic information present in metagenomic libraries has received little attention (Beja et al., 2000b; Liles et al., 2003). One approach to obtaining this phylogenetic information is to screen metagenomic libraries for 16S rRNA genes in order to identify clones that then can be used to explore the metabolic potential of targeted bacterial groups (Beja et al., 2001; Quaiser et al., 2002; Liles et al., 2003). It is also important to compare the phylogenetic make-up of metagenomic libraries to other measures of community structure, such as 16S rDNA clone libraries and fluorescence in situ hybridization (FISH), in order to determine if any of the dominant groups of bacteria are missing. Finally, the phylogenetic information in metagenomic libraries provides another view of community structure without the biases of PCR-dependent approaches (von Wintzingerode et al., 1997).

Only one study examined the similarity between the phylogenetic composition of a metagenomic library and published surveys of marine bacterial diversity (Beja et al., 2000a, 2001; Sabehi et al., 2003). Other studies have examined heterotrophic metabolism, including one study that identified chitinase genes of uncultured marine microbes (Cottrell et al., 1999). In soils, different types of metabolism have been targeted such as those involving alcohol oxidoreductases (Knietsch et al., 2003), lipases, amylases and nucleases (Rondon et al., 2000). Soil clones have also been obtained that produce broad-spectrum antibiotics (Gillespie et al., 2002) while other clones have hemolytic properties (Rondon et al., 2000). The data on the metabolic diversity revealed by metagenomic libraries are most valuable when viewed in the context of community structure. Phylogenetic information is needed together with estimates of metabolic potential in order to link specific members of the community to biogeochemical processes. However, the phylogenetic information present in metagenomic libraries has received little attention (Beja et al., 2000b; Liles et al., 2003). One approach to obtaining this phylogenetic information is to screen metagenomic libraries for 16S rRNA genes in order to identify clones that then can be used to explore the metabolic potential of targeted bacterial groups (Beja et al., 2001; Quaiser et al., 2002; Liles et al., 2003). It is also important to compare the phylogenetic make-up of metagenomic libraries to other measures of community structure, such as 16S rDNA clone libraries and fluorescence in situ hybridization (FISH), in order to determine if any of the dominant groups of bacteria are missing. Finally, the phylogenetic information in metagenomic libraries provides another view of community structure without the biases of PCR-dependent approaches (von Wintzingerode et al., 1997).

Only one study examined the similarity between the phylogenetic composition of a metagenomic library and published surveys of marine bacterial diversity (Beja et al., 2000a, 2001; Sabehi et al., 2003). Other studies have examined heterotrophic metabolism, including one study that identified chitinase genes of uncultured marine microbes (Cottrell et al., 1999). In soils, different types of metabolism have been targeted such as those involving alcohol oxidoreductases (Knietsch et al., 2003), lipases, amylases and nucleases (Rondon et al., 2000). Soil clones have also been obtained that produce broad-spectrum antibiotics (Gillespie et al., 2002) while other clones have hemolytic properties (Rondon et al., 2000).
Phylogenetic analysis of metagenomic clones bearing 16S rRNA genes from surface water of the Pacific Ocean identified several groups of Archaea and Bacteria that are common in marine bacterial communities, including a Euryarchaeota clone and clones related to SAR86, SAR116 and SAR11 bacteria, Roseobacter spp. and Cytophaga-like bacteria (Giovannoni and Rappé, 2000). However, no study has directly compared the phylogenetic composition of a metagenomic library and the microbial community in situ. It may be especially important to examine this issue in freshwaters where Gram-positive bacteria can be abundant (Glöckner et al., 2000; Zwart et al., 2002). These bacteria could be underrepresented in metagenomic libraries if DNA extraction from Gram-positive bacteria is less efficient.

Bacteria in rivers and other freshwater ecosystems have not been examined previously using metagenomic libraries. Bacteria in rivers play an important role in the consumption of organic matter transported from the terrestrial environment to the ocean (Amon and Benner, 1996; Opsahl et al., 1999; Hernes and Benner, 2002), and the make up of those communities likely has an impact on organic matter consumption (Covert and Moran, 2001). The diversity of bacteria in rivers may be especially high as they may contain mixtures of aquatic and terrestrial bacteria, including phylogenetically and metabolically diverse soil taxa (Liles et al., 2003; Topp, 2003). Microbes from soil may be introduced into waterways by runoff (Zaitlin et al., 2003) and mix with freshwater bacteria (Zwart et al., 2002). Less is known about riverine bacteria than their counterparts in other aquatic systems. A GenBank search in November 2003 for riverine bacterial and archaeal 16S rRNA genes returned approximately 1000 hits from studies in 14 rivers, compared with more than 17 000 sequences from marine environments and lakes.

In this study, we determined the coverage and composition of metagenomic and PCR libraries from the Delaware River. We also contrasted the composition of these libraries with community structure determined by FISH. In order to identify why the composition of libraries might differ, we tested the hypothesis that heteroduplex artefacts of the PCR step lead to overestimates of bacterial diversity (Thompson et al., 2002). We found that the metagenomic analysis and PCR libraries identified Actinobacteria, Cytophaga-like bacteria and beta-proteobacteria, which are potentially important members of the bacterial community in the Delaware River.

### Results

#### Overview of the libraries

The metagenomic library was comprised of 4608 clones with an average insert size of 40.5 ± 3.1 kb (n = 15). The library contained approximately 180 Mb of DNA, which is equivalent to 90 bacterial genomes, assuming two Mbp per bacterial genome (Button and Robertson, 2001). Eighty clones tested positive for 16S rRNA genes, which is consistent with the expected number of rRNA genes in the library, assuming a single 16S rRNA gene per bacterial genome. Bacteria isolated from aquatic and low-nutrient environments typically have only one or two rRNA operons (Button et al., 1998; Fegatella et al., 1998; Fogel et al., 1999; Klappenbach et al., 2000).

The compositions of the metagenomic library and PCR libraries were not significantly different at a high phylogenetic level based on analyses of coverage calculated using the LIBSHUFF tool. BLAST analysis was consistent with this result, revealing essentially the same major groups of bacteria in the two libraries (Table 1). Cytophaga-like bacteria, beta-proteobacteria and Actinobacteria were abundant, and the composition of communities in the Delaware River was similar to those found in marine environments.

### Table 1. Phylogenetic composition of the metagenomic and PCR libraries and structure of the actual community determined by FISH.

<table>
<thead>
<tr>
<th>Group</th>
<th>% of metagenomic clones</th>
<th>% of PCR clones</th>
<th>% of total prokaryotes detected by FISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha-proteobacteria</td>
<td>3</td>
<td>12</td>
<td>6 ± 2</td>
</tr>
<tr>
<td>Beta-proteobacteria</td>
<td>17</td>
<td>50</td>
<td>25 ± 5</td>
</tr>
<tr>
<td>Gamma-proteobacteria</td>
<td>4</td>
<td>0</td>
<td>3 ± 2</td>
</tr>
<tr>
<td>Epsilon-proteobacteria</td>
<td>0</td>
<td>2</td>
<td>ND</td>
</tr>
<tr>
<td>Cytophaga-like</td>
<td>54</td>
<td>13</td>
<td>17 ± 4</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>14</td>
<td>16</td>
<td>26 ± 5</td>
</tr>
<tr>
<td>Firmicutes</td>
<td>3</td>
<td>0</td>
<td>ND</td>
</tr>
<tr>
<td>Bdellovibrio</td>
<td>0</td>
<td>4</td>
<td>ND</td>
</tr>
<tr>
<td>Verrucomicrobiales</td>
<td>3</td>
<td>2</td>
<td>ND</td>
</tr>
<tr>
<td>Spirochaetaceae</td>
<td>1</td>
<td>0</td>
<td>ND</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>77</td>
</tr>
</tbody>
</table>

Composition of the metagenomic library is based on nucleotide sequences from 72 clones bearing 16S rRNA genes. Fifty-eight clones from the PCR library were sequenced and 500–1000 DAPI-positive prokaryotes were analysed with each FISH probe. ND, not determined.
comprised the largest fractions of the libraries while alpha-
proteobacteria, gamma-proteobacteria, Firmicutes, Verru-
comicrobiales and Spirochaetaceae were minor mem-
bers. However, closer examination of the relative abundance of different bacterial groups revealed noteworthy differences between the metagenomic and PCR libraries.

**Cytophaga-like bacteria in clone libraries**

The metagenomic and PCR libraries differed in the num-
ber and kinds of Cytophaga-like bacteria. One obvious
difference was the percentage of Cytophaga-like clones.
Only 13% of the clones in the PCR library belonged to the
Cytophaga-like group, compared with 54% for the
metagenomic library (Table 1). In addition, the variety of
Cytophaga-like bacteria was greater in the metagenomic
library than in the PCR library. More Cytophaga-like clus-
ters contained fosmid sequences than PCR sequences in
the phylogenetic tree of Cytophaga-like bacteria (Fig. 1).
The metagenomic library but not the PCR library con-
tained Cytophaga-like clones in the Chitinophaga and
Myroides groups (Fig. 1). One Cytophaga-like clone in the
PCR library (Sta2–97) was not similar to any of the
metagenomic clones.

Seventy percent of the Cytophaga-like bacteria were
associated with the Flavobacteriales group (Fig. 1) and
these clones could be grouped into three clusters (A–C).
A fourth cluster (D) was in the Flectobacillus group
(Fig. 1). Clusters A and B were related to different kinds
of Flavobacterium spp. and included some cultured bac-
teria. Cluster A included an isolate from the Elbe River
and cluster B included Flavobacterium xinjiangensi. The
average sequence identity in these two clusters with cul-
tured members was 96.6% ± 3.4% and 94.5% ± 6.1%
respectively. Clusters C and D were associated with the
Myroides and Flectobacillus groups, respectively, and
included bacteria with a higher degree of sequence simi-
arity (98.5% ± 0.5% and 99.5% ± 0.4% respectively), but
these clusters had no cultured members.

Some of the Cytophaga-like bacteria were represented
in both libraries. Half of the Cytophaga-like clones in the
PCR library were highly similar to clones in the metage-
nomic library, and every Cytophaga-like group with clones
from the PCR library also had representatives from the
metagenomic library. For example, cluster D was com-
prised of three PCR clones and four metagenomic clones
(Fig. 2). Cluster A included three metagenomic clones
and a highly similar clone from the PCR library. However,
similarity between Cytophaga-like bacteria in the two
libraries was restricted to these two clusters.

**Beta-proteobacteria in clone libraries**

The metage-
nomic and PCR libraries also differed in the number and
kinds of beta-proteobacteria. The PCR library had more
than twice as many clones in this group than the metage-
nomic library (50% versus 17% respectively) (Table 1).
The beta-proteobacteria clones in the two libraries belonged to five clusters (A–E). Clusters D and E were
comprised exclusively of clones from the PCR library
(Fig. 2). The degree of relatedness in the clusters made
up of only PCR clones and in cluster A was low, ranging
from 89% to 95% sequence identity. In contrast, some of
the beta-proteobacteria sampled by the two libraries were
similar. Clusters B and C, which included clones from both
libraries, were comprised of very closely related bacteria
with similarities of 96.4% ± 1.6% and 97.5% ± 3.1%
respectively.

**Other bacteria in the clone libraries**

In contrast to the differences outlined above for the
Cytophaga-like and beta-proteobacteria groups, Actino-
bacteria in the metagenomic and PCR libraries were quite
similar. The libraries had equivalent percentages of Acti-
obacteria (about 15%), and there was considerable over-
lap in the types of Actinobacteria in the two libraries. Of
the three clusters of Actinobacteria clones (A–C), each
included clones from both libraries (Fig. 3). Cluster B
included the most closely related clones with average
similarities of 96.4% ± 2.2%. The Actinobacteria clones in
clusters A and C were less closely related. Sequence
similarities in these clusters were 93.2% ± 4.1% and
92.9% ± 5.4% respectively.

Alpha-proteobacteria comprised a minor component of
the libraries, accounting for only 3% and 12% of the
cloned in the metagenomic and PCR clones respec-
tively (Table 1). The alpha-proteobacteria clones in the two
libraries were closely related to each other and to some
cultured alpha-proteobacteria (results not shown). Three
cloned from the PCR library and two from the metage-
nomic library belonged to a cluster that included cultured
Sphingomonas spp. One alpha-proteobacteria clone from
the PCR library belonged to the Rhodospirillales group.

The libraries also included clones belonging to five
groups of bacteria that can be present in aquatic sys-
tems but are typically minor members of the community.
These groups included Firmicutes and Spirochaetaceae
that occurred in only the metagenomic library and epsi-
on-proteobacteria and Bdellovibrio that were found
exclusively in the PCR library (Table 1). Clones belong-
ing to the Verrucomicrobiales group occurred in both
libraries.

**Library coverage and diversity**

Library coverage increased with the genetic distance, as
expected. At an evolutionary distance of 0.03 coverages
of the metagenomic and PCR libraries were 65% and 51%
Fig. 1. Phylogenetic relationships of selected Cytophaga-like bacteria as depicted in the ARB database tree constructed of sequences longer than 1000 bp (June 2002). Delaware River Cytophaga-like bacteria were added using the ARB parsimony quick add tool. Clones in the library of 16S rDNA amplicons are labelled PCR and 16S rDNA-bearing metagenomic DNA clones are labelled Fosmid. Values on the trapezoids indicate the number of sequences in the group. Labels A, B, C and D refer to clusters of closely related bacteria that are discussed in the text. The scale bar represents 0.1 substitutions per site.
Fig. 2. Phylogenetic relationships of selected beta-proteobacteria as depicted in the ARB database tree constructed of sequences longer than 1000 bp (June 2002). Delaware River beta-proteobacteria were added using the ARB parsimony quick add tool. Clones in the library of 16S rDNA amplicons are labelled PCR and 16S rDNA-bearing metagenomic DNA clones are labelled Fosmid. Values on the trapezoids indicate the number of sequences in the group. Labels A, B, C, D and E refer to clusters of closely related bacteria that are discussed in the text. The scale bar represents 0.1 substitutions per site.
Fig. 3. Phylogenetic relationships of selected Actinobacteria as depicted in the ARB database tree constructed of sequences longer than 1000 bp (June 2002). Delaware River Actinobacteria were added using the ARB parsimony quick add tool. Clones in the library of 16S rDNA amplicons are labelled PCR and 16S rDNA-bearing metagenomic DNA clones are labelled Fosmid. Values on the trapezoids indicate the number of sequences in the group. Labels A, B and C refer to clusters of closely related bacteria that are discussed in the text. The scale bar represents 0.1 substitutions per site.
Metagenomic analysis of freshwater bacteria

respectively (Fig. 4). This distance is approximately equal to 97% sequence identity, which is one approximation of species-level identity (Stackebrandt and Goebel, 1994). Coverage of the metagenomic library was higher than coverage of the PCR library at all levels of genetic distance with the greatest differences occurring at small genetic distance. Coverage decreased rapidly with decreasing genetic distance, and at a genetic distance of 0.01 the coverage of the metagenomic library was 46% while the coverage of the PCR library decreased even more to just 16%.

Indices of diversity were higher for the metagenomic library than the PCR library (Table 2). The Simpson index was 13.4 for the metagenomic library compared with 8.7 for the PCR library. The difference was not as substantial for the Shannon–Weaver index (Table 2). Similarly, measures of evenness were higher for the metagenomic library than the PCR library. Evenness calculated using the Shannon–Weaver index was 0.73 and 0.66 for the metagenomic library and PCR library respectively, while the Simpson index of evenness was 0.26 and 0.17 for the metagenomic and PCR libraries respectively.

We examined the possibility that diversity based on PCR clone library composition is overestimated because of artefacts of the PCR. Re-conditioned PCR was used to test for potential artefacts introduced when heterologous PCR products re-anneal and form heteroduplexes that are then subjected to mismatch repair upon cloning into E. coli (Thompson et al., 2002). Re-conditioning of the PCR amplicons reduced the number of RFLP types expected in a library of 80 clones from 50 in the standard PCR library to 40 in the library produced with re-conditioned PCR (Fig. 5). There was no difference between rarefaction curves of replicate PCR libraries produced with standard PCR (Fig. 5).

**Comparison of FISH and clone libraries**

Fluorescence in situ hybridization was used to assess the correspondence between clone library composition and bacterial community structure. Microscopic enumeration of bacteria hybridized with fluorescently labelled 16S rRNA directed probes revealed a community dominated by beta-proteobacteria, Actinobacteria and Cytophaga-like bacteria (Table 1). As expected for a freshwater system, Actinobacteria and beta-proteobacteria were the most abundant groups, making up about 25% of the community. Cytophaga-like bacteria were also prominent,

---

**Table 2.** Diversity of 16S rRNA genes in metagenomic and PCR clone libraries from the Delaware River.

<table>
<thead>
<tr>
<th>Library</th>
<th>Shannon–Weaver Diversity</th>
<th>Shannon–Weaver Evenness</th>
<th>Simpson Diversity</th>
<th>Simpson Evenness</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metagenomic</td>
<td>2.9</td>
<td>0.73</td>
<td>13.4</td>
<td>0.26</td>
<td>0.65</td>
</tr>
<tr>
<td>PCR</td>
<td>2.6</td>
<td>0.66</td>
<td>8.7</td>
<td>0.17</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Diversity indices were calculated using equations taken from Dunbar and colleagues (1999).
accounting for 17% ± 4% of the community. In contrast, alpha-proteobacteria and gamma-proteobacteria were considerably less abundant (less than 10%). Seventy-seven percent of all prokaryotes were identified with a group probe, and 69% ± 11% were detected with the general bacterial probe Eub338.

Bacterial community structure determined by FISH and inferred from the PCR library was surprisingly similar. The abundance of Cytophaga-like bacteria in the PCR library was within 25% of the abundance determined by FISH, while the abundance of gamma-proteobacteria differed by 3% or less and Actinobacteria differed by 38%, according to the two analyses (Table 1). In contrast, the abundance of beta-proteobacteria clones in the PCR library was twice the abundance of bacteria detected with the beta-proteobacteria FISH probe (Table 1). Three types of bacteria, including epsilon-proteobacteria, Bdellovibrio and bacteria belonging to the candidate division TM6, were present in the PCR library and would not have been detected with any of the FISH probes used in this study. Although we have no estimate of the abundance of bacteria in these groups, they likely accounted for some of the 23% of bacteria not detected by FISH (Table 1).

Discussion

It is necessary to examine whether the phylogenic composition of metagenomic libraries reflects that of the original microbial composition in order to effectively link community structure and function revealed by metagenomic analysis. If metagenomic libraries represent a biased sample of microbial diversity, they will not yield an accurate assessment of metabolic diversity. Assuming they are unbiased, large insert metagenomic libraries may be particularly powerful in examining the relationship between community structure and metabolic capacities. Large inserts increase the likelihood of obtaining a clone containing both ‘functional’ genes (coding for metabolic function) and genes with phylogenetic information, e.g. 16S rRNA genes. Therefore, it is necessary to know how the diversity of 16S rRNA genes sampled by large insert metagenomic libraries corresponds to actual community structure. Our main finding was that the composition of the metagenomic library at a high phylogenetic level was not statistically different from the PCR library, but closer examination revealed noteworthy differences between the libraries and the community structure determined by FISH. The metagenomic library sampled a broader diversity of Cytophaga-like bacteria than the PCR library, but it missed some groups of potentially important beta-proteobacteria. The metagenomic and PCR libraries sampled the same groups of Actinobacteria, but FISH analysis suggested that both types of libraries may have failed to adequately sample Actinobacteria.

Metagenomic analysis of Delaware River bacteria provided a view of Cytophaga-like bacterial diversity that was not possible with the PCR library and FISH analyses. The number of clones representing Cytophaga-like bacteria in the metagenomic library was threefold higher than the abundance determined by FISH and fourfold higher than in the PCR library. It was not surprising that the metagenomic library sampled a larger diversity of Cytophaga-like bacteria than the PCR library because previous studies had suggested that Cytophaga-like bacteria are underrepresented in 16S rDNA PCR libraries relative to the abundance of these bacteria detected by FISH (Cottrell and Kirchman, 2000; Kirchman et al., 2003). It was also not surprising that the metagenomic library had more clones from Cytophaga-like bacteria than would have been expected based on the FISH counts because the general FISH probe CF319a for Cytophaga-like bacteria does not recognize all Cytophaga-like bacteria (Weller et al., 2000). The abundance of Cytophaga-like bacteria according to the general FISH probe CFB560 for Cytophaga-like bacteria (O’Sullivan et al., 2002) was only 15% higher than that determined with CF319a. The denaturing gradient electrophoresis (DGGE) PCR primers used in this study appear to adequately sample Cytophaga-like bacteria, at least compared with FISH analysis (Castle and Kirchman, 2004), and these primers apparently recognize a broader diversity of Cytophaga-like bacteria than the EubA and EubB primers, as some Cytophaga-like bacteria detected by DGGE in the metagenomic library were not sampled by the PCR library.

The Cytophaga-like bacterial group includes a highly diverse collection of bacteria but not all of this diversity was present in our Delaware River sample. In fact, most of the Cytophaga-like bacteria were most closely related to just two genera of Flavobacteriales (Flavobacterium and Myroides) and a cluster in the Fllectobacillus group. This latter group is distinct from the Cytophaga–Flavobacterium cluster, which is synonymous with the Flavobacte-
rionales (Kirchman, 2002). The Delaware River bacteria in the Fllectobacillus group belong to the proposed PRD01a001B cluster of Cytophaga-like bacteria identified by Zwart and colleagues (2002). Nine of the 12 Delaware River sequences in the Fllectobacillus group were greater than 98% similar to the PRD01a001B clone from the Parker River, Massachusetts. The proposed PRD01a001B cluster, which is currently comprised solely of uncultivated bacteria, has also been found in Adirondack lakes, Lake Baikal and other freshwater environments (Zwart et al., 2002). However, this group of Cytophaga-like bacteria has not been detected in all freshwater environments; for example, it was not found in the Columbia River (Zwart et al., 2002).

The Delaware River beta-proteobacteria belonged to clusters widely distributed in freshwater systems. Three
groups of Delaware River beta-proteobacteria clones (B, D and E) were highly similar to bacteria in the freshwater beta-Proteobacteria clusters proposed by Zwart and colleagues (2002). Sequences in Delaware River beta-proteobacteria group B were 95–99% similar to the Rodoferax sp. in the Rodoferax sp. BAL47 cluster. Group D of Delaware Bay beta-proteobacteria corresponded to the Polynucleobacter necessarius cluster (88–99% sequence similarity to P. necessarius) and sequences in group E probably belong to the proposed LD28 cluster. Sequences in group D were 89–97% similarity to the LD28 clone from Lake Loosdrecht in the Netherlands (Zwart et al., 1998).

The clusters of freshwater beta-proteobacteria proposed by Zwart and colleagues (2002) were more prevalent in the PCR library than in the metagenomic library, suggesting that they may represent bacteria that are preferentially sampled by PCR libraries. Delaware River beta-proteobacteria group B included clones from both the metagenomic library and the PCR library, but Delaware River groups D and E were comprised of only clones from the PCR library. Different scenarios could lead to the overrepresentation of bacterial groups in PCR libraries compared with libraries constructed without a PCR step, including selective PCR amplification with general bacterial primers, differences in rRNA gene copy number or variation in DNA extraction efficiency (von Wintzingerode et al., 1999). Fluorescence in situ hybridization probes are needed to determine the abundance of these beta-proteobacteria, which appear to be in a variety of freshwater environments.

There was a high degree of similarity in the Actinobacteria sampled by the metagenomic and PCR libraries. All three groups of Actinobacteria identified in this study included clones from the metagenomic and PCR libraries. Our results indicate that the presumed thicker cell walls of these G+ bacteria do not interfere with the extraction of high molecular weight DNA from these bacteria. However, the overall efficiency of DNA extraction may be low as Actinobacteria were underrepresented in the libraries compared with their abundance determined by FISH. Although the libraries may be missing some Actinobacteria, they did recover clusters that are apparently widespread in freshwater systems. Delaware River Actinobacteria groups B and C, which correspond to the ACK-M1 and STA-30 clusters identified by Zwart and colleagues (2002), were identified in the metagenomic and PCR libraries. Clones in cluster B were 92–99% similar to the ACK-M1 clone and clones in cluster C were 92–97% similar to the STA-30 clone. Delaware River Actinobacteria group A, which included clones from both libraries, did not overlap with previously described clusters of freshwater Actinobacteria. None of the G+ bacterial groups we detected represented a terrestrial bacterial signal because of transport of soils into the river. Zwart and colleagues (2002) also did not find evidence for an important contribution of soil bacteria to bacterial communities in rivers.

Coverage in the metagenomic library was consistently higher than in the PCR library over a broad range of evolutionary distances, although substantially higher coverage by the metagenomic library was restricted to evolutionary distances less than 0.05. Differences in coverage were not the result of different sequence lengths; we analysed 1400 bp for the PCR clones, but only the highly variable V3 region of the 16S rRNA gene for the metagenomic clones. LIBSHUFF analysis conducted with the V3 region alone also revealed higher coverage by the metagenomic library (data not shown). Differences in coverage may reflect more comprehensive sampling of bacterial diversity by the metagenomic library.

Lower coverage by the PCR library was consistent with our hypothesis that PCR artefacts lead to overestimates of diversity. Polymerase chain reaction re-conditioning flattened the rarefaction curve, but the effect was only 20%, not as pronounced as would be expected if heteroduplexes were a substantial source of diversity in the PCR libraries. The higher diversity in the metagenomic library compared with the PCR library is also consistent with the idea that the PCR libraries do not overestimate bacterial diversity. In fact, they probably underestimate diversity, because they miss certain groups, such as Cytophaga-like bacteria (Cottrell and Kirchman, 2000). Additional factors such as cloning biases, variation in genome size (Fogel et al., 1999), rRNA gene copy number (Klappenbach et al., 2000) and ribosome content (Fegatella et al., 1998), could have resulted in larger differences than we observed between the two library approaches and FISH. Given the potential sources of differences the degree of similarity between the different library and FISH approaches was higher than expected.

These results and other suggest that the identities of prokaryotes in aquatic environments are no longer completely hidden in the microbial ‘black box’. With the aid of various culture-independent approaches, it is now clear that various proteobacterial groups, Cytophaga-like bacteria and Actinobacteria are the dominant groups of bacteria in surface waters of aquatic systems (Giovannoni and Rappé, 2000; Rappé and Giovannoni, 2003). Archaea make a substantial contribution to prokaryotic diversity in the ocean depths well below the euphotic zone (Massana et al., 1998; Karner et al., 2001). Although potentially these groups could contain many subgroups, at least in marine waters, it appears that only a few clades dominate, including alpha-proteobacteria belonging to the SAR11 group (Morris et al., 2002; Rappé et al., 2002) and Roseobacter spp. (Gonzalez and Moran, 1997). Cytophaga-like bacteria in Delaware cluster 2 (Kirchman et al.,
2003), the marine Actinobacteria group (Rappé et al., 1999) and Archaea in group I and group II (Fuhrman et al., 1992; Massana et al., 2000; Karner et al., 2001).

Dominant groups in freshwater are also becoming apparent. Our study revealed several widespread clusters of freshwater bacteria in the Delaware River that were identified in the metagenomic and PCR libraries. These clusters are obvious candidates for further investigation, including additional metagenomic analysis to explore their metabolic potential. It is unclear if clusters such as the P. necessarius and LD28 beta-proteobacteria clusters, which occurred in the PCR library but not the metagenomic library, are selectively amplified by the PCR or for some reason are missed by the metagenomic cloning. Additional examination by FISH could help resolve the numerical importance of these freshwater bacteria. Studies such as these will further advance our understanding of aquatic bacterial diversity as it evolves from a view of apparently intractable diversity towards what appears to be a manageable number of dominant bacterial groups.

**Experimental procedures**

**Sample collection**

A water sample was collected from a depth of 1 m in the Delaware River near Trenton, New Jersey (40°7.7'-N, 74°49.3'-W) in December 2001 and subsampled for the analyses performed in this study.

**Fluorescence in situ hybridization**

The sample for FISH analysis was fixed with 2% paraformaldehyde at 4°C for 12–18 h. Bacteria were then collected on 0.2 μm polycarbonate filters, rinsed twice with deionized water and the filters were stored at -20°C. Bacteria were detected with FISH probe Eub338 (Amann et al., 1999). We also tried a mixture of general probes for Bacteria (Daims et al., 1999) and found that it detected 70 ± 5% of DAPI positive bacteria compared with 69 ± 11% for Eub338 alone. The FISH probes for alpha-, beta- and gamma-proteobacteria were Alf968 (Neef, 1997; Glöckner et al., 1999), Bet42a and Gam42a (Manz et al., 1992) respectively. Cytophaga-like bacteria were assayed with probe CF319a (Manz et al., 1996) and CFB560 (O’Sullivan et al., 2002), and Actinobacteria were enumerated using probe HGC69a (Roller et al., 1994). Non-specific binding was assessed using a negative control probe (Kerner and Fuhrman, 1997). Hybridization of Cy3-labelled oligonucleotide probes was performed with portions of polycarbonate filters incubated in a hybridization buffer containing formamide to control the stringency of hybridization (Cottrell and Kirchman, 2000). Unbound probe was removed by washing the sample with a wash buffer at 48°C. The filter piece was then rinsed in water and 70% ethanol. The air-dried sample was mounted on a glass slide with a coverslip using an antifade mountant containing 2 μg ml⁻¹ DAPI. Probe-positive bacteria (500–1000 bacteria for each FISH probe) were enumerated using semiautomated fluorescence microscopy (Cottrell and Kirchman, 2003).

**Preparation of the bacterial size fraction**

Environmental DNA was extracted from the bacterial size fraction obtained by pumping river water (500 l) sequentially through a 1 μm nominal pore-size polypropylene string-wound filter (Cole Parmer) and a 0.8 μm polycarbonate filter (Nuclepore). Bacteria were collected from the filtrate by tangential flow filtration using a 0.1 μm hollow fibre filter (A/G Technology) and an Amicon DC10 gear pump. The sample was concentrated to 2 l, rinsed with a buffer (0.5 M NaCl, 0.1 M EDTA, 10 mM Tris pH 8.0), and stored frozen.

**Library construction using a fosmid vector**

Bacteria were collected from thawed bacterial concentrate by centrifugation in an SS34 rotor at 16 000 r.p.m. for 30 min. The bacteria were resuspended in 1 ml of the supernatant and then mixed with 1% Sea Plaque agarose in deionized water at 55°C (Stein et al., 1996). The molten mixture was drawn into a 1 ml syringe and allowed to solidify. The sample was equilibrated with buffer (10 mM Tris pH 8, 50 mM NaCl, 0.1 M EDTA, 1% Sarkosyl) and digested with 1 mg ml⁻¹ lysozyme at 37°C for 4 h. Buffer and lysozyme were replaced with 1% Sarkosyl in 0.5 M EDTA and the sample was digested with 1 mg ml⁻¹ proteinase K at 55°C for 18 h. Following the enzyme treatments, the sample was electrophoresed for 8 h at 23 V through 0.3% SeaKem Gold agarose in TAE buffer. DNA entering the gel, but larger than the 48 kb size standard was recovered from the gel by electrodereolution and concentrated by ethanol precipitation. The DNA was then sheared to 40 kb by 150 passages through a 200 μl pipette tip. The DNA was prepared for ligation using a blunt end repair kit (Epipcentre Technologies) following the manufacturer's protocol. The insert DNA was then size selected by electrophoresis on 1% Sea Plaque GTG agarose. DNA was recovered from the gel following treatment with Gelase enzyme (Epipcentre Technologies) digestion and ethanol precipitation. The DNA was then ligated to the pCC1FOS fosmid vector (Epipcentre Technologies) and packaged using phage protein extract following the manufacturer's recommended procedure. EPI300 host E. coli infected with recombinant phage were plated on LB media containing chloramphenicol. Clones were picked and sorted into 96-well microtitre plates.

**PCR library construction**

Three PCR libraries were constructed with the DNA isolated from the Delaware River. Two replicate libraries were constructed by cloning amplicons generated from separate PCR reactions using the general bacterial primers EubA and EubB (Lane, 1991). A third library was constructed using re-conditioned amplicons from the PCR reaction used to generate the first library. Re-conditioning was performed with three rounds of thermal cycling applied to amplicons diluted 10-fold in fresh PCR reagents (Thompson et al., 2002). The 25 μl PCR reactions included 0.5 μl, 1.2 μl and 2 μl additions of 10 mM deoxynucleoside triphosphate, 25 mM MgCl₂, 10 mM primer...
stocks respectively. Bovine serum albumin was added to a final concentration of 0.2 mg ml\(^{-1}\). Two and a half units of Taq polymerase (Promega) and a one-tenth volume of 10× buffer were added to each reaction. The thermal cycling conditions consisted of a touchdown series in which the annealing temperature decreased from 65°C to 55°C by 1°C per cycle, followed by 15 cycles at 55°C, each for 1 min. Each cycle included a 1 min denaturation step at 95°C and an extension step of 2.5 min at 72°C. Amplicons were cloned with a TOPO-TA cloning kit with pCR2.1-TOPO vector (Invitrogen) according to the manufacturer’s instructions.

### Screening PCR and metagenomic libraries

Clones were screened for insert size using M13 forward and reverse PCR primers, which flank the cloned insert, and those with full-length inserts were screened by restriction fragment length polymorphism (RFLP) analysis by digesting the M13 PCR amplicons with a mixture of Hhal and Rsal (New England Biolabs). Restriction patterns on 2% Metaphore agarose (FMC) agarose gels were compared manually. Clones with identical restriction patterns were grouped together into RFLP types. The 16S rRNA genes from one representative of each RFLP type in the first library were completely sequenced.

Pools of 96 fosmid clones were screened for 16S rRNA genes by DGGE of PCR amplicons generated with primers GC358F and 517R (Muyzer et al., 1995). Selected bands resolved on an 8% polyacrylamide gel containing a gradient of 25–55% denaturant (13.8–22% formamide and 10.5–23% urea) were re-amplified and sequenced. Phylogenetic classification was determined using BLAST and the ARB sequence analysis tool as described below. Bands with equal electrophoretic mobility were assigned the same phylogenetic classification.

### Sequence and library analysis

Nucleotide sequences were analysed using BLAST [version 2.1; National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST)] and the ARB package (Ludwig et al., 2004). Sequences were aligned with the ARB Fast Aligner version 1.03, adjusted manually with secondary-structure criteria. Phylogenetic relationships were determined using the Quick Add Parsimony tool and the 1000_pub_may02 phylogenetic tree in the ARB database (June 2002 release).

Sequence similarities were determined using the tools in ARB and rarefaction curves were calculated using Analytic Rarefaction 1.3 (http://www.uga.edu/strata/software/). Libraries were compared using the LIBSHUFF analysis tool (http://www.arches.uga.edu/~whitman/libshuff.html) (Singleton et al., 2001). Diversity indices were calculated using equations taken from Dunbar and colleagues (1999).

### Statistical analysis

Libraries were compared using LIBSHUFF analysis (Singleton et al., 2001). This approach compares libraries based on their coverage and the number of sequences in one library that are not found in the second library. This so-called ‘heterologous coverage’ is calculated over a range of evolutionary distances to obtain a coverage curve. The difference between the ‘heterologous coverage’ curve and the coverage curve of the first library is calculated using the Cramér-von Mises test statistic (Pettitt, 1982) and compared with the coverage curve calculated after shuffling sequences between the two libraries.

### Nucleotide sequence accession numbers

The nucleotide sequence data reported in this work are deposited in GenBank under accession numbers AY562245–AY562366.

### Acknowledgements

This work was supported by the US Department of Energy Microbial Genomes Program and the National Science Foundation.

### References


Neef, A. (1997) Anwendung der in situ-Einzelzel-


