Metagenomic Insights into the Evolution, Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem†‡

Seungdae Oh,1 Alejandro Caro-Quintero,2 Despina Tsementzi,1 Natasha DeLeon-Rodriguez,2 Chengwei Luo,2,3 Rachel Poretsky,1 and Konstantinos T. Konstantinidis1,2,3,*

School of Civil and Environmental Engineering,1 School of Biology,2 and Center for Bioinformatics and Computational Genomics,3 Georgia Institute of Technology, Atlanta, Georgia

Received 19 January 2011/ Accepted 5 July 2011

Lake Lanier is an important freshwater lake for the southeast United States, as it represents the main source of drinking water for the Atlanta metropolitan area and is popular for recreational activities. Temperate freshwater lakes such as Lake Lanier are underrepresented among the growing number of environmental metagenomic data sets, and little is known about how functional gene content in freshwater communities relates to that of other ecosystems. To better characterize the gene content and variability of this freshwater planktonic microbial community, we sequenced several samples obtained around a strong summer storm event and during the fall water mixing using a random whole-genome shotgun (WGS) approach. Comparative metagenomics revealed that the gene content was relatively stable over time and more related to that of another freshwater lake and the surface ocean than to soil. However, the phylogenetic diversity of Lake Lanier communities was distinct from that of soil and marine communities. We identified several important genomic adaptations that account for these findings, such as the use of potassium (as opposed to sodium) osmoregulators by freshwater organisms and differences in the community average genome size. We show that the lake community is predominantly composed of sequence-discrete populations and describe a simple method to assess community complexity based on population richness and evenness and to determine the sequencing effort required to cover diversity in a sample. This study provides the first comprehensive analysis of the genetic diversity and metabolic potential of a temperate planktonic freshwater community and advances approaches for comparative metagenomics.

Lakes of temperate geographical zones are anthropogenically influenced freshwater ecosystems that are directly impacted by and affect activities such as recreation and drinking water supply. Despite the importance of such lakes, however, the microbial communities associated with these ecosystems remain poorly characterized, primarily because most large-scale genomic surveys performed to date involved extreme environments (5, 41), the open ocean (10, 35, 42), or the human microbiome (31). A few surveys of lacustrine ecosystems have appeared more recently; however, these studies either focused on hypersaline and alpine lakes (8, 27, 34) or were restricted to the 16S rRNA gene level (18, 43) and thus had limited relevance to the gene content, functional significance, and genetic variability within temperate freshwater lake microbial communities. Therefore, the extent of novel genomic diversity within communities of temperate lakes as well as how these communities differ from their marine and soil counterparts remains essentially unknown. Advancing these issues will provide a framework for assessing the impact of human activities on the diversity and ecosystem resilience of temperate lakes. Further, previous 16S rRNA gene-based studies have suggested that freshwater lakes frequently harbor a remarkable number of novel taxa (18, 43), which warrants further investigation.

To provide new insights into these questions and broaden understanding of the complexity and metabolic potential of a planktonic freshwater microbial community, we sequenced several temporally distinct DNA samples from the community residing at a 5-m depth in the mesotrophic Lake Lanier near Atlanta, GA. Lake Lanier is roughly 150 km², encompasses 114 km of shoreline, and has a maximum depth of ~50 m. It is popular with boaters, jet skiers, and others, particularly around the summer holidays. Further, it constitutes the main source of drinking water for Atlanta, a metropolitan area of ~5 million people. Hence, the characterization of the microbial community of this lake has potential implications for human activity and health which may apply to other temperate freshwater lakes.

We obtained in excess of 10 Gb of community whole-genome shotgun (WGS) DNA sequence data from four independent lake samples, using both the Roche 454 FLX Titanium and the Illumina Genome Analyzer (GA) II platforms, and performed exhaustive comparisons of these data sets against many publicly available WGS data sets representing the breadth of habitats sampled today. The latter data sets included three open ocean (Sargasso Sea, Caribbean Sea, and Eastern Pacific, also identified as GS000_S13, GS018, and GS023, respectively), two coastal ocean (Newport Harbor, RI, and South Charleston, SC; also identified as GS008 and
GS014, respectively), one estuary (Chesapeake Bay, MD; also identified as GS012), and Lake Gatun, Panama (GS020); samples of the Global Ocean Survey (GOS) project (35); one sample from a depth of 4,000 m in the Pacific Ocean (20); and one farm soil sample (40). Comparisons between each of these data sets and the Lake Lanier data sets revealed specific trends that differentiate Lake Lanier microbial assemblages from those found in marine and soil environments and defined the functional potential and microbial community structure of Lake Lanier.

MATERIALS AND METHODS

Sample characterization and DNA extraction. Samples were collected from Lake Lanier, Atlanta, GA, below the Browns Bridge at four time points: three centered around a 27 August 2009 storm event and one during the fall mixing event in November 2009. A horizontal sampler (Wildco Instruments) was used to collect samples of planktonic microbial communities at a depth of 5 m. We restricted our sampling to 5 m because it represents a well-oxygenated, highly productive layer of the water column (Table 1) and is comparable to the depths sampled in previous large-scale ocean surveys (35). Further, this depth is within the epilimnion layer, which is fairly uniform in temperature and fully oxygenated during summer stratification. A total of 10 liters of water was filtered through 0.22-μm Sterivex filters (Millipore) using a peristaltic pump. Sterivex filters were stored at −80°C until DNA extraction. DNA extraction was performed as described previously (10), with minor modifications. In brief, Sterivex units were cracked open and filters were removed and placed in microcentrifuge tubes. Lysis buffer (50 mM Tris-HCl, 40 mM EDTA, and 0.75 M sucrose) was added, followed by the addition of 1 mg/ml lysozyme and incubation at 37°C for 30 min. Samples were subsequently incubated with 1% SDS, 10 mg/ml proteinase K, and 150 μg/ml RNase for 4 h at 80°C in a rotating hybridization oven. DNA was extracted from lysate with phenol and chloroform, precipitated with ethanol, and eluted in Tris-EDTA (TE) buffer. DNA yield was about 1.5 μg per liter of water filtered. About 5 μg of the DNA aliquot was sequenced, either using the Roche 454 GS-FLX-Ti (26 August 2009 sample) or the Illumina GA II (all four samples) sequencers available at the Emory University Genomics Facility and established protocols by the manufacturer.

Sequence annotation and functional assignment. All 454 or Sanger WGS sequences were individually annotated using the Metagenpe pipeline (28) for protein-coding genes and the rRNAscan-SE version 1.21 pipeline for RNA genes (http://selab.janelia.org/rRNAscan-SE); Illumina reads (~100 bp long) were first assembled into contigs, as described below, and then annotated using the Metagene pipeline. The annotated protein sequences were subsequently searched against the GenBank (4), Pfam release 24.0 (11,912 families) (14), KEGG (19), and Clusters of Orthologous Genes (COG) (38) databases, and their best match was used to infer their functional role. For Pfam searches, the HMMER 3.0 algorithm (http://hmmer.janelia.org) was used with an E value cutoff of 0.1, as previously suggested (14). For GenBank and COG searches, the cutoff was ≥30% amino acid identity over ≥70% of the length of the query protein. For KEGG, the cutoff was a ≥60 bit score. Protein sequences that belonged to mobile genetic elements, including plasmids, prophages, and transposons, were identified by their annotations, which were refined manually as needed. To reduce potential artifacts and make protein counts directly comparable between the different WGS data sets, all protein sequences annotated by the Metagene pipeline were clustered for each data set using the Blastclust algorithm (1) with the following parameters: X = 50 (similarity threshold), L = 0.5 (min. length coverage), and b = F (two genes required to be one sequence of a pair). The remaining parameters were at the default settings. The longest amino acid sequence from each resulting cluster was extracted to compile the nonredundant protein sequence list of the corresponding data set. For consistency purposes, the exact same clustering approach was applied to all data sets used in the study. The clustering approach reduced effects associated with variable species abundance when comparing gene stoichiometries between the disparate data sets since the similarity cutoff used to build the clusters encompasses the standards for species demarcation (15), and only one single sequence from each cluster was analyzed further. To maintain quantitative information for a specific function within each data set (e.g., a COG group or a Pfam model), the nonredundant protein sequences were queried against each database, and the number of resulting matches was normalized for the size of the data set by dividing by the total number of nonredundant proteins in the data set. The normalized counts were compared between WGS data sets. Comparison of total gene contents (i.e., including hypothetical genes and genes not present in COG or Pfam databases) and average sequence identities between WGS data sets were performed based on pairwise whole-data-set Blastp searches, using the nonredundant protein list of each WGS data set and a minimum cutoff of at least 70% amino acid identity over at least 50% of the length of the query protein. To analyze functions enriched in temperate freshwater samples, we combined all Lake Lanier data sets into one representative metagenome and compared it to the metagenomes from other environments. Because all Lake Lanier data sets were similar to each other (e.g., see Fig. 2), we used the 454 WGS data set, which contained longer read lengths, as a representative set for the remaining analyses (e.g., phylogenetic diversity and community complexity).

16S rRNA gene sequence analysis and phylogenetic affiliation of reads and assembled contigs. 16S rRNA genes were identified in a Sanger- or 454-generated data set using as described previously (10, 29, and 30). An Escherichia coli 16S rRNA reference gene sequence was searched against the WGS sequences using the Blastn algorithm (nucleotide level), and the matching 16S rRNA gene fragments in the WGS sequences were extracted using a Perl script. Blastn was run with the following settings: X = 150 (drop-off value for gapped alignment), q = −1 (penalty for nucleotide mismatch), and F = F (filter for repeated sequences); the rest of the parameters were at default settings. This approach can robustly detect divergent 16S RNA genes, down to the 60 to 65% nucleotide identity level, which represents the lowest 16S RNA gene sequence identity observed among bacterial and archaeal organisms, regardless of the reference sequence used (29). The retrieved 16S rRNA gene sequence fragments, when longer than 150 bp, were aligned to their nearest taxonomic affiliates using the ARB_EDIT4 tool in ARB (25) and the 16S rRNA ARB database from the Ribosomal Database Project (RDP; http://rdp.cme.msu.edu). Distance trees were generated from these alignments using the ARB_NT tool. Taxonomic assignments were verified by submitting sequences to the sequence match tool at the RDP database (6).

Lake Lanier 16S rRNA gene amplicons were PCR amplified from the same community DNA samples sequenced for metagenomic analysis using barcoded primers for the V1 to V3 region. The pooled amplicons were sequenced on the Roche 454 GS-FLX-Ti at the Emory University Genomics Facility.

Domain-level assignment of the protein-coding genes recovered in the 454 WGS data set were performed using the STRING database, as described previously (32). The phylogenetic affiliations of the assembled contigs were determined, when possible, based on the phylogenetic analysis of 16S rRNA genes or best
Blastp (1) match of universal protein-coding genes of the contig, as described previously (20, 29). Contig assembly from 454 WGS data is described below.

Assembly and community complexity comparison. To make direct comparisons of the assemblies of the Sanger- and 454-generated WGS data sets, Sanger-sequenced metagenomic data sets were transformed in silico into 454-like reads by trimming all read sequences with a Q-15 quality Phred quality score (13) for consistency and reducing the length of the resulting sequences to about 400 bp, corresponding to the average read length of the 454 reads used in the study. All 454 sequences were also trimmed using the same Q-15 quality standards. A subset of 100,000 reads per data set was then selected at random, provided that the distribution of the length of the reads matched well with the distribution of the reads of the Lake Lanier 454 WGS data set. Newbler 2.0 was used to assemble the 100,000 reads for each data set with default settings, except for the length overlap, which was set to 50, and the nucleotide identity threshold, which was set to 90. This analysis was not performed for the Lake Lanier Illumina data sets because of the shorter Illumina read lengths and the high similarity between all Lake Lanier data sets (for example, see Fig. 2).

The SOAPdenovo (23) and Velvet (44) assemblers were used to preassemble Illumina short reads into contigs using different k-mers (usually ranging from 21 to 31). We also evaluated other popular assemblers, such as the ALLPATHS (26) and ABySS (36). The combination of Velvet and SOAPdenovo was chosen because of its overall higher accuracy, computational efficiency, and complementarities of the two assemblers. We performed six independent assemblies for each Illumina sample, using k values of 21, 25, and 29 for the three SOAPdenovo runs and k values of 23, 27, and 31 for the three Velvet runs. The resulting contigs were merged into one data set, and Newbler 2.0 was used to assemble this data set into longer contigs, as described above. This hybrid protocol provided significantly longer contigs, with accuracy comparable with or higher accuracy than that of the contigs of Velvet or SOAPdenovo (C. Luo et al., unpublished data). Different combinations of k-mers, e.g., k values of 25, 31, and 35 for Velvet and k values of 29, 33, and 37 for SOAPdenovo, did not provide substantially different assemblies with the hybrid protocol (data not shown). The resulting contigs were annotated, and the annotated protein sequences were combined with those of the 454 WGS data set to create a composite, nonredundant protein list representing Lake Lanier, as described above.

Nucleotide sequence accession numbers. The WGS data from the Lake Lanier samples are available in GenBank under the accession numbers SRA023414.2 (454 data set), SRA029309.1 (S1), SRA029314.1 (S2), SRA029315.1 (S3), and SRA029316.1 (S4) (Illumina data sets).

RESULTS

Community structure of Lake Lanier. Phylogenetic analysis revealed that the overwhelming majority (~90%) of both the 16S rRNA gene fragments and the protein-coding sequences recovered in the Lake Lanier 454 WGS data set were bacterial and archaeal. The remaining ~10% of the total protein-coding sequences were of eukaryotic origin. All major bacterial groups commonly found in temperate freshwater ecosystems (18, 43) were present in Lake Lanier (Fig. 1). Proteobacteria accounted for about 37% of the 16S rRNA gene sequences recovered in the 454 data set that were classifiable at the phylum level (n = 447 sequences), revealing that this phylum, together with Actinobacteria (134 sequences; 32%) and Verrucomicrobia (62 sequences; 14%) dominated the bacterial fraction of the community (see Table S2 in the supplemental material). In contrast to the dominance of Alphaproteobacteria in planktonic open ocean communities (29, 33), Betaproteobacteria were relatively more abundant in Lake Lanier. Cyanobacteria, such as those affiliated with the Synechococcus genus, composed a smaller fraction of the community (41 sequences; ~9%). Similar results were obtained when we sequenced PCR-amplified 16S rRNA gene amplicons from the same DNA sample used for 454 sequencing (see Fig. S1 in the supplemental material) as well as 16S rRNA gene fragments recovered in the Illumina data sets from other time points (data not shown).

A few of the groups identified were represented by multiple identical 16S rRNA gene sequences within the same sample. These groups were also well represented among the assembled contigs (see Fig. S2 in the supplemental material), indicating that they were likely abundant in situ. The two most abundant populations were phylogenetically affiliated with an uncharacterized lineage of Betaproteobacteria of the Burkholderiales order (see Fig. S3 in the supplemental material) and a previously identified but uncultivated lineage of Actinobacteria, which is frequently found to be abundant in freshwater ecosystems (30). Our estimates, based on the level of coverage of the contigs by the WGS reads and assuming a genome size similar to the average genome size of the Lake Lanier community (see also the genome size section below), indicated that the genomic DNA derived from these two populations together constituted ~10% of the total DNA in the sample. However, these two populations represented an exception to the general patterns observed; the majority (>90%) of the populations identified were represented by only a couple of identical 16S rRNA gene sequences and/or short contigs (e.g., <10 kb long) of low coverage (representing <1% of the total DNA). Further, about one-third of the total WGS reads composing the 454 data set (1.2 million reads) could not be assembled (singletons) (Table 1), consistent with a diverse natural microbial community (see also the section on community complexity below).

About half of the recovered 16S rRNA gene fragments (n = 202) could not be assigned to a characterized family (Fig. 1), suggesting that they represent uncharacterized lineages. Consistent with the level of novel diversity revealed by analysis of 16S rRNA genes, only ~47% of the total proteins recovered in the assembly of the Lake Lanier WGS data set matched a protein in a completely sequenced microbial genome among those available in GenBank at the end of 2009 with an amino acid identity of 40% or higher (~1% did at a 90% nucleotide identity cutoff). These results reveal that the Lake Lanier microbial community harbors extensive novel phylogenetic diversity compared to the genomes in the public databases.

Lake Lanier functional gene content over time. The 454 WGS sample from Lake Lanier was collected 1 day prior to a strong summer storm that brought about 8 cm of rain to the lake and was followed by 10 days of sunny weather (wind speed of <10 km/h on average; temperature of water, 28°C ± 1°C). To assess the variability of the lake gene content over time, additional planktonic community samples were sequenced from this sample (sample Illumina.S1) as well as at 1 day (sample Illumina.S2), 1 week (sample Illumina.S3), and 3 months (during the November fall mixing; sample Illumina.S4) following the storm. The Illumina GA II analyzer was used to sequence these samples and yielded 2 to 3 GB of data per sample (100-bp paired-end reads). Analyses of the protein sequences recovered in the assembled contigs from the 454 and four Illumina metagenome data sets revealed that all Lake Lanier data sets were highly similar to each other, suggesting a relatively stable gene content pattern. For example, we were unable to statistically differentiate most of the Illumina data sets based on the relative abundance of protein sequences assignable to the Cluster of Orthologous Groups (COG) major functional categories (see Fig. S4 in the supplemental material). Similar results were obtained when gene content was accessed at a higher resolution, using the 4,870 protein families of COG. For example, fewer than 100 of a total of 4,870
FIG. 1. Lake Lanier microbial community structure based on 16S rRNA genes. All sequence fragments of 16S rRNA genes recovered in the Lake Lanier 454 metagenome were queried against the RDP database to identify their closest phylogenetic relative with a complete 16S rRNA gene sequence available. The tree was constructed using the ARB parsimony tool (25) based on the complete sequences of the 16S rRNA genes of all closest relatives (one relative per fragment recovered). Shading denotes taxa of the same phylum. Eukaryotic sequences were used as the outgroup. The numbers in parentheses refer to the number of individual sequences included in a cluster. Unassigned denotes sequences that were not possible to align robustly against the RDP database.
Gene content and phylogenetic comparisons against other habitats. Gene content comparisons against publicly available metagenomic data sets revealed that the Lake Lanier microbial community was similar to the communities of surface waters of the coastal and open ocean. For example, more than 80% of the total 11,912 Pfam protein models examined showed less than a 2-fold difference in abundance between the 454 Lake Lanier and South Charleston (coastal) WGS data sets (see Table S1 in the supplemental material). Therefore, the majority of genes and pathways previously reported to characterize planktonic marine microbial communities (10, 35, 42) are also present in the Lake Lanier community.

Comparisons against available metagenomic data sets from other environments revealed, as expected, that the gene content of the Lake Lanier 454 WGS data set was functionally more similar to that of the other lake evaluated (Lake Gatun, Panama) and the coastal and open ocean than to the deep ocean and the soil (most dissimilar) (Fig. 2). This pattern was reproducible regardless of the level of resolution achieved in the analysis, i.e., whether the gene content was characterized using the 20 broad functional categories of the COG database or the 11,912 protein families of the Pfam database (Fig. 3C and D). In contrast, when the communities were compared in terms of the presence and relative abundance of the bacterial families or phyla (based on 16S rRNA genes recovered in the WGS data sets), the Lake Lanier community was as different from the open ocean communities as it was from the soil community (Fig. 3A and B; all underlying data are available in Table S2 in the supplemental material). Furthermore, comparisons of the average amino acid identity of all nonredundant protein sequences shared between all WGS data sets used in the study revealed that the Lake Lanier proteins were not more related phylogenetically to their counterparts in the ocean relative to those in the soil sample. For example, the protein sequences shared between the Lake Lanier and the open ocean, coastal, and soil microbial communities all showed an average amino acid identity of ~50% (based on best Blastp matches). Thus, the apparent high similarity in metabolic potential (but not phylogenetic diversity) (Fig. 3A) of the various aquatic communities is attributable to (mostly) different taxa with genes encoding similar functions and not to higher overlap or genetic relatedness of the taxa composing the corresponding communities.

The Lake Lanier metagenomic data set we analyzed was significantly more similar to those of Lake Gatun and the Chesapeake Bay estuary than to open ocean samples in terms of both gene content and phylogeny (Student’s t test, P < 0.01; Fig. 2 and 3). For example, the Lake Lanier data set shared 42% of its nonredundant proteins (gene content) with the estuarine data set, and the average amino acid identity of the shared protein sequences was 58% (Lake Lanier Illumina data sets provided values within 2% of the values reported here). Slightly higher similarity was observed with the Lake Gatun data set (50% overlap in nonredundant proteins; shared protein sequences showed 62% average amino acid identity). In contrast, the coastal and open ocean data sets were each significantly less similar (Student’s t test, P < 0.01) to the Lake Lanier data set; e.g., we observed 39% overlap and 50% amino acid identity between Lake Lanier and the combined open ocean data set (three samples combined). This difference was even more pronounced when the data sets were normalized for their size (i.e., searchable sequence space). For instance, when the Lake Lanier data set was compared with a random collection of 137,000 protein sequences (from the 571,347 nonredundant proteins in the combined open ocean data set), which was similar in size to the Chesapeake Bay data set (137,170 nonredundant proteins), the fraction of shared protein sequences was 30% (47% average amino acid identity). It should be also noted that the salinity of the Chesapeake Bay sample was low (3.5) compared to that of other marine samples (25–35), consistent with the higher overall similarity of this sample to Lake Lanier and Lake Gatun samples (both <0.2).

Genome size explains some of the shifts in functional gene content. Comparisons of functional gene content based on COG categories revealed that the soil community had a relatively larger fraction of genes devoted to secondary metabolism, regulation, and transport than the Lake Lanier or the open ocean communities (Fig. 4). In contrast, the Lake Lanier metagenomic data sets were enriched in genes involved in informational (housekeeping) functions, such as ribosomal proteins, relative to the soil metagenomic data set. These trends echoed the trends with larger genome size noted previously for sequenced bacterial isolates (22), which prompted us to estimate the community average genome size (AGS) for each data set using the method of Raes and colleagues (32). This method estimates genome size based on the number of
homologs of 35 universal genes found in a metagenomic data set using the STRING database. We found that the soil community indeed had a substantially larger AGS than the Lake Lanier community (5.1 versus 2.3 Mbp, respectively; Table 2). Further, when we grouped all prokaryotic genomes available in GenBank into two groups, one that included genomes with a size similar to the AGS of soil (~5 Mbp) and one with genomes of similar size to the AGS of Lake Lanier (~2 Mbp), and performed a similar analysis to the one described above for the metagenomic data sets, we observed a similar distribution of genes in COG categories (Fig. 4, inset). In fact, we noted a strong correlation between the genome-derived and the soil and Lake Lanier WGS-derived distributions of COG categories, with a linear correlation coefficient greater than 0.64 (P < 0.001; all other pairs of habitats evaluated against the same genome-derived values provided coefficients smaller than 0.5). It is unlikely that the strong correlation observed was solely attributable to biases in the selection of genomes used in the analysis, since the ~5-Mbp genome group included both nonsoil and soil isolates, none of which originated from the same sample as the soil WGS data set. Similarly, the group of 2-Mbp genomes included only a few freshwater isolates. Thus, genome size appears to explain a large fraction (e.g., >60%) of the broad gene content shifts observed between the terrestrial (soil) and freshwater or marine samples evaluated. Consistent with the above results, the deep ocean community, which had a larger AGS (2.5 Mbp) than Lake Lanier and the open (surface) or coastal ocean (1.5 Mbp and 1.7 Mbp, respectively), showed an intermediate distribution of genes in COG categories between soil and Lake Lanier or surface ocean waters (Fig. 4, dark gray bars).

**Functions enriched in the lake community.** Although we performed all-versus-all comparisons, we primarily report on selected functions that were enriched in the Lake Lanier data sets relative to the coastal oceans and estuary, because these environments are more similar in terms of their physical characteristics and functional gene contents than the other environments included in this study (Fig. 3; all data are available in Table S1 in the supplemental material). To identify functions specific to freshwater relative to saltwater, we combined the five Lake Lanier data sets and compared the combined Lake Lanier data set against the two coastal GOS samples (Newport Harbor and South Charleston) and the estuarine sample from the Chesapeake Bay. Although more than 40% of Pfam proteins showed less than a 2-fold difference in abundance between the combined Lake Lanier data set and each coastal ocean data set, several differentially enriched functions in the lake or the coastal data sets were consistently noted in all
comparisons, and these were typically related to the inherent chemistry of the environments compared. The most pronounced difference we noted was the use of distinctive osmoregulators in each environment (Fig. 5). For instance, proteins involved in transport and regulation of sodium were abundant in saltwater communities (e.g., genes encoding the Na\(^+\)/H\(^+\) antiporter subunit and the sodium-sulfate symporter transmembrane region were more than 16 and 5 times more abundant in saltwater than in freshwater data sets, respectively) but were apparently replaced by potassium transporters (which are naturally more abundant than sodium salts) in freshwater communities. Many of the most differentially abundant proteins in the lake samples, however, had hypothetical or poorly characterized functions, such as the DUF820 and DUF1597 protein families. These findings highlight that additional dramatic differences in gene content probably differentiate freshwater- from saltwater-adapted microorganisms but remain obscured within uncharacterized hypothetical genes.

Additional differences between the Lake Lanier and coastal and estuarine habitats were observed. For example, the PEP-CTERM motif was about 3 to 20 times more abundant in Lake Lanier than in each coastal ocean sample (depending on the coastal sample considered). This C-terminal motif is thought to be a protein-sorting signal involved in the transport and export of proteins, perhaps for exopolysaccharide (EPS) synthesis, attachment to surfaces, or formation of biofilms, and is typically found in abundance in soil and sediment samples (16). Hence, the abundance of the PEP-CTERM motif in Lake Lanier possibly reflected the higher prevalence of particles and the particle-associated lifestyle relative to the ocean. In agreement with these interpretations, the turbidity of Lake Lanier (222 nephelometric turbidity units [NTU]; Table 1) was higher than that typically reported for open or coastal ocean waters (<100 NTU). Additional functions characterized the lake versus the coastal oceanic communities, albeit these functions were not as differentially abundant in the corresponding data sets as those noted above. For example, we noted a 2- to 4-fold-higher abundance of photosynthesis and accessory genes, e.g., genes encoding photosynthesis antenna proteins and BLUF domain proteins involved in light sensing (17), in the lakes than in the coastal or the open ocean. Similarly, the combined Lake Lanier data set contained 2 to 16 times more carbohydrate utilization genes (encoding binding, transport, and degradation enzymes) and 20% more oxidative phosphorylation genes than the coastal ocean data sets (see Fig. S5 in

![Figure 4](http://aem.asm.org/)

**TABLE 2. Average genome size estimates for selected WGS data sets**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Avg size (Mbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lake Lanier</td>
<td>2.2</td>
</tr>
<tr>
<td>Lake Gatun</td>
<td>1.8</td>
</tr>
<tr>
<td>CO, South Charleston</td>
<td>1.7</td>
</tr>
<tr>
<td>OO, Sargasso Sea</td>
<td>1.5</td>
</tr>
<tr>
<td>Deep ocean</td>
<td>2.5</td>
</tr>
<tr>
<td>Farm soil</td>
<td>5.1</td>
</tr>
</tbody>
</table>

* Average genome size estimates were performed for the prokaryotic fraction of the corresponding microbial communities, using the method of Raes and colleagues (32).
the supplemental material). In contrast, the lakes had relatively fewer proteorhodopsin (3, 9) and urea recycling genes than the ocean (see Table S1 in the supplemental material). These findings are consistent with a more streamlined and nutrient-scavenging lifestyle in the ocean than in a mesotrophic lake. Further, several genes that were overrepresented in the Lake Lanier metagenomic data set were attributed to plasmid stabilization system and putative addiction module component protein families, which were 4 times more abundant than those in the coastal ocean samples. We also noted a significant over-representation (2 to 4 times as many; \( P < 0.01 \)) of genes encoding phage proteins, such as viral helicases, phage capsid, tail, and viral inclusion proteins, and transposases in Lake Lanier relative to those of the South Charleston and Chesapeake Bay samples (see Fig. S6 in the supplemental material). These findings indicate a possible stronger selection for genomic plasticity, which is often enriched in highly dynamic environmental conditions (2, 32).

Our comparisons also revealed that the mode of motility correlated with the physical complexity of the environment (see Fig. S7 in the supplemental material). Environments with higher physical and structural heterogeneity, such as the soil and the deep sea (10), contained up to 2-fold more motility genes related to attachment (e.g., pilus and fimbral genes) than the surface ocean or the two lakes. On the other hand, the lakes and the surface ocean had more motility genes related to swimming. The lakes also had a slight overrepresentation (about 4% more) of flagellar and associated sensing (e.g., chemotaxis) genes compared to the surface ocean, although this difference was not statistically significant when we considered each Lake Lanier data set as an independent observation.

The Lake Lanier community is composed of sequence-discrete populations. We investigated whether the Lake Lanier community is composed of sequence-discrete populations, similar to what was reported previously for marine communities (21). We searched all assembled contigs of the 454 metagenomic data set that were longer than 5 kb against the individual WGS reads and produced fragment recruitment and coverage plots, as described previously (21, 35). In almost all cases examined (\( n > 100 \)), the contigs appeared to represent sequence-discrete populations in situ, i.e., only reads that shared at least 90% nucleotide identity to the reference contig were recovered, followed by a lack of reads with identities in the 80 to 90% nucleotide identity range (i.e., a sequence discontinuity). In many cases, reads of lower than 80% nucleotide identity to the reference contig were also recovered, and these apparently represented co-occurring populations that were closely related to but genetically distinct from the population represented by the reference contig (see Fig. 6 for an example). Thus, at least the abundant members of the lake community appear to form predominantly sequence-discrete populations that were closely related to but genetically distinct from the population represented by the reference contig (see Fig. 6 for an example). This, at least the abundant members of the lake community appear to form predominantly sequence-discrete populations. It should be noted, however, that it was not possible to fully assess organisms that showed low in situ abundances since such organisms were not typically represented by long-enough contigs or enough WGS reads in our data set.

The intrapopulation nucleotide divergence appeared to vary and depended on the group considered. For example, heterotrophic populations, such as the Burkholderia-like population
shown in Fig. 6, tended to show slightly lower intrapopulation diversity than the photosynthetic Synechococcus population (~99% and 98% average nucleotide identity among fully overlapping reads, respectively; the difference is significant by Student’s t test due to the high number of reads included in the calculation, \( n > 100 \)). Such patterns may reflect a longer period of coexistence of the genotypes composing the population (thus, more time available for mutations to occur) and/or less frequent population sweeps that purge sequence diversity in the latter population relative to the former one. The phylogeny of secY gene (preprotein translocase subunit) sequences recovered from the Lake Lanier 454 data set also supported the Burkholderia results observed in the coverage plots. For example, among the four Burkholderia secY sequences, one sequence (representing the less abundant population) showed lower sequence identity (<78%) to the other three sequences (representing the more abundant population), which shared at least 98% nucleotide identity among themselves (see Fig. S3 in the supplemental material). No case was observed in which the minimum intrapopulation sequence relatedness was lower than 90% nucleotide identity (it was often closer to 100%), similar to what was observed previously for marine populations (21).

**Comparisons of community complexity across habitats.** The prevalence of sequence-discrete populations within natural microbial communities provided the opportunity to perform fine-scale comparisons of community complexity (species richness and evenness) without the caveats associated with the use of 16S rRNA genes for such purposes (e.g., identical 16S rRNA gene sequences frequently characterize genetically distinct populations [7, 39]) and the errors of single-pass sequencing. To this end, we performed assemblies of comparable subsets of the various WGS data sets and compared the number of resulting unassembled (singletons) WGS reads and the number and average length of assembled contigs as a proxy for the complexity of the corresponding communities. The following parameters were used in the assembly, which encompassed the level of intrapopulation diversity and sequence discontinuity...
observed in the coverage plots (Fig. 6): 90% minimum nucleotide identity to merge contigs or individual WGS reads and a 50-bp minimum overlap. The analysis suggested that the highest complexity was observed, as expected, in the soil community, followed by that in the open ocean and the coastal and the lake communities (Fig. 7). For example, the 100,000 randomly selected WGS reads used in the assembly process resulted in ~4,000 contigs for the soil data set, while more than 80% of the reads remained unassembled (singletons). The same values for an equal number of reads used in the assembly for the Lake Lanier data set resulted in more than 10,000 contigs and fewer than 50% of the total reads remaining unassembled.

Further support for the higher diversity in terrestrial versus aquatic samples comes from analysis of the N50 of the soil data set assembly (i.e., the contig length such that using equal or longer contigs produces half the bases of the assembly), which was substantially smaller than that of the Lake Lanier data set (375 versus 500 bp, respectively). Based on randomly drawn subsets of the 100,000 WGS reads composing each data set, we also estimated that the N50 of the soil data set increased linearly ($R^2 = 0.97$) by ~10 bp when the number of reads used in the assembly increased from 10,000 to 100,000. Based on these estimates, the assembly of the soil data set would require ~1.25 million WGS reads in order to provide an N50 equal to that of the Lake Lanier data set (i.e., a 10-bp increase in N50 for every 100,000 soil WGS reads, taking into consideration that the N50 of Lake Lanier is longer by 125 bp). Thus, the soil sample appears to be about 10 to 15 times more complex than the Lake Lanier sample based on the N50 parameter. The lake and coastal microbial communities were characterized by comparable levels of complexity according to our analysis (compare the coefficient $\lambda$ of the regression analysis in Fig. 7).

**DISCUSSION**

Using a comparative metagenomics approach, we elucidated significant differences and similarities between Lake Lanier microbial communities and those from different habitats at a level that has not previously been attained for other freshwater habitats (8, 18, 27, 34, 43). We found that although there is considerable functional overlap between communities from freshwater and marine environments, the observed Lake Lanier community was not generally interchangeable with marine communities of coastal or open ocean waters. One important genomic adaptation that accounts for this ecosystem-level difference is the preferential use of potassium-based osmoregulators and transport systems in freshwater as opposed to sodium-based ones in saltwater environments. Many additional adaptations are
probably hidden in the numerous proteins of unknown function that were differentially present in the freshwater metagenomic data sets (Fig. 5). The Lake Lanier ecosystem harbors unique or deep-branching taxa compared to those of the marine ecosystems sampled to date (Fig. 3A and B), which also highlights the great diversity of microbial communities. These findings are consistent with previous results based on comparisons of 16S rRNA gene libraries, which indicated that salinity is the most important environmental parameter determining the degree of similarity between geographically isolated microbial communities (24). It also appears that there might be a gradient of genetic and functional similarity among microbial communities as one moves from freshwater lakes to estuaries to the coastal and the open ocean (Fig. 3). These findings also indicate that the higher functional similarity of Lake Lanier and saltwater communities relative to the terrestrial/soil community is probably attributable to convergent evolution for the functions and the genome size selected by a water-based lifestyle rather than extensive overlap in the exact taxa composing the corresponding communities. It is possible that these findings apply broadly to additional temperate freshwater ecosystems that show characteristics similar to those of Lake Lanier.

The average genome size of the community in particular appeared to explain a remarkably large portion of the broad gene content shifts between the freshwater, saltwater, and terrestrial metagenomic data sets compared here (Fig. 4). The soil WGS data set also contained more genes encoding transposases and phage integrases than the deep sea, and the deep sea had more of these genes than the coastal/open ocean or the lakes (see Fig. S6 in the supplemental material). These findings are consistent with previous results based on genomes of isolates (22) and the comparison of gene contents between deep and surface oceanic microbial communities (20), suggesting that larger genomes encode comparatively more mobile genes than medium or small genomes. Apparently, environmental selection for metabolic versatility and little penalty for slow growth, which presumably promote larger genome size and higher content of mobile elements in the genome (22), characterize terrestrial relative to aquatic ecosystems and the deep versus the surface ocean. Thus, genome size seems to reflect the ecological and environmental selection pressures acting upon the indigenous microbiota and represents an important parameter for describing natural communities and habitats.

Temperate freshwater lakes such as Lake Lanier are thought to be highly dynamic due to the large role of allochthonous nutrient inputs. This was evidenced by higher genomic plasticity and more genes mediating copiotrophic metabolism (e.g., r strategists) in the Lake Lanier metagenomic data set than in the ocean metagenomic data sets (e.g., K strategists) (Fig. 5; see also Fig. S5 in the supplemental material). The time series Illumina data also indicated that the above-described patterns do not represent a temporal effect of runoff or another environmental perturbation but rather constitute a stable signature of the Lake Lanier microbial community (e.g., Fig. 2). Finally, the freshwater communities examined here generally had lower species richness and evenness than the open ocean and the soil communities (most complex) (Fig. 7). In Lake Lanier, this was mostly attributable to a few (probably fewer than 20) populations that were substantially more abundant than the low-abundance (e.g., <0.1% of the total) populations composing the majority of the community (see Fig. S2 in the supplement).

The abundant populations were apparently favored by the prevailing in situ conditions at the time of sampling; we are currently following these populations over time in order to more fully understand what biotic or abiotic factors underlie their in situ abundance.

In general, most of our findings were reproducible regardless of whether the data sets were grouped by habitat type or analyzed as individual WGS data sets. For example, ~70% of the total 11,912 Pfam models analyzed showed less than 20% variation in relative abundance in individual pairwise comparisons between the Lake Lanier 454 data set and each of the three open ocean data sets. These findings are consistent with previous studies that reported high reproducibility of 454-based DNA sequencing of the same sample (37) or habitat (34) and high stability of the metabolic potential of natural freshwater microbial communities (11, 34), corroborating our assertion that the results reported here are robust and likely to be characteristic of temperate freshwater communities.

The identification of sequence-discrete populations in the Lake Lanier metagenomic data sets (Fig. 6) provides the means to examine with fine-scale resolution how natural microbial populations are impacted by human activities and major environmental perturbations. For example, we were able to identify some of the major microbial players and functions that could be targets for future work examining the response and resilience of these communities to environmental changes and human perturbations. Such studies will advance understanding of the importance of freshwater microbial communities for ecosystem and human health.

ACKNOWLEDGMENTS

We thank the personnel of the Emory University Genomics Facility for their assistance with sequencing of the Lake Lanier metagenomic samples.

This research was supported, in part, by the U.S. Department of Energy (award no. DE-FG02-07ER64389) and, in part, by the NSF (award no. 0967130).

REFERENCES