APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Mar. 2009, p. 1688–1696 Vol. 75, No. 6
0099-2240/09/$08.00+0 doi:10.1128/AEM.01210-08

Molecular Diversity of a North Carolina Wastewater Treatment Plant as Revealed by Pyrosequencing[▽]†

Nina Sanapareddy,[1] Timothy J. Hamp,[1] Luis C. Gonzalez,[1] Helene A. Hilger,[2] Anthony A. Fodor,[1]* and Sandra M. Clinton[3]

Bioinformatics Research Center,[1] Department of Civil and Environmental Engineering,[2] and Department of Biology,[3] University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, North Carolina 28223

Received 30 May 2008/Accepted 16 December 2008

We report the results of pyrosequencing of DNA collected from the activated sludge basin of a wastewater treatment plant in Charlotte, NC. Using the 454-FLX technology, we generated 378,601 sequences with an average read length of 250.4 bp. Running the 454 assembly algorithm over our sequences yielded very poor assembly, with only 0.3% of our sequences participating in assembly of significant contigs. Of the 117 contigs greater than 500 bp long that were assembled, the most common annotations were to transposases and hypothetical proteins. Comparing our sequences to known microbial genomes showed nonspecific recruitment, indicating that previously described taxa are only distantly related to the most abundant microbes in this treatment plant. A comparison of proteins generated by translating our sequence set to translations of other sequenced microbiomes shows a distinct metabolic profile for activated sludge with high counts for genes involved in metabolism of aromatic compounds and low counts for genes involved in photosynthesis. Taken together, these data document the substantial levels of microbial diversity within activated sludge and further establish the great utility of pyrosequencing for investigating diversity in complex ecosystems.

Although largely invisible in the urban landscape when they are functioning well, wastewater treatment plants are integral to the municipal obligation to protect public health, aquatic ecosystems, and the quality of life. At the heart of wastewater treatment plants is a process whereby a dense microbial consortium is employed to remove organic and nutrient contaminants. The microbes used to treat wastewater are a crucial tool in environmental protection. The current use of molecular techniques that do not require the isolation and cultivation of microorganisms (1, 33), including 16S rRNA analysis (6, 13, 20) and fluorescent in situ hybridization (8), have greatly expanded our understanding of wastewater microbial communities. Researchers have identified many bacteria of importance to wastewater treatment, including the bacteria involved in biological phosphorus removal (5, 16, 29), nitrifiers (8, 19, 25), denitrifiers (3, 12, 17), and methanogens (18, 36). Molecular techniques have also improved our understanding of fundamental processes such as nitrification and denitrification, as well as plant upsets, such as foaming (9, 24), which can decrease treatment efficiency.

In this paper, we apply recently developed pyrosequencing technology to probe the molecular diversity of the aerobic basin of a wastewater treatment plant in Charlotte, NC. In line with other studies of complex microbial communities (28, 32), we observed astounding levels of diversity. We found that substantial regions of the genomes of the most prevalent microbes in the wastewater treatment plant are poorly described by existing sequence databases. Our results demonstrate that despite recent technological advances that allow identification of microorganisms, the microbial population of wastewater treatment plants remains undersampled and inadequately characterized. Our results are a first step toward more complete molecular characterization of this important microbial community.

MATERIALS AND METHODS

The Mallard Creek Water Reclamation Facility is located in Charlotte, NC. The plant has an average daily inflow of 7.5 million gallons, and the wastewater is mostly domestic, with additional input from the University of North Carolina at Charlotte, University City Carolinas Medical Center hospital, and several industrial users. A schematic diagram of the flow through the plant is shown in Fig. S1 in the supplemental material. Influent raw wastewater is screened and sent through grit removal before it is routed to day tank equalization basins that distribute the flow among three primary clarifiers. Primary effluent enters anoxic basins, where it is joined by recycle flow from the aeration basins. Effluent from the anoxic basins enters aeration basins (solids retention time, ~8 days) and then flows to secondary clarifiers. Clarified effluent is routed to denitrification filters and then to UV disinfection before discharge into Mallard Creek.

The plant National Pollutant Discharge Elimination System permit requires the plant to meet a monthly 5-day test for carbonaceous biochemical oxygen demand of 4.2 mg/liter in the summer and 8.3 mg/liter in the winter months. Ammonia nitrogen (NH_3-N) levels must be below 1 and 2 mg/liter in the summer and winter, respectively. There are no other nitrogen or phosphorus limits. The total suspended solids are limited to a maximum of 30 mg/liter, and the pH must be between 6 and 9. Fecal coliform counts must be less than 200 CFU per 100-ml sample. These limits are routinely met by the plant unless there are extreme weather events or plant upsets. Wastewater entering the secondary treatment system was monitored over a 6-month period for filtered flocculated chemical oxygen demand, a good estimator of readily biodegradable soluble organics, and the values ranged from 40 to 75 mg/liter. The ammonia nitrogen concentrations in this same flow ranged from 12 to 24 mg/liter, with the concentration varying in part due to return flow from digested sludge dewatering.

On the morning of 20 March 2007 we collected a 50-ml sample from the aeration basin using a plastic dipper. At the time of sample collection, the temperature in the aeration basin was 18.5°C and the pH was 6.5. The sample was decanted to remove as much foam as possible before the liquid was trans-

* Corresponding author. Mailing address: Bioinformatics Research Center, Cameron 212, UNC Charlotte, 9201 University City Blvd., Charlotte, NC 28223. Phone: (704) 687-8214. Fax: (704) 687-6610. E-mail: anthony.fodor@gmail.com.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 29 December 2008.

ferred to a sterile tube. DNA was extracted from the sample using a Mo Bio UltraClean Water DNA kit. The sample tube was inverted several times to maximize homogeneity, and a 10-ml aliquot was removed and pipetted onto the provided filter (0.22 μm). The filtrate was discarded, and DNA was extracted from the membrane using the manufacturer's protocol. The final DNA extract was analyzed to determine its purity and concentration using a NanoDrop ND-1000 spectrophotometer. Approximately 100 μl of extracted DNA was concentrated in a vacuum centrifuge and resuspended in about 12 μl of molecular-grade biology water. The final sample concentration was 479 ng/μl as determined by a NanoDrop spectrophotometer. Preliminary analysis of the DNA using denaturing gradient gel electrophoresis indicated that there was substantial diversity in the observed bands, confirming that our DNA extraction was successful (data not shown). The sample was submitted to 454 Life Sciences for pyrosequencing by the 454-FLX platform. The methodology underlying pyrosequencing has been documented elsewhere (22).

The bioinformatics analyses used in this study are shown in Fig. S2 of and described in the supplemental material.

Nucleotide sequence and quality score accession number. Sequences and quality scores from our pyrosequencing run have been deposited in the NCBI short-read archive under accession number SRA001012.

RESULTS AND DISCUSSION

Our sequence set largely fails to assemble, although contigs that were generated from the assembly include many transposons as well as many genes encoding hypothetical proteins. Our pyrosequencing run yielded 378,601 sequences with a read length of 250.4 ± 29.1 (mean ± standard deviation). The distribution of sequence lengths was approximately normal, with a small left tail indicating some short reads (see Fig. S3 in the supplemental material). We attempted to assemble sequences in this data set using version 1.1.02 of the GS De Novo Assembler of the Genome Sequencer FLX data analysis suite with the default parameters applied. This assembly algorithm attempts to combine individual sequence reads into longer "contigs." Given that metagenomic data sets of complex ecosystems have been extremely resistant to assembly (23, 28), we expected to see very little assembly in our data set. The 454 sequence assembler defines a "large" contig as a contig that consists of at least 500 bp. Because our average sequence length was ~250 bp, this threshold could be achieved with overlap of a modest number of our sequences. Despite this, only 1,154 (or approximately 0.3%) of our reads were recruited into 117 contigs greater than 500 bp long (the sequences of these contigs are available in File S1 in the supplemental material). To assign possible functions to these contigs, we used the GenMark algorithm (4) to predict genes on our contigs and then performed a Blastp search of these predicted proteins against the pfam database. This method produces more assignments than other approaches, including those based on profile searches (see the supplemental bioinformatics methods for details). With an E-value cutoff of 0.01, this approach found matches for 75% (88/117) of our large contigs (see File S2 in the supplemental material). Of these matches, 22% (20/88) were to hypothetical proteins and 21% (19/88) were to transposases. The prevalence of transposase gene sequences in our assembled contigs suggests that transposons are much more strongly conserved across metagenomes than other genomic regions. The prevalence of contigs for hypothetical proteins shows that the function of many of the highly conserved regions of our metagenome is poorly understood.

This failure of the 454 assembly algorithm to assemble 99.7% of our sequence reads emphasizes the great diversity of the microbial community within the treatment plant. Because previous studies found a similar failure of assembly algorithms for metagenomic communities characterized by Sanger sequencing (28), as well as for simulated data sets created by sampling Sanger sequencing reads (23), we would not expect a significantly improved degree of assembly even if our sequence reads were longer.

The majority of taxa in the wastewater treatment plant cannot be classified at the genus level. In order to discover the 16S rRNA genes within our data set, we downloaded the 16S rRNA gene FASTA DNA sequences from version 9.52 of the Ribosomal Database Project (RDP) (7) and used these sequences to create a BLAST database. Using the Blastn algorithm, we asked which of our 378,601 query sequences could be found in this RDP database with an E-value of ≤0.01 (see supplemental bioinformatics methods for details). The resulting 648 sequences (available as File S3 in the supplemental material) were run through the RDP classification algorithm (34). The RDP classifier algorithm uses Bayesian statistics to assign taxa to 16S rRNA gene sequences. The output of this algorithm includes a confidence score, which ranges from 0 to 100, that indicates the degree of confidence that can be assigned to the classification based on the results of 100 bootstrap trials (see reference 34 for more details). The recommended threshold for assignment of a taxon by the RDP algorithm is a confidence score of ≥80. Because sequence reads as short as 90 bp have been shown to be long enough to accurately characterize taxa (15, 21), we anticipated that our results would not be substantially different even if we had a read length of more than 250 bp.

The classifications of the 148 16S rRNA sequences that could be assigned to a phylum with a confidence score of ≥80 are shown in File S4 in the supplemental material and are summarized in Fig. 1. In another paper (T. J. Hamp, W. J. Jones, and A. A. Fodor, submitted for publication), we show that these classifications of 16S rRNA sequences derived from the whole-genome wastewater sequence set are well correlated with results from PCR experiments targeting the 16S rRNA gene. At the phylum level, the observed taxa are dominated by the *Proteobacteria*, with ~70% of the classifiable taxa belonging to this category (Fig. 1, top panel). Moving from phylum to genus, fewer of the sequences can be classified with an RDP confidence score of at least 80%. At the genus level, nearly 60% of the sequences cannot be classified at an RDP threshold of 80, and, among the taxa that can be classified, there is no dominant taxon (Fig. 1). These data demonstrate the extraordinary microbial diversity of activated sludge and are consistent with reports for other complex environments (14, 28, 30). We note that the inability of the RDP algorithm to classify these sequences to taxa with high confidence is not primarily because our 16S rRNA sequences have never been observed previously. Figure 2 shows that many of the sequences with RDP scores of <80% (to the left of the vertical lines) have very high levels of identity with previously described sequences. These results demonstrate that for wastewater treatment plants, as is the case for other complex ecosystems, the accumulation of 16S rRNA sequences in public databases is vastly outpacing our ability to classify these sequences and that this problem becomes more pronounced as one moves from the

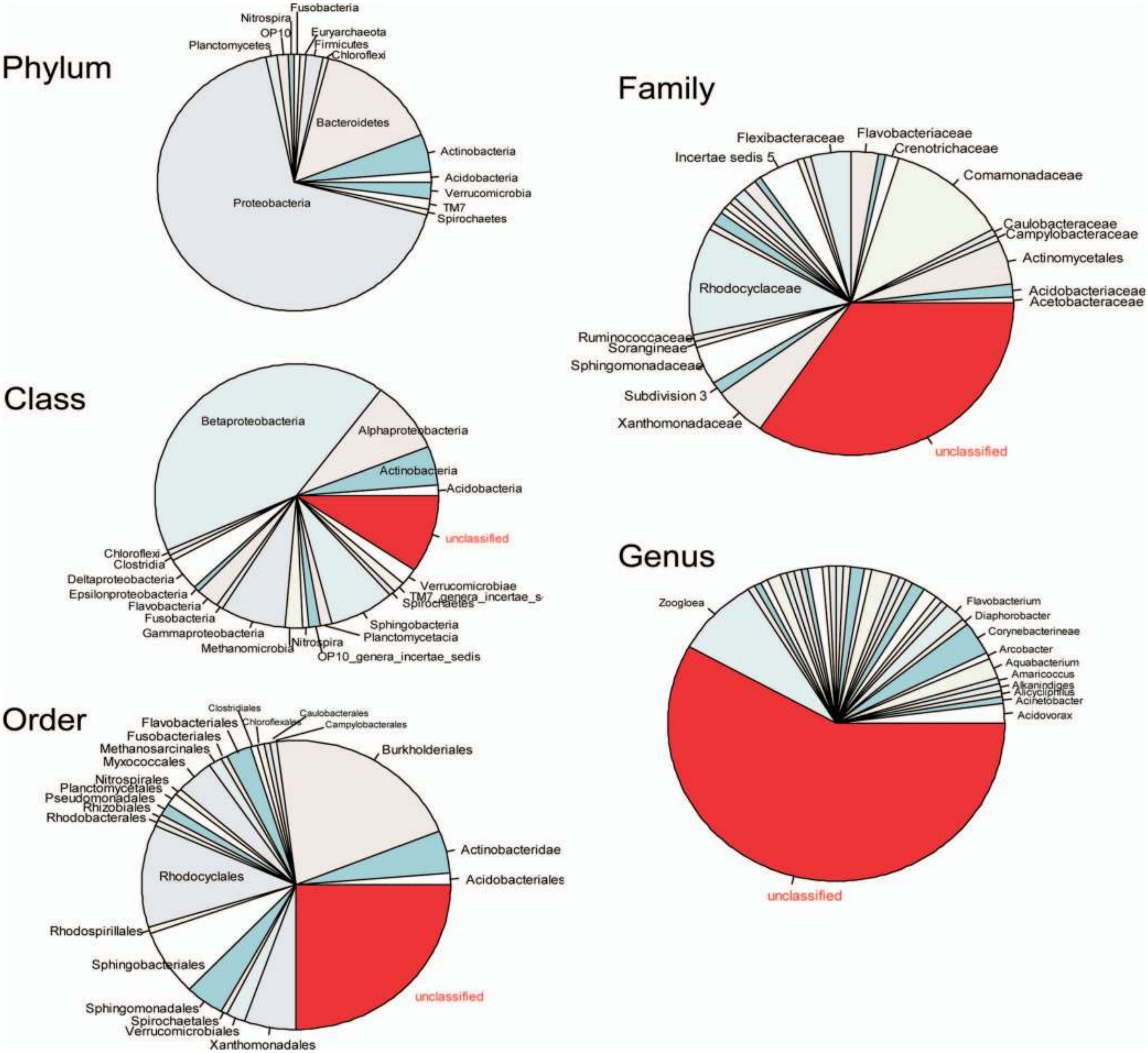

FIG. 1. Pie charts showing taxonomic assignments for 148 16S rRNA sequences in our data set that could be classified to the phylum level with RDP confidence scores of ≥80. At the phylum level, the Simpsons diversity index is 0.48.

phylum level to the genus level. Presumably, future annotation efforts will rectify this problem.

16S rRNA gene sequences from freshwater, soil, and other wastewater studies dominate our sequence set. For each of the 648 sequences in our pyrosequencing data set that matched sequences in the RDP 16S rRNA database (version 9.52) at an E-value cutoff of ≤0.01, we manually annotated where the corresponding RDP sequence was discovered. This was done by manual inspection of the GenBank records for these 648 sequences. The results of this annotation are shown in File S5 in the supplemental material and in Fig. 3. The *x* axis of Fig. 3 indicates our classification, while the *y* axis indicates the E-value with which the top hit from each of our query sequences matched the RDP database sequence. We found that while a large number of environments had at least one hit, if we restricted ourselves to environments with multiple hits at high stringency (i.e., low E-value), only three environments are well represented: freshwater, soil, and other wastewater studies (Fig. 3). While, of course, the low number of sequences for some of the other environments may simply reflect the low number of sequences from those environments in the RDP 16S rRNA database, there is a strikingly small number of sequences with high scores that are related to two 16S rRNA populations that are well represented in the database: marine and human. The relatively small number of human-derived 16S rRNA sequences observed is particularly interesting given the vast number of human microbes deposited

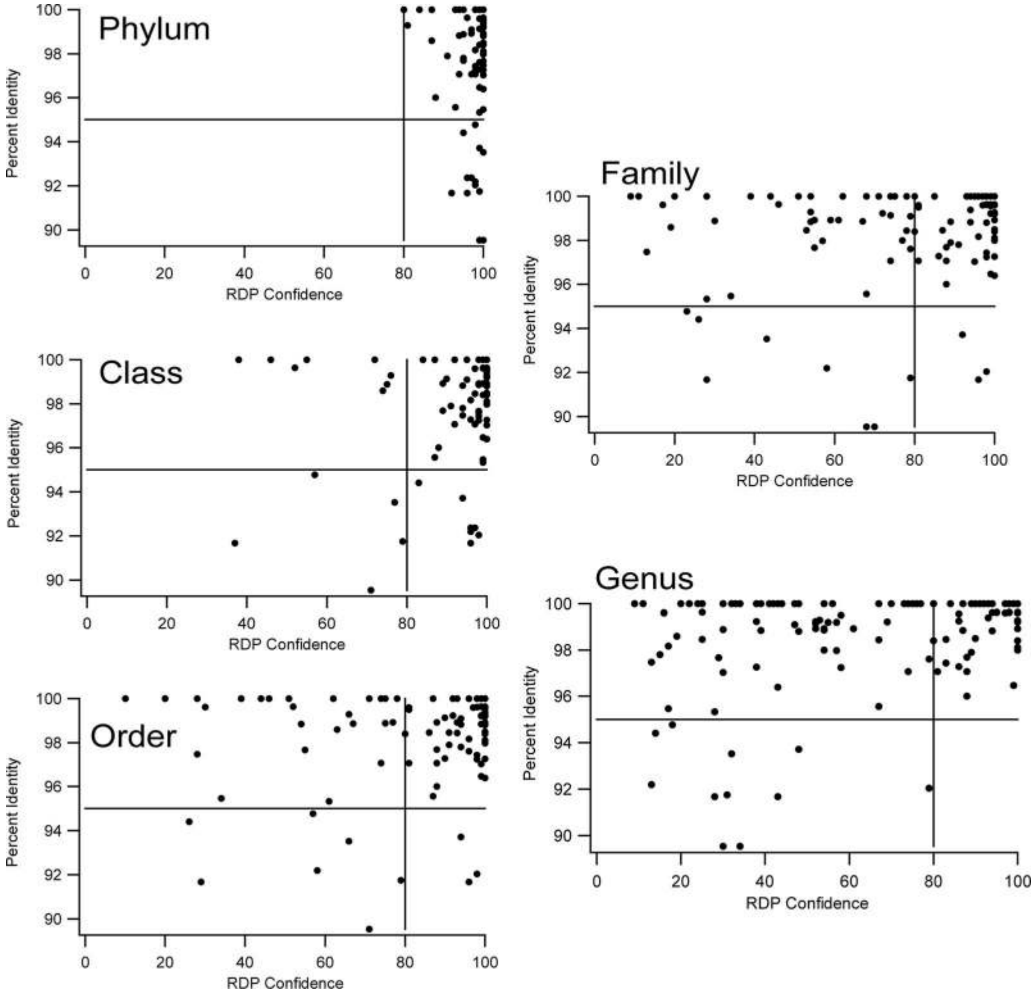

FIG. 2. Results obtained with the RDP classification algorithm for 148 16S rRNA sequences that can be assigned at the phylum level with a confidence score of ≥80. The *x* axis of each graph shows the confidence in assignments as reported by the RDP classification algorithm. The *y* axis of each graph shows the level of identity (expressed as a percentage) between our query sequence and the best Blastn hit in the RDP database (version 9.52). The horizontal and vertical lines indicate 95% sequence identity and an RDP confidence score of 80, respectively.

in the wastewater treatment plant each day. These results show that the environment within the wastewater treatment plant exhibits strong selection pressure against the microbes that are present in human feces.

Sequenced bacterial genomes are not well represented in the wastewater metagenome. When Blastn is used to compare our sequences to the nt database, only 34% of the sequences (73,274/378,601) match the nt database even at a relaxed threshold E-value cutoff of 0.01 (data not shown; see the supplemental bioinformatics methods). For the vast majority (over 98%) of the sequences that do match the nt database at this threshold, the best hit is to bacterial taxa (data not shown; see the supplemental bioinformatics methods for details). Since wastewater treatment plants are known to harbor many eukaryotes (27), this result likely reflects our DNA isolation strategy, which was designed to capture prokaryotic DNA, rather than the "true" ratio of prokaryotes to eukaryotes in the treatment plant.

As of November 2008, there were 772 complete bacterial genomes in the NCBI database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.fna.tar.gz). In order to explore how well these known genomes are represented in the treatment plant, we used Blastn to compare our wastewater sequences to the 1,442 assembled genome and plasmid sequences from the 772 sequenced bacteria. In order to eliminate spurious hits, we required that any hit matched at least 75 nucleotides in our query sequence (see the supplemental bioinformatics methods for more details). Because our average sequence length was 250.4 bp, this is not an overly conservative criterion. Using this criterion, only 20% (73,274/378,601) of our sequences matched any of the known bacterial genomes. This result again reflects the great diversity of organisms in the wastewater treatment plant and emphasizes a key challenge for genomics; despite the considerable effort that has been expended in microbial genome projects, the great majority of our sequence reads are not found in known genomes.

For the sequences that do match known genomes, we can determine how closely the sequenced genomes of cultivated organisms match the genomes present in our wastewater metagenome. We calculated for each of the 1,442 assembled se-

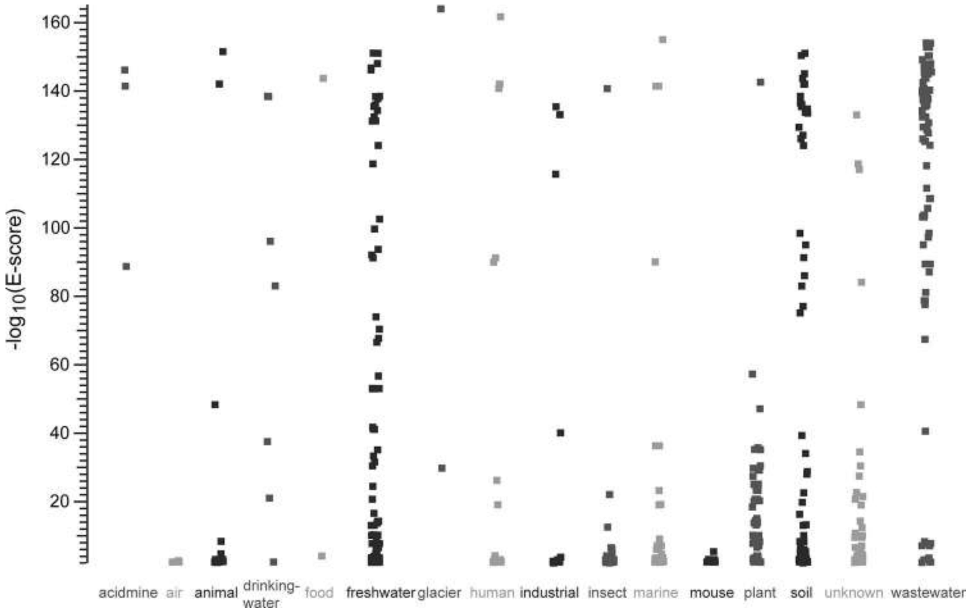

FIG. 3. Locations (as determined by manual annotation) and E-values of sequences from the 648-member pyrosequencing data set that matched the RDP 16S rRNA database at an E-value cutoff of 0.01.

quences from the 772 finished genome projects the number of nucleotides in that genome that have a Blastn match that aligns with at least one of our wastewater sequences. Dividing this number by the total length of each assembled sequence yielded the "fraction of the genome covered." Figure 4 shows that even for the bacterium with the most well-represented assembled genome, nitroaromatic compound degrader *Acidovorax* sp. strain JS42 (accession number NC_008782), only 25% of its genome sequence matched our wastewater metagenome. Table 1 shows that the fraction of the genome covered is similarly poor for the 10 genomes that recruited the most reads from our wastewater metagenome.

Figure 5 shows a recruitment graph for our wastewater treatment plant for the assembled *Acidovorax* sp. strain JS42 chromosome, which recruited the most sequences from our wastewater genome (Table 1; see File S6 in the supplemental

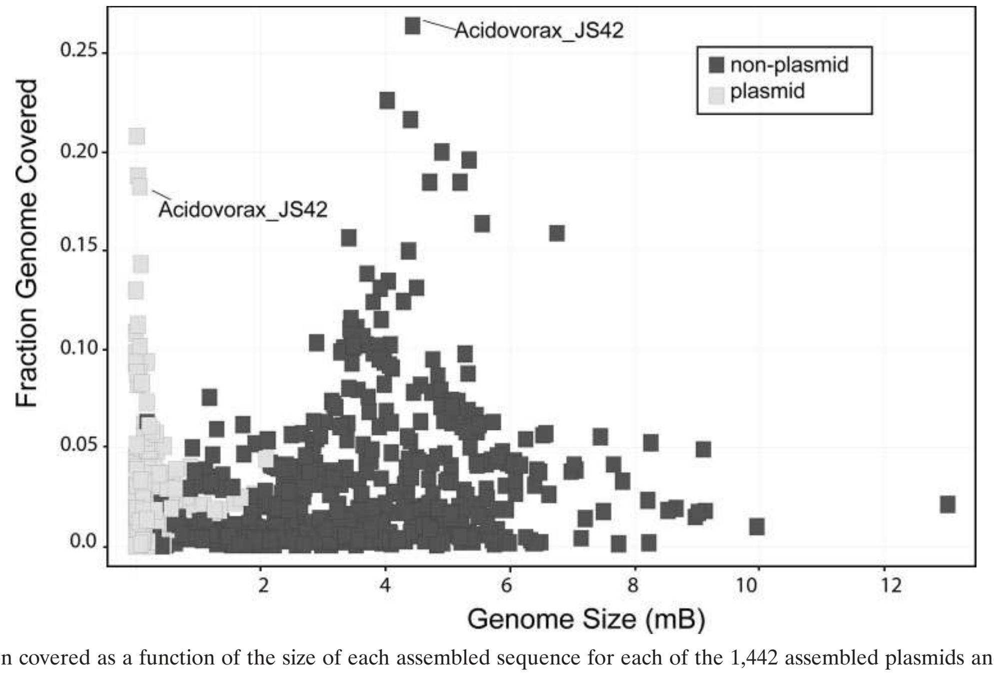

FIG. 4. Fraction covered as a function of the size of each assembled sequence for each of the 1,442 assembled plasmids and chromosomes in the NCBI datadase. The fraction covered is defined as the number of nucleotides in the assembled sequence that match at least one of our wastewater sequences divided by the total number of nucleotides in the assembled sequence.

TABLE 1. Top 10 assembled microbial genomes as sorted by the number of hits recruited from our wastewater metagenome[a]

No. of hits	Fraction of genome covered	Annotation
18110	0.26	gi 121592436 ref NC_008782.1 *Acidovorax* sp. JS42
17341	0.20	gi 120608714 ref NC_008752.1 *Acidovorax avenae* subsp. *citrulli* AAC00-1
17100	0.16	gi 160895450 ref NC_010002.1 *Delftia acidovorans* SPH-1
16800	0.20	gi 171056692 ref NC_010524.1 *Leptothrix cholodnii* SP-6
15752	0.23	gi 124265193 ref NC_008825.1 *Methylibium petroleiphilum* PM1
15695	0.22	gi 121602919 ref NC_008781.1 *Polaromonas naphthalenivorans*
15468	0.18	gi 91785913 ref NC_007948.1 *Polaromonas* sp. JS666
14735	0.16	gi 121607004 ref NC_008786.1 *Verminephrobacter eiseniae* EF01-2
13590	0.18	gi 89898822 ref NC_007908.1 *Rhodoferax ferrireducens* T118
11595	0.15	gi 119896292 ref NC_008702.1 *Azoarcus* sp. BH72

[a] A complete list of all assembled microbial genomes is shown in File S6 in the supplemental material.

material). The *x* axis shows the positions where sequence reads are mapped with Blastn against the *Acidovorax* genome. The *y* axis shows percent identity of a read compared to the matching subsection of the *Acidovorax* genome. Figure 5 shows sequences from two different sources: our 20 March aeration basin pyrosequencing run and the environmental sequence database from NCBI downloaded in June 2007, which at that time was largely dominated by sequences from the J. Craig Venter Institute's Global Ocean Sampling (GOS) expedition (28). We included the environmental sequence database because we wanted to assess how specific our wastewater treatment plant results were relative to other metagenomic sequencing databases.

The pattern shown in Fig. 5 is typical of nonspecific recruitment. For regions of the genome with conserved genes, both sources of sequence matched the genome, but the levels of identity were usually below 90%. For regions of the genome that are poorly conserved, such as the region encoding the putative transmembrane protein (Fig. 5), very few sequences from either source mapped to the genome. We observed similar patterns of nonspecific recruitment over a number of the genomes that recruited large numbers of sequence reads in our 20 March aeration basin data set (data not shown). In the Venter Institute's GOS survey, a similar pattern of nonspecific recruitment was observed for nearly every known microbial genome despite the presence of over 7,600,000 sequences in the data set (28). This result is one of the principal reasons that

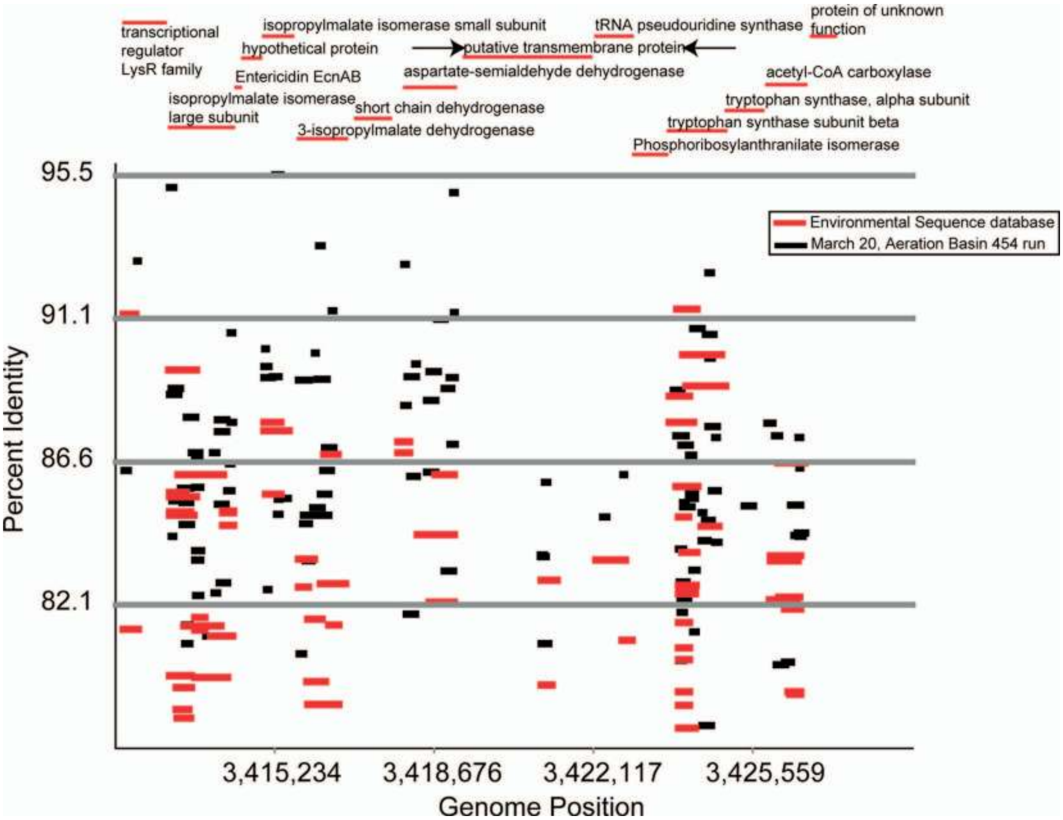

FIG. 5. Nonspecific recruitment against the *Acidovorax* sp. strain JS42 genome. BLAST hits with alignment lengths less than 75 nucleotides (for the 20 March run) or 250 nucleotides (for the environmental sequence database) were removed. Protein annotations are derived from the full NCBI core nucleotide report for the *Acidovorax* sp. strain JS42 genome (http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nuccore&id=121592436).

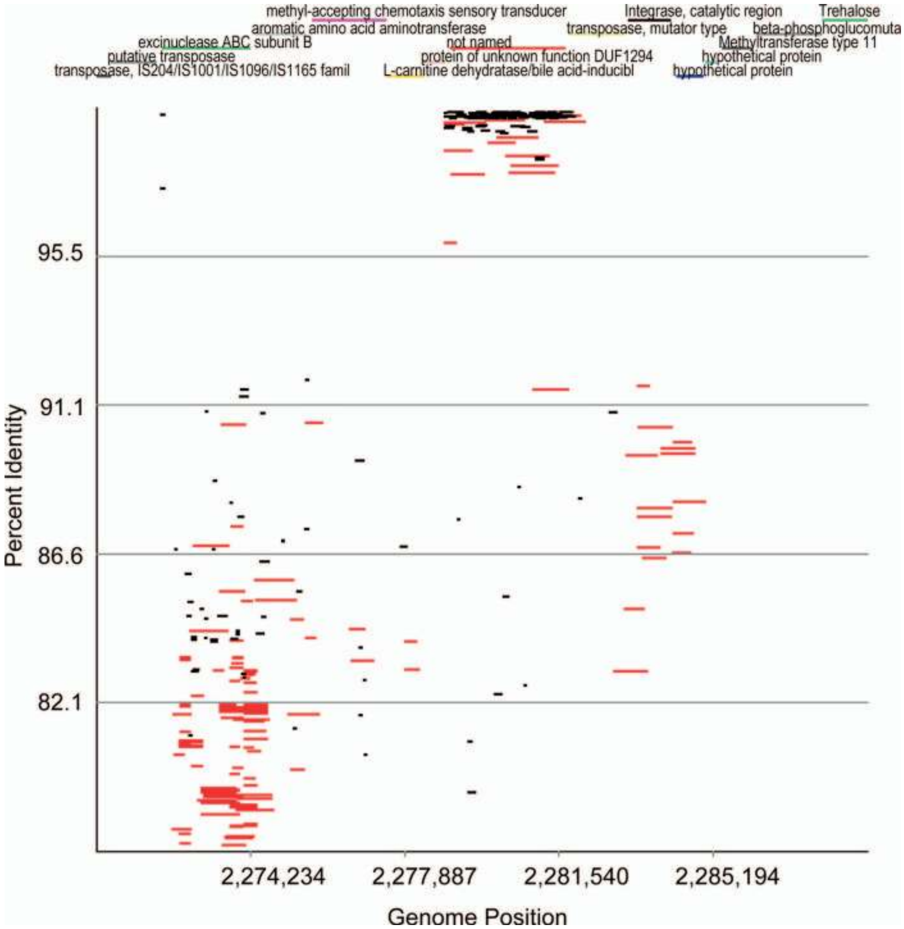

FIG. 6. Region involving a transposase from the JS42 genome that shows an exception to the pattern of nonspecific recruitment. For visualization, a small amount of random noise was added to the *y* axis (as otherwise most of the hits to the transposase region would be superimposed). The red sequences matching the transposase region are from the GOS (28).

the GOS study concluded that microbial diversity in the oceans is profound (28). Our results show that the genomes of the most abundant microbes in the wastewater treatment plant are largely uncharacterized. Moreover, the pattern of nonspecific recruitment shown in Fig. 5 suggests that even additional whole-genome shotgun sequencing would not improve the match between known genomes and the sequences observed in our metagenome.

The great diversity of our wastewater metagenome caused very few contigs to be assembled. Of the sequences that were joined as contigs, a substantial fraction involved transposases. We might expect, therefore, a different pattern of recruitment around transposons. Figure 6 shows a region of the *Acidovorax* genome around a transposase gene sequence with a stark exception to the pattern of nonspecific recruitment. A large number of sequences from our metagenome recruited to this region with a nearly perfect match. Interestingly, a number of marine sequences from the GOS (28) also matched the region around this transposase gene, suggesting that, unlike most genomic regions, parts of this transposon are conserved across a wide environmental space.

One genome of particular interest that has not yet been deposited as an assembled genome in the NCBI database is the genome of "*Candidatus* Accumulibacter phosphites," a taxon that dominates two recently sequenced lab-scale enhanced biological phosphorus removal sludges (11). Although the fully assembled genome was not available, we did perform a BLAST comparison of our wastewater metagenome to the largest assembled contigs from the "*Candidatus* Accumulibacter" project, and we observed nonspecific recruitment (data not shown). This suggests that the "*Candidatus* Accumulibacter phosphatis" taxon is not a dominant member of our North Carolina wastewater metagenome.

When mapped to protein space, the wastewater metagenome displays a distinct metabolic profile. By translating our nucleotide sequences in all six frames and mapping the translated sequences to known proteins, we can generate a distinct metabolic profile for our wastewater sequences. This approach, asking which genes a microbial community is capable of producing, has been successfully used to analyze the metabolic signatures of a number of metagenomic sequence sets (10, 31). To perform this analysis, we submitted our pyrosequencing data set for annotation on the SEED platform (2, 26). Within SEED, metabolic pathways are classified in a hierarchical structure in which all of the genes required for a specific task are arranged into subsystems. At the highest level of organi-

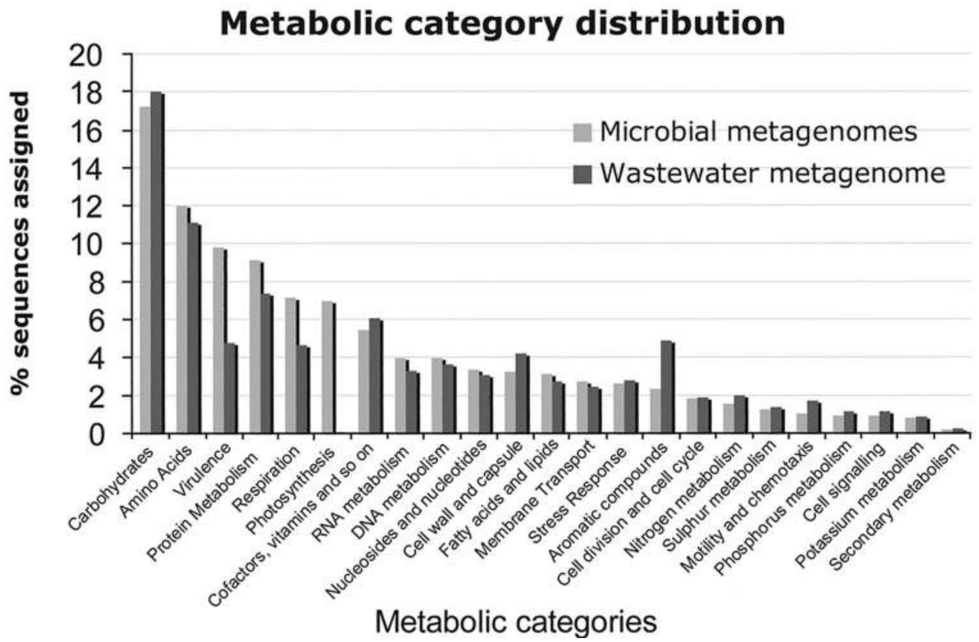

FIG. 7. Functional categories provided for our data set by the SEED server (http://www.theseed.org). The data for microbial genomes are averages for sequences gathered from multiple biomes (10).

zation, the subsystems include both catabolic and anabolic functions (for example, DNA metabolism), and at the lowest levels the subsystems are specific pathways (for example, the synthesis pathway for thymidine). Using the Blastx algorithm and an E-value cutoff of 0.001, the SEED database was able to assign ~60% of our sequences. The result of assigning these sequences to functional categories is shown in Fig. 7. For comparison, Fig. 7 shows the mapping to functional categories from a recently published survey of 1,040,665 sequences from 45 microbial metagenomes collected from nine distinct biomes (10). We note that compared to the "average" profile of these nine biomes, the wastewater treatment plant has a distinct metabolic signature. For example, compared to other biomes, the wastewater treatment plant contains almost no genes coding for proteins involved in photosynthesis. We would expect this as the primary energy source for the microbes at this treatment plant is the organic material being processed by the plant. In addition, genes involved in the degradation of aromatic compounds are expressed at a much higher rate in the wastewater treatment plant than in other metagenomic systems. Again, we might expect this given the nature of household and industrial wastes present in sewage. Finally, we note that the Mallard Creek Wastewater Treatment Plant has no additional biological nutrient removal facilities to treat phosphorus. Consistent with this, the percentage of sequences assigned to genes involved in phosphorus metabolism appears to be lower than that for genes involved in nitrogen metabolism in the activated sludge (Fig. 7).

Summary. We are at the beginning of a sequencing revolution. The 91 million base pairs in the sequence data described in this paper were generated from a single sequence run with a 454-FLX instrument generating over 6,000 bp of sequence per dollar. This is an approximately 10-fold-lower cost per base pair than Sanger sequencing, and moreover, this procedure eliminates the costly and time-intensive step of creating a bacterial clone library. As new sequencing technologies continue to be developed, we can expect both the cost and the experimental effort associated with metagenomic sequencing projects to drop exponentially.

Perhaps the most surprising result of our study is the pronounced conservation of transposases across widely different environments. While there is generally poor agreement between sequences from the GOS and known genomes (28) and between our wastewater genomes and known genomes (Fig. 4 and 5), there are a few regions of conservation involving transposons (Fig. 6) where there is a pronounced match between the metagenomes and the sequenced genomes. A substantial fraction of the contigs that could be assembled from our data set involved strongly conserved transposases. It is an open question why transposons have escaped the pronounced sequence mutability that mark nearly all of the rest of bacterial genomes.

Like the results of other metagenomic projects (28, 32, 35), our results point to the extraordinary diversity of microbial communities. Patterns of nonspecific recruitment to known genomes suggest that the structures of the genomes of the most abundant organisms in the wastewater treatment plant are unknown (Fig. 4 and 5). Despite the great diversity of microbes in the treatment plant, analysis at the protein level is surprisingly tractable, with the sequences from the treatment plant displaying a distinct metabolic profile consistent with what we would expect based on the plant's function (Fig. 7). This suggests that despite the great complexity of microbial communities, next-generation sequencing technology will be a useful tool for monitoring changes in microbial processes across time and space. As treatment requirements become more stringent and monitoring expands to address a broadening group of compounds of concern, probe-free sequencing will increase the rate at which key microbial groups can be identified and selected for to optimize contaminant removal.

REFERENCES

1. **Amann, R., H. Lemmer, and M. Wagner.** 1998. Monitoring the community structure of wastewater treatment plants: a comparison of old and new techniques. FEMS Microbiol. Ecol. **25:**205–215.
2. **Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko.** 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics **9:**75.
3. **Beline, F., J. Martinez, C. Marol, and G. Guiraud.** 2001. Application of the 15N technique to determine the contributions of nitrification and denitrification to the flux of nitrous oxide from aerated pig slurry. Water Res. **65:**2774–2778.
4. **Besemer, J., and M. Borodovsky.** 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. **33:** W451—W454.
5. **Bond, P. L., R. Erhart, M. Wagner, J. Keller, and L. L. Blackall.** 1999. Identification of some of the major groups of bacteria in efficient and nonefficient biological phosphorus removal activated sludge systems. Appl. Environ. Microbiol. **65:**4077–4084.
6. **Boon, N., W. De Windt, W. Verstraete, and E. M. Top.** 2002. Evaluation of nested PCR-DGGE (denaturing gradient gel electrophoresis) with group-specific 16S rRNA primers for the analysis of bacterial communities from different wastewater treatment plants. FEMS Microbiol. Ecol. **39:**101–112.
7. **Cole, J. R., B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity, and J. M. Tiedje.** 2007. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res. **35:**D169–D172.
8. **Coskuner, G., and T. P. Curtis.** 2002. In situ characterization of nitrifiers in an activated sludge plant: detection of *Nitrobacter* spp. J. Appl. Microbiol. **93:**431–437.
9. **de los Reyes, F. L., D. Rothauszky, and L. Raskin.** 2002. Microbial community structures in foaming and nonfoaming full-scale wastewater treatment plants. Water Environ. Res. **74:**437–449.
10. **Dinsdale, E. A., R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, L. McDaniel, M. A. Moran, K. E. Nelson, C. Nilsson, R. Olson, J. Paul, B. R. Brito, Y. Ruan, B. K. Swan, R. Stevens, D. L. Valentine, R. V. Thurber, L. Wegley, B. A. White, and F. Rohwer.** 2008. Functional metagenomic profiling of nine biomes. Nature **452:**629–632.
11. **Garcia Martin, H., N. Ivanova, V. Kunin, F. Warnecke, K. W. Barry, A. C. McHardy, C. Yeates, S. He, A. A. Salamov, E. Szeto, E. Dalin, N. H. Putnam, H. J. Shapiro, J. L. Pangilinan, I. Rigoutsos, N. C. Kyrpides, L. L. Blackall, K. D. McMahon, and P. Hugenholtz.** 2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat. Biotechnol. **24:**1263–1269.
12. **Gilbert, Y., Y. L. Bihan, and P. Lessard.** 2006. Acetylene blockage technique as a tool to determine denitrification potential of a biomass fixed on an organic media treating wastewater. J. Environ. Eng. Sci. **5:**437–442.
13. **Gilbride, K. A., and R. R. Fulthorpe.** 2004. A survey of the composition and diversity of bacterial populations in bleached kraft pulp-mill wastewater secondary treatment systems. Can. J. Microbiol. **50:**633–644.
14. **Huber, J. A., D. B. Mark Welch, H. G. Morrison, S. M. Huse, P. R. Neal, D. A. Butterfield, and M. L. Sogin.** 2007. Microbial population structures in the deep marine biosphere. Science **318:**97–100.
15. **Huse, S. M., L. Dethlefsen, J. A. Huber, D. M. Welch, D. A. Relman, and M. L. Sogin.** 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet. **4:**e1000255.
16. **Jeon, C. O., D. S. Lee, and J. M. Park.** 2003. Microbial communities in activated sludge performing enhanced biological phosphorus removal in a sequencing batch reactor. Water Res. **37:**2195–2205.
17. **Jones, W., P. Wilderer, and E. Schroeder.** 1990. Operation of a three-stage SBR system for nitrogen removal from wastewater research. J. Water Pollut. Control Fed. **62:**268–274.
18. **Jupraputtasri, W., N. Boonapatcharoen, S. Cheevadhanarak, P. Chaiprasert, M. Tanticharoen, and S. Techkarnjanaruk.** 2005. Use of an alternative Archaea-specific probe for methanogen detection. J. Microbiol. Methods **61:**95–104.
19. **Juretschko, S., G. Timmermann, M. Schmid, K.-H. Schleifer, A. Pommerening-Roser, H.-P. Koops, and M. Wagner.** 1998. Combined molecular and conventional analyses of nitrifying bacterium diversity in activated sludge: *Nitrosococcus mobilis* and *Nitrospira*-like bacteria as dominant populations. Appl. Environ. Microbiol. **64:**3042–3051.
20. **Layton, A. C., P. N. Karanth, C. A. Lajoie, A. J. Meyers, I. R. Gregory, R. D. Stapleton, D. E. Taylor, and G. S. Sayler.** 2000. Quantification of *Hyphomicrobium* populations in activated sludge from an industrial wastewater treatment system as determined by 16S rRNA analysis. Appl. Environ. Microbiol. **66:**1167–1174.
21. **Liu, Z., C. Lozupone, M. Hamady, F. D. Bushman, and R. Knight.** 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. **35:**e120.
22. **Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376–380.
23. **Mavromatis, K., N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides.** 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat. Methods **4:**495–500.
24. **Oerther, D. B., F. L. de los Reyes III, M. F. de los Reyes, and L. Raskin.** 2001. Quantifying filamentous microorganisms in activated sludge before, during, and after an incident of foaming by oligonucleotide probe hybridizations and antibody staining. Water Res. **35:**3325–3336.
25. **Otawa, K., R. Asano, Y. Ohba, T. Sasaki, E. Kawamura, F. Koyama, S. Nakamura, and Y. Nakai.** 2006. Molecular analysis of ammonia-oxidizing bacteria community in intermittent aeration sequencing batch reactors used for animal wastewater treatment. Environ. Microbiol. **8:**1985–1996.
26. **Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein.** 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. **33:**5691–5702.
27. **Pauli, W., K. Jax, and S. Berger.** 2001. Protozoa in wastewater treatment: function and importance, p. 203–252. *In* Biodegradation and persistence. Springer, Berlin, Germany.
28. **Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. **5:**e77.
29. **Seviour, R. J., T. Mino, and M. Onuki.** 2003. The microbiology of biological phosphorus removal in activated sludge systems. FEMS Microbiol. Ecol. **27:**99–127.
30. **Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc. Natl. Acad. Sci. USA **103:**12115–12120.
31. **Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, and J. C. Detter.** 2005. Comparative metagenomics of microbial communities. Science **308:**554–557.
32. **Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, and W. Nelson.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:**66.
33. **Wagner, M., A. Loy, R. Nogueira, U. Purkhold, N. Lee, and H. Daims.** 2002. Microbial community composition and function in wastewater treatment plants. Antonie van Leeuwenhoek **81:**665–680.
34. **Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole.** 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. **73:**5261–5267.
35. **Yooseph, S., G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J. M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. **5:**e16.
36. **Zheng, D.** 2000. Quantification of Methanosaeta species in anaerobic bioreactors using genus- and species-specific hybridization probes. Microb. Ecol. **39:**246–262.