Environ. Eng. Res. Vol. 14, No. 1, pp. 3~9, 2009
Korean Society of Environmental Engineers

Microbial Community Analysis using RDP II (Ribosomal Database Project II): Methods, Tools and New Advances

Erick Cardenas[1], James R. Cole[1], James M. Tiedje[1,2†], and Joonhong Park[2]

[1]*Center for Microbial Ecology, Michigan State University, East Lansing, Michigan, USA*
[2]*School of Civil and Environmental Engineering, Yonsei University, Seoul, Republic of Korea*

Received March 2009, accepted March 2009

Abstract

Microorganisms play an important role in the geochemical cycles, industry, environmental cleanup, and biotechnology among other fields. Given the high microbial diversity, identification of the microorganism is essential in understanding and managing the processes. One of the most popular and powerful method for microbial identification is comparative 16S rRNA gene analysis. Due to the highly conserved nature of this essential gene, sequencing and later comparison of it against known rRNA databases can provide assignment of the bacteria into the taxonomy, and the identity of its closest relatives. Isolation and sequencing of 16S rRNA genes directly from natural environments (either from DNA or RNA) can also be used to study the structure of the whole microbial community. Nowadays, novel sequencing technologies with massive outputs are giving researchers worldwide the chance to study the microbial world with a depth that was previously too expensive to achieve. In this article we describe commonly used research approaches for the study of individual microorganisms and microbial communities using the tools provided by Ribosomal Database Project website.

Keywords: Microbial community analysis, Ribosomal RNA, Microbial diversity

1. Introduction

Microbes are major players in the geochemical cycles, and important tools for industrial and environmental applications such as wastewater treatment, bioremediation, renewable energy, and medicine production. Because of the dimensions of the bacterial diversity, as much as 10^6 different species in a gram of soil,[1)] bacterial identification is critical for assessing and managing microbial processes in both natural and engineered conditions.

Bacterial identification has been traditionally done through cultivation-dependent methods such as metabolic and biochemical characterization of isolated strains, and also through microscopy. However, these cultivation-dependent and physiological assays can be time-consuming and results may change depending on the conditions used (e.g. temperature, pH, biological associations, etc). Additionally, a relative small fraction of the microbial diversity is cultivable using traditional methods (usually less than 0.1%),[2)] though recent advances has increase this to recoveries as high as 7.5%.[3)] To avoid the bias inherent to cultivation and isolation, molecular methods can be used to examine the microbial composition.

The most frequently used molecular method for bacterial identification is comparative 16S rRNA gene analysis. This technique takes advantage of the conserved nature of the 16S rRNA gene. This gene does not code for a protein but for a structural RNA part of the ribosome. Because ribosomes play an essential role in protein synthesis, this gene is ubiquitous in bacteria, highly conserved and it almost never horizontally transferred[4)] making it ideal for phylogeny reconstruction and identification.

More highly conserved regions in the ribosomal RNA gene sequence allow for the creation of (nearly) "universal" primers for the amplification of this gene from DNA extracted directly from natural environments. On the other hand, regions within the gene have increasing variation in sequence, reflective of evolutionary distance, and hence provides information that can be use for bacteria identification.

To identify the source of the sequences derived from environmental DNA, the sequences are compared with reference sequences from ribosomal RNA databases. This can be done through phylogenetic methods or classification methods. Phylogenetic

† Corresponding author
E-mail: tiedjej@msu.edu
Tel: +1-517-353-9021, Fax: +1-517-353-2917

methods cluster unknown sequences together with reference sequences using an alignment and a phylogeny reconstruction algorithm. Even though this is the preferred method, the computing power required greatly increases with increasing numbers of sequences and the results may differ depending on the phylogenetic algorithm used. On the other hand, classification methods sort the unknown sequences into a known taxonomic hierarchy by comparing features of the unknown sequence with those from references in the known taxonomy. Classification methods use either nearest-neighbor schemes or text-based Bayesian approaches. The first approach assigns a sequence into a taxon depending on the established classification of its closest relatives in the database. The second approach compares the "text features" of the sequences to find relatives with similar "text features". Classification methods are easier to interpret and faster for well understood groups.[5] Nowadays, classification methods for 16S rRNA gene analysis are becoming increasingly popular especially for environmental studies. Ribosomal RNA databases play a key role in this process by providing analysis tools, a standard taxonomy, and high quality sequences that can be used as references in the study of environmental sequences.

In this review article, we summarize a standard procedure for using the Ribosomal Database Project (http://rdp.cme.msu.edu/) in microbial community analysis. We also include recent status and features of "new" RDP services that allow using data from pyrosequencing for the study of both functional and phylogenetic gene analysis.

2. Recovery and Amplification of Genes Directly from Microbial Communities

Traditional studies of microbial communities involved isolation of their individual members. Since the microbial diversity is so large and bias in cultivation exists, microbial identification is now typically done by amplifying and sequencing the 16S rRNA gene directly from the community's DNA. The first step, DNA extraction, can be done with a variety of commercial kits that can be used on samples from water, soil, bioreactors and almost every type of environment. Usually the most challenging situation for obtaining high quality DNA occurs when the population density is very low or when chemicals that interfere with DNA processing are present, such as humic acids in soils.[6] After DNA is obtained, "universal" primers are used to amplify genes from all the members of the community with the polymerase chain reaction (PCR).

Because community DNA is amplified, a mixture of PCR products is obtained; thus creating a need to separate the individual products before sequencing. This is usually done by inserting individual products into a vector (e.g. plasmid) that is later inserted into a host cell, typically *Escherichia coli*, for amplification and sequencing (Fig. 1). Since all the hosts have identical genetic material but differ only in the PCR product they carry, the procedure is called clone library construction. By creating clone libraries and sequencing the 16S rRNA genes inserted in them, it is possible to study the microbial community structure without cultivating its members. A typical clone library will be composed of one or two 96-well plates (96 or 192 rRNA genes) per sample.

Given the relative small size of the 16S rRNA gene (~1500 bases), most of the gene can be sequenced using "single read" sequencing by the Sanger method.[7,8] The current version of the Sanger method is based on the polymerization of a DNA strand using fluorescent dye terminators. The terminators are dideoxynucleotides labeled with a fluorescent probe, one color for each of the four bases (A,

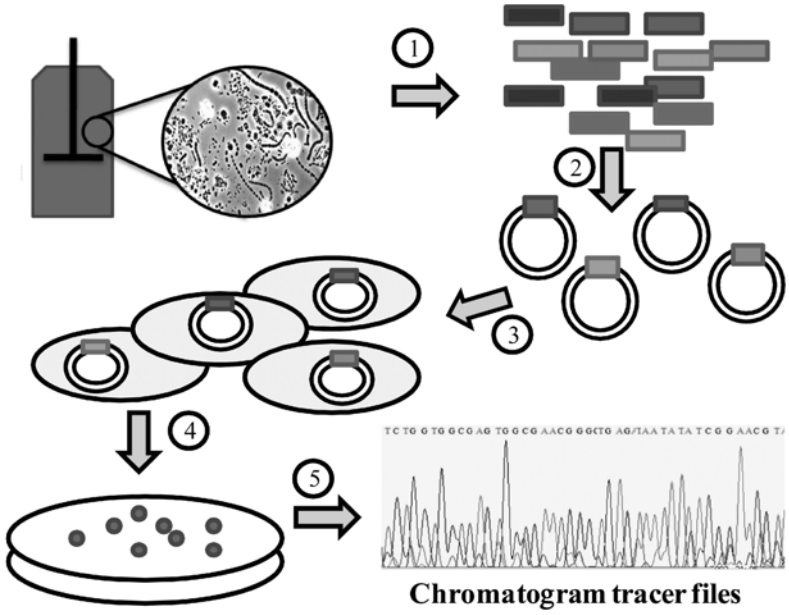

Fig. 1. Scheme of amplification and cloning of PCR product into bacterial hosts for sequencing. 1) DNA extracted from a community is amplified using primers for a given gene, 2) each PCR product is inserted into a single vector for amplification, 3) the vectors from the pool are then inserted into host cells, 4) the hosts are plated onto a selective medium that selects only for hosts carrying the product in the vector, 5) finally the vectors are extracted and sequenced.

C, G, and T). When the DNA polymerase extension is stopped by the terminators, a labeled nucleotide chain is generated. A population of chains can be later separated by electrophoresis, and because each chain is labeled with only one color, the sequence of the original DNA molecule can be determined.

Recently, new sequencing methods such as pyrosequencing have become available[9-11] (see Box 1). These new sequencing methods produce shorter reads than traditional Sanger sequencing method but in much larger numbers and at a much reduced cost per base. Shorter reads creates the need to target regions of the 16S rRNA gene that are the most informative for identification and not the complete gene itself. These hypervariable regions even when small as 100 bases are informative enough to accurately classify most sequences to the genus level.[5,12] The major advantage of these new sequencing technologies is the high throughput they provide, making possible to conduct large numbers of in-depth surveys of the microbial world. Of the new sequencing methods, pyrosequencing is rapidly becoming the most adopted method for microbial community analysis with 16S rRNA genes, while short-read methods, such as Illumina, are being used mostly for global expression studies.

Because the typical throughput of these technologies is very high (400,000 reads per run for pyrosequencing), multiplex methods are being used. In these methods, the genes are amplified with primers that are a combination of universal primers with four to six or more extra bases used as barcodes. A particular barcode is used to uniquely identify all sequences from one sample, so different samples can be mixed together and their sequence information later computationally separated; e.g. the 400, 000 sequences can be comprised of 80 libraries (samples) of 5000 sequences.

The first of these massive community surveys produced thousands of sequences per ocean sediment sample.[13] In one study of sediments close to hydrothermal vents, more than 750,000 sequences were recovered and yet the microbial diversity present was still not completely sampled.[14]

Box 1. Since its first publication in 1977, the Sanger method has been the gold standard for determining the sequence of nucleic acids. Advances in capillary electrophoresis, fluorescent dyes and automation allows one instrument to sequence up to 2.1 Megabases per day with average read lengths from 550 to 900 bases. Novel sequencing technologies based on different principles are nowadays providing much higher throughput but of shorter read lengths. Two examples of these technologies are pyrosequencing and Illumina sequencing. A single run of pyrosequencing generates up to 600 Megabases per day with average read length of 400 bases, while Illumina sequencing can generate 3 Gigabases of 36-base reads in a run which takes five days. The novel technologies are changing the way microbial communities can be studied providing a more comprehensive sampling of the microbial world.

3. Preprocessing and Quality Control

3.1. Processing of Sequences and Chromatogram Tracer Files

The output from sequencing is a chromatogram trace file. This file shows the signal for each nucleotide for every single position. Many programs are trained to read this trace file to assign a base to each position and provide an estimate of assignment accuracy (Q value). After each base is assigned, the complete sequence of the molecule can be determined. In the case of the RDP, either tracer files or sequence files can be used as input

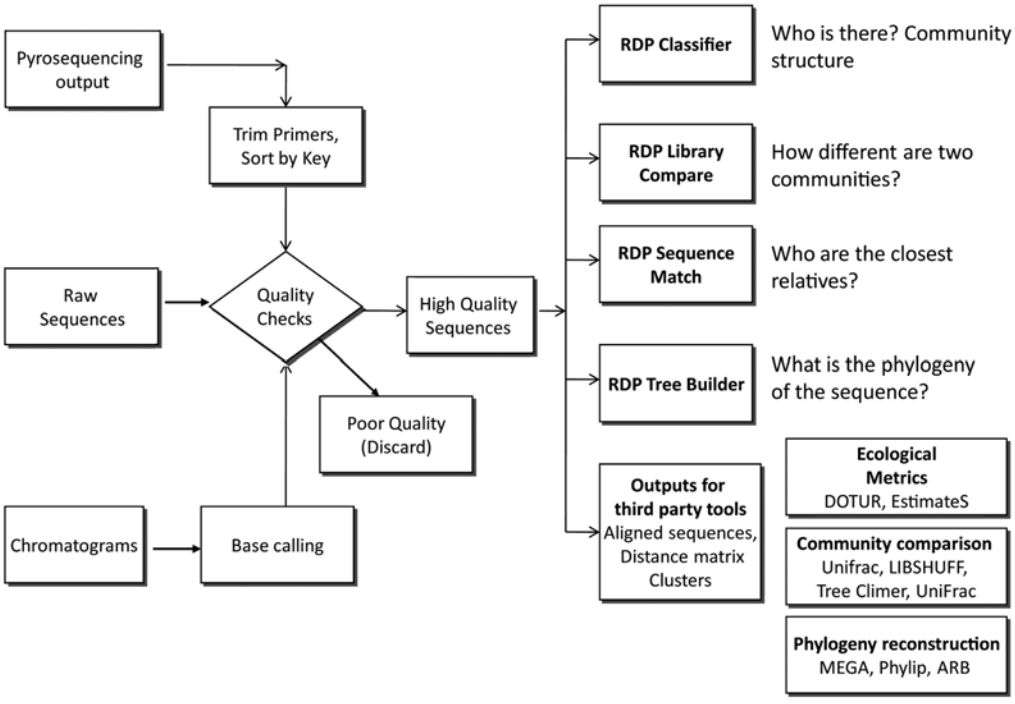

Fig. 2. Overview of the processing of sequence data by the RDP. Data entered can be either chromatogram files or sequence files. If chromatogram sequences are submitted, they will be used to assign bases to each position with a given quality standard.

for analysis (Fig. 2). In either case, users need to create a *my*RDP account (for free). The *my*RDP feature allows any scientist worldwide to upload sequences (sequences will not be submitted to any public databases such as GenBank or EMBL) or trace files to a private account that provides the same analysis features available for any public sequence such as classification into a taxonomic hierarchy, determination of similarity to closest relative, etc. If working with trace files, users need to submit the files to the RDP pipeline (part of the myRDP). The pipeline will convert the trace data into sequence data with quality information attached. The quality data as well as the location of each sample in the 96-well plate can be used later for quality control.

3.2. Quality Control

Quality control is key for the correct classification and diversity estimation, as artifact sequences tend to be unique. Amplification of genes from environmental DNA has the potential to create biases and artifacts.[15] Bias can be due to primer selection (as no single primer set is truly universal), to interference of the clone product with the cloning host,[16] and to difference in the size of the amplicons. Artifacts like heteroduplexes and chimeras are due to hybridization of incomplete PCR products. Different approaches can be used to deal with these problems. Chimeras can be detected using specialized software packages like Pintail[17] and Mallard[18] and web-applications like Bellerophone[19] and RDP's CHIMERA CHECK.[20] CHIMERA CHECK is a very popular tool that has been used to detect chimeric sequences in studies of microbial communities from petroleum-contaminated sediments,[21] phosphorus removal from wastewater treatment plants,[22] among others. Additionally, all public sequences are checked for artifacts using the Pintail application. This allows users to select a high quality database, and avoid problematic results due to unreliable reference sequences.

If working with trace files, the quality data information can be used to discard low quality sequences, and the spatial location of samples in the plate can be used to discover systematic errors in the sequencing process, e.g. low quality sequences coming from the same row could indicate a problem in a capillary of the sequencing instrument.

4. Who Is There? - Bacterial Classification

Bacterial classification is a form of identification that assigns unknown sequences to a position in a taxonomic hierarchy. The most common classification methods use a nearest-neighbor algorithm that assumes that the unknown sequence shares the same classification as its closest relatives. This approach is used by the Sequence Match feature of the RDP where the closest relatives "vote" on the classification of the unknown sequence using their own classification. If there is no consensus on the classification at the genus level, the next taxonomic levels is analyzed until a unanimous classification is reached.

The Classifier feature of the RDP assigns unknown sequences by calculating the frequency of 8-letter words and comparing these frequencies to those from a set that contains well-studied sequences from different regions of the bacterial Tree of Life.[5] This method is fast, consistent and powerful enough to correctly classify sequences as short as 100 bases long.[5,12]

As an example, this feature has been used to study the microbial community of an anaerobic enrichment that dechlorinated a PCB congener mixture.[23]

5. Finding Closest Relatives

Most studies of unknown sequences recovered from the environment benefit from finding close relatives to the unknown sequences. Relatives provide information about the phylogenetic association of the query sequences, and about their metabolic potential. The latter which is usually more useful than the name of the genus to which the sequence belongs, the best result from classification methods. This is especially important if a genus contains pathogenic and non-pathogenic members.

Several algorithms for finding close relatives exist with BLAST[24] being the most widely known. In the case of 16S rRNA sequences, the Seqmatch algorithm has been shown to be more accurate in finding the closest relative of unknown sequences.[20] This is because Seqmatch analyses the whole sequences while BLAST works on local alignments. As an example, Seqmatch has been used to find a close relative of a *Clostridium* strain that degrades cellulose in a thermophilic methanogenic bioreactor.[25] In another case, an isolate recovered from a trichloroethene-contaminated aquifer undergoing bioremediation was analyzed with the Seqmatch to find that the closest relative was *Bacillus anthracis*,[26] the causal agent of anthrax. Especially in such a cases where the identification presents health concerns, it is essential to use other methods to establish if the organism is a pathogen. Since 16S rRNA is a conserved molecule, its sequence alone cannot be reliably used for identification at the species level. In the case above, further analysis confirmed that the isolate was not *Bacillus anthracis*.[26]

6. Microbial Community Comparison

Several approaches can be used to compare microbial communities. The first approach is used in RDP's Library Compare. With this feature, two libraries are classified into RDPs taxonomy and the abundance of each taxon is compared along with its statistical confidence. This approach has the advantage of identifying which groups account for the differences between communities. This method is fast, does not require alignment of sequences, has a genus level resolution (RDP taxonomy's smallest taxa), and works better for well studied groups.

The second approach assigns sequences into operational taxonomic units (OTUs) based on their similarity, e.g. if OTUs are defined at 97% identity, sequences 97% identical or higher will be assigned to the same unit. These OTUs are then used to compare the communities based on their diversity components: richness and evenness, and by different diversity indices.[27,28] This approach requires sequences to be aligned in order to compare homologous positions and generate a matrix of distances between all sequences. It also requires a priori decision on the similarity

level that defines the OTUs. Typically "species" level has been defined at 97% sequence identity,[29] although new data suggest that a 98.5 to 99% sequence identity is more consistent with the DNA-DNA hybridization standard for species.[30-32] Diversity indices such as Chao1 and Ace can then be used to estimate the total number of species in the sample. Other metrics, such as Shannon's index and Evenness measure the distribution of species in the sample. However, two communities can have the same diversity by these measures but completely different compositions. Similarity between samples based on community composition can be calculated using the Sorensen[33] and Jaccard[34] indices.

A third approach uses phylogenetic information, statistical tests and Monte Carlo simulations. This approach is represented by LIBSHUFF,[35] ∫-LIBSHUFF36, UniFrac,[37,38] TreeClimber,[39] Analysis of Molecular Variance (AMOVA), and Homogeneity of Molecular Variance (HOMOVA).[39] However, these approaches are computationally difficult when the number of sequences is in the thousands. Additionally, results from methods that use phylogenetic trees in their analysis can be influenced by the length of the sequences used as well as the region of the 16S rRNA gene from which the sequence is derived. These methods, their specific approaches, with advantages and limitations have been recently reviewed.[40]

7. Aides for Using the Ribosomal Database Project

The adoption of new sequencing technologies is currently changing the way we study microbial communities. The high throughput sequencing technologies now available or becoming available creates the opportunity to massively survey microbial communities for identification of the more dominant organisms, even in very diverse communities, as well as study community structure and dynamics. The information obtained through sequencing is useful not only for community profiling but also for identification of its members, and for phylogenetic analysis. The main limitation of sequencing used to be cost, the limited output, and complexity of the clone library process. These limitations can largely now be bypassed with the new sequencing technologies; however, the very large numbers of sequences creates computational and technical issues in handling and interpreting the data.

The new limitations are partially technically because of the novel nature of the techniques. Issues such as error rates, data handling, quality control, and standard analysis methods are some of the new technical limitations.

To facilitate the adoption of these powerful new technologies for 16S rRNA analysis, a pipeline for the analysis of pyrosequencing surveys was developed by the RDP.[41] This pipeline (available at http://pyro.cme.msu.edu/pyro/) processes the raw sequences, checks their quality and separates them into their original samples by reading the barcode used in the amplification step. The pipeline uses a secondary-structure-based aligner to which the sequences can be compared. The use of a model for alignment instead of pairwise comparison reduces the speed for alignment, and provides a consistent alignment tool. In general, this pipeline facilitates the handling of big datasets, e.g 400,000 sequences, and provides tools for a consistent analysis.

The tools developed for the pipeline can also be used for massive surveys of functional genes of environmental relevance. The main difference in when dealing with functional genes is the alignment tool since the sequence (either nucleotide or protein) is relevant in contrast with ribosomal RNA genes where the secondary structure is the conserved feature. Some of the most popular aligner approaches use the programs CLUSTAL,[42] MUSCLE.[43] An alternative to these aligners is to use a protein model for alignment, such a Hidden Markov model, in the same way we use a secondary structure model for the ribosomal genes. Pfam (http://pfam.sanger.ac.uk/) and the Functional Gene Database/ Repository (http://fungene.cme.msu.edu/) use Hidden Markov Model (HMM) search programs to retrieve sequences that fit a given protein model from the public databases such as GenBank, EMBL, etc.

Pyrosequencing together with a HMM was used to study the diversity of biphenyl dioxygenase, a gene association with polychlorinated biphenyl (PCBs) degradation.[44] A short variable region of the *bph* gene associated with substrate specificity was targeted for massive parallel sequencing. The results revealed new clusters of Rieske non-heme iron dioxygenase genes not detectable by the previously standard clone library approach.[44] In the near future, when longer sequences will be available from the new technologies, this type of survey has the potential to rapidly and more comprehensively sample the diversity nature has produced, to provide information that corresponds with functional diversity, and to provide the data for probe design that can then be used to recover genes or operons of proteins with novel or important functions.

8. Conclusions

The study of microbial communities is essential in the understanding of the processes microbes mediate. For this purpose comparative 16S rRNA gene analysis is one of the most powerful method currently available to study the microorganisms in their natural or managed environments. The information retrieve from sequencing this gene can be used for microbe classification, community structure determination, and phylogenetic analysis. For any of these applications, ribosomal RNA gene databases play a key role by providing the most current sequences, taxonomic information, and analysis tools. These features can be used to create a high quality, consistent, and replicable analysis.

Novel and cheaper sequencing technologies are revolutionizing the biological fields and comparative 16S rRNA gene analysis is also being changed by them. With the new methodologies, many of the restrictions are disappearing allowing researcher worldwide, including those in small laboratories, to discover and characterize the microbial diversity in microbial communities of their interest.

New databases and tools are being developed to meet the data analysis challenge necessary to realize the potential that the new sequencing technologies provide. The new analysis tools promise

to also revolutionize the study of functional genes with massive surveys that reveal more of the functional diversity present in the microbial world.

Acknowledgement

This work was supported by Korea Science and Engineering Foundation (KSEF) through World Class University Project (grant # : R33-2008-000-10076-0).

References

1. Dykhuizen, D. E., "Santa Rosalia revisited: why are there so many species of bacteria?" *Antonie Van Leeuwenhoek*, **73**, 25-33 (1998).
2. Skinner, F. A., Jones, P. C., and Mollison, J. E., "A comparison of a direct- and a plate counting technique for the quantitative estimation of soil micro-organisms," *J. Gen Microbiol.*, **6**, 261-271 (1952).
3. Janssen, P. H., Yates, P. S., Grinton, B. E., Taylor, P. M., and Sait, M., "Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia," *Appl. Environ. Microbiol.*, **68**, 2391-2396 (2002).
4. Miller, S. R., Augustine, S., Olson, T. L., Blankenship, R. E., Selker, and J., Wood, A. M., "Discovery of a free-living chlorophyll d-producing cyanobacterium with a hybrid proteobacterial/cyanobacterial small-subunit rRNA gene," *Proc. Natl. Acad. Sci. USA*, **102**, 850-855 (2005).
5. Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R., "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Appl. Environ. Microbiol.*, **73**, 5261-5267 (2007).
6. Persoh, D., Theuerl, S., Buscot, F., and Rambold, G., "Towards a universally adaptable method for quantitative extraction of high-purity nucleic acids from soil," *J. Microbiol. Methods.*, **75**, 19-24 (2008).
7. Sanger, F., Nicklen, S., and Coulson, A. R., "DNA sequencing with chain-terminating inhibitors," *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467 (1977).
8. Maxam, A. M., and Gilbert, W., "A new method for sequencing DNA," *Proc. Natl. Acad. Sci. USA*, **74**, 560-564 (1977).
9. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M., "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, **437**, 376-380 (2005).
10. Nyren, P., Pettersson, B., and Uhlen, M., "Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay," *Anal. Biochem.*, **208**, 171-175 (1993).
11. Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M., "Accurate multiplex polony sequencing of an evolved bacterial genome," *Science*, **309**, 1728-1732 (2005).
12. Liu, Z., DeSantis, T. Z., Andersen, G. L., and Knight, R., "Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers," *Nucleic. Acids. Res.*, **36**, e120 (2008).
13. Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J., "Microbial diversity in the deep sea and the underexplored "rare biosphere"," *Proc. Natl. Acad. Sci. USA*, **103**, 12115-12120 (2006).
14. Huber, J. A., Welch, D. B. M., Morrison, H. G., Huse, S. M., Neal, P. R., Butterfield, D. A., and Sogin, M. L., "Microbial population structures in the deep marine biosphere," *Science*, **318**, 97-100 (2007).
15. Kanagawa, T., "Bias and artifacts in multitemplate polymerase chain reactions (PCR)," *J. Biosci. Bioeng.*, **96**, 317-323 (2003).
16. Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., Bork, P., and Rubin, E. M., "Genome-wide experimental determination of barriers to horizontal gene transfer," *Science*, **318**, 1449-1452 (2007).
17. Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J., "At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies," *Appl. Environ. Microbiol.*, **71**, 7724-7736 (2005).
18. Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J., "New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras," *Appl. Environ. Microbiol.*, **72**, 5734-5741 (2006).
19. Huber, T., Faulkner, G., and Hugenholtz, P., "Bellerophon: a program to detect chimeric sequences in multiple sequence alignments," *Bioinformatics*, **20**, 2317-2319 (2004).
20. Cole, J., Chai, B., Farris, R. J., Wang, Q., Kulam, S. A., McGarrell, D. M., Garrity, G. M., and Tiedje, J. M., "The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis," *Nucleic. Acids. Res.*, **33**, D294-296 (2005).
21. Allen, J. P., Atekwana, E. A., Duris, J. W., Werkema, D. D., and Rossbach, S., "The microbial community structure in petroleum-contaminated sediments corresponds to geophysical signatures," *Appl. Environ. Microbiol.*, **73**, 2860-2870 (2007).
22. Kong, Y., Xia, Y., Nielsen, J. L., and Nielsen, P. H., "Structure and function of the microbial community in a full-scale

enhanced biological phosphorus removal plant," *Microbiology*, **153**, 4061-4073 (2007).
23. Bedard, D. L., Bailey, J. J., Reiss, B. L., and Jerzak, G. V., "Development and characterization of stable sediment-free anaerobic bacterial enrichment cultures that dechlorinate aroclor 1260," *Appl. Environ. Microbiol.*, **72**, 2460-2470 (2006).
24. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic. Acids. Res.*, **25**, 3389-3402 (1997).
25. Shiratori, H., Ikeno, H., Ayame, S., Kataoka, N., Miya, A., Hosono, K., Beppu, T., and Ueda, K., "Isolation and characterization of a new Clostridium sp. that performs effective cellulosic waste digestion in a thermophilic methanogenic bioreactor," *Appl. Environ. Microbiol.*, **72**, 3702-3709 (2006).
26. Moss, E., Microbial community structure in a trichloroethylene aquifer during Toluene stimulated bioremediation, Ph.D. dissertation, Michigan State University (2004).
27. Schloss, P. D., and Handelsman, J., "Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures," *Appl. Environ. Microbiol.*, **72**, 6773-6779 (2006).
28. Schloss, P. D., and Handelsman, "J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness," *Appl. Environ. Microbiol.*, **71**, 1501-1506 (2005).
29. Rossello-Mora, R., and Amann, R., "The species concept for prokaryotes," *FEMS. Microbiol. Rev.*, **25**, 39-67 (2001).
30. Konstantinidis, K. T., and Tiedje, J. M., "Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead," *Curr. Opin. Microbiol.*, **10**, 504-509 (2007).
31. Stackebrandt, E., and Ebers, J., "Taxonomic parameters revisited: tarnished gold standards," *MICROBIOLOGY TODAY*, **33**, 152-155 (2006).
32. Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M., "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities," *Int. J. Syst. Evol. Microbiol.*, **57**, 81-91 (2007).
33. Sorensen, T., "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons," *Biol. Skr.*, **5**, 1-34 (1984).
34. Jaccard, P., "Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences," *Naturelles*, **37**, 547-579 (1901).
35. Singleton, D. R., Furlong, M. A., Rathbun, S. L., and Whitman, W. B., "Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples," *Appl. Environ. Microbiol.*, **67**, 4374-4376 (2001).
36. Schloss, P. D., Larget, B. R., and Handelsman, J., "Integration of microbial ecology and statistics: a test to compare gene libraries," *Appl. Environ. Microbiol.*, **70**, 5485-5492 (2004).
37. Lozupone, C., and Knight, R., "UniFrac: a new phylogenetic method for comparing microbial communities," *Appl. Environ. Microbiol.*, **71**, 8228-8235 (2005).
38. Lozupone, C., Hamady, M., and Knight, R., "UniFrac--an online tool for comparing microbial community diversity in a phylogenetic context," *BMC Bioinformatics*, **7**, 371 (2006).
39. Martin, A. P., "Phylogenetic approaches for describing and comparing the diversity of microbial communities," *Appl. Environ. Microbiol.*, **68**, 3673-3682 (2002).
40. Schloss, P. D., "Evaluating different approaches that test whether microbial communities have the same structure," *ISME J.*, **2**, 265-275 (2008).
41. Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., and Tiedje , J. M., "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis," *Nucl. Acids. Res.*, **37**, D141-145 (2009).
42. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G., "Clustal W and Clustal X version 2.0," *Bioinformatics*, **23**, 2947-2948 (2007).
43. Edgar, R. C., "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic. Acids. Res.*, **32**, 1792-1797 (2004).
44. Iwai, S., Chai, B., Sul, W. J., Cole, J. R., Hashsham, S. A., and Tiedje, J. M., "Exploring environmental biphenyl dioxygenase genes by clone libraries and pyrosequencing," *Twelveth International Symposium on Microbial Ecology (ISME-12)*, Cairns, Australia (2008).