Single Cell Whole Genome Amplification of Uncultivated Organisms

Mircea Podar, Martin Keller, and Philip Hugenholtz (✉)

Contents

Abstract Whole genome amplification of single cells is emerging as a powerful technique for accessing the genomes of individual members of microbial communities without the complication of identifying the source of sequence data posed by shotgun sequencing of environmental samples (metagenomics). This method holds particular promise for the molecular unveiling of uncultivated organisms that comprise the bulk of the microbial diversity and functionality on our planet.

1 Introduction

Over the last two decades, it has become increasingly apparent that microbial diversity is massively undersampled by culture collections, and consequently coverage of the tree of life by reference genome sequences is also highly incomplete (Hugenholtz 2002; Rappe and Giovannoni 2003). This limitation greatly impedes our understanding

M. Podar and M. Keller
Oak Ridge National Laboratory, Biosciences Division, Oak Ridge, TN, USA

P. Hugenholtz
Microbial Ecology Program, DOE Joint Genome Institute, Walnut Creek, CA, USA
e-mail: phugenholtz@lbl.gov

Microbiol Monogr, doi:10.1007/7171_2008_10

of microbial evolution and ecology. Directed or fortuitous culturing efforts have managed to provide axenic or enriched cultures for some major phylogenetic gaps (Preston et al. 1996; Zhang et al. 2003; Zoetendal et al. 2003; Cho et al. 2004; Konneke et al. 2005), most of which immediately become the focus of genome sequencing projects (Liolios et al. 2006). To really accelerate genomic coverage of microbial diversity, however, methods not involving cultivation will need to be used. Metagenomics, the application of high-throughput sequencing to DNA extracted directly from environmental samples (Riesenfeld et al. 2004; Tringe and Rubin 2005), is a promising route to bypass the cultivation bottleneck, but has the drawback that multiple genomes are being sequenced simultaneously, often at very low coverage, which must be subsequently deconvoluted in silico. Therefore, metagenomic studies rarely result in complete or near-complete genomes, with the exception of communities highly enriched in one or more populations (Tyson et al. 2004; Garcia Martin et al. 2006; Strous et al. 2006). Whole genome amplification and sequencing of single cells has recently been demonstrated as a viable approach (Zhang et al. 2006), providing direct access to the full spectrum of microbial diversity without the multispecies complications of metagenomics. This opens the possibility for genomic characterization of any microbial member of a community that can be selected on the basis of taxonomic, morphological, or physiological criteria. As a result, acquiring genomic data for lineages encompassing the entire microbial tree of life appears feasible for the first time, regardless of cultivation success or abundance in the community for any specific type of organism.

Single cell genomics opens exciting opportunities to study a variety of aspects of microbial ecology and evolution. Understanding the degree of genomic variation between cells in natural populations not related by direct clonal descent will be important for defining the core genome vs. the accessory genes and ultimately getting better metrics for defining and distinguishing ecotypes, populations, and species or their equivalents. Since a coherent theoretical framework for microbial ecology is still lacking and important issues such as the definition or existence of bacterial species, and the extent and impact of horizontal gene transfer in microbial communities, genomic data from as many microbial lineages as possible will provide a better picture of the genetic and functional diversity in the microbial world and the evolutionary and environmental forces that shape genomes and communities.

In this chapter, we discuss the pros and cons of single cell genomics with particular reference to its recent application to members of candidate phylum TM7, a major lineage of the bacterial domain thus far lacking any pure-culture representatives and genome sequences.

2 Cell Isolation Methods

In most cases, a prerequisite to microbiological characterization of an organism has been the clonal expansion in the laboratory of a single cell, isolated from an environmental source. The resulting population, genetically uniform, serves for formal taxonomic

description and further characterization of its physiology, genomics, and other biological aspects. Traditionally, obtaining that initial cell progenitor has involved techniques as simple as streaking on solid media or dilution to extinction. Such approaches are not broadly applicable to single cell genomics, as they offer little control over the types of organisms that are isolated and require active growth and formation of (micro)colonies or cultures. More discriminatory techniques such as micromanipulation with microcapillaries or optical tweezers allow separation of specific cells based on morphology or, in some cases, certain physiological characteristics (e.g., autofluorescence) but have a low throughput and relatively restricted use. Such isolation techniques are the only viable options for subsequent genomic characterization if cells can grow at least to low densities on defined media in the laboratory.

2.1 Microdroplet Capture

A promising new approach to isolate microbial cells from the environment and test cultivation conditions in the laboratory involves encapsulation of individual cells into agarose microdroplets (Zengler et al. 2005). The microdroplets are then incubated under defined conditions that may allow formation of microcolonies, which can be detected microscopically or, using a higher throughput, by flow cytometry. If an appropriate selection can be designed (morphology, physiological property, or taxonomic staining by fluorescence in situ hybridization, see below), microdroplets containing microcolonies of interest can be isolated. When the cells are viable and dividing, a fraction of the microcolony can be further propagated in culture, while a reduced number of cells can be used directly for genomic characterization.

The above approaches, while suitable for cells that can be cultured in the laboratory, cannot be used for the isolation and genomic characterization of microbes that will not divide and establish a clonal population. The vast majority of microbes, while likely not intrinsically "unculturable" in the laboratory, fall into this category. Many major bacterial and archaeal lineages (phyla or divisions) still have no cultured representatives (Hugenholtz 2002) and all other phyla contain lower level taxa that are known solely on the basis of ribosomal RNA sequences obtained by culture-independent molecular surveys. Single cell genomics has great potential for characterizing such uncultured groups. In general, the most straightforward approach to identify and isolate single cells relies on the unique information available for those organisms, their ribosomal RNA sequences.

2.2 Micromanipulation

Antisense probes (oligonucleotides labeled with fluorescent small molecules) have been used for many years to detect and quantify specific microorganisms in environmental samples (Amann et al. 2001). The procedure, known as FISH (fluorescence in

situ hybridization) involves the design of the antisense probe to target a region (usually 18–24 nt) of the ribosomal RNA sequence (usually the small subunit) that can confer the desired specificity (from species specificity to phylum or even domain) followed by chemical synthesis of the oligonucleotide and end labeling with a fluorophore that can be detected using UV, lasers, and specific filters. Software such as ARB (Ludwig et al. 2004) have probe design algorithms that facilitate the process and identify potential problems in terms of specificity, probe dimerization, and other artifacts that may occur. Studies on the in situ accessibility of ribosomal RNAs to FISH probes over the length of the rRNA molecule provide an additional guideline for probe design (Amann et al. 2001). Databases such as probeBase (Loy et al. 2007) catalog the sequences and characteristics of optimized probes, providing a valuable resource to the community. A large number of fluorophores emitting at different wavelengths are available commercially, which enable application of multiple probes with different specificities to be resolved on the basis of color combinations.

One of the limitations of FISH-directed cell isolation is that cells must be fixed and permeabilized and therefore are not viable and cannot potentially be used for establishing cultures. A variety of cell fixation protocols have been used, not all compatible with subsequent extraction and use of the nucleic acids for further characterizations. Formaldehyde, in particular, crosslinks DNA to protein (Speit et al. 2000), which severely impairs downstream applications. Milder fixation reagents such as ethanol have been used successfully and are compatible with enzymatic steps in DNA extraction protocols. Even so, fixation, hybridization conditions, probe specificity, and fluorescent signal strength are parameters that require optimization on a case-by-case basis. Once fluorescently labeled, the simplest approach to identify and isolate specific cells involves an epifluorescence microscope and a micromanipulator fitted with a microinjector. The main difficulty is contamination with nontarget cells and the tedious, low-throughput process. Nevertheless, this approach has been used successfully to isolate *Methanothermobacter thermoautotrophicus* as a test organism, followed by the isolation of a soil crenarchaeote (Kvist et al. 2007).

2.3 Fluorescence Activated Cell Sorting (FACS)

Another approach widely used to separate microorganism is flow cytometry, a fluorescence-based cell characterization technique that enables rapid analysis of entire cell populations on the basis of single cell characteristics (size, shape, cell count) as the individual cells are passed in front of an intense light source (laser or laser diode). When coupled with cell sorting based on specific fluorescent signal (FACS), a large number of cells from complex environmental samples can be rapidly analyzed and the ones containing the FISH probe can be isolated. A difficulty in using this approach is ensuring the sterility of the procedure and avoiding contamination of the separated cells with nontarget organisms and DNA. A secondary limitation is the low phylogenetic resolution of ribosomal RNAs due to their strong evolutionary conservation (Woese 1987); for example, 16S rRNA cannot typically

resolve strains (genotypes). Therefore, if more than one cell is sorted, multiple genotypes may well be represented, complicating assembly of subsequent shotgun data. Using this approach, we have recently tested the feasibility of isolating from soil samples individual cells as well as small cell batches (5–100 cells) representing the uncultured TM7 phylum, after labeling with fluorescent oligonucleotides broadly specific for that group of bacteria (Podar et al. 2007). When sorting multiple cells, the different batches contained a varying degree of contamination with nontarget organisms or exogenous DNA; however, the TM7 cells appeared phylogenetically coherent (same species at least; Fig. 1) despite the broad specificity of the 16S rRNA-targeted FISH probe used. For example, a *Pseudomonas* cell was coisolated with four TM7 cells, resulting in significant contamination of the target TM7 genomes. The low frequency of the target cells in the soil community that served for that experiment (0.02%) indicates that the approach can be used to identify and isolate minor representatives of the community, "invisible" to other community genomic approaches. Obvious objectives for further improving the technique are finding ways to reduce the instances of contamination.

2.4 Microfluidics

Rapid developments in laser technology and nanotechnology have led to the miniaturization of flow cytometry and the development of an integrated microfabricated cell sorter using multilayer soft lithography (Fu et al. 2002). The integrated cell sorter is incorporated with various microfluidic functionalities, including peristaltic pumps, dampers, switch valves, and input and output wells, to perform cell sorting in a coordinated and automated fashion with extremely low fluidic volumes. Owing to the simple fabrication process and inexpensive materials, these devices can be disposable, thereby eliminating any cross-contamination from previous runs, which make them attractive for further downstream molecular methods requiring extreme sterility to avoid contaminating DNA. This benefit might compensate for the much lower sorting speed compared to conventional FACS machines; however, in practice it appears that removal of exogenous DNA entrained with the target cell is difficult even after repeated flushing of the microfluidic chamber perhaps because of the high surface tension in nanoliter volumes (Marcy et al. 2007). The cell sorting device can be directly combined with chemical or enzymatic reactions, such as cell lysis and DNA amplification. This approach has recently been used to characterize a different ecotype of TM7 phylum bacteria from the human subgingival microbiota ((Marcy et al. 2007); Fig. 1). Large rod-shaped cells, resembling oral TM7 previously characterized by FISH (Ouverney et al. 2003), were separated and lyzed in the microchamber, followed by DNA amplification in nanoliter scale volume to confirm identity and provide access to the genome of the isolated cell. Cross-contamination with exogenous DNA from other cell types present in the community still needs to be addressed; however, this approach allows selective isolation of viable cells based on morphology, which can then be analyzed genomically or transferred to culture media for propagation.

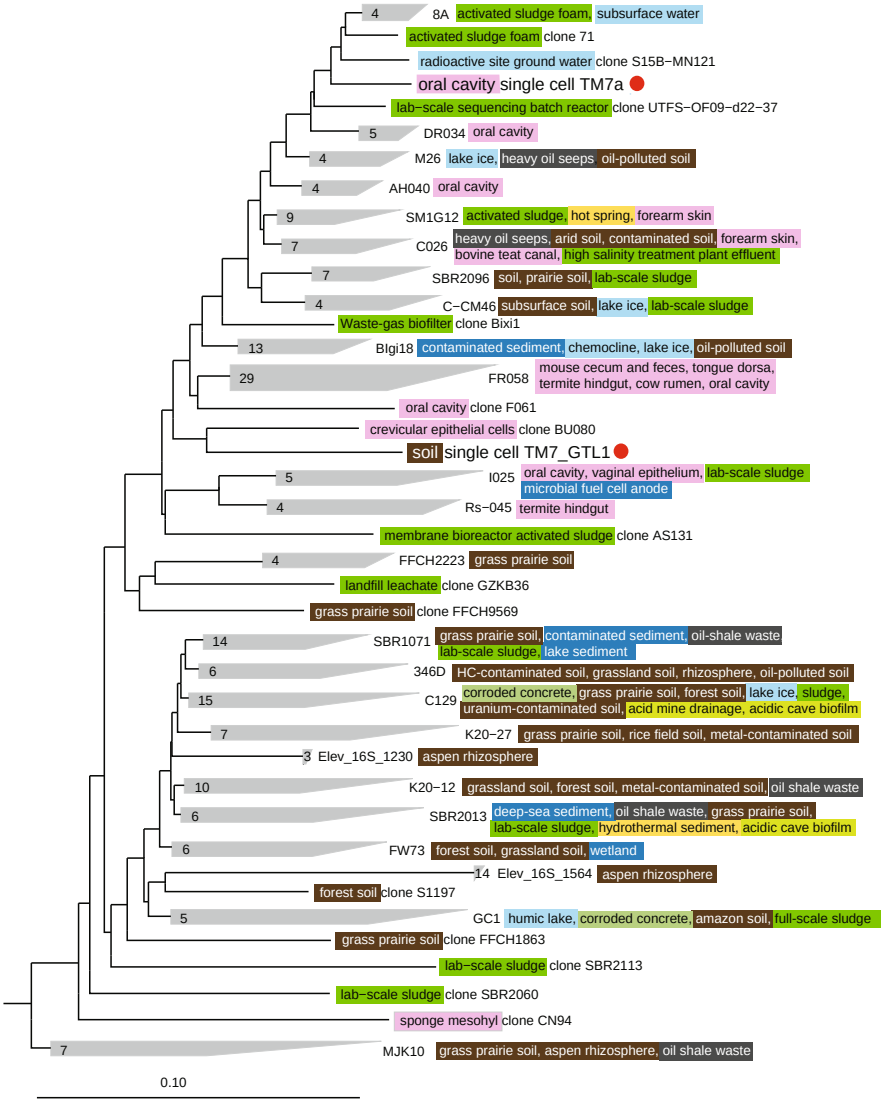

Fig. 1 A 16S rRNA gene tree of the bacterial phylum TM7 currently only comprising sequences obtained directly from the environment. Because of the large number of such environmental clone sequences, much of these data have been compressed into clusters named after the oldest clone representative, with the number indicating the number of clones in a given cluster. The habitats from which the clones were derived are listed and color-coded. The phylogenetic locations of amplified TM7 cells from soil (Podar et al. 2007) and the human oral cavity (Marcy et al. 2007) are indicated by red dots. The two ribotypes are ~9% divergent from each other. Data was obtained from the greengenes database (DeSantis et al. 2006)

3 DNA Isolation and Amplification from Single Cells

The amount of DNA present in a bacterial cell (~10^{-15} g) is presently well below the minimum amount necessary for cloning or direct sequencing. Amplification by factors of 10^6–10^9 is therefore required in order to generate nanogram to microgram quantities that can be used in standard genomic characterization protocols. Additional difficulties also arise because of handling of a single or a few chromosome copies. Very large DNA molecules are unstable when released from cells into solution and subject to single and double stranded breaks due to hydrodynamic shearing, nucleases, and chemical breakdown. While embedding of cells in gels, gentle lysis and handling, and inclusion of nuclease inhibitors can minimize such processes, they still occur but are averaged out in large DNA molecule populations that are obtained from bacterial cultures containing a large number of cells. When dealing with one or a few molecules, however, every single breakage event will result in the loss of information around that break point since the amplification enzymes may not be able to process them. Additional research will be required to understand the stability of whole chromosomes during cell lysis and enzymatic amplification and to determine whether genome sequencing to closure from a single molecule is feasible on a routine basis.

The current methods for DNA isolation from single bacterial cells include a cell treatment with lysozyme followed by alkaline denaturation (Zhang et al. 2006). Because the resulting DNA amount is very low, maintaining the volumes to a minimum is important in order to increase effective template concentration. Using conventional pipetting systems, reactions in several hundred nanoliters have been performed, which has resulted in amplification of single molecules to levels that allowed product visualization after gel electrophoresis (Hutchison and Venter 2006). Recent developments in the field of microfluidics allow, however, cell lysis and DNA amplification in much smaller volumes (tens of nanoliters) and therefore achieve an effective concentration of the template several orders of magnitude higher than using conventional pipetting (Hong et al. 2004). Such microfluidic chips that allow selection of target cell in addressable chambers followed by lysis and DNA amplification have been successfully used to generate sufficient DNA for partial genomic sequencing of uncultured TM7 bacteria from the human subgingival crevice (Marcy et al. 2007). The number of reports documenting the amplification of single molecules from uncultured organisms in small volumes is still small, and therefore factors such as efficient cell lysis and contamination minimization have to be investigated further.

Several different techniques have been published for random amplification across the whole genome, earlier work being done with Taq DNA polymerase (Hawkins et al. 2002). Multiple displacement amplification (MDA) is the approach that has been most successfully used for microbial genomic amplification and is based on using a highly processive phage polymerase with strand displacement capability, the phi29 DNA polymerase. Initial experiments using random hexamer

oligonucleotides, circular phage genomic DNA, or plasmids as templates have resulted in 10,000-fold amplification by a multiple-primed rolling circle (RCA) mechanism (Dean et al. 2002). Sufficient DNA to perform cloning and sequencing was obtained in several hours at 30°C starting with a bacterial colony or a phage plaque. Commercial versions of the enzyme and optimized kits and protocols for MDA are available (e.g., GenomePlex by Sigma-Aldrich, REPLI-g by Qiagen, GenomiPhy by Amersham, RepliPHI by Epicentre), and are valuable for a wide range of medical, forensic, and microbiological applications in which the biological material is very limited. A recent study (Pinard et al. 2006) compared the available methods to evaluate the bias introduced during bacterial whole genome amplification, starting with 25 ng template DNA. The MDA reaction with phi29 polymerase vastly surpassed the amplification with Taq in both yield and absence of bias. However, it should be noted that the starting material was still much higher than that normally available when the reaction is performed using one or a few cells as DNA source.

There are several major difficulties when amplifying a single DNA molecule that have to be overcome before the technique can be robustly used to sequence whole microbial genomes. First, while phi29 is highly processive, it will still dissociate from the template on average after 10–20 kb (Blanco et al. 1989). Therefore, multiple primers hybridized to random positions along the chromosome can ensure broad amplification. However, because only a limited number of such primers can be positioned on the chromosome without affecting the reaction, a limited number of regions will be amplified from each individual chromosome, introducing a bias that is manifested when only one or a few initial templates are used, but averaged out in nanogram-amount template reactions. In the case of microbes for which clonal populations are available, a solution can be to perform multiple separate amplifications using single cells followed by pooling of the individual products prior to further characterization. When analyzing uncultured microbes from natural communities, however, populations are usually not clonal and therefore individual cells likely contain genomic differences. Reactions on multiple, separately collected cells could therefore allow the retrieval of a larger amount of "pangenomic" information for a given species, which can be valuable from ecological and evolutionary perspectives. However, if the goal is to assemble a complete genome, such heterogeneity may hinder genome finishing efforts, particularly if cells have been identified using 16S rRNA-targeting probes which may result in pooled cells with large amounts of genomic divergence due to the extreme conservation of the 16S rRNA molecule. Collecting cells that occur together in filaments or microcolonies will have a higher likelihood of being clonal copies. In analyzing an uncultured soil TM7 bacterium, we pooled five cells that were genetically related (identical small subunit ribosomal RNA sequence) and used them as template for a two-stage MDA in an effort to minimize the bias (Podar et al. 2007). On the basis of the assembled contigs, there did not appear to be a high frequency of single nucleotide polymorphisms; however, because the amplification was still biased and we retrieved only a relatively small fraction of the genome (~20%), evaluation of gene order and overall chromosomal conservation was difficult.

3.1 Chimerism

Another current limitation of MDA is the formation of chimeric structures, which result in fragmented genes, and difficulties in assembling large genomic contigs. Published data (Zhang et al. 2006) indicate that hyperbranched structures generated during the strand displacement amplification are resolved after ligation into vectors and cloning in *E.coli* as chimeric, rearranged sequences. To remove hyperbranched structures, Zhang and colleagues used an S1 nuclease to cut the junctions of branched DNA molecules, and constructed a 3-kb sequencing library. This library showed a significant improvement in comparison to the library constructed without S1 nuclease treatment (Zhang et al. 2006). However, the remaining chimeras still limited the quality of genome assemblies. An improved assembly with longer contigs was obtained by computationally splitting these chimeric sequences at their junction points based on a reference genome (Zhang et al. 2006). A new study (Lasken and Stockwell 2007) suggests, however, that most rearrangements are caused during the amplification step and that sequencing methods that would bypass the need for cloning may not necessarily avoid this problem. Computational tools and extensive sequence analysis appear effective in detecting and eliminating or flagging such rearrangements and, while not ideal, they will be a useful compromise until a biochemical solution is found.

A final difficulty in amplifying single bacterial chromosomes that should be mentioned is the high risk of contamination. Separating and handling single bacterial cells is technically challenging and there are numerous sources of potential contamination throughout the process, not to mention that when working with complex communities, avoiding isolation of more than one type of cell is very difficult. Not only are other cells potential contaminants but free DNA can be present in the sample, attached to cells, as well as in the reagents. For example, we have noted that *Delftia*, a common contaminant of laboratory reagents, appears to be present in some batches of MDA kits. Since any DNA present can be amplified, special precautions have to be taken to prevent introducing it into the reaction by ultraclean procedures and reagents. Even then, contamination usually occurs and can only be dealt with by recognizing and computationally removing the sequences that belong to the unwanted organism, using phylogenetics and other sequence characteristics (Marcy et al. 2007; Podar et al. 2007). For example, we used GC content binning to separate sequences that belong to a *Pseudomonas* contaminant from the bulk of the data representing the soil TM7 bacterium. The high GC content sequences contain genes with very high similarity values to other *Pseudomonas* genes and phylogenetically distant from organisms that appear to be more related to TM7 than Proteobacteria (Cyanobacteria, Chloroflexi, Firmicutes). As more refined methods to bin sequences from organisms based on sequence composition appear (McHardy et al. 2007), such computational approaches should improve the efficiency of filtering single cell genomic datasets for contaminant sequences.

In addition to single cell genome amplification, MDA has been used also to amplify environmental genomic DNA for shotgun sequencing. Low-biomass

samples from highly contaminated soils yield DNA amounts that have limited use for direct, native analysis and screening. Using MDA, Abulencia and colleagues (Abulencia et al. 2006) recovered sufficient DNA from several low-biomass communities, which allowed diversity analysis as well as construction of genomic libraries for shotgun sequencing. While bias is difficult to control and evaluate in such cases, simply getting access to the genomes of uncultured organisms that inhabit such extreme environments should provide valuable information from both fundamental (ecological, physiological, and evolutionary) and practical (bioremediation) perspectives.

4 Sequencing and Genome Assembly

The strategy that has been applied almost exclusively so far for sequencing microbial genomes has been shotgun end sequencing of genomic DNA libraries using end chain terminator chemistry ("Sanger sequencing"). Over the last decade or so, tremendous improvements in this approach have been made, from long reads that approach 1,000 nucleotides, reduced cost associated with high-throughput platforms in specialized sequencing centers to efficient algorithms that allow assembly of the individual reads. Nevertheless, the time and effort to complete a genome project is still too long, primarily because of steps prior to the actual sequencing (e.g., genomic library construction, plasmid preps), and still too expensive ($8–$10k) to apply to more than a few organisms at a time in a standard laboratory or project. In the last several years a number of novel sequencing strategies have been explored that have the potential to trigger a new revolution in microbial genomic sequencing. One such approach, pyrosequencing, based on detection of the pyrophosphate released by the polymerase during copying of the sequencing template, was proposed two decades ago, but has only recently reached the stage of being applicable to genomic sequencing, primarily through an efficient parallel bead array–based technology developed by 454 Life Sciences (Ronaghi 2001). While the throughput of 454 sequencing is much higher than that of the traditional capillary Sanger sequencing (~100 Mbp vs. 3–4 Mbp per day per machine), the sequence reads are significantly shorter (~0.2 kbp vs. 1 kbp). The shorter reads present challenges for genome assembly, particularly in repetitive regions; however, a higher read depth coverage can address some of the limitations, and special assembly algorithms have been developed for the short sequence reads (Sundquist et al. 2007). An advantage of 454 sequencing is that genomic libraries do not need to be constructed, an important factor in sequencing genomes that have unusual sequence composition and are hard to clone into *E. coli*, or when the amount of starting DNA is limited such as in single cell genomics. Bypassing the cloning step in the latter case can avoid some of the recombinants that form in *E. coli* and are responsible for chimerism. However, it has been recently shown that a significant proportion of those chimeras form during amplification and not necessarily during cloning (Lasken and Stockwell 2007). Shorter reads will, in turn, make it more difficult to identify recombination points in the sequence. No

direct comparisons have yet been made in sequencing genomes starting from one or a few cells using Sanger and pyrosequencing.

While not yet widely used as a prime approach for de novo sequencing of microbial genomes, 454 can be used efficiently in combination with Sanger sequencing (Goldberg et al. 2006) as well as for genome resequencing or sequencing of closely related strains. We have used both sequencing approaches separately in sequencing the two different members of the TM7 phylum. For the soil TM7, ~20,000 Sanger sequence reads were generated from a small insert genomic clone library, resulting in combined a contiguous sequence (contig) length of ~1.8 Mbp after initial assembly. Using an additional chimera detection strategy followed by annotation, taxonomic binning, and a second round of contig analysis, we eliminated sequences likely belonging to a *Pseudomonas* coisolate and further reduced the presence of recombinant chimeric sequences. The total length of the contigs assigned to the TM7 bacterium was ~0.6 Mbp, encoding approximately 600 genes. On the basis of functional gene categories accumulation curves as well as inferences based on universally present genes in bacterial genomes, the coverage of the soil TM7 was estimated to be of ~0.2×, which suggests a genome size of at least 3 Mbp. Using a sliding window analysis we detected sharp fluctuations in the coverage depth along the contigs, reflecting the known bias in amplifying the genomes of single cells as a result of the limited number of primer annealing events. However, analysis of the frequency of different types of functional gene categories represented in the soil TM7 genomic data suggested that the bias is not related to preferential amplification of certain types of genes.

Pyrosequencing using the 454 technology was applied in the case of three individual TM7 cells isolated from the human oral microbial community by microchip fluidics (Marcy et al. 2007). As is the case with other single cell separation techniques including flow cytometry, contamination is still a problem and manifested in the microfluidic chips separation by the presence of small amounts of hitchhiking DNA from *Leptotrichia*, a recognized member of the oral community. While no cloning was involved, genomic sequence analysis revealed, as with the soil TM7, chimeric sequences and bias in genome coverage. A different assembly strategy had to by used, specifically designed to handle the short reads generated during pyrosequencing, namely the 454 Newbler assembler and the Forge whole genome shotgun assembler (D. Platt, unpublished). Overall, from the cell that yielded the highest quality sequence, 1,474 genes on 288 contigs were obtained. The gene prediction software, Fgenesb (http://www.softberry.com/) was used to predict genes, and has been shown to be perform relatively well on simulated metagenomic datasets (Mavromatis et al. 2007), followed by analysis using the IMG/M database platform (Markowitz et al. 2008).

The two described projects aimed at genomic analysis of organisms from the TM7 phylum employed different samples and different experimental and computational strategies. While overall the two studies encompass the same concepts of single cell isolation, genome amplification, and sequencing, it will be important in future to do side-by-side comparisons of contamination likelihood, amplification efficiency, and the quality of the resulting DNA sequences when applying flow cytometry and chip microfluidics to the same environmental sample.

5 From Genomes to Biology

While traditionally selection of microorganisms for genome sequencing is based on prior knowledge of an organism's physiology, ecology, or other biological aspects, the TM7 phylum was targeted because of its novel phylogenetic position and lack of cultivated representatives. Not surprisingly, little is known about the numerous uncultured bacterial phyla highlighted by culture-independent surveys (Hugenholtz 2002); however, TM7 has received somewhat more attention, providing a small amount of physiological and ultrastructural data with which to compare genome data. TM7 bacteria are known to inhabit a wide range of environments, from soils, water, and activated sludge, to termite guts ((Hugenholtz et al. 1998), Fig. 1). They have also been found in the human oral cavity and positively correlated with mild periodontitis (Brinig et al. 2003; Kumar et al. 2003). There is microscopic evidence that members of the TM7 phylum are morphologically diverse, and large conspicuous sheathed TM7 filaments often have epiflora (Hugenholtz et al. 2001; Thomsen et al. 2002; Xia et al. 2008). Electron microscopy also revealed that the cell envelope of a sheathed filament TM7 morphotype from activated sludge is ultrastructurally indistinguishable from Gram-positive bacteria, raising the possibility that TM7 represents a third Gram-positive bacterial phylum (Hugenholtz et al. 2001). If correct, TM7 may be able to help shed light on the evolutionary origins of the Gram-positive cell envelope (i.e., monophyletic vs. paraphyletic origin). There has been a sustained effort to obtain a glimpse into the physiology of TM7 bacteria in the absence of pure cultures, revealing that bacteria belonging to the TM7 phylum are capable of protein hydrolysis (Xia et al. 2007) and can uptake a number of carbon substrates under aerobic and anaerobic conditions (Thomsen et al. 2002). TM7 bacteria from soil have been shown to form microcolonies on a membrane support (Ferrari et al. 2005), although formal species description and continuous cultivation have not yet been reported for any member of this phylum. Genomic information may in fact facilitate cultivation attempts as has been recently demonstrated for bacteria from an acidophilic community (Tyson et al. 2005).

As is usually the case when genomic information first becomes available for members of a microbial phylum, it provides an opportunity to analyze evolutionary relationships with other organisms using sequence data other than the traditional ribosomal RNA. While rRNA trees are remarkably consistent with genome trees (Wolf et al. 2002), the latter can provide greater resolution because they incorporate a larger amount of phyogenetically informative data. For example, on the basis of comparative analysis of concatenated conserved marker genes, the bacterial phylum Acidobacteria appears to be specifically related to the Deltaproteobacteria (Ciccarelli et al. 2006), a relationship that was not detected by 16S rRNA gene comparisons. Previous analyses of environmental rRNA sequences have shown no specific affiliation of TM7 to any other bacterial phylum; indeed, this was the basis for classifying TM7 as a candidate phylum in the first place (Hugenholtz et al. 1998). Using concatenated ribosomal protein sequences, we were able to show that TM7 is most likely a sister group of the Chloroflexi (green nonsulfur bacteria, Fig. 2). This may indicate that the Gram-positive cell envelope has arisen more than

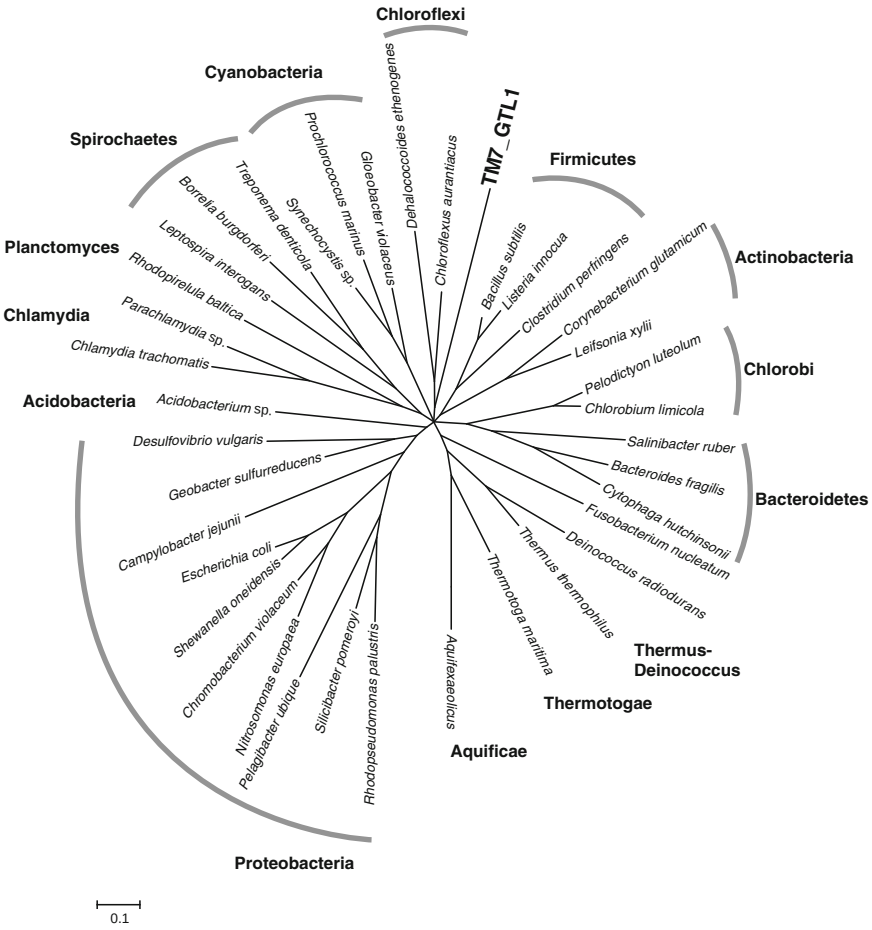

Fig. 2 Phylogenetic reconstruction of the relationships between bacterial phyla (divisions) based on concatenated ribosomal protein sequences. Numbers indicate bootstrap support, and branch-points with less than 70% bootstrap support were collapsed. Note the apparent relationship between TM7 and the Chloroflexi. Note also that the oral TM7 could not be added to this tree owing to an absence of many of the ribosomal proteins used in the analysis. Figure adapted from (Podar et al. 2007)

once since TM7 is not monophyletic with the Gram-positive phyla, the Actinobacteria, and Firmicutes, although further sequenced representatives of TM7 will be required to verify this inference. The Chloroflexi have unusual cell membranes (Hugenholtz and Stackebrandt 2004), and it will be interesting to further investigate the putative relationship to TM7 in this light when more genomic data become available. However, as it is often the case in reconstruction of the interphyla relationships within bacteria, the use of other phylogenetic markers leads to a relatively unresolved tree (Fig. 3).

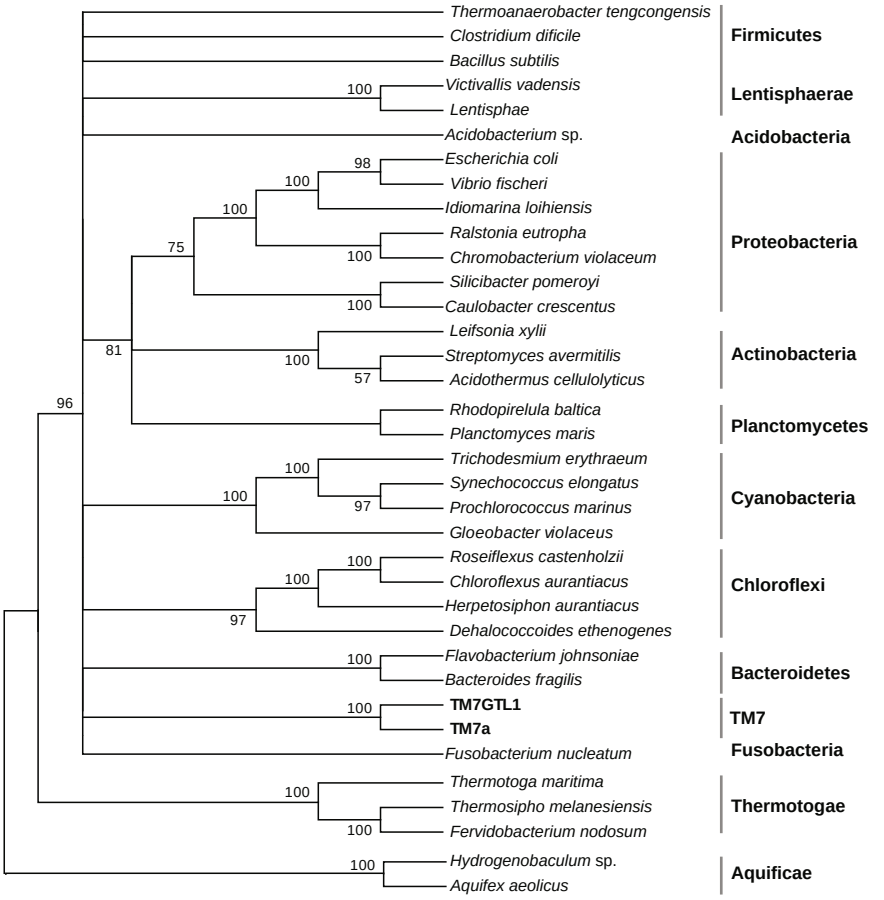

Fig. 3 Consensus dendrogram of the relationships between bacterial phyla (divisions) based on a concatenation of two conserved single copy marker genes: DNA gyrase (subunit A) and recA. Both genes were represented in the soil and oral TM7 datasets, allowing a comparative analysis of the two genomes beyond the 16S rRNA gene (Fig. 1). Branchpoints with less than 50% bootstrap support were collapsed. Note the monophyly of the TM7 genotypes but lower inter and intraphylum resolution as compared to the ribosomal proteins tree (Fig. 2), including loss of a specific relationship between TM7 and the Chloroflexi

Since we did not recover complete genomic sequences for either the soil or the human oral TM7 species, comprehensive metabolic reconstructions were not feasible. Nevertheless, there are certain aspects of the physiology and ecology of these organisms that can be inferred with reasonable confidence. Both TM7 genotypes contain genes involved in type IV pilus assembly, which has been implicated in twitching motility and biofilm formation, behavior that is important in colonizing either soil interstitial spaces or the tooth surface. A distinctive feature of the soil

TM7 appears to be an overrepresentation of restriction–modification systems, DNA repair, and antibiotic or heavy-metal transporters. These represent potentially important adaptations to an environment that is characterized by a very high diversity of microorganisms, competition, and predation as well as microbial chemical warfare. Unfortunately, the overlap of the two TM7 gene inventories was insufficient for inferring potential metabolic similarities and differences, owing presumably to the low and random genome coverage in both cases. However, as the remaining methodological hurdles are overcome in whole genome amplification and sequencing of single cells, a more complete picture of the metabolic repertoire of the TM7 phylum will emerge. More broadly speaking, it is clear that microbial ecology and evolution are entering a golden age and with the aid of powerful tools such as single cell genomics we will come to know the organisms with which we share the planet.

References

Abulencia CB et al. (2006) Environmental whole-genome amplification to access microbial populations in contaminated sediments. *Appl Environ Microbiol* 72:3291–3301

Amann R, Fuchs BM, Behrens S (2001) The identification of microorganisms by fluorescence in situ hybridisation. *Curr Opin Biotechnol* 12:231–236

Blanco L, Bernad A, Lazaro JM, Martin G, Garmendia C, Salas M (1989) Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. *Symmetrical mode of DNA replication. J Biol Chem* 264:8935–8940

Brinig MM, Lepp PW, Ouverney CC, Armitage GC, Relman DA (2003) Prevalence of bacteria of division TM7 in human subgingival plaque and their association with disease. *Appl Environ Microbiol* 69:1687–1694

Cho JC, Vergin KL, Morris RM, Giovannoni SJ (2004) Lentisphaera araneosa gen. nov., sp. nov, a transparent exopolymer producing marine bacterium, and the description of a novel bacterial phylum, Lentisphaerae. *Environ Microbiol* 6:611–621

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287

Dean FB et al. (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA* 99:5261–5266

DeSantis TZ et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72:5069–5072

Ferrari BC, Binnerup SJ, Gillings M (2005) Microcolony cultivation on a soil substrate membrane system selects for previously uncultured soil bacteria. *Appl Environ Microbiol* 71:8714–8720

Fu AY, Chou HP, Spence C, Arnold FH, Quake SR (2002) An integrated microfabricated cell sorter. *Anal Chem* 74:2451–2457

Garcia Martin H et al. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* 24:1263–1269

Goldberg SM et sal. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc Natl Acad Sci USA* 103:11240–11245

Hawkins TL, Detter JC, Richardson PM (2002) Whole genome amplification--applications and advances. *Curr Opin Biotechnol* 13:65–67

Hong JW, Studer V, Hang G, Anderson WF, Quake SR (2004) A nanoliter-scale nucleic acid processor with parallel architecture. *Nat Biotechnol* 22:435–439

Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol 3:REVIEWS0003

Hugenholtz P, Stackebrandt E (2004) Reclassification of Sphaerobacter thermophilus from the subclass Sphaerobacteridae in the phylum Actinobacteria to the class Thermomicrobia (emended description) in the phylum Chloroflexi (emended description). *Int J Syst Evol Microbiol* 54:2049–2051

Hugenholtz P, Goebel BM, Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* 180:4765–4774

Hugenholtz P, Tyson GW, Webb RI, Wagner AM, Blackall LL (2001) Investigation of candidate division TM7, a recently recognized major lineage of the domain Bacteria with no known pure-culture representatives. *Appl Environ Microbiol* 67:411–419

Hutchison CA III, Venter JC (2006) Single-cell genomics. *Nat Biotechnol* 24:657–658

Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* 437:543–546

Kumar PS, Griffen AL, Barton JA, Paster BJ, Moeschberger ML, Leys EJ (2003) New bacterial species associated with chronic periodontitis. *J Dent Res* 82:338–344

Kvist T, Ahring BK, Lasken RS, Westermann P (2007) Specific single-cell isolation and genomic amplification of uncultured microorganisms. *Appl Microbiol Biotechnol* 74:926–935

Lasken RS, Stockwell TB (2007) Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol* 7:19

Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res* 34:D332–D334

Loy A, Maixner F, Wagner M, Horn M (2007) probeBase–an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Res* 35:D800–D804

Ludwig W et-al. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* 32:1363–1371

Marcy Y et al. (2007) Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci USA* 104:11889–11894

Markowitz VM et-al. (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* 36:D534–D538

Mavromatis K et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* 4:495–500

McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* 4:63–72

Ouverney CC, Armitage GC, Relman DA (2003) Single-cell enumeration of an uncultivated TM7 subgroup in the human subgingival crevice. *Appl Environ Microbiol* 69:6294–6298

Pinard R et al. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 7:216

Podar M et al. (2007) Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl Environ Microbiol* 73:3205–3214

Preston CM, Wu KY, Molinski TF, DeLong EF (1996) A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc Natl Acad Sci USA* 93:6241–6246

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

Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 38:525–552

Ronaghi M (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Res* 11:3–11

Speit G, Schutz P, Merk O (2000) Induction and repair of formaldehyde-induced DNA-protein crosslinks in repair-deficient human cell lines. *Mutagenesis* 15:85–90

Strous M et al. (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* 440:790–794

Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. PLoS ONE 2:e484

Thomsen TR, Kjellerup BV, Nielsen JL, Hugenholtz P, Nielsen PH (2002) In situ studies of the phylogeny and physiology of filamentous bacteria with attached growth. *Environ Microbiol* 4:383–391

Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* 6:805–814

Tyson GW et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43

Tyson GW, Lo I, Baker BJ, Allen EE, Hugenholtz P, Banfield JF (2005) Genome-directed isolation of the key nitrogen fixer Leptospirillum ferrodiazotrophum sp. nov. from an acidophilic microbial community. *Appl Environ Microbiol* 71:6319–6324

Woese CR (1987) Bacterial evolution. *Microbiol Rev* 51:221–271

Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees and the tree of life. *Trends Genet* 18:472–479

Xia Y, Kong Y, Nielsen PH (2007) In situ detection of protein-hydrolysing microorganisms in activated sludge. *FEMS Microbiol Ecol* 60:156–165

Xia Y, Kong Y, Thomsen TR, Nielsen PH (2008) Identification and ecophysiological characterization of epiphytic protein-hydrolyzing saprospiraceae (Candidatus Epiflobacter spp.) in activated sludge. *Appl Environ Microbiol* 74:2229–2238

Zengler K et al. (2005) High-throughput cultivation of microorganisms using microcapsules. *Methods Enzymol* 397:124–130

Zhang H et al. (2003) Gemmatimonas aurantiaca gen. nov., sp. nov., a gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum Gemmatimonadetes phyl. nov. *Int J Syst Evol Microbiol* 53:1155–1163

Zhang K et al. (2006) Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* 24:680–686

Zoetendal EG, Plugge CM, Akkermans AD, de Vos WM (2003) Victivallis vadensis gen. nov., sp. nov., a sugar-fermenting anaerobe from human faeces. *Int J Syst Evol Microbiol* 53:211–215