Classification: PHYSICAL SCIENCES: Applied Physical Sciences and BIOLOGICAL SCIENCES: Microbiology

Dissecting Biological Dark Matter: Single Cell Genetic Analysis of TM7, a Rare and Uncultivated Microbe from the Human Mouth

Yann Marcy,[1,5] Cleber Ouverney,[2] Elisabeth M. Bik,[3] Tina Lösekann,[3] Natalia Ivanova,[4] H. Garcia Martin,[4] E. Szeto,[4] Darren Platt,[4] Philip Hugenholtz,[4] David A. Relman,[3] and Stephen R. Quake[1,*]

1 Dept of Bioengineering, Stanford University and Howard Hughes Medical Institute, Stanford, CA 94305

2 Dept of Biological Sciences, San Jose State University, San Jose, CA 95192

3 Depts of Microbiology & Immunology, and of Medicine, Stanford University, Stanford, CA 94305; Veterans Affairs Palo Alto Health Care System, Palo Alto, CA 94304

4 Department of Energy Joint Genome Institute, Walnut Creek, CA 94598

5 Present address: Genewave SAS, XTEC Ecole Polytechnique 91128 Palaiseau, France

* To whom correspondence should be addressed. Email: quake@stanford.edu

18 Text pages, 4 Figures, 1 Table

118 words in abstract. 26,880 characters in manuscript.

Abstract We developed a microfluidic device that allows isolation and genome amplification of individual microbial cells, thereby enabling organism-level genomic analysis of complex microbial ecosystems without the need for culture. This device was used to perform a directed survey of the human subgingival crevice and isolate bacteria with rod-like morphology. Several isolated microbes had a 16S ribosomal RNA sequence that placed them in candidate phylum TM7, which has no cultivated or sequenced members. Genome amplification from individual TM7 cells allowed us to sequence and assemble more than 1,000 genes, providing the first insights into the physiology of members of this phylum and the first genetic analysis of any uncultivated minority member of a microbial community.

Introduction

Earth contains enormous microbial diversity; microbes colonize a wide variety of environmental niches creating complex ecosystems and communities. Despite the marvelous progress in microbiology over the past century, we have only scratched the surface of this microbial world: it has been estimated that fewer than 1% of bacterial species have been axenically cultured, and fewer than half of the recognized bacterial phyla have cultivated representatives (1). This can be viewed as biology's "dark matter" problem: just as astronomers have been able to infer indirectly the existence of a large amount of as-yet-undetected mass in the universe, microbiologists are only able to estimate microbial diversity by techniques such as comparative 16S ribosomal RNA gene analysis (2), community DNA hybridization efficiency (3), and metagenomic gene inventories (4). While these techniques are useful, the cell, which is the ultimate unit of biological organization, is lost as a distinct informational entity.

There have been two general approaches to this problem. The first is to work on simple communities that contain only a few microbial species, in which case genome sequences can be reconstructed computationally after sequencing bulk DNA purified from the community (5). The second approach has been to isolate individual cells by fluorescence activated cell sorting (FACS), micromanipulation, or serial dilution, followed by genomic DNA amplification with techniques such as multiple strand displacement amplification (MDA) (6,7). The latter approach has been successfully used to perform genomic analysis of the cultivated and abundant marine bacterium *Prochlorococcus* MIT9312 (7). However, this approach remains difficult for two primary reasons: the confidence needed to assert the presence of single cells in microliter

volumes, and the meticulous reagent cleaning and sample handling required to suppress background amplification in microliter volume MDA (6). Those hurdles become even greater when complex environmental samples are used. The number of species present requires substantial reagent consumption and expensive post amplification screening, and the probability of contamination is much higher due to the presence of free DNA. Neither approach has been validated with a complex ecosystem.

We designed and fabricated a microfluidic chip to address these limitations. This device provides the ability to perform parallel isolation of single bacteria by steering them to any one of 8 individually addressable chambers, followed by lysis and amplification of their individual genomes in 60 nanoliter volumes. By using nanoliter volumes, the specific template concentration is increased by three orders of magnitude, as previously suggested (8,9). To demonstrate the potential of this approach in microbial ecology, we performed a selective survey of microbes found in the human subgingival crevice, followed by whole genome amplification and high throughput sequencing. The 16S ribosomal RNA gene-based phylogeny of several of these microbes placed them within the candidate phylum TM7, for which no cultivated or sequenced members exist (12), thereby providing the first genetic information about oral representatives of the TM7 phylum.

Results and Discussion

The microfluidic strategy for microbe isolation and genome amplification (Figure 1) was validated on *Escherichia coli*. More than two dozen amplifications on single *E. coli* cells were performed, with a success rate of >90%. Subsequent PCR analysis of 10 genomic loci distributed over the *E. coli* chromosome showed that the amplification achieved excellent coverage and was able to amplify sequences with equal efficacy independent of their location on the genome. (Supplementary figure 5) Control experiments with only culture fluid in the chamber showed no significant amplification.

We then demonstrated the ability to select, isolate and amplify the genomes of single bacteria from the human oral microbiota. The number of species in the human mouth is estimated to be about 700 (11,23). Due to the challenges of removing intact biofilm samples, rather than performing a comprehensive survey of this complex community our purpose was instead to target an unexplored phylum and a relatively rare subset of the oral microbiota, TM7. By selecting microbial cells with a rod-like morphotype, we expected to enrich for the candidate phylum TM7 (13,14). Little is known about the TM7 lineage. Based on comparative analyses of 16S ribosomal RNA genes, it is one of a number of prominent “candidate” bacterial phyla lacking any cultivated representatives, but comprising greater than 50 phylotypes (1). Ribosomal RNA gene sequences from the TM7 phylum have been found in a variety of habitats ranging from deep-sea hydrothermal vents to the healthy human mouth (12-14); in addition, sequence types within this phylum have been associated with chronic periodontitis in humans (13-14). Fluorescence *in situ* hybridizations specific for TM7 showed that 0.7-1.9% of the subgingival microbiota belongs to the TM7 phylum (13). A significant subset of this phylum has a peculiar morphology, characterized by long, thick

filaments (up to 50 x 4 μm), making these cells good candidates for a morphotype-based selection (12,13).

In order to identify the amplified genomes of the isolated rod-like cells, we performed PCR on the 16S ribosomal RNA gene using primer sequences conserved across most species of the bacterial domain. Positive results for 16S rDNA PCR were obtained for 34 of 35 captured, single cells**.** After gel purification, 30 of these amplicons were directly sequenced; 28 of these gave unique sequences that were compared against the NCBI database using BLAST (24). Fig. 2 shows a phylogenetic tree based on 16S ribosomal RNA gene sequences of most recognized bacterial phyla, with annotations for isolates from the present survey. The 28 sequences from this study are associated with 5 different bacterial phyla, with most sequences located in the phylum *Fusobacteria* and specifically related to the genus *Leptotrichia*.

We identified 4 members of the phylum TM7 from the amplified cells, of which 3 were closely related to a known oral TM7 clone (>99.6%, Genbank accession AY144355) (14) and a fourth clone related to a more distant lineage in the phylum (97.3%, AY134895) (25). To verify that the genome of a unique sequence type was amplified, the 16S rRNA amplicon of one TM7 sample (TM7a) was cloned, and 24 clones were sequenced; 23 of the 24 clones had >99.5% sequence identity to the directly sequenced PCR product. To provide insight into the biology of the TM7 phylum and to investigate the ability to recover whole genome sequences from single uncultivated cells, we used the amplified genomic DNA from this sample for pyrosequencing and genome assembly. The resulting genome sequence data set was loaded into the IMG/M database (19) to facilitate comparative analysis.

The assembly of TM7a genomic sequence resulted in the generation of 3,245 genes and gene fragments distributed across 1,825 scaffolds, totaling 2.86 MB. Genome size estimates based on approaches such as the Lander-Waterman equation (26) or the characterization of known, conserved, single-copy genes (27) rely on random sampling of the genome. Single cell amplification introduces a bias in read sampling such that we could not reliably estimate TM7 genome size. The assembly was fairly fragmented, with only 60% of the genes on multi-gene scaffolds. This suggested that there is multiple representation of some genes in the assembly, and that the actual number of sampled genes in TM7a is somewhat smaller. If one applies a more conservative filter and only includes genes from large contigs (defined as those having 3 or more genes), then one is left with 1,474 genes on 288 scaffolds; this is probably a better estimate of the number of unique sampled genes in TM7a. Approximately 43% of genes were assigned a predicted function based on homology to published sequences, and 44% of the genes were mapped to clusters of orthologous groups (COG) (Table 1). We tested the validity of the assembly by choosing 5 regions of the genome with an average size of 1 kb, designed PCR primers, and successfully amplified all 5 regions from aliquots of the amplified TM7a genomic DNA (Supplementary figure 6).

Sequence similarity-based mapping showed that most of the TM7a genes are not closely related to genes from representatives of any known phyla. For example, 80% of the predicted TM7 proteins have less than <60% sequence identity to proteins from other sequenced organisms (Fig. 3). Using this approach, a full third (33%) of the TM7 genes have less than 30% protein sequence identity to genes from any known phylum. This result is consistent with other cases of genome sequencing in previously-uncharacterized

phyla. For example *Rhodopirellula baltica* was the first sequenced representative of the *Planctomycetes* phylum and 89% of its proteins have <60% identity to proteins from other known organisms; 20% have no matches with >30% identity. In contrast, a survey of 13 bacterial species in phyla with multiple sequenced representatives showed that on average only 15% of the proteins have <60% identity to proteins in other organisms, and 3% are unassigned at the 30% cutoff (Supplementary figure 7).

Although the majority of genes in the TM7a assembly are only distantly related to genes found in other organisms, there is a minority with relatively high sequence similarity (>60% identity) to genes found in members of the classes *Bacilli*, *Clostridia*, or *Fusobacteria*. The presence of these genes may be the result of extensive lateral transfer between species in the mouth, as has been postulated for other oral bacteria (16), or may be due to the presence of contaminating DNA in our samples – perhaps from free DNA that entered the microfluidic amplification reactor with the TM7 cell, either in solution or bound to the cell membrane. If the presence of these genes was due to contaminant DNA, one would expect them to cluster together by organism in the assembly. The data show that in many cases the opposite is true: genes with putative relationships to disparate organisms assemble onto the same contig. The TM7a assembly does contain at least some exogenous DNA: examination of the raw sequencing reads shows that more than 40 reads assembled into the TM7a 16S ribosomal rRNA gene sequence, while 4 reads assembled onto a separate small contig with the 16S rRNA gene sequence belonging to *Leptotrichia* species. Extrapolating from the ratio between these raw reads, we estimate that the proportion of *Leptrichia* contamination is less than 10%. Since it is difficult to

assign a more precise estimate, one avenue of analysis is to interpret the TM7a assembly as a metagenome that is highly enriched for a TM7 bacterium.

We also sequenced a second TM7 cell, TM7b, with an identical 16S r RNA gene sequence to TM7a, that had been isolated on a separate day on a separate chip. Ten Mb of sequence data were obtained and assembled into ~15,000 small contigs, none larger the 4 kb. This was not enough to provide a complete assembly, but represents a sampling of the genome. These sequence data were analyzed with whole genome Vista (wgVista) (21) as an independent confirmation of the TM7a genome assembly and to facilitate identification of bona fide TM7 genes. The results are shown in Fig. 4. In a global alignment, the vast majority of TM7b sequences could be mapped to contigs in TM7a with a sequence identity exceeding 70%. As a control experiment, we also aligned the TM7 genome sequences to *Fusobacterium nucleatum* (the only sequenced organism in the phylum Fusobacteria to which *Leptotrichia* belongs) and *Chloroflexus aurantiacus* (the sequenced organism with closest 16S rRNA gene sequence to TM7 in figure 2). Neither of the latter demonstrated substantial sequence identity to the TM7b sequence assembly. Sequencing multiple representatives of a novel phylum is therefore a useful approach for identifying *bona fide* target phylum genes in metagenomic samples containing exogenous DNA, which may be an unavoidable limitation associated with amplification of single cells removed from multi-species samples.

Metabolic analysis of TM7 was performed by pooling sequence data from TM7a and TM7b, along with data from a third TM7 cell (TM7c). TM7c assembled into 474 kB and 632 genes, but was not used as an independent reference since a sample handling error during sequencing caused commingling with genomic DNA from TM7a. We

performed binning of the metagenome based on similarities between the three TM7 samples and phylogenetic markers by selecting contigs that have phylogenetic marker genes on very long branches. Based on the presence of recognizable signature genes, the oral TM7 cells are predicted to be capable of a range of common metabolic processes, such as glycolysis (3-phosphoglycerate kinase, phosphoglycerate mutase triosephosphate isomerase and pyruvate kinase), the TCA cycle (succinyl-CoA synthetase), nucleotide biosynthesis (dihydroorotate dehydrogenase, uridylate kinase, guanylate kinase, aerobic-type ribonucleoside diphosphate reductase, thymidylate synthase) and some amino acid biosynthesis and salvage pathways (cysteine synthase, glycine hydroxymethyltransferase). We identified several genes coding for glycosyl hydrolase family enzymes distantly related to alpha-amylases and oligo-1,6-glucosidases, suggesting that oral TM7 cells may be capable of utilizing oligosaccharides as growth substrates. Arginine is another potential growth substrate, due to the presence of genes from the arginine deiminase pathway (arginine deiminase, ornithine carbamoyltransferase and carbamate kinase). We also identified genes for ABC transporters that are likely responsible for oligopeptide uptake, suggesting that TM7 cells may be capable of using other amino acids as well.

It is an open question whether these bacteria have attributes associated with virulence, and might be capable of contributing to oral disease. We noted the presence of genes for type IV pilus biosynthesis, including one with similarity to that which encodes the *Vibrio vulnificus* type IV pilin (28). While type IV pili may facilitate adherence of bacteria to epithelial cells, and contribute to biofilm formation, in Gram-positive cells, type IV pili have been shown to be responsible for an unusual communal form of gliding

motility (29). TM7 cells from a sludge bioreactor appeared to have typical Gram-positive cell envelopes by electron microscopy (12). Therefore, if the TM7 are Gram positive, their type IV pili may be involved in gliding motility.

We also investigated genes that might participate in cell envelope biosynthesis and found a gene predicted to encode a novel sortase, distantly related to those of *Firmicutes* and *Actinobacteria*, and a gene predicted to encode a UDP-N-acetylmuramyl tripeptide synthetase related to those of the bifidobacteria, suggesting a specific relationship of the TM7 cells to the Gram-positive lineages (Supplementary figure 8). Interestingly, in bifidobacteria the latter enzyme is predicted to add an atypical amino acid (ornithine or lysine instead of the more common diaminopimelate) to the growing peptidoglycan chain producing an A4alpha/beta type peptidoglycan. This peptidoglycan type has been implicated in chronic granulomatous inflammation (30) and may serve as a virulence factor for oral TM7.

In conclusion, we have isolated single bacterial cells from a complex human microbial community and sequenced their DNA to provide the first genetic insights into the TM7 phylum. The cell selection process described here used morphology as the basis for selection of the targeted bacteria. It would also have been possible to achieve the same results from an unbiased survey of the environmental sample; this simply would have required processing a larger number of cells. Since the cells were isolated from a complex bacterial biofilm with no manipulation other than pipetting and dilution, many environmental microbial ecosystems should be amenable with this technique. (32) We predict that as genomes from the microbial dark matter are sampled using techniques

such as single cell amplification, a much richer tapestry of microbial evolution will emerge.

Materials and Methods

Microfluidic chip fabrication: Microfluidic chips (figure 1) were fabricated as described previously (10) using the “push up” geometry with the following adjustments. The flow molds contained two layers, one for feeding lines and valves (SPR220 7 μm high), and one for the reaction chambers (SU8 2025, 25 μm high). The control molds contained two layers: one layer for hydration channels under the reaction chamber (SU8 2015, 10μm high) and one for the control lines (SU8 2025, 25μm high).

Sample Collection and Isolation. Samples were collected from periodontal pockets by scraping subgingival tooth surfaces of a healthy individual (male, 40 years) after 5 days without tooth brushing. These biofilm specimens were dispersed, suspended and washed twice in 1xPBS buffer, and re-suspended in 1xPBS 0.2% Tween® 20 before loading onto the chip. The chip was placed on an optical microscope and the sample was pumped through a sorting channel. When a single rod-shaped cell or a filament with the appropriate morphology (13) was visually detected in front of each processing unit, an isolation valve was closed and the cell was examined with a higher magnification. If the cell satisfied the selection criteria, the sorting valve was opened and the cell was pumped into the sorting chamber. Otherwise, the isolation valve was reopened and another cell was selected. This operation was repeated for 7 processing units of the chip; the eighth unit was used for a negative control having only suspension fluid inside. The chip also

contains an independent processor with a separate, non-addressable input that was filled with a mixture of lysed cells as a positive control. Every template chamber was then carefully checked for the number of bacterial cells and a high magnification image was recorded for every cell (Fig. 2). Out of 42 processing units (6 chips) used, 35 contained only one visible cell or filament.

Cell Lysis and Whole Genome Amplification (WGA). Lysis, neutralization, and WGA were performed with the REPLI-g kit (Qiagen) using the recommended protocol except for on-chip WGA, for which the reaction mix was supplemented by 0.2% Tween® 20 and one additional volume of polymerase. Once all the chambers were loaded with cells, an hour-long lysozyme treatment was applied using 1xPBS with 0.2% Tween® 20 and 100 Units/µl of lysozyme (Epicentre). This was performed by taking advantage of the gas-permeability of PDMS to dead-end fill the feeding lines with the lysis buffer (Fig. 1C), and by opening the feeding valve to push the contents of the sorting chamber into the lysis chamber (Fig. 1D). Lysis and DNA denaturation reagents were allowed to incubate for 30 min. During this time, the feeding lines were washed first with air then with the neutralization buffer (Fig. 1E). After completion of the lysis, the feeding valve was reopened, and neutralization buffer was pushed into the unit via dead-end filling of the neutralization chamber (Fig. 1F). After 15-20 min, washing of the feeding line was repeated, this time with the WGA reaction mix (Fig. 1G). The feeding valve was reopened and the reaction mix was used to dead-end fill the reaction chamber. With each WGA reaction isolated by closed valves, the chip was placed on a hotplate set at 32°C. The on-chip amplification took place for 10 to 16 h after which samples were retrieved from the chip. The amount of amplified DNA after this step was estimated to be about 50

ng. A second, off-chip amplification was performed with the REPLI-g kit in order to obtain micrograms of DNA, the amount required for sequencing.

16S ribosomal RNA gene amplification, cloning, and sequencing. 16S ribosomal RNA gene PCR was performed on amplified genomic DNAs using broad-range bacterial primers 8FM (5' AGAGTTTGATCMTGGCTCAG 3') (adapted from ref 31) and 1391R (5'GACGGGCGGTGTGTRCA 3') (adapted from ref 22). These primers amplify approximately >90% of the full-length bacterial 16S ribosomal RNA coding sequence. PCR mixtures were composed of 1x PCR buffer II (Applied Biosystems, Foster City, CA), 1.5 mM $MgCl_2$, 0.05% Triton X-100, 20 mM tetramethylammonium chloride, 0.1 mM concentrations of each deoxyribonucleoside triphosphate, 0.4 μM concentrations of each primer, 2.5 U of AmpliTaq DNA polymerase (Applied Biosystems), and 1 μl of amplified DNA in a final volume of 50 μl. PCRs included 5 min at 95°C, 35 cycles of 30 sec at 94°C, 30 sec at 55°C and 90 sec at 72°C, followed by 8 min at 72°C. PCR reactions were sequenced (Geneway, Hayward, CA) directly after purification from agarose gel using the QIAquick Gel Extraction Kit (Qiagen, Valencia, CA), or after cloning using the TOPO-TA cloning kit (Invitrogen).

Genome Sequencing and Assembly. Pyrosequencing (454 Life sciences, CT) was performed on randomly-amplified genomic material from three TM7 cells, named TM7a, TM7b and TM7c. Each sequencing run yielded between 10 and 39 MB of raw data composed of ~100 bp reads; the reads were assembled using the 454 Newbler^{TM} assembler and Forge whole genome shotgun assembler (D. Platt, unpublished). An initial assembly treating the coverage as a classic Poisson distribution indicated that the coverage of these genomes was quite uneven and that some regions were not joined due

to either excess or very low coverage. The data were reassembled with Forge using "metagenomic assumptions". In this configuration, the assembler relaxes the Poisson depth assumption which allows for much deeper coverage and exploration of low-coverage, less-certain overlaps between reads. All single-read, more highly error-prone contigs were excluded from the assembly. Genes were predicted on contigs greater or equal in length to an average Sanger read (750 bp) using fgenesb as previously described (18), then loaded into the Integrated Microbial Genomes with Microbiomes (IMG/M) system (19) to facilitate comparative analysis.

Acknowledgments

This work was supported by National Institutes of Health Director's Pioneer Awards (to S.R.Q. and D.A.R.) and by NIH 1RO1 HG002644-01A1 (S.R.Q. and Y.M.). T.L. was supported by the Stanford Dean's Postdoctoral Fellowship/Aaron Fund. This work was performed in part under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

References

(1) Rappe MS, Giovannoni SJ 2003 Annu Rev Microbiol.;57:369-94.
(2) Schmidt TM, Relman DA (1994) *Method Enzymol* 235:205-22.
(3) Palmer C, Bik EM, Eisen MB, Eckburg PB, Sana TR, Wolber PK, Relman DA, Brown PO. 2006 Nucleic Acids Res. 34(1):e5.
(4) Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. Community structure and metabolism through reconstruction of microbial genomes from the environment. 2004 Nature. 2004 Mar 4;428(6978):37-43
(5) Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO. 2004 Science. Apr 2;304(5667):66-74.
(6) Raghunathan A, Ferguson HR Jr, Bornarth CJ, Song W, Driscoll M, Lasken RS. 2005 Appl Environ Microbiol. Jun;71(6):3342-7.
(7) Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM. 2006 Nat Biotechnol. Jun;24(6):680-6.
(8) L. McBride, M. Lucero, M. Unger, H.R. Nassef, and G. Facer, US Patent application US 2005/0019792A1 (2005).
(9) Hutchison CA 3rd, Smith HO, Pfannkoch C, Venter JC. 2005 Proc Natl Acad Sci U S A Nov 29;102(48):17332-6.
(10) Thorsen T, Maerkl SJ, Quake SR. Science. 2002 Oct 18;298(5593):580-4
(11) Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, Levanos VA, Sahasrabudhe A, Dewhirst FE. J Bacteriol. 2001 Jun;183(12):3770-83.
(12) Hugenholtz P, Tyson GW, Webb RI, Wagner AM, Blackall LL. 2001 Appl Environ Microbiol. Jan;67(1):411-9
(13) Ouverney CC, Armitage GC, Relman DA. 2003 Appl Environ Microbiol. Oct;69(10):6294-8.
(14) Brinig MM, Lepp PW, Ouverney CC, Armitage GC, Relman DA. 2003 Appl Environ Microbiol, Mar;69(3):1687-94
(15) Kapatral V, Anderson I, Ivanova N, Reznik G, Los T, Lykidis A, Bhattacharyya A, Bartman A, Gardner W, Grechkin G, Zhu L, Vasieva O, Chu L, Kogan Y, Chaga O, Goltsman E, Bernal A, Larsen N, D'Souza M, Walunas T, Pusch G, Haselkorn R, Fonstein M, Kyrpides N, Overbeek R. 2002 J Bacteriol. 2002 Apr;184(7):2005-18.
(16) Mira A, Pushker R, Legault BA, Moreira D, Rodriguez-Valera F. 2004 BMC Evol Biol. Nov 26;4:50.
(17) Goldberg SM, Johnson J, Busam D, Feldblyum T, Ferriera S, Friedman R, Halpern A, Khouri H, Kravitz SA, Lauro FM, Li K, Rogers YH, Strausberg R, Sutton G, Tallon L, Thomas T, Venter E, Frazier M, Venter JC. 2006 Proc Natl Acad Sci U S A. Jul 25;103(30):11240-5
(18) García Martín, H., N. Ivanova, V. Kunin, F. Warnecke, K. Barry, A.C. McHardy, C. Yeates, S. He, A. Salamov, E. Szeto, E. Dalin, N. Putnam, H. 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. Nature Biotechnology 24(10):1263-1269.

(19) Markowitz, V.M., N.N. Ivanova, K. Palaniappan, E. Szeto, F. Korzeniewski, A. Lykidis, I. Anderson, K. Mavrommatis, V. Kunin, H. García Martín, I. Dubchak, P. Hugenholtz, N.C. Kyrpides. 2006. An experimental metagenome data management and analysis system. Bioinformatics 22(14):e359-367.

(20) Kapatral, V., I. Anderson, N. Ivanova, G. Reznik, T. Los, A. Lykidis, A. Bhattacharyya, A. Bartman, W. Gardner, G. Grechkin, L. Zhu, O. Vasieva, L. Chu, Y. Kogan, O. Chaga, E. Goltsman, A. Bernal, N. Larsen, M. D'Souza, T. Walunas, G. Pusch, R. Haselkorn, M. Fonstein, N. Kyrpides, and R. Overbeek. 2002. Genome Sequence and Analysis of the Oral Bacterium Fusobacterium nucleatum Strain ATCC 25586. J. Bacteriol. 184:2005-2018.

(21) Frazer K.A., Pachter L., Poliakov A., Rubin E.M., Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004 Jul 1;32 (Web Server issue):W273-9

(22) Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. (1985) Proc Natl Acad Sci U S A 82(20):6955-9.

(23) Aas, J. A., B. J. Paster, L. N. Stokes, I. Olsen, and F. E. Dewhirst (2005) J. Clin. Microbiol. 43:5721-5732.

(24) Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 29:2994-3005.

(25) Paster B.J., Russell M.K., Alpagot T., Lee A.M., Boches S.K., Galvin J.L. and Dewhirst F.E. (2002) Ann. Periodontol. 7 (1), 8-16.

(26) Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics. 1988 Apr;2(3):231-9.

(27) Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P. Prediction of effective genome size in metagenomic samples. Genome Biol. 2007 Jan 15;8(1):R10

(28) Paranjpye RN, Strom MS, A Vibrio vulnificus type IV pilin contributes to biofilm formation, adherence to epithelial cells, and virulence. Infect Immun., 2005, 73(3), 1411-1422

(29) Varga JJ, Nguyen V, O'Brien DK, Rodgers K, Walker RA, Melville SB. Type IV pili-dependent gliding motility in the Gram-positive pathogen Clostridium perfringens and other Clostridia. Mol. Microbiol. 2006, 62(3), 680-694

(30) Simelyte E, Rimpilainen M, Zhang X, Toivanen P. Role of peptidoglycan subtypes in the pathogenesis of bacterial cell wall arthritis. Ann Rheum Dis. 2003 Oct;62(10):976-82.

(31) Edwards U, Rogall T, Blocker H, Emde M & Böttger EC (1989) Nucleic Acids Res. 17: 7843–7853.

(32) Kolenbrander, P. E., R. N. Andersen, D. S. Blehert, P. G. Egland, J. S. Foster, and R. J. Palmer, Jr. 2002. Communication among Oral Bacteria, Microbiol. Mol. Biol. Rev. 66:486-505.

Figure captions:

Figure 1. A. Photograph of a single cell isolation and genome amplification chip capable of processing 9 samples in parallel. To visualize the architecture, the channels and chambers have been filled with blue food coloring, and the control lines to actuate the valves have been filled with red food coloring (scale bar 5 mm). B. Schematic diagram of a single amplification unit. The feed line is used to bring reagents into the chambers when the Vr valve is open, and to the waste when the Vw valve is open. The Vin valve allows deposition of a single bacterium into the sorting chamber. The lysis (3.5 nl), neutralization (3.5 nl), and reaction chambers (50 nl) are used in sequence and are separated by individual valves Vl, Vn, and Vr. Valve Vout allows recovery of the amplified genomic material from the chip into an individual microfuge tube. C. After a cell is trapped in the chamber, the feed line is filled with lysis buffer. D. The lysis buffer is used to push the cell into the lysis chamber. E. While the lysis buffer is mixing with the cell solution by diffusion, the feed line is flushed. F. Neutralization buffer is loaded into the feed line and used to push the cell lysate into the neutralization chamber. G. While the neutralization reaction is mixing by diffusion, the feed line is flushed. H. The WGA reagents are loaded into the feed line and used to push the neutralized cell lysate into the reaction chamber. I. The amplification reaction proceeds in a closed system comprising sorting, lysis, neutralization, and reaction chambers.

Figure 2. Left. Phylogenetic tree showing bacterial phyla based on 16S ribosomal RNA gene analysis (adapted from (1)). Green text indicates that at least one member of the

phylum has been cultivated, while different shades of blue indicate the number of genome sequencing projects in a particular phylum that were completed or in progress as of May, 2006. Red numbers and percentages indicate the results of our single cell survey of the human subgingival crevice, in which filamentous bacteria with rod-like morphotypes were isolated, lysed and their genomes amplified. Right, optical micrographs of the 4 TM7 cells which were isolated in this survey**.**

Figure 3. Phylogeny mapping of the genes in the TM7a assembly using IMG/M shows that the majority of TM7a genes are unlike those of any previously sequenced organism. Column D indicates the superkingdom: A=Archaea, B=Bacteria, E=Eukarya, V=Virus. "No. of Genomes" is the number of genomes available for comparison in each phylum. "No. of hits 30%" is the number of TM7a genes with at least 30% sequence identity to a member of the indicated phylum. "Histogram 30%" is a histogram representing the relative proportion of TM7a genes with at least 30% identity to genes in each phylum. "No. of hits 60%" and "Histogram 60%" represent the same analysis, but based on genes with at least 60% sequence identity.

Figure 4. The TM7b assembly has much higher sequence similarity to the TM7a assembly than to *Fusobacterium nucleatum* or *Chloroflexi aurantiacus.* Mapping was performed using whole genome Vista (Ref); the horizontal axis reflects the position along the TM7b assembly; the vertical axis reflects the similarity of 100 bp fragments from the

test genome to TM7b. Pink shading indicates >70% sequence identity. Top: *Chloroflexi aurantiacus*, Middle: *Fusobacterium nucleatum*, Bottom: TM7a.

Table 1. Statistics characterizing TM7a assembly and annotation, derived from IMG/M.

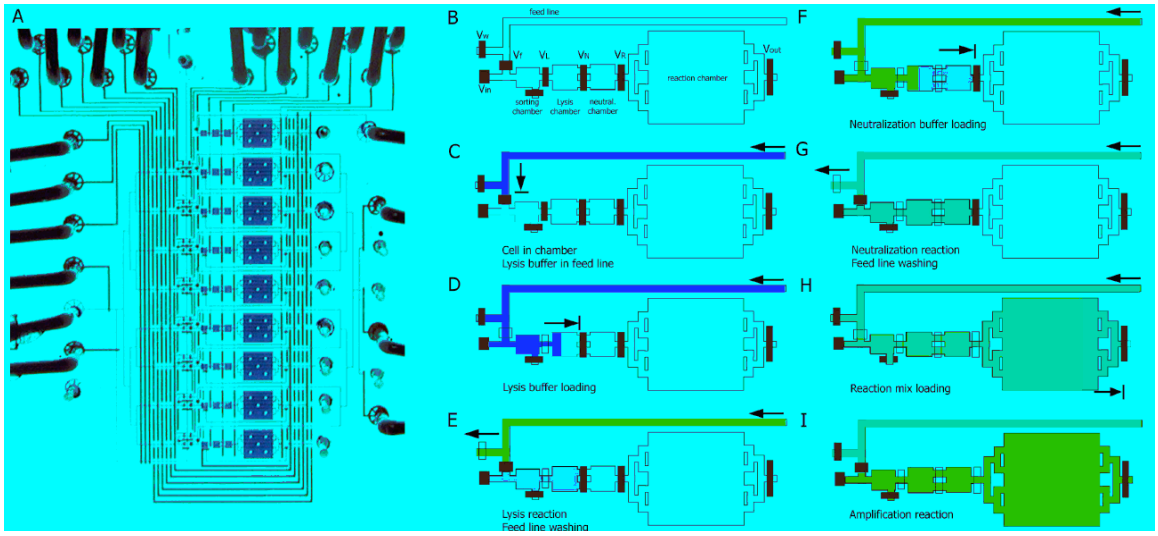

Figure 1, Marcy et al

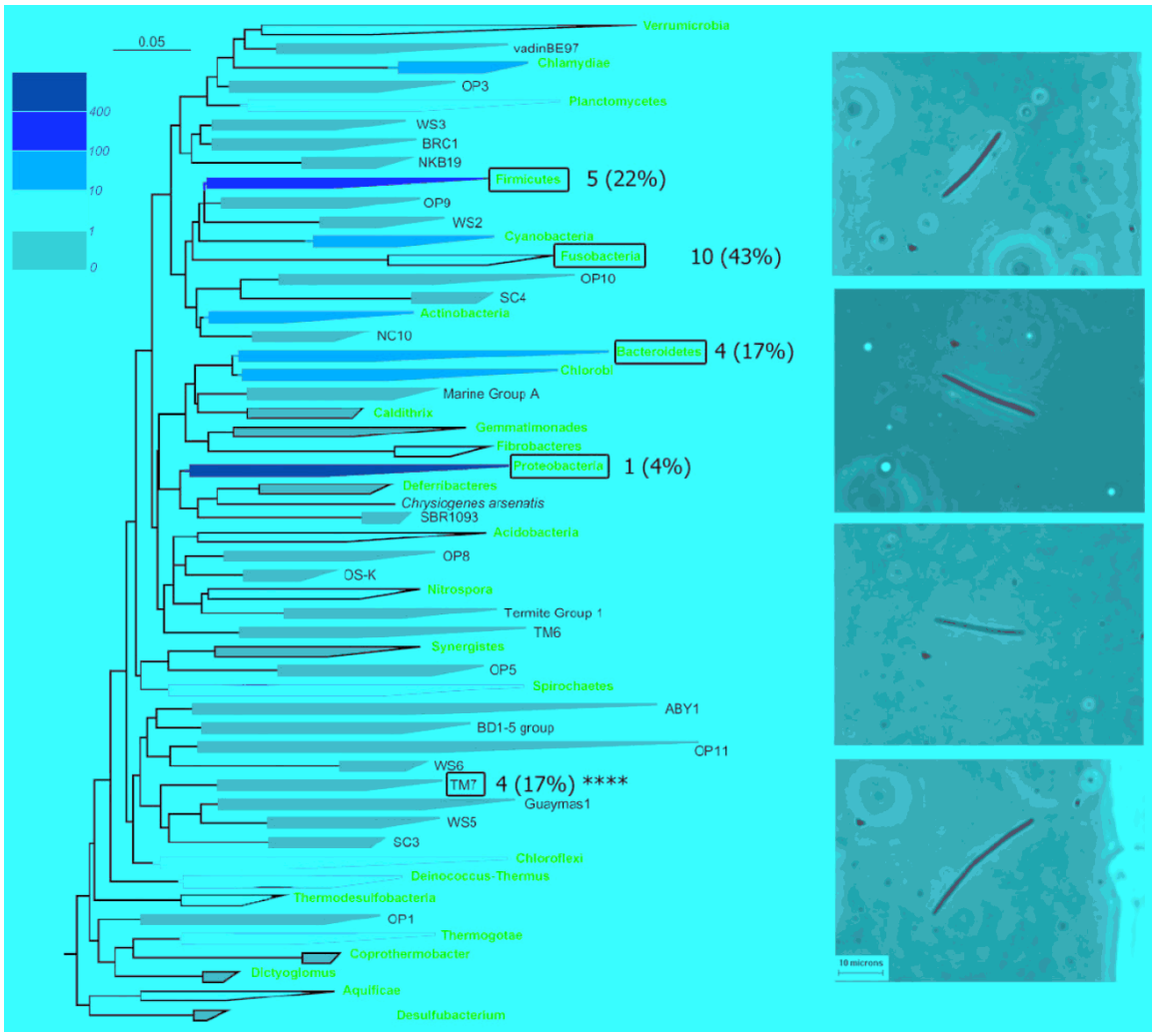

Figure 2, Marcy et al

D	Phylum/Class	No. Of Genomes	No. Of Hits 30%	Histogram 30%	No. Of Hits 60%	Histogram 60%
A	Crenarchaeota	6	2		-	
A	Euryarchaeota	4	39		13	
B	Acidobacteria	1	10		2	
B	Actinobacteria	1	106		30	
B	Aquificae	1	5		1	
B	Bacteroidetes	2	55		16	
B	Chlamydiae	11	5		-	
B	Chlorobi	10	10		4	
B	Chloroflexi	3	31		6	
B	Cyanobacteria	15	33		8	
B	Deinococcus-Thermus	4	7		-	
B	Bacilli	89	356		157	
B	Clostridia	22	395		141	
B	Mollicutes	17	11		1	
B	Fusobacteria	1	491		206	
B	Planctomycetes	2	1		-	
B	Alphaproteobacteria	87	35		7	
B	Betaproteobacteria	54	32		9	
B	Deltaproteobacteria	17	66		15	
B	Epsilonproteobacteria	21	76		48	
B	Gammaproteobacteria	159	135		37	
B	Magnetococcus	1	1		-	
B	Spirochaetes	9	36		11	
B	Thermotogae	1	8		2	
E	Alveolata	1	2		-	
E	Fungi	1	9		1	
V	Retro-transcribing viruses	54	6		6	
V	dsDNA viruses, no RNA stage	1	5		-	
-	Unassigned	-	1192		2439	

Figure 3, Marcy et al

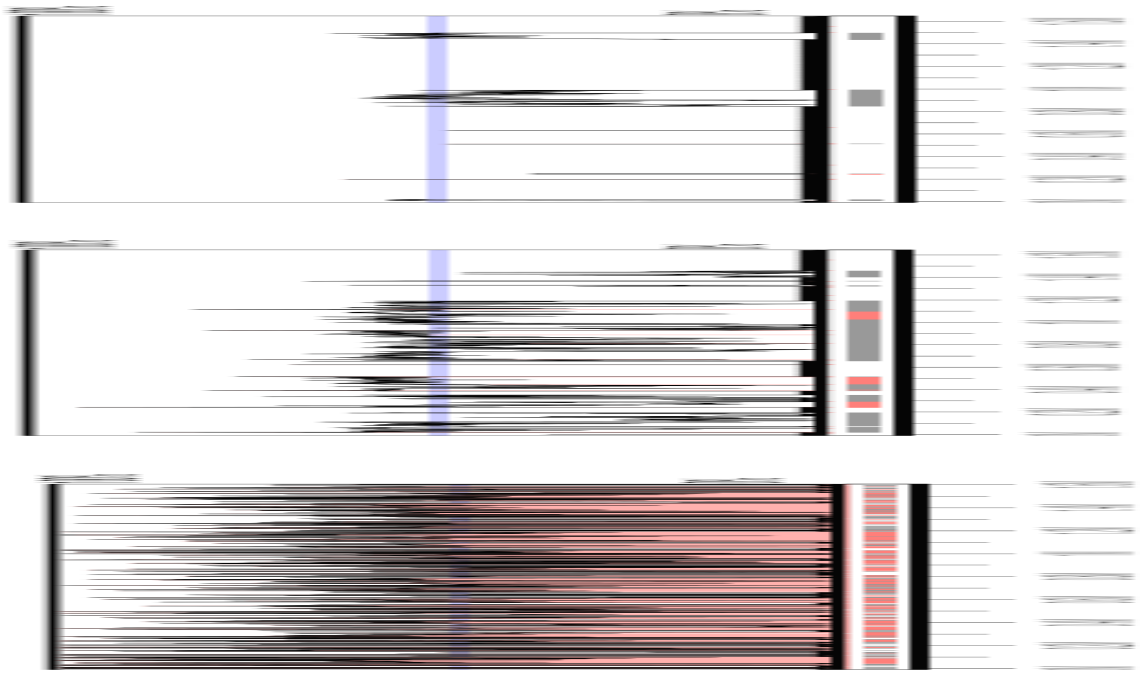

Figure 4, Marcy et al

	Number	% of Total
DNA, total number of bases	2,864,887	100.0%
DNA coding number of bases	1,160,954	40.5%
DNA G+C number of bases	981,862	34.3%
DNA scaffolds	1,825	100.0%
Genes total number	3,245	100.0%
Protein coding genes	3,160	97.4%
Genes with function prediction	1,389	42.8%
Genes without function prediction	1,771	54.6%
Genes assigned to enzymes	530	16.3%
Genes connected to KEGG pathways	400	12.3%
Genes not connected to KEGG pathways	2,760	85.1%
Genes in COGs	1,422	43.8%
Genes in Pfam	1,221	37.6%

Table 1, Marcy et al