Section 2

Microbial-Community Sequencing

14

Ribosomal Database Project II: Sequences and Tools for High-Throughput rRNA Analysis

J.R. Cole* (colej@msu.edu), B. Chai, R.J. Farris, Q. Wang, S.A. Kulam, D.M. McGarrell, G.M. Garrity, and **J.M. Tiedje**

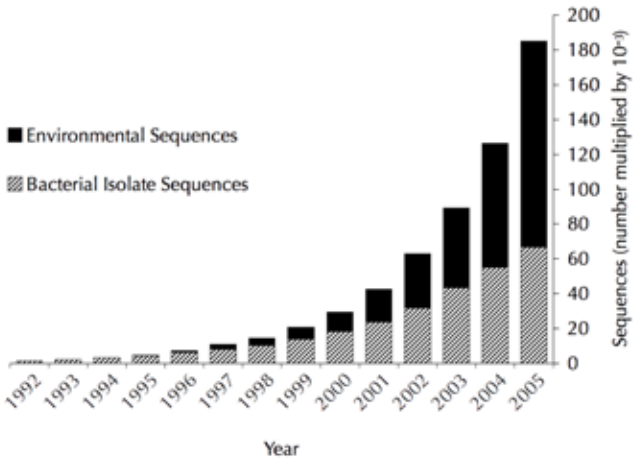
Michigan State University, East Lansing, MI

The Ribosomal Database Project II (RDP) provides data, tools, and services related to ribosomal RNA sequences to the research community (Cole et al., 2005. Nucleic Acids Res 33:D294). Through its website (http://rdp.cme.msu.edu), RDP-II offers aligned and annotated rRNA sequence data and analysis services. These data and services help discovery and characterization of microbes important to energy, biogeochemical cycles and bioremediation.

The RDP-II databases are updated monthly with data obtained from the International Nucleotide Sequence Databases (GenBank/EMBL/DDBJ). As of October 2005 (Release 9.31), RDP-II maintains 184,990 aligned and annotated bacterial small-subunit rRNA gene sequences (Figure 1). These sequences are available in several subsets, including higher quality near-full-length sequences (72,540), sequences from environmental samples (118,450), from in-culture bacterial isolates (66,540), and sequences from bacterial species type strains (4,445). The latter are of special interest, because species type-strains serve as archetype and link rRNA-base phylogeny with bacterial taxonomy. These type-strain sequences cover about 60% of the validly named bacterial species.

High-throughput environmental rRNA projects routinely produce hundreds to thousands of sequences per sample. The RDP-II tools have been redesigned to accommodate this trend towards large-scale rRNA sequencing efforts. Among the tools offered in RDP Release 9, Hierarchy Browser allows users to rapidly navigate through the RDP sequence data, RDP Classifier provides a rapid taxonomic placement of one to hundreds of user sequences, Sequence Match (completely re-implemented for Release 9) is more accurate than BLAST at rapidly finding closely related rRNA sequences, and the new version of Probe Match finds probe and primer binding sites using a more efficient algorithm and enables users to skip partial

Figure 1. Increase in number of publicly available bacterial small-subunit rRNA sequences.



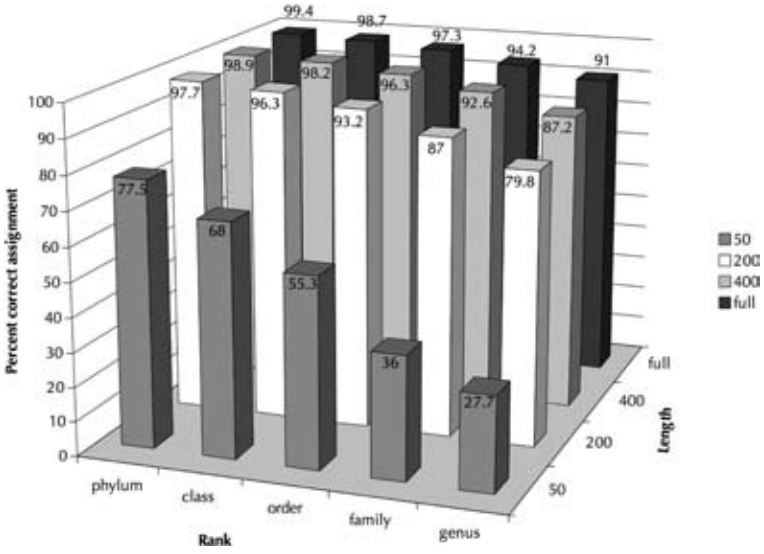
sequences missing the target region. The newly developed tool, Library Compare, combines the RDP Classifier with a statistical test to flag taxa differing significantly between 16S rRNA gene libraries.

The RDP Classifier was trained using information for approximately 5,486 bacterial species type strains (and a small number of other sequences representing regions of bacterial diversity with few named organisms) in 896 genera. One recent study found up to 5% of rRNA sequences examined contained sequencing artifacts when examined using a novel method (Ashelford et al., 2005. Appl Environ Microbiol 71:7724). This study included most of our training set sequences but found potential artifacts in only three, underscoring the importance of using well-characterized sequences for rRNA comparisons.

For the near full-length and 400 base partial rRNA sequences, the RDP Classifier was accurate down to the genus level (Figure 2), while even with 200 base partial sequences the RDP Classifier was accurate down to the family levels. The RDP Classifier did not perform well for partial sequences of length 50, likely due to insufficient features provided by such short partial sequences.

Over 60% of the full-length sequences misclassified at the genus level were more similar to one or more sequences derived from members of other genera than to other sequences within the same genus. Although some of these may be due to mistakes in sequence provenance, it seems likely that many of these "misclassifications" represent divergence in the nomenclatural taxonomy from the underlying (rRNA based) phylogeny. We are collaborating with the Taxomatic Project (Lilburn & Garrity 2004. Int J Syst Evol Microbiol 54:7) to help correct these discrepancies.

Figure 2. RDP Classifier accuracy by query size. (Exhaustive leave-one-out testing.)



* Presenting author

15

Community Genomics as the Foundation for Functional Analyses of Natural Microbial Consortia

R.J. Ram[1], V.J. Denef[1], I. Lo[1], N.C. VerBerkmoes[2], G. Tyson[1], G. DiBartolo[1], E.A. Allen[1], J. Eppley[1], B.J. Baker[1], M. Shah[2], R.L. Hettich[2], M.P. Thelen[3], and **J.F. Banfield**[1]* (jill@seismo.berkeley.edu)

[1]University of California, Berkeley, CA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; and [3]Lawrence Livermore National Laboratory, Livermore, CA

The objective central to our DOE Genomics:GTL project is develop methods for the study of microbial communities in their natural environments. In order to circumvent barriers associated with cultivation and limitations that arise from studies of isolated organisms, we are developing methods for genomic and functional characterization of microbial communities *in situ*. Genomic data provide the information needed to monitor activity as organisms assemble to form consortia and respond to perturbations in their environments. As our approach relies upon the availability of relatively comprehensive genomic information for the dominant community members, our efforts focus on chemoautotrophic biofilms with low species richness that grow in the subsurface in association with a dissolving metal sulfide deposit. The self-sustaining microbial communities that populate acid mine drainage systems are particularly amenable to high-resolution ecological analyses. In addition to their utility as model systems, these communities are important targets for study because microbially promoted dissolution of metal sulfides causes acid mine drainage, a major environmental problem associated with energy resources. It is also a process that underpins bioleaching-based metal recovery and coal desulfurization and mercury removal. Samples are collected from a range of locations within the Richmond mine at Iron Mountain, Redding, CA.

To date, extensive community genomic data have been obtained from an air-solution interface biofilm and a subaerial biofilm. Although both biofilms are dominated by *Leptospirillum* group II species that appear clonal at most loci, detailed analyses reveal significant heterogeneity in gene content in certain genomic hot spots. Variability in gene content occurs both within and between populations. Furthermore, some large genomic blocks exhibit anomalously high sequence identity, suggesting very recent lateral transfer of genomic regions between the populations. Similar observations have been made for the archaeal populations.

In the first study of its type, genomic data from one biofilm were used to characterize the proteome of a similar biofilm using mass spectrometry-based proteomics. Over 2,000 proteins were confidently identified, allowing initial insights into partitioning of function and the environmental challenges faced by the microbial community (Ram et al. 2005). As it is impractical to acquire extensive genomic data from every site sampled, we have used this and subsequent datasets to evaluate the influence of genome sequence dissimilarity on the efficiency of peptide and protein detection. Specifically, we are testing the likelihood of cross detection of proteins from closely related microbial species and evaluating the limitations imposed by sequence divergence for detection of proteins from closely related strains.

We developed an end member random substitution model that considers the average amino acid dissimilarity, the average peptide length, the fraction of uniquely detectable peptides, and predicted protein length. The model predicts a preferential loss of shorter proteins and a sigmoid relationship between the fraction of proteins detected and the average amino acid dissimilarity. Because amino acid substitutions do not occur randomly, our analyses also make use of genomic data from closely

related strains and species. To date, results suggest high protein detection rates using genomic data for proteome analysis of closely related strains, significant discrimination of proteins from different species, and highlight the importance of identifying unique peptides for distinguishing proteins from strain variants and closely related species.

For abundant proteins with regions of no peptide coverage, we PCR amplified the corresponding gene directly from the environmental sample and determined that there were very few cases where amino acid changes caused a peptide to be undetected, even in cases where abundant proteins had very divergent coding regions at the nucleotide level (up to 7%). This result is generally consistent with predictions from modeling.

For the environmental *Leptospirillum* group II populations and a new isolate, gene variants often have >20 nucleotide changes but few amino acid substitutions. Environmental samples tend to be dominated by populations with a single sequence type for most genes. Variants at each locus segregated independently among different populations, suggesting that there may be a few ancestral strains that have recombined to produce the extant populations. Although the bacterial groups appear to be shaped by recombination, DNA transfer among closely related organisms appears less common than in coexisting archaea, where we have determined that homologous recombination represents a major force in population dynamics.

Research sponsored by the Genomics:GTL program, Office of Biological and Environmental Research, U.S. Department of Energy.

Reference

1. Ram, R.J., VerBerkmoes, N.C., Thelen, M. P., Tyson, G.W., Baker, B.J. Blake, R.C. II, Shah, M., Hettich, R.L. and Banfield, J.F. (2005) "Community proteomics of a natural microbial biofilm," *Science*, 308, 1915-1920.

16

Environmental Whole-Genome Amplification to Access Microbial Diversity in Contaminated Sediments

Denise L. Wyborski[1,3], Carl B. Abulencia[1,3], Joseph A. Garcia[1], Mircea Podar[1], Wenqiong Chen[1,3], Sherman H. Chang[1], Hwai W. Chang[1], Terry C. Hazen[2,3], and Martin Keller[1,3]* (mkeller@diversa.com)

[1]Diversa Corporation, San Diego, CA; [2]Lawrence Berkeley National Laboratory, Berkeley, CA; and [3]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Low biomass samples from nitrate and heavy metal contaminated soils yield DNA amounts which have limited use for direct, native analysis and screening. Multiple displacement amplification (MDA) using ϕ 29 DNA polymerase was used to amplify whole genomes from environmental, contaminated, subsurface sediments. By first amplifying the gDNA, biodiversity analysis and genomic DNA library construction of microbes found in contaminated soils were made possible. We extracted DNA from samples with extremely low cell densities from a Department of Energy contaminated site. After amplification, SSU rRNA analysis revealed relatively even distribution of species across several major phyla. Clone libraries were constructed from the amplified gDNA and a small subset of clones was used for shot gun sequencing. BLAST analysis of the library clone sequences, and COG analysis, showed that the libraries were diverse and the majority of sequences had sequence identity to known proteins. The libraries were screened by DNA hybridization and sequence analysis for

* Presenting author

native histidine kinase genes. 37 clones were discovered that contained partial histidine kinase genes, and also partial, associated response regulators and flanking genes. Whole genome amplification of metagenomic DNA from very minute microbial sources enables access to genomic information that was not previously accessible.

17

Domestication of Uncultivated Microorganisms from Soil Samples

Annette Bollmann[1]* (a.bollmann@neu.edu), Lisa Ann Fagan[2], Anthony Palumbo[2], **Kim Lewis**[1] and Slava Epstein[1]

[1]Northeastern University, Boston, MA and [2]Oak Ridge National Laboratory, Oak Ridge, TN

The majority of microorganisms from natural environments cannot be grown in the lab by standard cultivation techniques. The diffusion chamber-based approach (Kaeberlein et al. 2002, Science 296:1127) is an alternative method to grow microorganisms in a simulated natural environment. We observed (Nichols et al., unpubl.) that continuous cultivation in diffusion chambers of species initially incapable of growing in Petri dish eventually leads to production of their cultivable variants, and thus to their domestication. Here we use this approach to cultivate and domesticate microorganisms from subsurface soil samples from the FRC site (borehole FW111, area3). The microorganisms were extracted from the soil samples and incubated in diffusion chambers. The incubations took place in containers filled with wetted soil. After several weeks of incubation, the content of diffusion chambers was inoculated into a new generation of diffusion chambers, to the total of three such generations. In parallel, the chamber-grown material was inoculated into Petri dishes to monitor the process of domestication and subsequent isolation of growing microorganisms.

Phylogenetic analysis based on 16SrRNA sequences revealed clear differences between species diversity in Petri dishes inoculated with soil material as compared to Petri dishes inoculated with material from the diffusion chambers. The isolates obtained by standard approaches belong to *Alpha-* and *Gamma-Proteobacteria, Actinobacteria* and *Verrucomicrobia*. Petri dishes inoculated with diffusion chambers-grown material contained bacteria from the phyla, *Alpha-*, *Beta-*, and *Gamma-Proteobacteria, Actinobacteria, Firmicutes,* and CFBs. Interestingly, several isolates could only be obtained from material passaged through multiple diffusion chambers, and were never observed in the earlier generations of the chambers or in Petri dishes inoculated directly by environmental samples. Several of the isolates are close related to bacteria found in different experiments with molecular methods (North et al. 2004, AEM 70:4911; Reardon et al. 2004, AEM70:6037; Fields et al. 2005, FEMS Microb. Ecol 53:417).

In conclusion passaging environmental microorganisms through the diffusion chamber leads to the domestication of many previously uncultivated microorganisms, which enables their isolation in pure culture. The diffusion chamber-based domestication does select for specific microorganisms, the associated biases are different from those of traditional culture techniques. This provides access to cultures of at least some microorganisms previously known only by their molecular signatures.

18

Metagenomic Analysis of Microbial Communities in Uranium-Contaminated Groundwaters

Jizhong Zhou[1,6,7]* (jzhou@ou.edu), Terry Gentry[1], Chris Hemme[1,6,7], Liyou Wu[1], Matthew W. Fields[2,7], Chris Detter[3], Kerrie Barry[3], David Watson[1], Christopher W. Schadt[1], Paul Richardson[3], James Bristow[3], Terry C. Hazen[4,7], James Tiedje[5], and Eddy Rubin[3]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Miami University, Oxford, OH; [3]DOE Joint Genome Institute, Walnut Creek, CA; [4]Lawrence Berkeley National Laboratory, Berkeley, CA; [5]Michigan State University, East Lansing, MI; [6]University of Oklahoma, Norman, OK; and [7]Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

Due to the uncultivated status of the majority of microorganisms in nature, little is known about their genetic properties, biochemical functions, and metabolic characteristics. Although sequence determination of the microbial community 'genome' is now possible with high throughput sequencing technology, the complexity and magnitude of most microbial communities make meaningful data acquisition and interpretation difficult. Therefore, we are sequencing groundwater microbial communities with manageable diversity and complexity (~10-400 phylotypes) at the U.S. Department of Energy's Natural and Accelerated Bioremediation Research (NABIR)-Field Research Center (FRC), Oak Ridge, TN. The microbial community has been sequenced from a groundwater sample contaminated with very high levels of nitrate, uranium and other heavy metals and pH ~3.7. Sequence analysis of this groundwater sample based on a 16S rDNA library revealed 10 operational taxonomic units (OTUs) at the 99.6% cutoff with >90% of the OTUs represented by an unidentified γ-proteobacterial species similar to *Frateuria*. Additional OTUs were related to a β-proteobacterial species of the genus *Azoarcus*. Three clone libraries with different DNA fragment sizes (3, 8 and 40 kb) were constructed, and 50-60 Mb raw sequences were obtained using a shotgun sequencing approach. The raw sequences were assembled into 2770 contigs totaling ~6 Mb which were further assembled into 224 scaffolds (1.8 kb-2.4 Mb). Preliminary binning of the scaffolds suggests 4 primary groupings (2 *Frateuria*-like γ-proteobacteria, 1 *Burkholderia*-like β-proteobacteria and 1 *Herbaspirillum*-like β-proteobacteria). Genes identified from the sequences were consistent with the geochemistry of the site, including multiple nitrate reductase and metal resistance genes. Despite the low species diversity of the samples, evidence of strain diversity within the identified species was observed. Analysis with functional gene arrays containing ~23,000 probes designed based on these community sequences as well as genes important for biogeochemical cycling of C, N, and S, along with metal resistance and contaminant degradation suggested that the dominant species could be biostimulated during *in situ* uranium reduction experiments at the FRC. These results also suggest that the dominant species could play a direct or indirect role in the bioremediation of uranium.

* Presenting author

19

Understanding Phage-Host Interactions Using Synergistic Metagenomic Approaches

Shannon J. Williamson* (shannon.williamson@venterinstitute.org), Douglas B. Rusch, Shibu Yooseph, Aaron Halpern, Karla Heidelberg, Cynthia Pfannkoch, Karin Remington, Robert Friedman, Marvin Frazier, Robert Strausberg, and J. Craig Venter

J. Craig Venter Institute, Rockville, MD

Marine bacteriophages (viruses that infect bacteria) are the most abundant biological entities on our planet. Interactions between phages and their hosts impact several important biological processes in the world's oceans from horizontal gene transfer to the cycling of essential nutrients. Interrogation of microbial metagenomic data collected as part of the Sorcerer II Expedition (http://www.sorcerer2expedition.org) has revealed an unexpectedly high abundance of phage sequences. Analysis of these data resulted in observations that cast new light on the nature of environmental phage-host interactions. For example, we found that the site-site distribution of phage closely related to the cyanomyophage P-SSM4 is almost identical to that of the dominant population of *Prochlorococcus* present in our samples. Studies have shown that P-SSM4 can infect divergent strains of *Prochlorococcus* (Sullivan et al. 2005), yet the co-occurrence of the P-SSM4-like phage and its host over a wide geographic area suggests a steady-state infection of the dominant *Prochlorococcus* population. It is interesting to note that the cyanomyophage P-SSP7, which is highly specific for one high light-adapted strain of *Prochlorococcus*, is largely absent from our data, suggesting the scarcity of its host at the times of sampling. As phage infection can greatly influence the clonal composition of host cell communities, our observation suggests that this particular phage may control the abundance and distribution of perhaps one of the most dominant components of picophytoplankton in oligotrophic oceans. Furthermore, environmental factors appear to influence the occurrence of temperate versus lytic phage in a hypersaline environment. Tailed phage that are members of the family *Siphoviridae* are often characterized as temperate and therefore have the ability to establish silent infections, otherwise known as lysogeny, with their hosts. By far, the greatest proportion of siphophage sequences in our dataset originated from a terrestrial hypersaline pond on Floreana Island in the Galapagos. Assembly of these sequences from the hypersaline pond resulted in the recovery of a complete phage genome approximately 50kb in length. The presence of an integrase gene confirms that this phage is indeed temperate and its genomic architecture appears to be conserved with respect to other temperate siphophage genomes. The predominance of this phage genome to the exclusion of others implies that propagation of phage through the lysogenic pathway of infection is favored over lytic replication in response to productive challenges stemming from environmental pressures. Finally, cluster analysis of viral peptide sequences revealed the presence of seven viral clusters, each containing hundreds of host-derived proteins of varying metabolic function, including certain proteins involved in the processes of photosynthesis and photoadaptation, phosphate stress and acquisition, and carbon metabolism. Distribution of the proteins within these seven clusters is geographically diverse, suggesting that viral acquisition of host metabolic genes is a more abundant and widespread phenomenon than previously recognized. In summary, our study indicates that metagenomic analysis of coincident viral and microbial sequence data provides a unique opportunity to explore phage-host interactions on a global scale.

20

The Impact of Horizontal Gene Transfer: Uniting or Dividing Microbial Diversity?

Rachel J. Whitaker[1,2] (rwhitaker@nature.berkeley.edu), Dennis Grogan[3], Mark Young[4], and Frank Robb[5]

[1]University of California, Berkeley, CA; [2]University of Illinois, Urbana, IL; [3]University of Cincinnati, Cincinnati, OH; [4]Montana State University, Bozeman, MT; and [5]University of Maryland, College Park, MD

Comparative genomics has provided evidence for horizontal gene transfer (HGT) events occurring from the twigs all the way down the trunk of the tree of life. The adaptive significance of horizontal transfer events to the trajectory of microbial evolution depends upon both the frequency at which genes are exchanged and the action of natural selection upon transferred genes. For example, if horizontal transfer of homologous genes among closely related individuals within a population occurs more frequently than natural selection, gene transfer provides a cohesive force within a lineage, increasing the level of neutral diversity within a population by allowing natural selection to act at the level of the gene rather than the genome. If, on the other hand, horizontal transfer of homologous sequence is a rare, periodic selection events will purge neutral diversity and drive diversification of independent clonal 'ecotypes'. Multilocus sequence analysis and population genomics reveal that the frequent transfer of homologous sequences has a major impact on speciation and adaptive evolution in several Bacterial and Archaeal species. Recently, community and comparative genomics have revealed large differences in gene content between very closely related individuals, suggesting that the rapid movement of non-homologous sequences in and out of microbial chromosomes also drives microbial evolution. Here again, the rate of movement of novel non-homologous sequences and the selective regime acting upon a natural population will determine whether these genes are transient, neutral byproducts of mobile extra chromosomal elements or confer adaptive advantage. Ultimately, determining the impact of gene transfer on microbial evolution and the adaptive consequences of horizontally transferred genes will depend upon placing these genomes into well-defined environmental context.

Highly structured biogeographic populations provide a unique biological framework in which to place genome dynamics in geological and evolutionary context. For example, high-resolution multilocus sequence analysis of 130 *Sulfolobus islandicus* strains cultured from geothermal regions in North America, the Kamchatka peninsula, and Iceland revealed that *S. islandicus* is the dominant cultivable *Sulfolobus* species in the Northern Hemisphere and that at least five endemic populations of *Sulfolobus* are isolated from one another by geographic distance. In this highly structured system, evolutionary events are likely to have occurred in each of five geothermal communities independently, and may be correlated with the unique geologic history of their individual locale. Genome sequencing of strains isolated from these endemic populations will allow quantification of rates of horizontal gene transfer relative to nucleotide divergence between isolated populations, and calibration of rates of HGT with the geologic history of isolated geothermal sites. In addition, this novel system may allow the correlation of differences in gene content among closely related individuals to the selective regime of each unique geothermal environment.

* Presenting author