Environmental Genomics

65

Whole Community Proteomics Study of an Acid Mine Drainage Biofilm Reveals Key Roles for "Hypothetical" Proteins in a Natural Microbial Biofilm

Jill Banfield*[1] (jill@eps.berkeley.edu), Rachna J. Ram[1], Gene W. Tyson[1], Eric Allen[1], Nathan VerBerkmoes[2,3], Michael P. Thelen[1], Brett J. Baker[1], Manesh Shah[3], Robert Hettich[3], and Robert C. Blake II[4]

[1]University of California, Berkeley CA; [2]University of Tennessee-Oak Ridge National Laboratory, Oak Ridge, TN; [3]Oak Ridge National Laboratory, Oak Ridge, TN; and [4]Xavier University of Louisiana, New Orleans, LA

We are studying relatively simple, low diversity microbial communities associated with extremely acidic, metal-rich mine drainage system to develop an understanding of adaptation, evolution, and the linkages between microbial activity and environmental geochemistry. In order to assess the genetic potential of an entire natural biofilm sample we partially reconstructed the genomes of five dominant organisms (Tyson et al. 2004). Comparative genomic analyses have revealed the structure of each population and provided insights into the processes that create and remove genome heterogeneity. The genomes contain large blocks dominated by genes that encode hypothetical proteins. These are regions inserted in one species or strain relative to others and are inferred to be of phage origin. Phage insertion appears to be the most rapid process leading to strain heterogeneity, and possibly diversification. Within strain populations, gene order is largely retained, and insertions or loss of single genes are rare events. Bacterial populations are dominated by a single clonal type. Homologous recombination is a key force shaping the genomes of archaeal populations in the system, but is rare between different species.

We have characterized the protein complement of a natural microbial community similar to that studied genomically in order to determine which genes are expressed and functionally important. By combining mass spectrometry-based "shotgun" proteomics with community genomics we confidently identified at least 1700 proteins from five dominant species. Proteins involved in protein refolding and response to oxidative stress were abundant, indicating that damage to biomolecules is a key challenge for survival. We validated more than 400 hypothetical proteins, a small subset of which are encoded within blocks of genes apparently acquired by lateral transfer. Entire operons encoding expressed, novel, lineage-specific proteins may be important for acid, metal and radical tolerance. 26% of the detected *Leptospirillum* group II proteins were hypothetical. An extracellular fraction was dominated by a novel protein shown to be a cytochrome central to iron oxidation and AMD formation. Sequencing of DNA encoding cytochrome regions for which peptides were not recovered revealed two amino acid substitutions. Using the strain variant sequence, 100% peptide coverage of the mature protein was achieved. Thus, an iterative genomic and proteomic approach analyses enable detailed in situ analyses of activity within natural microbial consortia.

66

Application of High Throughput Microcapsule Culturing to Develop a Novel Genomics Technology Platform

Martin Keller[1]* (mkeller@diversa.com), Karsten Zengler[1], Marion Walcher[1], Carl Abulencia[1], Denise Wyborski[1], Sherman Chang[1], Imke Haller[1], Trevin Holland[1], Fred Brockman[2], Cheryl Kuske[3], and Susan Barns[3]

[1]Diversa Corporation, San Diego, CA; [2]Pacific Northwest National Laboratory, Richland, WA; and [3]Los Alamos National Laboratory, Los Alamos, NM

Project Description

The overall goal of this project is to demonstrate the combination of high-throughput cultivation in microcapsules, which gives access to previously uncultivated microorganisms, with genome sequencing from one to a few microcolony-containing microcapsules. This will allow direct access to physiological and genomic information from uncultured and/or difficult-to-culture microorganisms. This approach is fundamentally different than characterization and/or assembly of shotgun or BAC clones derived from community DNA or RNA. The units of analysis in our approach are living, pure microbial cultures in microcapsules, as opposed to the disassembled mixture of small fragments of genomes and cellular networks that have lost their biological context in studies using community nucleic acids. It is envisioned that the microcapsule based, high-throughput cultivation method will also be combined with Transcriptomics and Proteomics technology in the future.

Project achievements

A high-throughput cultivation method based on single cell encapsulation in microcapsules in combination with flow cytometry has been applied to two soil samples. It has been demonstrated that microorganisms belonging to a variety of bacterial phyla, such as *Acidobacteria*, *Gemmatimonadetes* and candidate division TM7, were growing in the microcapsules and subsequently formed microcolonies.

A fluorescent in situ hybridization (FISH) method has been optimized to selectively target and sort encapsulated microcolonies of interest using fluorescence activated cell sorting.

A whole-genome amplification technique has been employed to acquire a sufficient mass of DNA from targeted, encapsulated microcolonies (*E. coli*) to generate libraries for shotgun sequencing of entire genomes. The whole-genome amplification has been optimized so that DNA from as few as two cells can be amplified routinely. The genome coverage of amplified FISH-targeted microcolonies was evaluated using microarrays. Microarray data demonstrated a genome-coverage of 97% to 99% without high levels of bias towards certain genes. Subsequently, genomic libraries have been constructed from these amplified DNA samples. Almost 300 clones of each library have been sequenced. Sequence analysis confirmed that 70% of the clones originated from *E. coli* DNA. Additional gene sequences were affiliated with certain beta-*Proteobacteria* typically found as experimental contaminants.

This work was supported by the Office of Science (BER), U.S. Department of Energy, Grant No. DE-FG02-04ER63771.

* Presenting author

67

Environmental Bacterial Diversity from Communities to Genomes

Janelle R. Thompson[1,2]*, Silvia G. Acinas[1], Vanja Klepac-Ceraj[1,2], Sarah Pacocha[1,2], Chanathip Pharino[1], Dana E. Hunt[1], Luisa A. Marcelino[1], Jennifer Benoit[1,2], Ramahi Sarma-Rupavtarm[1], Daniel L. Distel[3], and Martin F. Polz[1] (mpolz@mit.edu)

[1]Massachusetts Institute of Technology, Cambridge, MA; [2]Woods Hole Oceanographic Institution, Woods Hole, MA; and [3]New England Biolabs, Beverly, MA

We are studying the patterns of diversity among co-occurring coastal bacterioplankton from the level of the entire community to the individual genome. Our goal is to advance the understanding of structure-function relationships in microbial assemblages addressing questions including: What is the range of genomic diversity encompassed by functionally similar populations in specific environmental contexts? What mechanisms govern selection and diversification of natural microbial populations? Using environmental 16S ribosomal RNA gene sequences (ribotypes) as a proxy for bacteria, we have shown that despite a high diversity in the environment, the majority of organisms fall into closely related clusters (<1% 16S rRNA divergence) (1). Such microdiverse sequence clusters are hypothesized to represent functionally-differentiated populations, which arise by selective sweeps (2) and persist because competitive mechanisms are too weak to purge diversity from within them (1).

To examine this hypothesis we quantitatively estimated the genomic diversity within one 16S rRNA microdiversity cluster (*Vibrio splendidus)*. Quantitative PCR analysis (3) over an annual cycle indicated that *V. splendidus* was consistently present as a member of the coastal bacterioplankton community. Vibrio strains were isolated from representative months and the majority were identified as *V. splendidus.* Determination of sequence diversity of a universally distributed protein-coding gene (Hsp60) among all *Vibrio* isolates showed high heterogeneity but confirmed the monophyly of the *V. splendidus* strains. Still greater heterogeneity was revealed when the number of unique genotypes among strains was assayed by pulse field gel electrophoresis (PFGE), moreover, the PFGE analysis provided evidence that a large proportion of genotypes are differentiated by insertions and deletions of large genome fragments. In a set of 12 *V. splendidus* strains genome sizes ranged from 4.5 to 5.6 Mb with only weak correlation of genome size difference to Hsp60 sequence divergence (R = 0.37).

The high degree of heterogeneity among the *V. splendidus* genomes suggest that the average environmental concentration of individual genotypes is astoundingly small. To illustrate this, we divided the QPCR-based estimates of population size of *V. splendidus* in samples taken in Aug 03, Sept 03, and Oct 03 (1,890, 600, and 640 cells/ml, respectively) by the Chao-1 statistical estimates (4) for the number of Hsp60 alleles (125, 94 and 279) and genotypes (465, 553 and 901) in those same samples. The result suggests that unique Hsp60 alleles occurred in the monthly samples at average concentrations of 2 to 15 cells per ml (or at a frequency of 0.3 to 1%) while unique genotypes were present at ~10-fold lower frequency (average concentration for all samples estimated at <1 cell per ml).

The observed pattern of co-existing diversity suggests that purging of genotypes from the population is rare compared to processes introducing variation and that therefore variation persists because it is either favored (e.g., by balancing selection or resource specialization) or neutral. We present ecological considerations to suggest much of the observed diversity may in fact be neutral in an environmental context. Such observations of extreme genomic heterogeneity among closely related individuals have significant implications for the assembly of genome sequences from environmental samples. In addition, if similar patterns of diversity are common to other bacterial populations caution should be exercised in interpreting the extent to which gene complements or even metabolic traits of individual

isolates may reflect the overall properties of populations. Indeed our results suggest that not only the gene content, but also quantitative abundance and dynamics of individual traits should be considered when evaluating the ecological significance of differences among coexisting genotypes.

References:

1. S. G. Acinas *et al.*, *Nature* **430**, 551-554 (2004).
2. F. M. Cohan, *Annu. Rev. Microbiol.* **56**, 457–487 (2002)
3. J. R. Thompson *et al.*, *Appl. Environ. Microbiol.* **70**, 4103-4110 (2004).
4. J. B. Hughes, J. J. Hellmann, T. H. Ricketts, B. J. M. Bohannan, *Appl. Environ. Microbiol.* **67**, 4399-4406 (2001).

68

Distribution and Variation of *Prochlorococcus* Genotypes Across Multiple Oceanic Habitats

Adam C. Martiny* (martiny@mit.edu), P. K. Amos Tai, Anne W. Thompson, and Sallie W. Chisholm

Massachusetts Institute of Technology, Cambridge, MA

The cyanobacterium *Prochlorococcus* is very abundant in oligotrophic regions of the world's oceans, constituting up to 50% of the cells in the euphotic zone. Thirty-two cultures have been isolated and based on these, a phylogenetic tree has been constructed showing the presence of 6 clades, four low light adapted and two high light adapted. That opens up two questions: (i) To what extent do these cultures represent the phylogenetic space of *Prochlorococcus*? (ii) Are the patterns of genotypic diversity within a given clade similar in field samples collected from different geographical locations?

We are using the sequence of the intergenic transcribed spacer region between the small and large subunit rRNA (ITS) as a neutral marker for genetic variation to describe the diversity of *Prochlorococcus* in field samples. We have cloned and sequenced 1200 ITS fragments from the North Pacific subtropical gyre (Hawaii Ocean Time Series), Sargasso Sea (Bermuda-Atlantic Time Series) at three depths – 25m, 80m and 160m and an upwelling region off the coast of Mexico at two depths (60m and 130m). The phylogenetic analysis showed a high frequency of sequences belonging to the 9312-clade from the 25m samples and 80m — consistent with the finding that the 9312 'ecotype' numerically dominates the upper euphotic zone in many oceanic environments (Zinser et al, in prep.). A comparison using Mantel test of the microdiversity within the 9312 clade at 25m and 80m revealed that the populations were significantly different between these two depths. In addition, a significant amount of "yet to be cultured" diversity was discovered at both 80 and 160m depth including new lineages as well as sub-lineages within the 6 known clades. In a sample collected from a sub-oxic zone off the coast of Mexico, we found a group affiliated to the low light adapted 9313 ecotype, but forming an independent lineage not yet seen in other samples. A future goal is to target such new lineages, amplify their genome using the approach described by Zhang et al. (this meeting) and thereby expand our view on the *Prochlorococcus* physiology and evolution through comparative genome analysis..

* Presenting author

69

From Perturbation Analysis to the Genomic Regulatory Code: the Sea Urchin Endomesoderm GRN

Paola Oliveri*[1] (poliveri@caltech.edu), Pei-Yun Lee[1], Takuya Minokawa[2], Joel Smith[1], Qiang Tu[1], Meredith Howard[1], David McClay[3], and Eric H. Davidson[1]

[1] California Institute of Technology, Pasadena, CA; [2] Tohoku University, Asamushi, Aomori, Japan; [3]Duke University, Durham, NC

The sea urchin endomesoderm gene regulatory network (GRN) is the most comprehensively understood regulatory apparatus for control of spatial and temporal gene expression in any complex developmental system. It contains almost 50 genes, mainly encoding regulatory proteins. It was initially constructed on the basis of spatial and temporal gene expression data, interpreted through a large scale, systematic, perturbation analysis in which expression of each gene was taken out or otherwise altered, and the effects on all other relevant genes measured quantitatively and with high sensitivity. The GRN provides the essential, overall, "transformation function" by which can be solved the causal relations between the genomic regulatory code that is hardwired into the DNA sequence, and the observed events of spatial and temporal gene expression. That is, it specifies the key *cis*-regulatory inputs into regulatory genes, and their key outputs terminating at other regulatory genes. Hence it is directly testable at the *cis*-regulatory level. In the last year, we identified by computational and experimental methods, and isolated, over a dozen of the central *cis*-regulatory nodes of the GRN. We then require that in gene transfer experiments that these genomic fragments display the same responses to the appropriate perturbations as do the endogenous genes in the whole embryo; and that when the genomic target sites for the relevant inputs are mutated, that the *cis*-regulatory constructs behave in the expected ways. This system wide task will be completed this year, and to date the results indicate that the perturbation analysis indicated the real encoded linkages with perhaps surprising accuracy. In addition: we have demonstrated for the first time that knowledge of the GRN can be used to reengineer the process of development; we have found ways to identify the modular "kernels" of the GRN, which consist of multiple genes recursively "wired" to one another and which are evolutionarily resistant to change; we have developed a new theory of logic processing within *cis*-regulatory modules, as a start on formalization of the genomic regulatory code; we have enlarged our knowledge of the GRN and are adding into it all regulatory genes encoded in the genome that are expressed in the appropriate time and place, so that it will approximate a complete regulatory treatment; and we have begun to extend GRN analysis to later processes.

 * Presenting author