DOE/SC-0071

New Horizons in Genomics

U.S. Department of Energy Joint Genome Institute

Santa Fe, New Mexico

March 30-April 1, 2003

Prepared for the Joint Genome Institute
U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research

JGI Production Genomic Facility
2800 Mitchell Drive
Walnut Creek, CA 94598

Prepared by
Human Genome Management Information System
1060 Commerce Park
Oak Ridge National Laboratory
Oak Ridge, TN 37830
Managed by UT-Battelle, LLC
For the U.S. Department of Energy
Under contract DE-AC05-00OR22725

Contents

New Horizons in Genomics

The completion of the international Human Genome Project is one of the most significant milestones in the history of biology. It marks the beginning of an era that promises profound insights into the molecular functioning of all forms of life. Systems biology, the impact of organisms on each other and on the earth's environment, and fields of biological investigation yet to be identified surely will be influenced by the discoveries and technologies of genomics.

As significant as it is, however, the elucidation of its genomic sequence obviously is only one step along the path to understanding how an organism is built and its actions are governed. The genome, which has been described as a "parts list," helps us characterize the basic elements involved in nearly all biological processes. The next major task for genomics is to begin drafting an "operating manual" that will tell us how the parts work together in their development and functioning and how they interact with their environment. These insights will substantially further our understanding of systems biology and the mechanisms of evolution.

At this pivotal moment in the history of genomics, the DOE Joint Genome Institute is pleased to sponsor this meeting on "New Horizons in Genomics." The meeting's aim is to share information about the progress and promise of the various genome projects completed, under way, or planned; the Department of Energy's Microbial Genome and Genomes to Life programs; and visions of the ways in which genomics and its associated technologies may transform biological sciences in the 21st Century. The meeting will include invited speakers as well as talks from abstracts in the areas of the human genome, comparative and evolutionary genomics, microbial genomics, systems biology, functional genomics, and genomic technologies.

On behalf of the DOE Joint Genome Institute, we welcome you to Santa Fe and look forward to your participation in and contributions to the meeting.

DOE Joint Genome Institute

Sequencing Resources

1

Plasmidomics: Cloning Naturally Occurring Plasmids for Sequencing and Functional Analysis

Anne Marie Erler[1], Patricia Sobecky[2], Gary L. Andersen[1], and **Peter Agron**[1] (argon1@llnl.gov)

[1]Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA; and [2]School of Biology, Georgia Institute of Technology, Atlanta, GA

Plasmids are nonessential, stably maintained, extra-chromosomal DNA molecules commonly found in microbes. They imbue a great variety of traits on their hosts including resistance to toxic metals, increased ultraviolet radiation resistance, use of alternative food sources such as organic waste from human activity, virulence and antibiotic resistance. Therefore, nature has "pre-packaged" genes encoding traits of great interest, and the ability to take advantage of this cloning work would be of great value. A simple approach has been developed that potentially allows any circular plasmid to be established in *Escherichia coli*, thus facilitating sequencing and functional analysis. In vitro transposition was used to introduce a selectable marker as well as a plasmid replicon that is functional in *E. coli*. This way, circular plasmids that recombine with the donor molecule will replicate in the heterologous host when introduced by transformation regardless of the natural ability to do so. Three transposon donors, each tailored for different experimental approaches, were constucted. A small archaeal plasmid was established in *E. coli*, and testing with the recombinants provides strong evidence that this plasmid cannot normally replicate in the bacterial host. In addition, an approximately 60-kb uncharacterized plasmid from a South Carolina salt marsh bacterium was established in *E. coli* and was subsequently shown to have the natural ability to replicate in this host. Copy number amplification in *E. coli* easily produced quantities of this plasmid for direct DNA sequencing, which revealed high similarity to restriction-modification systems, suggesting one putative function for this plasmid in the environment.

2

Shotgun Library Utilization for Sequencing Projects at the JGI

Chris Detter (detter2@llnl.gov), **Eileen Dalin**, Jamie M. Jett, Doug Smith, Jenna Morgan, Hope Tice, Saima Shams, Corey Chinn, Eddy Rubin, and Paul M. Richardson

U.S. Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598 USA

The U.S. Department of Energy Joint Genome Institute has produced shotgun libraries for sequencing projects from sources including cosmids, fosmids, BACs and whole genomes including numerous microbes. Other large genomes that have been subcloned for sequencing include *Fugu rubripes*, *Ciona intestinalis*, *Phanerochaete chrysosporium*, *Thalassiosira pseudonana*, *Populus trichocarpa*, and *Xenopus tropicalis*. Standard subclones for these projects have been produced by random shearing of DNA followed by size selection (typically 3 kb inserts) and ligation into pUC18. Most of these libraries produce adequate assemblies with few uncaptured gaps. However, some problematic projects require more specialized libraries to span gaps and produce longer-range contiguity. A major need in the Production Sequencing process, especially for larger whole genome shotgun projects, is the use of medium insert

libraries that span repeat sequences. Having paired end information from clones that span repeats greatly improves sequence contiguity in whole genome assemblies. For these reasons, our R&D group focused on developing a robust and reproducible 8-10 kb library construction protocol that could be scaled for use in our high-throughput sequencing facility.

Here, we will describe the construction, sequencing, and analysis of medium insert (8-10 kb) libraries used for sequencing at the Joint Genome Institute. The libraries were constructed from randomly sheared BAC or whole genomic DNA that was size selected and cloned into a kanamycin resistant low-copy plasmid (pCUGI21blu). To date, more than 100 BACs, as well as several microbial and larger whole genomes, have been successfully subcloned using these methods. Individual subclones produced by this method can be sequenced without modification of the high-throughput production process. Analysis of initial sequencing results indicates the inserts are stable and have a narrow distribution around the expected size. These libraries have been useful for assembling BAC clones spanning highly repetitive human sequence. Specific protocols and results will be described.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, under Contracts No. W-7405-Eng-48, No. DE-AC03-76SF00098, and No. W-7405-ENG-36.

3

BAC Libraries for Whole-Genome Sequencing, Comparative Genomics and Haplotype Analysis

Pieter J. de Jong (pdejong@mail.cho.org), Baoli Zhu, Mikhail Nefedov, Chung-Li Shu, Yuko Yoshinaga, and Kazutoyo Osoegawa

Children's Hospital and Research Center Oakland (CHRCO)

Bacterial Artificial Chromosome Libraries are enjoying a continuing diversification of applications in sequencing, in functional genomics and diagnostics. BACs are used as the start for clone-based shotgun sequencing. BAC-ends provide scaffolding information for whole-genome sequence assembly. BACs also permit prioritized sequencing of chromosomal regions or gene families, hence being at the starting point for comparative genomics. Increasingly, BACs are used for gap filling in difficult regions, for haplotype analysis and for functional complementation. We will present an overview of the current strategies for BAC cloning and the growing number of species represented by BACs. About 70 BAC libraries have been prepared in our laboratory for an equivalent number of species and strains. To resolve difficult genomic regions, BACs are cloned with presumed minimal bias from sheared genomic fragments, resulting in "sheared" BACs from *Drosophila melanogaster*, mouse (C57BL/6J) and human DNA.

Work performed under the auspices of various funding sources, including USDA, NIH, NSF and DOE.

4

Using Transposons to Resolve Repeats in BAC Clones

Jamie Jett[1] (jmjett@lbl.gov), Jeremy Schmutz[2], Eileen Dalin[1], Jane Grimwood[2], Corey Chinn[1], Doug Smith[1], Susan Lucas[1], Chris Detter[1], Paul Richardson[1], and Eddy Rubin[1]

[1]US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598 USA; [2]The Stanford Human Genome Center; and Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305 USA

At the U.S. Department of Energy Joint Genome Institute, standard 3 kb shotgun libraries from tiling path BAC clones were created and sequenced for human Chromosomes 5, 16, and 19. For the most part, assemblies were complete with few uncaptured gaps. However, there were a number of individual BAC clones that stood out as problematic and unfinished due to highly repetitive regions composed of both tandem repeats and duplications within the clone. Chromosome 16 stood out as being especially repetitive.

Medium-size insert (8-10 kb) libraries were created to span problematic regions in an effort to capture unique sequence at both ends to help close gaps and flank repeat regions. Several hundred specifically selected BAC clones were sheared, size selected to 8-10 kb, subcloned and sequenced to a 10x depth. Following this process, many BAC clones were assembled to provide contiguous sequence. There were, however, a small number of important clones with unresolved assembly problems caused by direct tandem and duplication repeats that exceed the medium-subclone size. These problematic regions required an alternative method to verify large repeats (>16 kb) and identify possible unique sequence in the middle of each 8 kb subclone.

Using paired end sequence, ~500 medium-sized insert clones were selected from a total of 21 individual BAC clones for a transposon-based DNA sequencing strategy. Transposons, artificial transposable elements that integrate into the interior of DNA, allow for directed sequencing within a targeted subclone. The insertion of a transposon within internal regions of the medium-size clone provided sequencing priming sites throughout the targeted sequence for initiating a set of bi-directional di-deoxy ladders.

Here, we will describe DNA isolation, transposon insertion, sequencing, and analysis of bi-directional sequencing of internal regions within the 8-10 kb plasmids. To date, the paired bi-directional sequence information has assisted in the finishing of 9 BAC clones with 12 more in process.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, under Contracts No. W-7405-Eng-48, No. DE-AC03-76SF00098, and No. W-7405-ENG-36.

5

Single Molecule DNA Sequence Profiling In Zero-Mode Waveguides Using γ-Phosphate Linked Nucleotide Analogs

Jonas Korlach (jk109@cornell.edu), Michael Levene, Mathieu Foquet, Stephen W. Turner, Harold G. Craighead, and Watt W. Webb

Cornell University, School of Applied Physics, Clark Hall, Ithaca, New York 14853

We show that nucleotides with a fluorescent label attached to the γ-phgosphate constitute a superior class of compounds for successful observation of single molecule DNA polymerase activity in zero-mode waveguides. In contrast to base-labeled analogs (Levene et al., Science 299: 682), the fluorophore is cleaved from an incorporated base upon polymerization, providing a low background necessary for detection of successive enzymatic turnovers. The enzyme is not inhibited, permitting replacement of all four bases with spectrally separable analogs.

Using these analogs, single molecule DNA synthesis by the highly processive DNA polymerase from phage ϕ29 was measured in zero-mode waveguides. These nanostructures drastically reduce the observation volume to

tens of zeptoliters, thereby enabling single fluorophore detection at high concentrations (>1μM) where polymerase-mediated DNA synthesis is fast and processive. Incorporation is detected by longer residence times of analogs in the active site (~1-10ms), compared to diffusion (<100μs). One-base sequence patterns on synthetic sequences were obtained. Strategies for efficient excitation of spectrally separable analogs involving FRET pair coupled dyes for de novo single DNA molecule sequencing are presented.

DoE DE-FG02-99ER62809 & NCRR-NIH P41-RR04224.

6

Complete Direct Sequencing of BAC, Phage and Microbial Genomes using ThermoFidelase, Fimer and D-Strap Technologies

S. Kozyavkin (serg@fidelitysystems.com), O. Shcherbinina, V. Shakhova, N. Pavlova, A. Morocho, V. Karamychev, Y. Malykh, A. Pavlov, N. Polouchine, A. Malykh, and A. Slesarev

Fidelity Systems, Inc., 7961 Cessna Avenue, Gaithersburg, MD 20879-4117
http://www.fidelitysystems.com

We have developed tools and workflow for direct genomic DNA sequencing that eliminates the need in subcloning and production of shotgun libraries, minimizes the number of sequencing reactions and dramatically accelerates the assembly of complete sequence. Using our approach we have completed sequencing of many BAC, phage and microbial genomes.

A core component of the procedure is the use of genomic DNA as a template in a robust sequencing reaction. The addition of ThermoFidelase 2 with its unique combination of topoisomerase and DNA binding activities is used to shorten the cycles of denaturation and primer annealing. The dramatic increase in specificity, quality and yield of priming from megatemplates is achieved by using Fimers (modified oligonucleotides with proprietary SUC modifications) instead of regular primers and multiplying the number of thermal cycles. The third element of new strategy, D-Strap is based on Fimer design that targets evolutionary conserved elements. The advantages of direct genomic sequencing include elimination of cloning artefacts and library or PCR cross contamination that are extremely important for production sequencing of closely related organisms, non-biased complete and low coverage of the genome that results in significant savings on data processing. In addition, the use of D-Strap Fimers in multiple projects contributes to cost savings and has a potential for the fastest way of complete sequencing of closely related species.

We have completed five types of BAC projects starting from full BAC shotgun, skimming shotgun, whole mouse genome shotgun, cDNA sequence and more recently, composite (reference) human genome. In all cases robust sequencing reactions on BAC DNA using ThermoFidelase and plates of Fimers were used for finishing. The projects were somewhat different in workflow organization. The new requirements for individual's genome sequencing (NO ERRORS) are much stricter than current (Bermuda) requirements for composite human genome project (1 error per 10 kb). The no errors goal was achieved by complete direct sequening of BAC provided by Dr. Larionov (NCI). The assembly of Fimer directed reads was straightforward, completely automated and took less than a minute. The results indicated that TAR method of BAC production and direct complete BAC sequencing using ThermoFidelase and plates of Fimers can be scaled up for NO ERRORS sequencing of genomes of many individuals for a fraction of cost and effort compared to that of shotgun or PCR based methods.

Shotgun method of sequencing 200 kb phage genomes is unproductive because of the under representation of significant portion of phage sequences in shotgun libraries and the up front cost of library production. In contrast, Fimer directed sequencing of phage DNA turned out to be a straightforward method fully compatible with numerous modifications in phage genomic DNA. We have completed NO ERRORS sequencing of a number of phages. In addition, we have designed and validated a set of D-Strap Fimers that target weakly conserved phage genes. The developed

set of D-Strap Fimers is useful for initiation of phage genome sequencing and multi locus sequence typing of phages.

A high fraction of unfinished microbial genomes and long time required for their finishing clearly illustrates deficiencies of whole genome shotgun method. ThermoFidelase and Fimer based direct microbial DNA sequencing was used by us and other teams to accelerate finishing of shotgun projects. Our recent goals have been concentrated on optimization and scale up of direct microbial sequencing. Last year we have completed and published Whole Genome Direct sequencing of *M. kandleri* AV19. Using D-Strap method we have sequenced another isolate of this species. Our recent work on finishing Lacto genomes provides us with an opportunity to compare deep shotgun vs direct genome sequencing mehtods. Our experience indicates that deep shotgun and increased complexity of the shotgun libraries does not automatically give a complete genome sequence. Many gaps, misassemblies, low quality regions and uncertain repeat structure are all present. At the same time the data processing load on assembly, contamination clean up and chimera detection increases exponentially.

We expect that minimizing shotgun coverage and implementation of direct genome sequencing into production setting is a viable alternative to the current practice of deep shotgun drafting of microbial genomes that can produce complete sequences without escalating project costs. In collaboration with JGI we have demonstrated that direct sequencing of microbial genomes using ThermoFidelase and plates of Fimers is compatible with 96 capillary sequencers. The two remaining items needed for complete direct sequencing of microbial genomes that does not rely on shotgun clones are the ability to sequence genomic DNA samples of poor quality and resolution of sequence repeats that are longer than read length. We have found the solution to both problems by developing a new method of direct genomic sequencing based on capture of Sanger fragments. To increase the yield of capture we have developed novel biotin amidites with super-long linkers that meet additional requirements on minimal charge density and optimized rigidity and flexibility. We will present the examples of using biotinilated Fimers with super-long linkers for elimination of background fluorescent noise and long repeat resolution in direct microbial genome sequencing.

Supported in part by DOE (DE-FG02-98ER82577, 00ER83009).

7

Isolation of Exceptional Chromosomal Regions to Close the Gaps in the Draft Human Genome Sequence

S.-H. Leem, N. Kouprina, and **V. Larionov**
(larionov@mail.nih.gov)

Laboratory of Biosystems and Cancer; National Cancer Institute, NIH, Bethesda, MD 20892

The Human Genome Project has entered a final phase during which the sequence must be completed, corrected and finalized. During this phase of the project, sequence gaps must be closed and the overall quality of the sequence improved. To carry this out, it will be necessary to collect additional sequence data. It is not clear whether all of the sequences missing from the draft human genome sequence are represented in the bacterial libraries. Recent works on sequencing of *Plasmodium* and *Dictyostelium* genomes has shown that a high AT content of the chromosomes prevents the construction of large insert BAC libraries. The gaps between contigs in the draft of human genome may also arise from chromosomal regions that are not present in the *Escherichia coli* libraries used for DNA sequencing because they can not be cloned efficiently, if at all, in bacteria. To address this question, ten gap regions from human chromosomes 5, 16 and 19 were recovered in yeast as circular YAC/BACs with a selective TAR cloning method. Further analysis of the gap isolates revealed two types of sequences: a) those that were unstable both in YAC and BAC forms, b) those that were stable in yeast but toxic for bacterial cells. Sequencing of these exceptional regions required non-standard approaches. Some clones were analyzed using a BAC direct sequencing strategy. Other clones were

sequenced in YAC forms to avoid transfer of YAC clones to bacterial cells, where they undergo deletions and rearrangements. This work helped to close last four gaps on chromosome 19. Closing the gaps on chromosome 16 is in progress. In summary, this work and other reports indicate that alternative cloning systems and hosts may be critical to complete the final phase of the Human Genome Project. One of such approaches, TAR cloning in yeast, will allow for rapid and selective isolation of targeted regions of the human genome that can not be verified or completed using clones generated and propagated in *E. coli*.

8

pFOS-LA: A Modified Vector for Production of Random Shear Fosmid Libraries

J. Longmire[1] (longmire@telomere.lanl.gov), N. Brown[1], S. Malfatti[2], Jack Meeks[3], and Patrick Chain[2]

[1]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM; [2]Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA; and [3]Section of Microbiology, University of California, Davis, CA

The standard fosmid cloning vector pFOS-1 was modified to allow simple and rapid cloning of 40 kb fragments generated by random shearing. The modified vector is termed pFOS-LA. A double-stranded oligonucleotide was ligated into the *Hind*III site of pFOS-1 to provide two, unique blunt end sites (*Pml*I and *Swa*I). Cloning in the modified vector proceeds as follows: Target DNA is homogeneously sheared to 40 kb, end repaired (to make blunt ends), dephosphorylated (to prevent chimeric inserts) and ligated into either the *Swa*I or *Pml*I site of the modified vector. Ligation products are packaged, infected and plated using methods developed for pFOS-1 (Kim et al., [1992] Nucleic Acid Research 20: 1083-1085). The modified vector retains advantages afforded by pFOS-1 including single copy origin of replication, double cos site design, high cloning efficiency and ability to make partial digest libraries if so desired. Distinct advantages of the modified vector include 1) A more random cloning approach resulting in libraries with potentially better sequence representation; and 2) Starting DNA can be smaller in size since it does not have to be partially digested prior to cloning. The relaxed requirement for very high molecular weight target DNA allows cloning of DNA samples that are already sheared to 30-40 kb upon isolation. Such DNA would be difficult or impossible to clone by partial digestion.

We have used pFOS-LA to construct a random shear library for the 9.5 Mb genome of the microbe *Nostoc punctiforme*. Cloning 0.2 ug of *N. punctiforme* genomic DNA yielded a library containing 95,000 independent clones with inserts averaging 41.3kb (by fingerprint analysis) and provides >400-fold coverage. As will be discussed, a portion of this library is currently being used to finish sequence the genome of *N. punctiforme*.

It is anticipated that the modified vector will be especially advantageous for mapping and sequencing genomes that are difficult to clone by partial digestion, such as the genomes of many bacterial species where restriction sites are not evenly distributed or where these sites are blocked by methylation. Given the high cloning efficiency and unbiased cloning strategy, this vector should be useful for making deep coverage, high representation "metagenome" libraries from complex microbial communities.

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Los Alamos National Laboratory.

9

Using YACs to Close Gaps at the JGI

Jenna Morgan (jlmorgan@lbl.gov), Duncan Scott, Joel Martin, Tijana Glavina, Susan Lucas, Chris Detter, Paul Richardson, and Eddy Rubin

U.S. Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598 USA.

Yeast Artificial Chromosomes (YACs) have been used for over a decade to construct representative large insert DNA libraries for several genome sequencing efforts (e.g. human, mouse, rice, *Arabidopsis*). YACs confer several advantages over alternative methods for constructing physical maps of large genomes. The YAC vector allows the ligation of insert DNA fragments of over 1,000 kilobases. YAC clones are propagated in yeast as stable artificial chromosomes, and can maintain sequences that are unstable in prokaryotic-based cloning systems. In some cases (e.g. *Dictyostelium discoideum*), YACs may represent the only viable method for the construction of large insert libraries.

YACs present unique challenges in practice, however. The larger inserts tend to lose integrity during transmission in yeast, suffering from deletions and chromosomal rearrangements. Although frequency of occurrence varies among libraries, YACs often contain chimeric inserts. Additionally, YACs are present in single copies and are very similar in size to endogenous yeast chromosomes, thus complicating YAC DNA isolation. In part for these reasons, the JGI has relied almost exclusively on more “user friendly” cosmid and BAC libraries for sequencing its chromosomes 5, 16, and 19. In general, this approach has been extremely successful; however, as the Human Genome Project draws to a close, there are a few regions of the chromosomes that are not spanned by sequenced BAC clones. These regions, referred to as “clone gaps” or “Type 3” gaps, require alternative approaches for closure.

Here we present our strategy for using YACs to close these Type 3 gaps in human chromosomes 5, 16, and 19. The challenge to increase low YAC DNA yield was tackled by scaling up standard YAC isolation protocols. The optimization of pulsed-field gel conditions to separate YACs ranging in size from 150kb to 1700kb was the most onerous task. To date we have successfully closed several Type 3 gaps in each of our human chromosomes using YAC subclone libraries.

Recent advances in YAC cloning technology have greatly improved the efficiency of the YAC shotgun sequencing effort. The transformation-associated recombination (TAR) cloning strategy developed by Larionov et al. has produced gap-spanning YAC clones without chimerism and without the need for the construction of an entire library of random clones. Recently we produced a subclone library that was used to close the last remaining clone gap on chromosome 19 using a circular YAC clone. Currently, work is in progress to produce circular YAC clones for each of the remaining clone gaps in chromosomes 5 and 16.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, under Contracts No. W-7405-Eng-48, No. DE-AC03-76SF00098, and No. W-7405-ENG-36.

10

Helix-Hairpin-Helix Motifs to Create Processive, Hyperstable and Inhibitor-Resistant Enzymes

Alexei Slesarev (alex@fidelitysystems.com), Andrey Pavlov, Nadya Pavlova, and Sergei Kozyavkin

Fidelity Systems, Inc.

We have developed a technology for creating DNA metabolizing enzymes with the enhanced processivity, outstanding stability, remarkable tolerance to high salt concentrations and resistance to chemical and biological inhibitors. We demonstrate the potential of this technology on DNA polymerases. Our method consists in creating chimeras composed of core polymerase domains or entire unmodified enzymes fused with different helix-hairpin-helix (HhH) domains derived

from topoisomerase V (TopoV) of *Methanopyrus kandleri* (1,2). HhH is a widespread motif involved in sequence-nonspecific DNA binding. There are 24 such motifs in Topo V (3), most of them being dispensable for the activity of TopoV, yet their removal greatly affects the stability and salt concentration range of TopoV. We demonstrate that different HhH cassettes fused with either NH2-terminus or COOH-terminus of the Stoffel fragment of Taq polymerase or with Pfu polymerase increase the resistance of the DNA polymerase activity to high salt concentrations, up to 0.5 M NaCl or 1.8 M KGlu. The processivity of chimeric polymerases increases and depends on the structure of HhH attached to the catalytic domains. Anions play a major role in the inhibition of DNA polymerases. Chimeras are more thermostable than their unmodified counterparts and show no loss of the activity after incubation at 100°C for at least 1 hour. Moreover, the chimeras are able to extend primers at least at 105°C. Our approach to raise the salt tolerance of polymerases and their stability also allows for cycle sequencing and PCR at high salt concentrations and at temperatures inaccessible for other DNA polymerases.

This work was supported by grants from DOE and NIH.

1. Andrey R. Pavlov, Galina I. Belova, Sergei A. Kozyavkin, and Alexei I. Slesarev. PNAS 2002 99: 13510-13515

2. Slesarev, et al. PNAS 2002 99: 4644-4649

3. Galina I. Belova, Rajendra Prasad, Sergei A. Kozyavkin, James A. Lake, Samuel H. Wilson, and Alexei I. Slesarev. PNAS 2001 98: 6015-6020

11

Efficient Isothermal Amplification of Single DNA Molecules

Stanley Tabor (tabor@hms.harvard.edu) and Charles Richardson

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115

We are developing DNA polymerases for use in DNA sequencing and amplification applications. We will describe our progress in developing a very efficient isothermal amplification system that is based on the replication machinery of bacteriophage T7. This system is capable of amplifying single DNA molecules, increasing the amount of DNA more than a trillion-fold in a 30 min reaction. Amplification is nonspecific. The template can be circular (e.g. plasmid or BAC DNA) or linear (e.g. genomic DNA). The products are linear double-stranded DNA fragments that average several thousand base pairs in length. The reaction requires the T7 DNA polymerase, the T7 helicase/primase complex (T7 gene 4 protein), and single-stranded DNA binding protein. The reaction requires no exogenous primers, using the inherent priming activity of the T7 primase. It is critical to remove all contaminating DNA from the reaction mixture, since all DNA present will be amplified. We have developed a successful strategy to cleanse the reaction mixture of contaminating DNA by pretreatment with Micrococcal nuclease.

We will present our progress on the use of this system in the following applications:

1. The preparation of plasmid and BAC DNA templates for DNA sequencing.

2. The whole genome amplification of rare DNAs, such as hard-to-culture microorganisms and purified chromosomes.

3. An extremely sensitive and rapid assay to determine the total amount of DNA in a sample, with a linear range of over 13 orders of magnitude.

4. The sequencing of haplotypes by the amplification of DNA from single chromosomes.

12

Improved DENS: Finishing Without Custom Primers—From Human to Microbes

Olga Chertkov, Marie-Claude Krawczyk, Mira Dimitrijevic-Bussod, David Bruce, Mark Mundt, Paul Gilna, Norman Doggett, and **Levy Ulanovsky** (levy@anl.gov)

Los Alamos National Laboratory

DENS (Differential Extension with Nucleotide Subsets) is primer walk sequencing without custom primer synthesis. DENS largely eliminates the cost of custom primer synthesis—several dollars, compared to less than a dollar for the rest of the expenses (per lane) combined. DENS works by converting a short primer (selected from a pre-synthesized library of 1440 octamers with 2 degenerate bases each) into a longer one on the template at the intended site only. DENS starts with a limited initial extension of the octamer primer at 20° C in the presence of only 2 of the 4 possible dNTPs. The primer is extended by 5 bases or longer at the intended priming site, which is deliberately selected, as is the two-dNTP set, to maximize the extension length. The subsequent cycle-sequencing at 60° C accepts the primer extended at the intended site, but not at alternative sites where the initial extension (if any) is generally short. We have now automated all labor-intensive steps in DENS and have employed this as part of our finishing strategy to improve low quality targets. Our current throughput is 16,000–20,000 DENS reactions per month. Much of human chromosome 16 and several bacterial genomes have been finished using > 140,000 DENS reactions with the success rate rising from ~40% to ~80%. DENS accounts for more than half of finishing at LANL.

Systems Biology

13

New Methods and Models for Genomic Systems Biology

George Church[1] (church@arep.med.harvard.edu), Martin Steffen[1], Wayne Rindone[1], Matt Wright[1], Daniel Segre[1], Dennis Vitkup[1], Jake Jaffe[1], Rob Mitra[2], Jay Shendure[1], Greg Porreca[1], Vincent Butty[1], and Jun Zhu[1]

[1]Harvard Medical School, 200 Longwood Ave, Boston, MA 02115; and [2]Washington University

Understanding biological experiments will increasingly benefit from system models, not just subsystems, but comprehensive, genome-wide analyses. Genotype + environment yields phenotype. New methods allow us to cost-effectively "overdetermine" each of these components enabling studies of mechanism, optimality and bioengineering. New technologies include single-molecule sequencing with polymerase colonies (polonies) to assess alternative RNA splicing and DNA haplotyping. Polonies may also offer a path toward a Mbp per $ sequencing goal conducive to expanded diagnostics and environmental monitoring. New computational approaches include "expression coherence" for combinations of transcription elements and "Minimization of Metabolic Adjustment" (MoMA) to model proliferation of mutants. See http://arep.med.harvard.edu/

14

Stepping up the Pace of Discovery

Marvin E. Frazier
(marvin.frazier@science.doe.gov)

Office of Science, Office of Biological and Environmental Research, U.S. Department of Energy

The systems biology revolution is proceeding along multiple pathways as science agencies and the private sector adopt strategies suited to their particular needs and cultures. During the past 3 years, the DOE Office of Science held 15 workshops involving some 500 scientists to advise the department on how it should contribute to meeting the systems biology challenge and to determine related technological needs, potential applications, and societal considerations. The result was the development of the new DOE Genomes to Life (GTL) program. A central focus of GTL is environmental microbial biology, and its key goal over the next 10 to 20 years is to achieve a basic understanding of thousands of microbes and microbial systems in their native environments. This focus demands that we address huge gaps in knowledge, technology, computing, data storage and manipulation, and systems-level integration.

The GTL program has several distinguishing features, including (1) strategies for unprecedented levels of comprehensive data collection using emerging high-throughput technologies, tightly coupled with (2) advanced computing, mathematics, algorithms, and data-management technologies; (3) a unique focus on microbial organisms and systems possessing capabilities for possible solutions to energy and environmental challenges; and (4) implementation of new research and management models that link facilities dedicated to production-scale systems biology data generation and analysis in a teaming environment for a large community of individual investigators.

The DOE Office of Science Joint Genome Institute (JGI) has played and will continue to play a key role in GTL by sequencing microbial and other important genomes. Consistently, roughly half the genes found in diverse microbial species are of unknown function, and half of those have not been observed previously, suggesting that the number of essentially novel genes could eventually range into the tens of millions. From this perspective alone, DNA sequencing clearly is and will remain for some time the most economical method for gene discovery. Discovery of new and novel genes and pathways that can aid DOE in its missions of energy security, bioremediation, and carbon management remains an important aspect of GTL. JGI thus will remain a key component of the GTL program and OBER facilities portfolio. In addition, JGI has a broader role to play in a variety of projects that will push forward the frontiers of science. DNA sequencing should be used, for example, to help understand and characterize the diversity of life on our planet, to address fundamental questions in biology and the evolutionary processes, and to develop new methods for understanding how protein structure affects functionality and efficiency.

Knowledge is power—but only if you use it. The vast amount of information contained in the hundreds of genomes sequenced and the thousands of future sequencing projects offers an unprecedented opportunity for understanding complex biological systems and our environment. This opens exciting new avenues to solve some of the most urgent problems in healthcare, national security, agriculture, energy, the environment, and industry. Addressing these challenges expeditiously demands that we take bold steps to achieve a new, much faster, and more efficient pace of biological discovery.

Analysis Results: Functional Genomics

15

The Celltech/MRI ENU Mutagenesis Program for Identifying Genes Controlling Immune Function in the Mouse

M. Brunkow[1] (Mary.Brunkow@sea.celltechgroup.com), M. Appleby[1], K. Staehling-Hampton[1], J. Gilchrist[2], P. Charmley[1], F. Ramsdell[1], J. Bouck[1], T. Britschgi[1], A. Snell[1], T. Howard[1], M. McEuen[1], P. Tang[1], S. Proll[1], B. Paeper[1], P. Tittel[1], G. Carlson[2], and R. Schatzman[1]

[1]Celltech R&D, Inc., Bothell, WA 98021; and
[2]McLaughlin Research Institute, Great Falls, MT 59405

A major focus of the Celltech/MRI Mutagenesis Consortium is the use of ENU mutagenesis in mice to identify novel, clinically relevant targets in the areas of lymphocyte biology, inflammation and autoimmunity. The approach involves a three-generation recessive screen, and we are focusing on phenotypes which mimic desired clinical responses (e.g., suppressed inflammatory response), thus improving our chances of directly identifying relevant therapeutic targets. We have implemented a number of in vitro screens including activation of T- and B-cells, as well as T-dependent and T-independent inflammatory responses. These are carried out on peripheral blood lymphocytes, and have the advantage of being relatively high throughput and requiring only small volumes of blood, thus affording us the opportunity to perform a number of challenges on a single sample. The in vitro screens have been coupled with a complementary set of more complex in vivo screens based on classic pharmacologic models of inflammation and immune response (e.g., graft-versus-host response and colitis). In the past 2-1/2 years, over 100 phenodeviants have been identified and entered into the mapping process. We have developed an integrated laboratory / informatic pipeline which enables rapid identification of a candidate interval and the genes contained within, as well as efficient tracking of gene testing results, capturing both exon sequence and gene expression data. The development and utilization of informatics tools has proven critical in the effective management of the program. Another important aspect of the program is the ongoing process of new screen development to ensure as broad an interrogation of the immune system as possible. Specific lessons learned from the mutations identified so far will be discussed in more detail.

16

Nucleotide- or Amino Acid-Coded Mass Tagging for Functional Genomics and Proteomics

Sheng Gu, Songqin Pan, Tom Hunter, Haining Zhu, Fadi Abdi, John Engen, E. Morton Bradbury, and **Xian Chen** (chen_xian@lanl.gov)

Bioscience Division, Los Alamos National Laboratory

Mass spectrometry (MS) is a promising tool for rapid, accurate, and sensitive analyses in both areas of functional genomics and proteomics, but critical advances are needed to further increase its specificity and accuracy for the large-scale analyses of biomolecules at the genomic level. To address these cutting-edge issues, with systems biology in mind, we have developed a novel MS-based technique of mass tagging with stable isotopes for post-genomic studies. Our strategy of

nucleotide- or amino acid-specific mass tagging in DNA or protein molecules provides a much more sensitive and accurate way of molecular labeling than radiological or chemical labeling. In addition to mass-to-charge ratio (m/z) in MS spectra, the use of these stable-isotope labels for tagging biological molecules in a sequence-specific way have dramatically enhanced the specificity, accuracy, sensitivity, and throughput of the MS-based technology for functional genomics and proteomics analyses.

We have extended the applications of our technology of mass tagging to quantitative proteomics, de novo peptide sequencing, direct detection of post-translation modifications and low abundant membrane proteins, and protein-protein interactions. In practical cases, we have investigated the differential protein expression involved in p53-induced apoptosis of cancer cells systematically using our quantitative mass tagging strategy that will be generally applicable for quantitative proteomics of any disease cells. A dozen of papers describing our technology have been published in the leading journals.

17

Understanding the Biology of *Brucella melitensis* from Genome to Proteomes

Vito G. DelVecchio[1] (vimb@aol.com), Cesar V. Mujer[1], Mary Ann Wagner[1], Michel Eschenbrenner[1], Sue Hagius[2], and Phil Elzer[2]

[1]Institute of Molecular Biology and Medicine, University of Scranton, Scranton, PA 18510-4625; and [2]Department of Veterinary Science, Louisiana State University AgCenter, Baton Rouge, LA 70803

Brucellae are pathogenic gram-negative bacteria that cause brucellosis, a chronic infectious disease in humans characterized by undulant fever, arthritic pain, and neurological disorders. Brucellosis frequently causes abortion and sterility in domesticated animals such as cattle, sheep, and goats. Based on pathogenicity and host specificity, six species are found within this genus: *B. abortus*, *B. canis*, *B. melitensis*, *B. neotomae*, *B. ovis*, and *B. suis*. In addition, strains from marine mammals have been isolated and are tentatively referred to as members of *B. maris*. The genome of *B. melitensis* has been sequenced, annotated, and analyzed. It consists of two circular chromosomes of 2,117,144 bp and 1,177,787 bp that have been predicted to encode for 3198 ORFs. Sequence analysis confirmed that *B. melitensis* has the ability to survive and grow in aerobic, microaerophilic or anaerobic conditions. Although typical virulence factors and pathogenicity islands are absent, adhesins, invasin, and hemolysins are present.

Genomic data cannot designate which theoretical ORFs are active and thus cannot provide a definitive description of the ultimate biological potential of an organism. To obtain a functional overview, a global proteomics study of the laboratory-grown virulent strain 16M was initiated. So far, 937 proteins representing 269 ORFs were identified using 2-D gel electrophoresis and peptide mass fingerprinting. The two circular chromosomes of *B. melitensis* are functionally active and the locations of ORFs identified at the protein level are evenly distributed in each chromosome. A comparison of strain16M proteome with that of Rev1 revealed significant differences in the expression of several proteins affecting iron metabolism, sugar binding, protein biosynthesis and lipid degradation. To enhance these MALDI-TOF studies, SELDI-TOF was used to again pinpoint the differences between strains Rev1 and 16M. In general, genomic and proteomic information will eventually result in better insight to biomarker discovery, rapid identification and diagnostics, and aid in future vaccine development.

18

Finishing of Human Chromosome 16 Reveals Extensive Segmental Duplications

Norman Doggett[1] (doggett@lanl.gov), Cliff Han[1], Mark Mundt[1], Gary Xie[1], Robert Sutherland[1], David Bruce[1], Levy Ulanovsky[1], Jane Grimwood[2], Jeremy Schmutz[2], Susan Lucas[3], Laurie Gordon[3], Joel Martin[3], and JGI Staff[3]

[1]Center for Human Genome Studies, Los Alamos National Laboratory; [2]Stanford Human Genome Center; and [3]DOE Joint Genome Institute

The minimal tiling path of sequenced clones covering chromosome 16 consists of 636 BACs (169 Caltech and 467 RP11), 75 cosmids, 9 PACs, 5 YAC derived subclones (including subcloned cosmids from half YACs for each telomere) 4 P1 clones, 3 fosmids and 3 PCR fragments. These provide essentially complete coverage of the ~79 Mb of euchromatin. 685 clones are currently finished (93.5%) and the remainder are active in finishing (32 exist as phase 2 ordered accessions). There are currently 8 clone gaps. Four of the clone gaps are small, with a total combined size of less than 100 Kb. Four clone gaps occur in complex segmental duplication regions and are estimated to be small but have not been reliably sized. We have discovered a high level of intrachromosomal duplications during the mapping and sequencing of this chromosome. To help us overcome the complexities of assembling the correct sequence over the most complex of these segmental duplications, we have drafted over 400 additional BAC, cosmid and fosmid clones specifically targeted at duplications and finished close to 100 redundant clones. These efforts allowed us to produce sequence contigs representing a single haplotype across many segmental duplications. We find that 7.8 Mb (~10% of the chromosome) consists of intra-chromosomal duplicated sequence. This is significantly higher than the estimate of 3.4% made by the public consortium effort, based on the analysis of the draft sequence of the human genome. Intrachromosomal duplications occur in 109 duplication blocks along the chromosome. The largest of these segmental duplications is 520,022, 423,731, and 424,145 bp, and these contain many smaller duplications. Many duplications contain known and predicted genes. The Polycystic Kidney Disease 1 (PKD1) gene for example is duplicated or partially duplicated as 5 copies on chromosome 16. The nuclear pore complex interacting protein (NPIP) is copied 23 times on chromosome 16 and displays greater sequence divergence of its exons than its introns which provides an indication of positive selection acting at this locus. We will present further detailed sequence and evolutionary analysis of the complete set of intrachromosomal duplications.

Supported by the U.S. DOE under contract No. W-7405-ENG-36.

19

The Molecular Basis for Metabolic and Energetic Diversity

Timothy Donohue[1] (tdonohue@bact.wisc.edu), Jeremy Edwards[2], Mark Gomelsky[3], Jonathan Hosler[4], Samuel Kaplan[5], and William Margolin[5]

[1]Bacteriology Department, University of Wisconsin-Madison; [2]Chemical Engineering Department, University of Delaware; [3]Department of Molecular Biology, University of Wyoming; [4]Department of Biochemistry, University of Mississippi Medical Center; and [5]Department of Microbiology and Medical Genetics, University of Texas Medical School at Houston

Our long-term goal is to engineer microbial cells with enhanced metabolic capabilities. As a first step, this team of scientists and engineers seeks to acquire a thorough understanding of energy-generating processes and genetic regulatory networks of the photosynthetic bacterium, *Rhodobacter sphaeroides*. The ability to capitalize on the metabolic activities of this versatile bacterium was increased by the completion of the *R. sphaeroides* genome sequence at the DOE-supported Joint Genome Institute. The *R. sphaeroides* Genomes to Life Consortium is deciphering important energy-generating activities of this bacterium and studying the assembly and operation of energy generating machines. The long term goals of these efforts

are to acquire the information needed to design microbial machines that degrade toxic compounds, remove greenhouse gases, or synthesize biodegradable polymers with increased efficiency. At the March 2003 workshop, we will provide a progress report on our analysis of the metabolic capabilities of this facultative microorganism.

In particular, we will report on activities in the following areas. 1. The identification of proteins that are central to growth via respiration and the utilization of solar energy by photosynthesis. 2. The formulation of a first generation metabolic map and new software tools that will aid future analysis of the pathways and regulatory networks of this bacterium. 3. Microscopic imaging techniques that will allow us to visualize the organization of the photosynthetic apparatus and the assembly of key bioenergetic molecular machines. In this poster, we hope to illustrate why this cross-disciplinary, systems approach to the analysis of energy generation by this facultative bacterium can provide new insights into fundamental aspects of energy generation by this photosynthetic organism.

20

Biomarker Discovery for *Brucella melitensis* Wild Type and Vaccine Strains using SELDI-MS Technology

Michel Eschenbrenner[1] (eschenbrenm2@scranton.edu), Mary Ann Wagner[1], Frank Estock[1], Cesar V. Mujer[1], Sue Hagius[2], Philip Elzer[2], and Vito G. DelVecchio[1]

[1]Institute of Molecular Biology and Medicine, The University of Scranton, Scranton, Pennsylvania 18510; and [2]Department of Veterinary Science, Louisiana State University AgCenter, Baton Rouge, Louisiana 70803

The Gram-negative bacteria, *Brucella*, are responsible for brucellosis, a zoonotic disease afflicting various domesticated animals and humans. Their importance as potential bioterrorism agents requires the need for a quick and efficient identification system. Surface-Enhanced Laser Desorption/Ionization (SELDI) technology allows selective protein capture from crude extracts. SELDI-MS was used to compare the wild type 16M from the vaccine strain Rev 1. The proteins were bound on different protein chips, and their respective spectra were compared. Seven putative biomarkers, with molecular masses ranging from 6.5 to 85.2 kDa, were identified for each strain. Two protein peaks were specifically detected in Rev 1 using normal phase chips and five protein peak differences were observed between 16M and Rev 1 using weak cation exchange chips.

21

Multi-Species Comparative Sequence Analysis of a 365 kb Interval on Human Chromosome 21 Surrounding SIM2

Kelly A. Frazer[1] (kelly_frazer@perlegen.com), Kazutoyo Osoegawa[2], Mark F. Doherty[1], Michael Jenn[1], Xiyin Chen[1], Pieter J. de Jong[2], and David R. Cox[1]

[1]Perlegen Sciences, 2021 Stierlin Court, Mountain View, CA 95051; and [2]Children's Hospital and Research Center, Oakland, CA 94609

The rate of evolution varies widely in different regions of a genome within a species as well as for orthologous sequences between species. Thus, when performing cross-species sequence comparisons it is not possible to choose a standard threshold criteria of "functional" conservation that is applicable across the entire human genome for distinguishing between sequences that are conserved due to constraints from those that are conserved because of shared ancestry. We previously performed a three-way comparative analysis of human, mouse and dog DNA across a 6-Mb 21q22 region. This study suggested that comparing the sequences of multiple species is a powerful empiric means of distinguishing actively conserved sequences from sequences conserved due to shared ancestry. We have expanded this study to identify and compare the distribution of conserved human-horse, human-cow, human-pig, human-dog, human-cat, and human-mouse elements

within a 365-kb interval in human 21q surrounding the single-minded 2 (*SIM2*) gene.

High-density arrays representing 365-kb of human chromosome 21 sequences were hybridized with orthologous horse, cow, pig, dog, cat, and mouse DNA to identify evolutionarily conserved human sequences. Approximately 15.8% (57,482 bp) of the human sequence analyzed was identified as conserved, of which ~28.3% (16,258 bp) is found in humans and only one of the six mammalian species, ~45.5% (26,157 bp) is found in humans and between two to five of the mammalian species, and ~26.2% (15,067 bp) is found in humans and all six of the mammalian species analyzed. These data suggest: 1. A significant fraction of the human DNA sequences that are evolutionarily conserved will not be identified by human-mouse sequence comparisons. 2. A comprehensive comparative analysis of the human genome for the identification of functional elements will require that it be compared with the genomic sequences of multiple mammals.

22

The Microbial Proteome Project: A Database of Microbial Protein Expression in the Context of Genome Analysis

Carol S. Giometti (csgiometti@anl.gov) and Gyorgy Babnigg

Biosciences Division, Argonne National Laboratory

Using complete genome sequences to predict the proteins expressed by a cell does not provide an accurate assessment of the relative abundance of proteins under different environmental conditions. In addition, genome sequences do not define the subcellular location, biomolecular and cofactor interactions, or covalent modifications of proteins that are critical to their function. Therefore, analysis of the protein components actually produced by cells (i.e., the proteome) in the context of genome sequence is essential to understanding the regulation of protein expression. As the number of complete microbial genome sequences increases, vast amounts of genome and proteome information are being generated. In parallel with the proteome analysis of numerous microbial systems, we are developing methods for managing and interfacing the diverse data types generated by both genome and proteome studies as part of Argonne's Microbial Proteome Project. The goal is to provide users with a highly interactive database that contains proteome information in the context of genome sequence in formats conducive to data interrogations pertinent to biological questions. To achieve that goal, we are developing and maintaining three World Wide Web-based databases: Proteomes2, ProteomeWeb, and GelBank. The Proteomes2 database (http://proteomes2.bio.anl.gov) is a password-protected site that provides DOE project collaborators with access to the experimental details for approximately 1,000 samples from seven different microbes (*Shewanella oneidensis, Geobacter sulfurreducens, Prochlorococcus marinus, Methanococcus jannaschii, Pyrococcus furiosus, Rhodopseudomonas palustris*, and *Deinococcus radiodurans*) and links each sample with multiple protein patterns. ProteomeWeb (http://ProteomeWeb.anl.gov) is an interactive public site that provides the identification of expressed microbial proteins, links to genome sequence information, tools for mining the proteome data, and links to metabolic pathways. GelBank currently includes the complete genome sequences of approximately 90 microbes and is designed to allow queries of proteome information. The database is currently populated with protein expression patterns from the Argonne Microbial Proteomics studies and will accept data input from outside users interested in sharing and comparing proteome experimental results.

This research is funded by the United States Department of Energy, Office of Biological and Environmental Research, under Contract No. W-31-109-ENG-38.

23

Comparative Mapping and Sequencing of Syntenically Homologous Segments of Human Chromosome 19 Across Multiple Vertebrate Species Including Chicken

L.A. Gordon[1] (gordon2@llnl.gov), **M. Tran-Gyamfi**[1], **R. Nandkeshwar**[1], M. Groza[1], M. Christensen[1], E. Fields[1], P. Butler[1], M. Wagner[1], I. Ovcharenko[2], A. Aerts[3], K. Kadner[3], J. Smith[4], R. Crooijmans[5], M. Groenen[5], S. Lucas[3], and L. Stubbs[1]

[1]Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore CA; [2]Lawrence Berkeley National Laboratory, Berkeley, CA; [3]D.O.E. Joint Genome Institute, Walnut Creek, CA; [4]Department of Genomics and Bioinformatics, Roslin Institute, Roslin, U.K.; and [5]Department of Animal Sciences, Wageningen Agricultural University, Wageningen, The Netherlands

Cross-species comparison of syntenically homologous, conserved sequence provides biologically relevant insight into the human genome and evolutionary processes. It facilitates the identification of low-copy or rarely expressed genes, signals the presence of otherwise difficult to detect non-coding regulatory elements, and sheds light on ancestral genome organization, lineage-specific chromosomal rearrangements and mechanisms of gene evolution. We previously mapped and sequenced human chromosome 19 (HSA19) - related homology segments in mouse (*Genomics* 74:129-141, 2001; *Science* 293:104-111, 2001). While comparisons between the two mammalian species are proving extraordinarily helpful, biological understanding is substantially enhanced by comparing sequences from additional reference species at informative evolutionary distances. To this end we have mapped and are sequencing HSA19-related regions from a third, evolutionarily more distant vertebrate, the chicken.

As homology breakpoints had not been previously detailed in chicken we designed overgo and PCR probes wherever protein-translated HSA19 gene sequences identified well-conserved (60-95%) chicken ESTs. Probes for over 100 gene loci were hybridized successfully to three BAC libraries, one from *Gallus domesticus* and two from *Gallus gallus*. Clones identified by hybridization were restriction digested and assembled into maps to assess clonal integrity and overlap, facilitate contig extension, identify homology breaks and generate efficient sequencing tiling paths. Contigs homologous to HSA19p13.3 and p13.1 located on chicken chromosome 28 (GGA28), as well as islands of HSA19q homology scattered throughout the chicken genome, have been successfully characterized and 170 clones submitted for sequencing at the JGI. Preliminary analyses of 80 clones yields sequence-based identification of additional syntenic orthologs and provides high resolution detail of homology segment breakpoints and rearrangements.

As expected, chicken sequence exhibits much higher levels of conservation relative to mouse and human than, for instance, that of the evolutionarily more remote puffer fish, *Fugu rubripes*, recently sequenced at the JGI (*Science* 297:1301-1310, 2002). While linkage groups as a whole are well conserved in chicken, interruptions and rearrangements in synteny at the level of gene-to-gene resolution are pervasive. Comparisons of homology breakpoints between the three species suggest presumptive ancestral genome arrangements; in at least one case mouse and chicken share gene order that is not preserved in human, while in other cases disruptions in synteny can be attributed to breaks and rearrangements in mouse. These data are facilitating the annotation of HSA19 while shedding intriguing light on the mechanisms that drive genome evolution and vertebrate speciation.

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

24

Isolation of DNA Binding Proteins from Nuclear Extracts by Biomolecular Interaction Analysis (BIA)-Based Ligand Fishing

Christopher A. Hack[1] (cahack@lbl.gov), Michael Murphy[1], Shirin Fuller[1], Lior Pachter[2], Dario Boffelli[1,3], Sharon Doyle[1], Paul Richardson[1], and Eddy Rubin[1,3]

[1]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA; [2]Department of Mathematics, University of California, Berkeley, CA 94720, USA; and [3]Department of Genome Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Genome Sequencing efforts are producing large quantities of data, but proper interpretation of these data is difficult without some understanding of the proteins encoded by the genomic DNA and the cis-regulatory elements which oversee their expression. Of particular interest are transcription factor proteins that regulate gene expression and the cis-regulatory elements to which they bind. The identification of transcription factors and their cis-regulatory binding sites will facilitate more comprehensive analysis of specific protein expression patterns, adding key information to the process of decoding the genome. We have developed an assay that uses Biomolecular Interaction Analysis (BIA)-based ligand fishing as a means to characterize the interactions of transcription factors that bind DNA enhancer elements, with the goal of identifying isolated protein binders by downstream analysis. The Surface Plasmon Resonance (SPR) biosensor in a BIA instrument allows for real-time monitoring of interactions between proteins and their binding partners. Double-stranded oligonucleotides of DNA matching sequence in regions immediately upstream of the human apolipoprotein A (apoA) promoter were captured in turn to the surface of BIA sensor chips. Nuclear extracts from human liver cell line HepG2 were passed over the captured oligonucleotides, and the interaction of DNA-binding proteins was monitored with the SPR biosensor. (Oligonucleotides and nuclear extracts were prepared as described by D. Boffelli et al., Science, Vol. 299, pp. 1391-4, (2003).) Protein binding to highly conserved regions of DNA was observed to be significantly stronger than binding to regions of DNA that showed greater sequence divergence between hominoids and Old World monkeys. In addition, the BIA process is non-destructive, allowing recovery of bound proteins for downstream analysis and identification by silver-stained PAGE and/or mass spectrometry. Protein bands were observed from samples eluted from conserved regions of DNA and analyzed by silver-stained PAGE, indicating the presence of one or more transcription factors specific to binding sites on the conserved oligonucleotides. Such specific bands were not observed on PAGE analysis of samples eluted from non-conserved regions, further supporting the hypothesis that the conserved regions of genomic DNA are functionally important as regulators of gene expression. Further work to characterize the eluted transcription factors by mass spectrometry is in progress.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, under Contracts No. W-7405-Eng-48, No. DE-AC03-76SF00098, and No. W-7405-ENG-36.

25

Differential Expansion of Homologous Zinc Finger Gene Clusters Located on Human Chromosome 19q13.2 and Mouse Chromosome 7 (Mmu7)

Aaron T. Hamilton[1] (hamilton28@llnl.gov), Mark Shannon[1,3], Laurie Gordon[1,2], Elbert Branscomb[2], and Lisa Stubbs[1,2]

[1]Biology and Biotechnology Research Program (BBRP), Lawrence Livermore National Laboratory; [2]DOE Joint Genome Institute; and [3]Applied Biosystems

In comparative studies on regions of the mouse genome that are syntenically homologous to Human chromosome 19, we discovered one conserved region of HSA19q13.2 and mouse chromosome 7 (Mmu7) that contained a cluster of zinc finger genes (encoding transcriptional regulators) including multiple probable orthologous pairs of genes. However there appear to be very different duplication histories in the expansion of the cluster, resulting in unequal numbers of zinc-finger genes (21 human, 10 mouse) through differential duplication of ancestral genes such that strictly orthologous relationships have not been maintained. For example, one human gene (ZNF235) is related to six mouse genes which have apparently arisen by duplication events since the divergence of the two lineages, while a single mouse ZNF gene (Zfp61) remains as the single mouse "homolog" for ten recently duplicated human genes. We have developed a hypothesis to explain the phylogenetic history of this gene cluster and have studied the divergence of expression patterns for the duplicated genes. Planned experimental manipulation of individual zinc-finger (ZNF) gene expression will reveal potential target genes that may be regulated by the ZNF genes in the cluster, allowing comparisons between recently-duplicated ZNF genes and also an inter-species assessment of the functional conservation of orthologs. We have also begun to investigate how differences in expression patterns between paralogous zinc-finger genes are reflected in diverging structures of the duplicated regulatory elements of the ZNF genes. Because mammalian genomes contain hundreds of zinc-finger transcriptional regulators, many of which are of the same KRAB-ZNF type as those in the cluster we surveyed, such changes in gene number and expression patterns have implications for the study of mechanisms for tissue-specific gene regulation and for the analysis of the origin of genetic diversity on which natural selection acts. As this region is sequenced for other species the data for the cluster in each will be added to the comparative analysis. An overview and progress reports on these aspects of the project will be presented.

26

Genome Construction and Analysis in *Rhodobacter sphaeroides* 2.4.1

Samuel Kaplan (Samuel.Kaplan@uth.tmc.edu), Madhusudan Choudhary, Ronald C. Mackenzie, Jung Hyeob Roh, and William E. Smith

Microbiology & Molecular Genetics, University of Texas Health Science Center at Houston

Rhodobacter sphaeroides 2.4.1 is a free-living facultative photosynthetic member of the α-3 Proteobacteria. This organism is capable of displaying a diverse array of growth modes, reflecting its very substantial metabolic potential. Our interest in this organism has focused on its ability to transition from aerobic to anaerobic photosynthetic growth and on the structure and function of its complex genome.

The J.G.I. completed the high throughput genome sequence of *R. sphaeroides* 2.4.1 in October of 2001, resulting in 195 contigs, and these together with our own genome "skimming" project of chromosome II, enabled us to provide the compete physical assembly of chromosomes I (C-I) and chromosome II (C-II) which will be presented. Using the genome sequence data, and taking into consideration the third position bias of this high G+C organism (68.81%) we and members of the DOE-sponsored *R. sphaeroides* Microbial Cell Project Team, together with the Affymetrix Corp. constructed a Gene Chip. The "Chip" consists of probe sets for 4292 orf's, 47-rRNA and tRNA genes and 394 intergenic regions,

which for most part have been reassigned as orf's following genome assembly.

Analysis of the transcriptome in our laboratory has proceeded along several parallel lines involving analyses of transcriptome expression under standard growth conditions, and employing the use of mutant organisms known to possess alterations in gene expression. In order to validate the results of the Gene Chip experiments we have developed standardized protocols for RNA isolation and cDNA development. We have performed all experiments in triplicate and have routinely obtained R values (Pearson Coefficient) of 0.980 or better. We have followed the ratios for the expression of all of the ribosomal proteins from each replicate to each other replicate, which is predicted to be 1.00 and which is revealed to be on average 0.983.

Because of our long-term interest in the aerobic to photosynthetic transition, we have focused the first of our transcriptome analyses on the expression of genes involved in photosynthesis, genes involved in taxis and flagellar assembly for which there are numerous duplicate and triplicate representatives, as well as generalized gene expression showing patterns of change under these growth conditions. We shall provide both the data derived form each of these sets of experiments as well as summary results which are more easily viewed and reveal for the first time a global picture of gene expression in specific, multi-dimensional regulatory systems of *R. sphaeroides*.

This work has been supported by the DOE Grant OBER DE-FG02-01ER63232 and USPHS Grant GM15590.

27

Characterization of an Imprinted Domain Located in Human Chromosome 19q13.4/ Proximal Mouse Chromosome 7

Joomyeong Kim (kim16@llnl.gov), Anne Bergmann, Angela Kollhoff, and Lisa Stubbs

Genomics Division, Biology and Biotechnology Research Program, L-441, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94551

For a subset of mammalian autosomal genes, the two parental alleles are not functionally equivalent due to genomic imprinting. Imprinting involves inactivation of one allele, depending upon the parental origin. In early studies, we located one imprinted gene, Peg3 (paternally expressed gene 3), to human chromosome 19q13.4. We have isolated and characterized 5 additional imprinted genes from the 1MB-genomic intervals surrounding human and mouse PEG3, including Zim1 (imprinted Zinc-finger gene 1), Zim2, Zim3, Usp29 (Ubiquitin-specific processing protease 29), and Znf264. We are currently studying the potential regulatory mechanism controlling the imprinting and expression of these six genes using comparative genomics approaches. Based on our preliminary results, we predict that one region, the surrounding region of the first exon of Peg3, might be responsible for the imprinting of a whole domain. Sequence comparison of the regions derived from human, mouse and cow revealed the presence of one evolutionarily conserved sequence motif that is repeated multiple times within the first intron of Peg3 in all three mammals. DNA mobility shift and chromatin immunoprecipitation (ChIP) assays clearly demonstrated that this motif is an in vivo binding site for the Gli-type transcription factor YY1. The Peg3 YY1-binding sites are methylated only on the maternal chromosome in vivo, and ChIP assays confirmed that YY1 binds specifically to the unmethylated paternal allele of the gene. Promoter, enhancer and insulator assays with deletion constructs of sequence surrounding the YY1-binding sites indicate that the region functions as a methylation-sensitive insulator that may influence the imprinted expression of Peg3 and

neighboring genes. Our current study is the first report demonstrating the involvement of YY1 in methylation-sensitive insulator activity and suggests a potential role of this highly conserved protein in mammalian genomic imprinting.

28

Comparative Analysis of Syntenic Genomic Sequences

Jonathan E. Moore[1] and **James A. Lake**[1,2] (lake@mbi.ucla.edu)

[1]Molecular Biology Institute, University of California, Los Angeles; and [2]Departments of Molecular, Cell, and Developmental Biology, and Human Genetics, University of California, Los Angeles

Comparative analyses of genomic sequence holds great promise for the identification of genes, their structure, and various regulatory elements. We have developed a gene- and putative-regulatory- element-finder that utilizes a method called pattern filtering. Pattern filtering optimally filters the evolutionary signals of the conserved functional elements from the stochastic noise of mutation, allowing the reliable determination of biological elements. In tests of pattern filtering's ability to predict coding regions in the 200-kb CD4 regions of human and mouse, our methods achieve a correlation coefficient per nucleotide of 98.6%, well above that of any gene-finder of which we are aware. In addition, our methods show conserved regions which do not code for proteins, which are assumed to be regulatory elements or genes of untranslated-RNAs. We are applying these methods to syntenic sequences in order to identify novel genetic and functional elements.

29

Proteomic Profiles of *Rhodopseudomonas palustris*

Nathan C. Verberkmoes[1], Caroline S. Harwood[2], Loren J. Hauser[1], Dale A. Pelletier[3], and **Frank W. Larimer**[3] (larimerfw@ornl.gov)

[1]Graduate School of Genome Science and Technology, University of Tennessee, Oak Ridge, TN; [2]Department of Microbiology, University of Iowa, Iowa City, IA; and the [3]Center for Molecular and Cellular Systems, Oak Ridge National Laboratory, Oak Ridge, TN

We recently described (VerBerkmoes, et al., *J. Proteome Research* 1:239-252, 2002,) a comprehensive method for proteome analysis that integrates both intact protein measurement ("top-down") and proteolytic fragment characterization ("bottom-up") mass spectrometric approaches, capitalizing on the unique capabilities of each method. This approach is being applied to proteomic profiling of the anoxygenic photobacterium *Rhodopseudomonas palustris*. Multiple physiological states, i.e., aerobic heterotrophic growth, anaerobic heterotrophic growth, anaerobic photoheterotrophic growth, and anaerobic phototrophic growth, are being profiled. In addition, profiles of mutants defective in major assembly and regulatory processes are being profiled. The proteomic profiles are also being used to enhance the annotation of the genome: a significant number of "genes of unknown function" have been authenticated, and their cellular localization and physiological response are now known. Over 25% of the proteins profiled represent the "unknown" class.

30

Molecular Comparisons of Gene Homologs in Primates

N. Kouprina, V. N. Noskov, J. C. Barrett, and **V. Larionov** (larionov@mail.nih.gov)

Laboratory of Biosystems and Cancer, National Cancer Institute, NIH, Bethesda, MD 20892

Transformation-Associated Recombination (TAR) cloning allows selective isolation of a desired chromosomal region or gene from complex genomes. The method exploits a high level of recombination between homologous DNA sequences during transformation in the yeast *Saccharomyces cerevisiae*. We investigated the effect of nonhomology on the efficiency of gene capture and found that up to 15% DNA divergence did not prevent efficient gene isolation. Such tolerance to DNA divergence greatly expands the potential applications of TAR cloning for comparative genomics. We efficiently and accurately isolated primate gene homologs using a TAR vector containing a human gene targeting sequences. Complete copies of the breast cancer BRCA1 (80 kb) and a major determinant of cerebral cortical size, the gene ASPM, (70 kb) were isolated from chimpanzee, gorilla, orangutan and rhesus macaque genomes, sequenced and compared to corresponding human DNA sequences. Such comparison allowed to follow the gene evolution in great apes and explain a high frequency of intragenic rearrangements in BRCA1 in human population. Because the entire isolation procedure of a gene homolog from several primates could be accomplished in approximately 2 weeks, TAR cloning is a powerful tool for comparative genomics.

31

Elucidating the Role of Two Mammalian Telomerase-Associated Protein Components in vivo—TERT and VPARP

Yie Liu[1] (liuy3@ornl.gov), Bryan E. Snow[2], Wen Zhou[3], Natalie Erdmann[2], Karuna Chourey[1], Marla Gomez[1], Murray O. Robinson[3], and Lea Harrington[2]

[1]Functional Genomics Group, Life Sciences Division, Oak Ridge National Laboratory, TN 37831-6445; [2]Ontario Cancer Institute/Amgen Institute, Department of Medical Biophysics, University of Toronto, 620 University Avenue, Toronto, Ontario M5G 2C1 Canada; and [3]Amgen Inc., 1840 DeHavilland Drive, Thousand Oaks, CA 91320

Telomeres are DNA-protein complexes localized on the end of each chromosome, the function of which is to cap and protect chromosomes against degradation or fusion. Telomeres thus play an essential role in the control of genomic stability. Although telomeres are lost during the aging process in most human somatic cells, telomeres are maintained in germ line cells due to the expression of telomerase, which catalyzes the addition of telomeres and replenishes telomere loss during cell division. Eukaryotic telomerase contains a telomerase reverse transcriptase (TERT) and an RNA template component that together comprise its catalytic core; several other associated factors, of which only a few have been identified, are also known to be essential. We used a gene targeting approach to generate embryonic stem cells and mice lacking TERT or one telomerase associated proteins, VPARP, in order to determine the role of these proteins in vivo.

ES cells lacking mTert lose telomerase activity and show progressive telomere shortening, leading to end-to-end fusions and genetic instability. ES cells heterozygous for mTert knockouts also showed a progressive loss of telomeric DNA; however, despite an average telomere length similar to mTert null ES cells, no genetic instability was observed and a minimal amount of telomeric DNA can be detected at all chromosome ends. Taken together with previous studies, these finding suggest it is the presence of a subset of criti-

cally short chromosome ends, and not a shorter average telomere length per se, that herald the onset of genetic instability.

VPARP-deficient mice are viable and fertile. Furthermore, there is no detectable change in telomerase activity or telomere length in early passages of Vparp-deficient ES cells and tissues from early generation deficient mice. Since VPARP is also localized to the mitotic spindle, we examined microtubule and spindle architecture, chromosome stability and chromosome segregation in VPARP deficient mice. These data will also be presented.

32

Noncoding Deletion Present in Van Buchem Patients Removes Essential Regulatory Elements Required for Bone-Specific Expression of BMP-Antagonist Sclerostin

Gabriela G. Loots[1] (ggloots@lbl.gov), Michaela Kneissel[2], Mary Brunkow[3], Jessie Chang[1], Dmitriy Ovcharenko[1], Ingrid Plajzer-Frick[1], Veena Afzal[1], and Edward M. Rubin[4]

[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; [2]Novartis Pharma, Basel, Switzerland; [3]Celltech Inc., Bothell, WA, USA; and [4]DOE, Joint Genome Institute, Walnut Creek, CA, USA

Sclerosteosis is a generalized progressive bone overgrowth disorder due to the loss of function of the SOST gene product sclerostin. Van Buchem disease is a similar skeletal disorder characterized by milder sclerosteosis-like phenotypes and is associated with the presence of a 52 kb deletion (VBDel) located ~35kb downstream of the SOST transcript and ~10kb upstream of the MEOX1 gene on human chromosome 17p21. Human-Mouse comparative sequence analysis revealed several highly conserved noncoding elements present in the VBDel suggesting that Van Buchem disease is caused by the removal of essential SOST-specific regulatory elements. Using in vitro BAC-recombination techniques we have engineered a ~160kb human BAC containing the SOST and MEOX1 transcripts by removing the ~52kb intergenic region absent in patients suffering from Van Buchem Disease. We have generated several lines of transgenic animals carrying either the wildtype human SOST BAC or the VBDel modified BAC. Following the expression pattern of the endogenous sost mouse gene, we have investigated the expression pattern of the human transgenes in these two types of transgenic animals. Similar to the murine SOST expression, the human SOST transcript from the wildtype BAC is predominantly expressed in the mineralized bones of fetal, neonatal and adult mice, as well as in the apical ectodermal ridge of the developing embryo. Transgenic animals carrying the modified VBDel BAC fail to express the human SOST transcript in mineralized bone, while the embryonic expression of this transgene is unaffected. Using comparative sequence analysis and transient transgenic technology we have testes all the evolutionarily conserved noncoding elements present in the VBDel for the potential to drive expression *in vitro* and *in vivo*. Our findings suggest that Van Buchem disease is caused by a regulatory mutation that diminishes osteoblast-specific expression of the BMP-antagonist sclerostin.

33

Evolutionary Analysis of Enzymatic Functions and Metabolic Pathways

N. Maltsev, **E. Marland** (marland@mcs.anl.gov), A. Rodrigez, D. Sulakhe, R. Krishnamurthy, L. Ulrich, and P. Anumula

Mathematics and Computer Science Division, Argonne National Laboratory

Bioinformatics group at Argonne National Laboratory is developing an integrated computational environment WIT3 for high-throughput analysis of the genomes, metabolic reconstructions and evolutionary analyses of metabolic networks. It includes the following components: a) databases containing sequence, metabolic, and chemical data, b) GADU—an automated pipeline with the scalable backend for high-throughput analysis of the genomes. GADU utilizes distributed computing technology (Globus) and DOE Science Grid and ANL computational resources for analysis of biological data c) rule-based knowledge base for evolutionary analysis of enzymes, and d) tools and algorithms for analysis of protein families developed by our group (e.g. PhyloBlocks, PhE-B, SVMMER). Analysis of 106 prokaryotic genomes is available via WIT3.

34

Analysis of Novel *Deinococcus radiodurans* Mutants following Whole Genome Transcriptome Analysis

Vera Yu. Matrosova[1] (vmatrosova@usuhs.mil), **Marina V. Omelchenko**[1] (omelchen@ncbi.nlm.nih), Amudhan Venkateswaran[1], Min Zhai[1], Mathias Hess[1], Elena K. Gaidamakova[1], Kira S. Makarova[2], Jizhong Zhou[3], and Michael J. Daly[1] (mdaly@usuhs.mil)

[1]Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, Tel: 301-295-3750; [2]National Center for Biotechnology Information, NIH, Bethesda, MD; and [3]Oak Ridge National Laboratory, Oak Ridge, TN

Deinococcus radiodurans R1 (DEIRA) is a Gram-positive aerobic bacterium with an extraordinary resistance to ionizing radiation. Molecular mechanisms underlying this phenotype remain poorly understood. To define the repertoire of DEIRA genes responding to acute irradiation (15 kGy), transcriptome dynamics were examined in cells representing early, middle, and late phases of recovery using DNA microarrays covering ~94% of its predicted genes. At least at one time point during DEIRA recovery, 832 genes (28% of the genome) were induced and 451 genes (15%) were repressed two-fold or greater. All genes were classified according to general expression patterns. Genes induced in the early phase of recovery (displaying a recA-profile) included those involved in DNA replication, repair, recombination, cell wall metabolism, cellular transport, and many encoding uncharacterized proteins. To test if uncharacterized genes implicated by transcriptional profiling contribute to its resistance phenotype, DEIRA mutants were constructed and characterized.

35

Functional Annotation of Human Genes by Gene-Driven Chemical Mutagenesis in Mice

E. J. Michaud[1,2] (michaudejiii@ornl.gov), C. T. Culiat[1,2], Z. Liu[3], K. Krylova[3], F. W. Larimer[1], K. T. Cain[1], D. J. Carpenter[1], L. L. Easter[1], C. M. Foster[1], A. W. Gardner[1], K. J. Houser[1], L. A. Hughes[1], M. Kerley[1], T.-Y. S. Lu[1], R. E. Olszewski[1], I. Pinn[1], G. D. Shaw[1], S. G. Shinpock[1], A. M. Wymore[1], M. L. York[1], E. J. Baker[1], J. R. Snoddy[1], D. K. Johnson[1,2], and E. M. Rinchik[1,2,4]

[1]Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831; [2]The University of Tennessee Oak Ridge National Laboratory Graduate School of Genome Science and Technology, Oak Ridge, TN 37830; [3]SpectruMedix, 2124 Old Gatesburg Road, State College, PA 16803; and [4]Department of Biochemistry, Cellular, and Molecular Biology, University of Tennessee, Knoxville, TN 37996

The availability of the complete DNA sequence of the mouse genome, coupled with the development of high-throughput methods for rapid detection of single-nucleotide polymorphisms (SNPs), have made it practical to consider genome-wide, gene (sequence)-driven approaches to mouse germline mutagenesis. Such gene-driven strategies allow one to perform whole-genome mutagenesis, and then screen for alterations in any pre-selected gene(s). To complement embryonic stem-cell-based gene-driven mutagenesis resources, such as gene-trap libraries and banks of N-ethyl-N-nitrosourea (ENU)-mutagenized ES cells, we have been generating a cryopreserved bank of DNA, tissues (for RNAs and proteins), and sperm from 4,000 C57BL/6JRn mice that each carry a unique load of paternally induced ENU mutations. This ORNL Cryopreserved Mutant Mouse Bank (CMMB) is a source of induced, heritable SNPs in both regulatory regions and coding sequences of virtually every gene in the genome. High-throughput Temperature Gradient Capillary Electrophoresis (TGCE) is used to identify mutations by heteroduplex analysis in pre-selected genes in the CMMB DNA panel, and mutant stocks will be recovered by in vitro fertilization or intracytoplasmic sperm injection from the parallel bank of frozen sperm. Thus, the CMMB will provide mouse models of a wide range of altered proteins for phenotypic, gene/protein-network, and structural biology-type analyses. We will present progress on (i) production of the 4,000-member CMMB (now completed); (ii) methods used for mutation screening by high-throughput TGCE; (iii) our current estimate of the per-base-pair mutation frequency in the CMMB; and (iv) reconstitution of mutant stocks.

Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, and by the Office of Biological and Environmental Research, U.S. DOE, under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC.

36

The Genome of Marine *Synechococcus sp.* Strain WH8102

Brian Palenik[1] (bpalenik@ucsd.edu), Bianca Brahamsha[1], Jay McCarren[1], Eric Allen[1], Eric Webb[5], John Waterbury[5], Fred Partensky[4], Alexis Dufresne[4], Frank Larimer[2], Miriam Land[2], Ian Paulsen[3], and Patrick Chain[6]

[1]Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]The Institute for Genomic Research, Rockville, MD; [4]Station Biologique, Roscoff, France; [5]Woods Hole Oceanographic Institution, Woods Hole, MA; and [6]Lawrence Livermore National Laboratory, Livermore, CA

Cyanobacteria in the open oceans are major contributors to carbon fixation on a global scale. The sequencing and analysis of the genome of marine *Synechococcus sp.* strain WH8102 shows for the first time that these organisms are highly adapted to their oligotrophic marine environment, with relatively small compact genomes and reduced regulatory machinery. WH8102, for example, utilizes more sodium-dependent transporters than a model freshwater cyanobacterium. It also appears to have adopted strategies for conserving limited iron stores by using nickel and cobalt in some enzymes. In contrast to other marine cyanobacteria, however,

WH8102 appears to be more of a generalist, possibly due to its novel ability among cyanobacteria to swim toward nutrient patches. This microorganism is predicted to transport dissolved organic nitrogen (DON) and phosphorus (DOP) sources that are likely present but have been largely ignored to date in phosphorus and nitrogen cycling of oligotrophic environments. The genome of WH8102 appears to have been greatly influenced by horizontal gene transfer, likely through phages. The genetic material contributed by horizontal gene transfer appears to include multiple glycosyltransferases. These may help the cell change its surface glycosylation and thus evade detection by grazers and/or phages. Horizontal gene transfer may have also contributed the genetic material that was used to develop the novel form of swimming motility seen in this strain and closely related cyanobacteria.

37

Genomes to Proteomes to Life: Application of New Technologies for Comprehensive, Quantitative and High Throughput Microbial Proteomics

Richard D. Smith (rds@pnl.gov), James K. Fredrickson, Mary S. Lipton, David G. Camp, Gordon A. Anderson, Ljiljana Pasa-Tolic, Ronald J. Moore, Margie F. Romine, Yufeng Shen, Yuri A. Gorby, and Harold R. Udseth

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352

At present our understanding of biological processes is substantially incomplete; e.g. we do not know with good confidence all the biomolecular players in even the most studied pathways and networks in microbial systems. It is clear that many important signal transduction proteins will be present only at very low levels (~ hundreds of copies per cell) and will provide extreme challenges for current characterization methods. There is also a growing recognition of the limitations associated with gene expression (e.g. cDNA array) measurements. Increasing evidence indicates that the correlation between gene expression and protein abundances can be low, and that the correlation between gene expression and gene function is even lower. Thus, global protein characterization (proteomic) studies actually complement gene expression measurements.

Successes in genome sequencing efforts have provided an informatic foundation for high throughput proteomic measurements to broadly identify large numbers of proteins and their modification states with high confidence, as well as to measure their abundances. The challenges associated with making useful comprehensive proteomic measurements include identifying and quantifying large sets of proteins that have relative abundances spanning more than six orders of magnitude, that vary broadly in chemical and physical properties, that have transient and low levels of modifications, and that are subject to endogenous proteolytic processing. Additionally, proteomic measurements should not be significantly biased against e.g. membrane, large or small proteins. A related need is the ability to rapidly and reliably characterize protein interactions with other biomolecules, particularly their multi-protein complexes. The combined information on protein complexes and the changes observed from global proteome measurements in response to a variety of perturbations is essential for the development of detailed computational models for microbial systems and the eventual capability for predicting their response e.g. to environmental changes and mutations.

We report on development and application of new technologies for global proteome measurements that are orders of magnitude more sensitive and faster than existing technologies. The approaches are based upon the combination of nano-scale ultra-high pressure capillary liquid chromatography separations and high accuracy mass measurements using Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. Combined, these techniques enable the use of highly specific peptide ‘accurate mass and time’ (AMT) tags. This new approach avoids the throughput limitations associated with other mass spectrometric technologies using tandem mass spectrometry (MS/MS), and thus enables fundamentally greater throughput and sensitivity for proteome measurements. Additional new

developments have also significantly extended the dynamic range of measurements to approximately six orders of magnitude and are now providing the capability for proteomic studies from very small cell populations, and even single cells. A significant challenge for these studies is the immense quantities of data that must be managed and effectively processed and analyzed in order to be useful. Thus, a key component of our program involves the development of the informatic tools necessary to make the data more broadly available and for extracting knowledge and new biological insights from complex data sets.

The development of this new technology is proceeding in concert with its applications to a number of microbial systems (initially *Shewanella oneidensis* MR1, *Deinococcus radiodurans* R1, and *Rhodopseudomonas palustris*) in collaboration with leading experts on each organism. This research is providing the first comprehensive information on the nature of expressed proteins by these systems and how they respond to mutations in the organism or perturbations to its environment. Initial studies applying these approaches have demonstrated the capability for automated high-confidence protein identifications, broad and unbiased proteome coverage, and the capability for exploiting stable-isotope (e.g. ^{15}N) labeling methods to obtain high precision relative protein abundance measurements from microbial cultures. These initial efforts have demonstrated the most complete protein coverage yet obtained for a number of microorganisms, and have begun revealing new biological understandings.

Finally, it is projected that the AMT tag approach can also be extended to the characterization of the proteomes of much more complex microbial communities.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC06-76RLO 1830.Rd., Germantown, MD 20874.

38

An Integrated Approach to Functional Annotation of Mammalian Genomic Sequence

Lisa Stubbs[2] (stubbs5@llnl.gov), Xiaochen Lu[1], Joomyeong Kim[2], Aaron Hamilton[1], Nagarajan Lakshmanan[1], Sha Hammond[1], Eddie Wehri[1], Matt Groza[1], Thomas Gulham[1], Mary Tran[2], Tim Harsch[1], Laurie Gordon[2], and Art Kobayashi[2]

[1]Genome Biology Division, Lawrence Livermore National Laboratory and [2]D.O.E. Joint Genome Institute, 7000 East Avenue, L-441, Livermore CA 94550

With the availability of finished human sequence, high quality draft sequence from mouse, rat, and Fugu, and the genomes of vertebrates from other evolutionary branches on the way, we are now in possession of powerful tools for a full functional description of all the genes, regulatory sequences and other functional elements in the human genome. Comparative alignment with sequences from divergent vertebrate genomes has proved to be a powerful tool for distilling out that small fraction of the human genome with critical, evolutionarily conserved functions. The differences between related genomes can also be very revealing, especially when they can be linked to species-specific aspects of biology. We are focused on both the conservation and change of protein-coding genes and the regulatory networks that control their transcription in vertebrate evolution.

Effectively mining conserved elements from alignments of multiple, complex genomes is itself a daunting task, but one for which excellent computational tools have been developed in recent years. Once similarities and differences have been distilled from complex genomes computationally, however, the task of confirming predictions about the functions of these sequences remains a significant experimental challenge. To meet that challenge, we have begun to assemble a suite of experimental tools to test functional predictions regarding conserved human sequences in a high-throughput manner. We have focused on testing the validity of predicted genes and regulatory elements in human chromosome 19

(HSA19), and especially gene-rich chromosome that has recently been finished by JGI teams (J. Grimwood et al., in preparation). To add extra depth to HSA19 genome comparisons, we are generating sequence from related regions of the chicken genome, which because of its position in the evolutionary tree permits particularly informative comparisons.

We are integrating verification of predicted gene structures and the functional testing of candidate regulatory sequences in cell culture with high-throughput methods for determining gene expression in sectioned mouse and human tissues. Our goal is to produce a fully annotated version of HSA19 sequence with all transcription units, promoters and enhancers verified experimentally, with novel genes archived as full-length sequences in expression vectors, with cell-type specific expression patterns determined in both human and mouse, and with lineage-specific conservation and change in coding and non-coding elements fully documented for future study. This project is integrated with related studies ongoing in the laboratory of Barbara Wold, California Institute of Technology, involving a major effort to integrate HSA19 genes into global regulatory networks using microarray expression technology, to develop tools for automated analysis of in situ images, and to test regulatory elements in vivo using high-throughput transgenic methods. Although these studies are focused on specifically on the 60 Mb and 1400 genes of HSA19, the tools we are developing should be extrapolated easily to functional annotation of any complex genome. [Related abstracts including details of specific aspects of this project and closely integrated programs will also be presented; see abstracts by L. Gordon et al.; S. Hammond, N. Lakshmanan et al.; A. Hamilton et al.; and J. Kim et al.].

This work was performed under the auspices of the U. S. Department of Energy, Office of Biological and Environmental Research by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

39

Comparative in silico Proteomes of Two *Brucella* Species

Mary Ann Wagner (wagnerm2@uofs.edu), Michel Eschenbrenner, Frank Estock, Cesar V. Mujer, and Vito G. DelVecchio

University of Scranton, Institute of Molecular Biology and Medicine, Scranton, PA 18510

Recent publications announcing completion of both the *Brucella melitensis* 16M and *B. suis* 1330 genomes provided invaluable information regarding the genomic potential of these important pathogens. As two whole genomes have been sequenced and annotated, researchers are now able to perform in silico comparative genomic analysis of these two organisms. Various in-house and web-based software packages were used to compile theoretical values of physical properties for all proteins of each organism, as determined by the primary annotations for each genome. These properties include pI, Mr, presence of secretory signal sequences, presence of transmembrane domains and grand average of hydropathy (GRAVY) scores. Theoretical parameters were evaluated with respect to species and chromosome of origin. A related member of the alpha proteobacteria, *Agrobacterium tumefasciens* C58, was subject to the same analyses. Chromosome II-encoded proteins of both *B. melitensis* and *B. suis* were found to have a greater average GRAVY score than those of chromosome I, indicating that chromosome II-encoded proteins are generally more hydrophobic than those of chromosome I. Thus, proteins arising from the two different chromosomes have different overall physical characteristics, and monochromosomic origin of both replicons is therefore not favored. In addition to genomic characteristics, such as GC content, it is suggested that overall characteristics of potential proteins may also provide a means to evaluate possible horizontal gene transfer events of large, protein-coding segments of DNA.

40

Signatures for the Detection, Identification and Characterization of Microbial Pathogens

P. Scott White (scott_white@lanl.gov), Lance Green, Murray Wolinsky, Tom Brettin, David Torney, and John Nolan

Los Alamos National Laboratory

The need for robust nucleic acid-based signatures has intensified with the recent focus on the development of tools for biothreat reduction. The availability of whole genome sequence data from pathogens and their neighbors makes it possible to develop signatures using a comparative genomics approach with appropriate levels of resolution and power of exclusion for each typing application.

Single nucleotide polymorphisms, or SNPs, are an abundant source of variation that can be used as signatures, and are amenable to a wide variety of scoring methods and platforms. Furthermore, the use of DNA sequence variation as signatures allows for technology-independent scoring and databases.

We will describe a DNA signature design pipeline that we are developing that makes use of whole genome sequence data. Using comparative genomics tools, targets for candidate signatures are determined, sequence data from the appropriate samples are collected, then phylogenetic analyses distill the signature to a highly informative subset of the total genetic variation discovered.

In addition, we will also describe a high throughput SNP scoring capability that we have recently developed. The method combines robust SNP scoring assays with a flow cytometry platform (i.e. no electrophoresis), and provides rapid scoring of numerous SNPs simultaneously (via multiplexing), with very high serial throughput rates. We will show examples of signature and assay design and implementation using *Bacillus anthracis* and influenza virus sequences.

By combining carefully designed, DNA/RNA sequence-based signatures with rapid typing it is possible to address many of the current and future surveillance, forensic, and clinical diagnostic needs.

41

Oligonucleotide-Directed Single Base DNA Alterations in Mouse Embryonic Stem Cells

Kyonggeun Yoon[1] (kyonggeun.yoon@mail.tju.edu), O. Igoucheva[1], V. Alexeev[1], and E. A. Pierce[2]

[1]Department of Dermatology and Cutaneous Biology, Jefferson Medical College; and [2]F.M. Kirby Center for Molecular Ophthalmology, University of Pennsylvania School of Medicine

We have investigated the use of single-stranded oligodeoxynucleotides (ODN) to introduce specific single-base alterations into endogenous genes in mouse ES cells. The primary advantage of this approach is the ability to introduce a specific base change into a gene of interest in a single step. We have recently demonstrated that ODN can be used to introduce targeted single base changes into the genomic DNA of mouse ES cells at approximately 0.01%. If oligonucleotides were to be used for gene targeting, how can we make it more practical? Low rates of homologous recombination, on the order of 10-5, were overcome by the ingenious use of selectable markers in gene targeting vectors. However, it has been difficult to devise a general selection strategy, because positive and negative selections used in the gene targeting vectors cannot be incorporated into ODN. We hypothesized that cells competent in ODN-mediated alteration of one gene might be also be competent in alteration of other gene. Based on this concept, we developed a selection strategy to identify cells that have undergone a gene modification by the use of two ODNs, one targeting a gene of interest and the other targeting a defective selectable marker gene that manifests a phenotypic change upon gene alteration. Our results indicate that if two oligonucleotides are present within the nucleus of a “repair-competent” cell, then dual

targeting events could possibly occur with a relatively high frequency. Thus, the absolute frequency remains the same level, but the probability of finding cells with the desired gene alteration is increased by first selecting cells according to the phenotypic change. Such selected ES cells could in turn be used to create accurate mouse models of inherited diseases.

42

Microarray-Based Functional Analysis of the Radiation-Resistant Bacterium, *Deinococcus radiodurans*

Jizhong Zhou[1] (zhouj@ornl.gov), Yongqing Liu[1], Dorothea Thompson[1], and Michael Daly[2]

[1]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831; and [2]Department of Pathology, Uniformed Services University of the Health Sciences, Bethesda, MD 20814

Deinococcus radiodurans (DEIRA) is a bacterium best known for its extreme resistance to the lethal effects of ionizing radiation, but the molecular mechanisms underlying this phenotype remain poorly understood. To define the repertoire of DEIRA genes responding to acute irradiation (15 kGy), transcriptome dynamics were examined in cells representing early, middle, and late phases of recovery using DNA microarrays covering ~94% of its predicted genes. At least at one time point during DEIRA recovery, 832 genes (28% of the total predicted genes) were induced and 451 genes (15%) were repressed two-fold or greater. The expression patterns of the majority of the induced genes resemble the previously characterized expression profile of *recA* following irradiation. DEIRA *recA*, which is central to genomic restoration following irradiation, is substantially up-regulated upon DNA damage (early phase) and down-regulated before the onset of exponential growth (late phase). Many other genes were expressed later in recovery, displaying a growth-related pattern of induction. Genes induced during the early phase of recovery included those involved in DNA replication, repair, recombination, cell wall metabolism, cellular transport, and many encoding uncharacterized proteins. Most striking was the observation that metabolic functions, in particular, appear to play crucial roles in DEIRA's recovery from acute radiation. Collectively, the microarray data suggest that DEIRA cells efficiently coordinate their recovery by a complex network that involves the regulation of multiple cellular functions. Components of this network include a predicted novel ATP-dependent DNA ligase, which appears to functionally replace the repressed NAD-dependent DNA ligase, and metabolic pathway switching that could prevent additional genomic damage elicited by metabolism-induced free radicals.

Bioinformatics and Computational Biology

43

Predicting Genes in Prokaryotic Genomes: Are "Atypical" Genes Derived from Lateral Gene Transfer?

John Besemer[1] (jbesemer@biology.emory.edu), Yuan Tian[2], Mark Borodovsky[2], and John Logsdon[1]

[1]Department of Biology, Emory University, Atlanta, Georgia; and [2]Schools of Biology and Biomedical Engineering, Georgia Institute of Technology, Atlanta, Georgia

Algorithmic methods for gene prediction have been developed and successfully applied to many different prokaryotic genome sequences. As the set of genes in a particular genome is not homogeneous with respect to DNA sequence composition features, the GeneMark.hmm program utilizes two Markov models representing distinct classes of protein coding genes denoted "typical" and "atypical." Atypical genes are those whose DNA features deviate significantly from those classified as typical and they represent approximately 10% of any given genome. In addition to the inherent interest of more accurately predicting genes, the atypical status of these genes may also reflect their separate evolutionary ancestry from other genes in that genome. We hypothesize that atypical genes are largely comprised of those genes that have been relatively recently acquired through lateral gene transfer (LGT). If so, what fraction of atypical genes are such bona fide LGTs? We have made atypical gene predictions for all fully completed prokaryotic genomes; we have been able to compare these results to other "surrogate" methods of LGT prediction. In order to validate the use of atypical genes for LGT detection, we are building a bioinformatic analysis pipeline to rigorously test each of the gene candidates within an explicit phylogenetic framework. This process starts with gene predictions and ends with a phylogenetic reconstruction of each candidate. From the set of bona fide LGTs that we have identified, we will be able to determine the LGT parameters to which our gene finding programs are most sensitive (i.e. time scale of transfers, phylogenetic distance from transfer source, etc.). We are utilizing this pipeline to estimate the extent and pattern of LGT in a selection of genomes, both complete and nearly complete, with the long term goal of analyzing all such sequences.

44

VISTA Comparative Genomics at LBNL

N. Bray, O. Couronne, **I. Dubchak** (ildubchak@lbl.gov), L. Pachter, A. Poliakov, D. Ryaboy, and E. Rubin

Genome Sciences Department, Lawrence Berkeley National Laboratory; and University of California, Berkeley

The VISTA Web server (http://www-gsd.lbl.gov/vista) is an integrated set of software tools for comparing two or more genomic sequences. The server consists of two autonomous modules—one for alignment of long genomic sequences, and one for the visualization and identification of conserved elements (Dubchak et al. 2000; Mayor et al. 2000). The VISTA server currently uses AVID, a global alignment program (Bray et al., 2002) that works by first finding maximal exact matches between two sequences using a suffix tree, and then recursively identifies the best anchor points based on the length of the exact

matches and the similarity in their flanking regions.

High quality draft human and mouse genomic sequences have been aligned using a computational strategy where mouse sequence contigs are anchored on the human genome by local alignment matches and then globally extended (Couronne et al., 2003) Alignments on the whole-genome scale can be visualized using an interactive tool Vista Genome Browser accessible at the gateway web site http://pipeline.lbl.gov. Vista Genome Browser is an applet that allows for displaying results of comparative sequence analysis in a VISTA format on the scale of whole chromosomes.

The computational strategy of anchoring sequence contigs from one species onto a base genome sequence assembly of a second species by local alignment matches and then globally aligning these contigs to candidate regions is also implemented for user-submitted sequences at another VISTA server http://pipeline.lbl.gov/cgi-bin/GenomeVista. This server assists in finding candidate orthologous regions for a submitted sequence from any species on either the human or mouse genome sequence assembly, and provides detailed comparative analysis.

Bray, N., Dubchak, I., and Pachter, L. (2003) AVID: A Global Alignment Program. Genome Res. 13:97

Couronne O., Poliakov A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter L, Dubchak, I. (2002) Strategies and Tools for Whole Genome Alignments, 2003. Genome Res., 13:73

Mayor, C., Brudno, M., Schwartz,J.R., Poliakov,A., Rubin, E.M., Frazer, K.A., Pachter, L.S., Dubchak, I. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length. Bioinformatics, 16: 1046

Dubchak, I., Brudno, M., Pachter, L.S., Loots, G.G., Mayor, C., Rubin, E.M., Frazer, K.A. (2000) Active conservation of noncoding sequences revealed by 3-way species comparisons. Genome Res., 10: 1304

45

Cell Cycle Regulation Model Construction Using Trainable Neural Networks

E. Mjolsness[1] (emj@uci.edu), T. Vinogradova[2], C. Hart[2], and B. Wold[2]

Institute for Genomics and Bioinformatics, University of California, Irvine; and [2]Caltech Division of Biology

We use trainable neural network models to combine the yeast ChIP-chip transcription factor binding data of Lee et al. with the yeast cell cycle microarray expression data of Cho et al., arriving at hypotheses for a small core network involved in transcriptional regulation in the cell cycle. The stages of analysis may be outlined as (a) finding a reliable clustering of the expression time course data, (b) finding a robust set of genes whose expression class is predictable based on binding data, (c) finding a robust set of regulators most involved in this prediction for each class, and (d) optimizing a small, trainable neural network model of transcriptional regulation using the foregoing steps. The number of free parameters of the model is almost as great as the amount of data available to constrain it, so the method is near the boundary of current feasibility. However, network structures emerge robustly and a modification to the form of the assumed dynamics can be proposed based on the fits to existing data.

46

Fast Alignment & Analysis of Multiple Genomes

Gary R. Montry (montry@spsoft.com) and Don A. North

Southwest Parallel Software Thoughtware

We present a program which uses the sensitive Smith-Waterman alignment algorithm, but which is faster than BLAST, to align multiple genomes. We also present a novel viewer which allows the user to select and view multiple genome alignments of highly-conserved regions. We show performance comparisons between BLAST and our high-speed Smith-Waterman core.

47

Engineering Tools to Characterize the Coding Regions of the Genome

Michael B. Murphy (mbmurphy@lbl.gov), Shirin Fuller, Sharon A. Doyle, Paul M. Richardson, and Eddy Rubin

DOE Joint Genome Institute, Walnut Creek, CA 94598

With genome sequencing efforts producing vast amounts of data, attention is now turning towards unraveling the complexities encoded in the genome: the protein products and the cis-regulatory sequences that govern their expression. Understanding the spatial and temporal patterns of protein expression as well as their functional characteristics on a genomic scale will foster a better understanding of biological processes from protein pathways to development at a systems level. Currently, the main bottlenecks in many proteomics initiatives, such as the development of protein microarrays, remain the production of sufficient quantities of purified protein and affinity molecules or probes that specifically recognize them. Methods that facilitate the production of proteins and high affinity probes in a high-throughput manner are vital to the success of these initiatives. We have developed a system for high-throughput subcloning, protein expression and purification that is simple, fast and inexpensive. We utilized ligation-independent cloning with a custom-designed vector and developed an expression screen to test multiple parameters for optimal protein production in *E. coli*. A 96-well format purification protocol was also developed that produced microgram quantities of pure protein. These proteins were used to optimize SELEX (Systematic Evolution of Ligands by Exponential Enrichment) protocols that use a library of DNA oligonucleotides containing a degenerate 40mer sequence to identify a single stranded DNA molecules (aptamers) that bind their target protein specifically and with high affinity (low nanomolar range). Aptamers offer advantages over traditional antibody-based affinity molecules in their ease of production, regeneration, and stability, largely due to the chemical properties of DNA versus proteins. These aptamers were characterized by surface plasmon resonance (SPR) and were shown to be useful in a number of assays, such as western blots, enzyme-linked assays, and affinity purification of native proteins.

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, under Contracts No. W-7405-Eng-48, No. DE-AC03-76SF00098, and No. W-7405-ENG-36.

48

Computational Analysis of Gene Deserts in the Human Genome

Marcelo A. Nobrega[1], **Ivan V. Ovcharenko**[1], Gabriela G. Loots[2], and Edward M. Rubin[3]

[1]Life Sciences Division, Lawrence Berkeley National Laboratory, MS 84-171, Berkeley, CA 94720; [2]Genomics Division, Lawrence Livermore National Laboratory, 7000 East Ave., Livermore CA 94550; and [3]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598

The sequencing and annotation of the human genome revealed a non-uniform gene distribution across different chromosomes, implying on the existence of vast genomic segments devoid of identifiable genes, or gene deserts. As an initial step to characterize these noncoding sequences and derive insights into their evolution and possible biological function, we computationally identified all the gene deserts present in the human genome and compared them with other average-sized intergenic regions as well as with homologous genomic segments from the mouse and the pufferfish *Fugu rubripes*, organisms whose genomes have recently been sequenced to completion. Our analysis revealed that gene deserts correspond to ~10% of the total size of the human genome, ranging in size between 600kb and ~3Mb. Using various computational approaches we compared the density of repetitive elements, GC content, SNP density and human-mouse conservation between gene deserts and non-desert intergenic regions. We observed a wide distribution for each measured parameter among different gene deserts, without a distinct signature shared by the majority of these noncoding regions. On average, we found the age of the repetitive elements in gene deserts to be younger than that of any other genomic fractions that we analyzed, possibly reflecting a higher incidence of deletions that swipe out older repeats in gene deserts than in other parts of the genome. The vast majority of human gene deserts are represented by corresponding gene deserts in the mouse genome and about half of these gene deserts carry sequences that are conserved in the *Fugu rubripes* genome. Interestingly, genes involved in various aspects of embryonic development flank most of the gene deserts containing fugu fish conservation, suggesting that these regions are embedded with transcriptional regulatory elements. Here, we will depict a "functional" gene desert, carrying several gene regulatory elements that could only be identified by comparing human, mouse and fugu sequences. We also outline an ongoing strategy for generating genetically engineered mice carrying deletions of gene deserts that will test the in vivo function of these large segments of noncoding DNA.

49

Decoding Transcriptional Regulation in the Human Genome

Ivan Ovcharenko[1] (ivovcharenko@lbl.gov), **Roded Sharan**[2], Asa Ben-Hur[3], Eddy Rubin[4], and Richard M. Karp[2]

[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [2]International Computer Science Institute, 1947 Center St., Suite 600, Berkeley, CA 94704; [3]Department of Biochemistry, B400 Beckman Center, Stanford University, CA 94305; and [4]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598

Transcriptional gene regulation in the human genome is complex in nature. Spatial and temporal gene expression patterns are defined by a combinatorial interplay of several transcription factors binding to promoter region of a gene. While microarray experiments unveiled tissue- and conditions-specific patterns of gene expression for many genes, there is still a lack of knowledge of the underlying sequence motifs that induce the observed gene expression patterns. We present a novel method for detecting cis-regulatory modules in the promoter regions of human genes using genome-scale alignment of the human and mouse genomes. Initially, transcription factor binding sites (TFBS) were identified in the promoters of all annotated RefSeq human transcripts based on more than 400 TFBS profiles catalogued in the TRANSFAC database. From an overwhelming number of predicted TFBS the majority of which are false positives, we extracted only those that are aligned and conserved in human and mouse, using the

rVista tool (http://nemo.lbl.gov/rvista/index.html). New statistical measures were developed for pinpointing TFBS that are enriched in the promoters of a group of genes of interest compared to the background set. A novel hashing algorithm and appropriate statistical tests were devised to identify groups of TFBS that tend to co-occur in the promoters of interest. We applied our method to find regulatory modules related to cell-cycle and stress response. On the cell cycle data our algorithm identified several relevant TFs, including E2F, and seven cis-regulatory modules that are statistically significant. The sets of genes containing each of the modules were verified by checking for coherence of their expression patterns. Roughly half of the identified sets of genes were found to be significantly coherently expressed. On the stress response data about half of the detected gene sets fell predominantly into well-defined functional sub-categories.

50

Mining the Frequency Distribution of Transcription Factor (TF) Binding Sites in Promoters of Suppressed and Enhanced Genes During Human Adaptive Response to Ionizing Radiation

Leif E. Peterson[2] (peterson@bcm.tmc.edu), Ilkay Altintas[3], Bertram Ludaescher[3], Terrence Critchlow[1], Andrew J. Wyrobek[1], and Matthew A. Coleman[1]

[1]Biology & Biotechnology Research Program, L-452, Lawrence Livermore National Laboratory, Livermore CA, 94551; [2]Department of Medicine, Baylor College of Medicine, Houston, TX. 77030; and [3]San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093

Motivation: Through the Human Genome Sequencing Project a wealth of information has been gained at the nucleotide level. With the advent of DNA-based microarrays the amount of data available for interpretation is quickly becoming daunting. A starting point for discovery is to better link genomic biology approaches with bioinformatics to identify and characterize eukaryotic promoters. For example, microarray experiments use various cluster analysis algorithms to identify genes that share similar patterns of gene expression profiles that are predicted to be co-regulated as part of an interactive biochemical pathway. Further identification and characterization of DNA consensus sequences, regulatory elements that regulate the responsive genes could provide a valuable understanding of the genetic and biochemical mechanisms of the cell and should provide powerful biological indicators of genetic susceptibilities for tissue and genetic damage. To clearly identify these co-regulated groups of genes, we describe scalable computational workflow approaches that use web-based molecular biology tools and schemas for carrying out a variety of tasks such as hierarchical clustering, comparison of DNA sequences, and identification of transcription factor binding sites, comparison of clustered promoters and visualization of compiled data.

Results: We start an example workflow using cluster analysis of Affymetrix U95A array results for human adaptive response to ionizing radiation. Replicate gene expression data were based on lymphoblastoid cell lines derived from a radiosensitive “non-adapter” and two “adapters.” Exon mapping was performed to extract promoters (3kb upstream) for a cluster of mRNAs enhanced in only the non-adapter and a cluster of mRNAs enhanced in only the adapters. A search of transcription factor (TF) binding sites in the extracted promoters was then ran, resulting in a frequency distribution of all identified TF binding sites across all promoters. We then clustered the promoters based on frequency of TF binding sites. The resulting cluster image display shows a discernable pattern in TF binding site frequency that can be further mined for relevance to regulatory control of expression. There are many other possible ways to use microarray data for exploring relationships between co-expressed genes within and between species in order to infer co-regulation—and some of these will be discussed.

51

A Scalable Visual Data Analysis Pipeline Framework Supporting Large-Scale Bioinformatics Research

Dong-Guk Shin[1] (shin@engr.uconn.edu), Ravi Nori[2], Jae-Guon Nam[1], Jeffrey Maddox[1], and Hsin-Wei Wang[1]

[1]Computer Science and Engineering Department, University of Connecticut, Storrs, CT 06269; and [2]CyberConnect EZ, LLC, Storrs, CT 06268

One key challenge in supporting large scale bioinformatics research is developing a computational environment in which scientists can easily use various types of bioinformatics resources available in diversified platforms and locations. Such resources include databases, data available through third party web sites, files downloadable from ftp sites, analysis programs, format conversion routines, commonly usable scripts, high performance computing facilities, etc. A novel software framework has been developed that effectively harness all these diversified resources for ease of use by the bioinformaticists. This framework establishes a clear division of labor between the support core whose primary function is to develop and maintain computational resources and the scientists who uses such resources to conduct various bioinformatics analysis tasks. The support core configures the environment by interlinking various available resources. The scientists reap the benefits of the support core's resource integration. This framework relies on a distributed architecture allowing computational resources to be scattered around LAN, WAN and Internet. This framework also liberally adopts a visual iconic solution in user interface design and this easy-to-use feature presents a great potential for significantly improving scientists' research productivity.

This novel software framework has been deployed and is currently under use for supporting a large scale bioinformatics activity. It is being used to develop a set of semi-automated pipelines designed to aid scientists in selecting non-redundant clones from the NIH cDNAs consortia library. The analysis steps of this pipeline include use of the publicly available EST repository, dbEST, elimination of redundant EST sequences by using a pre-built UniGene information, conducting pair-wise Blast sequence comparisons, and finally use of non-redundant clones selection heuristics.

Another use of the framework is to do microarray data analyses. A series of modular pipelines have been developed. One module addresses the quality control issue using global and intensity methods. One module is responsible for normalization of raw data produced from image analysis. Several modules are responsible for conducting expression level analysis including clustering and promoter analyses. The first two modules encapsulate use of conventional statistical packages in the pipeline. The expression level analysis modules include use of public/commercially available microarray data analysis packages as well as custom developed visualization programs.

We have also tested the feasibility of using the framework on a high speed sequence assembly work at JGI. This pipeline module includes use of RepeatMasker and Blast (running on a Linux cluster using MPI) in tandem. In an attempt to further demonstrate the scalability of this visual pipeline framework, we are planning a closer collaboration with JGI and LLNL in which LLNL's supercomputers are fully accessed through the easy-to-use visual analysis pipeline interface.

This work was supported in part by DOE SBIR Phase II Grant No. DE-FG02-99ER82773.

52

JGI Human Chromosome 19 Annotation

Astrid Terry (terry7@llnl.gov), Laurie Gordon, Ivan Ovcharenko, Andrea Aerts, Uffe Helsten, Wayne Huang, Isaac Ho, Victor Solovyev, Duncan Scott, Steve Lowry, Olivier Couronne, Sam Rash, Paramvir Dehal, Inna Dubchak, Lisa Stubbs, and Dan Rokhsar

Computational Genomics, DOE Joint Genome Institute, 2800 Mitchell Dr, B400, Walnut Creek, CA 94598

The JGI has been mandated to finish and annotate human Chrs 5, 16, and 19. The final finished sequence for Chromosome 19 (Chr 19) was received in the middle of February and an automated pipeline for generating annotation has been developed. Gene models are built in many different ways: using experimentally known human mRNAs, EST/protein-seeded GeneWise, GenomeScan, FgenesH and GrailEXP. Annotators choose the best automated models using a hierarchy of evidence. Additionally, finished sequence is reviewed for evidence of any single base indels (we have corrected multiple finishing errors this way) and 5'/3' UTRs are extended by spliced ESTs or compatible redundant mRNAs. The syntenic mouse mRNA libraries augment the human mRNA libraries, using human exons when available. Alternative splicing is only reported if supported by at least nearly complete mRNA. High quality evidence is manually reviewed if no models can be created in the automated system. So far there are roughly 100 problematic loci, with various issues. Some of the challenges on Chr 19 include large gene families with known gene structure lacking extensive human mRNA/EST evidence and tandemly duplicated genes. Using expected gene structure and known properties, custom models are built for genes and pseudogenes, which are common in gene families. Web based interfaces allow annotators to view a predicted peptide's properties and aid in putative function assignment based on pre-computed alignments of homology and domains. Using Chr19 as a model system, the other chromosomes are expected to follow shortly.

53

Request Handling Web Application Using JAVA Struts: Separation of Presentation and Transaction/Data Layer

Qing Zhang (qzhang@lbl.gov), Nate Slater, Heather Kimball, Ivan Ovcharenko, Susan Lucas, Jan-Fang Cheng, and Eddy Rubin

DOE Joint Genome Institute, Walnut Creek, CA 94598

DOE Joint Genome Institute (JGI) is currently soliciting requests to sequence genomic regions of strong scientific value. To qualify for this program, the sequencing regions need to be contained in individual BACs, cosmids, or fosmids (from any organism). The applicant must also provide the clone for sequencing; however, JGI may screen available libraries to identify appropriate clones. The goal of this program is to focus on issues requiring long stretches of genomic sequence, and not to sequence small DNA fragments.The reviews will be conducted every two months by a panel of biologists who are very familiar with the DOE's missions and research programs. All approved regions will be listed on the status page. All sequence reads will be generated using either a MegaBACE or a ABI3730 instrument. The raw/assembled sequences and analysis results will be provided on our ftp site.

The request can be submitted through our web application, which is developed in Jakarta Struts framework. Struts provides an open-source framework for creating web applications that easily separate the presentation layer and allow it to be abstracted from the transaction/data layers. Here we present step-by-step illustration of the request handling web application.

54

Target Selection in *Ciona* Whole Genome Enhancer Screening: Algorithm and Visualization

Qing Zhang[1] (qzhang@lbl.gov), David N. Keys[1], Buying-in Lee[1], Mike Levine[2], and Paul Richardson[1]

[1]DOE Joint Genome Institute, Walnut Creek, CA 94598; and [2]Department of Molecular and Cellular Biology, University of California at Berkeley, Berkeley, CA 94720

To characterize gene regulatory network, we used electroporation assays to screen genomic DNA fragments for tissue specific enhancer activities in *Ciona intestinalis*. The *Ciona* genome is one of the smallest of all chordate genomes and *Ciona* tadpole represents the most simplified chordate body plan.

We designed the methodology of selecting targets for *Ciona* enhancer screening. In this computational approach, the forward and reverse sequencing reads are connected to create virtual clone sequences. This large pool of virtual clone sequences is then used to generate a blastable database. BLAST analysis aligns these clone sequences to large genomic DNA segments. We designed the algorithm to select the minimum tiling path clones to cover these genomic segments. Gaps are considered and artificial clones for filling gaps are suggested by the algorithm. A web application is set up for running the selection behind the scene. An index page is automatically updated once the tiling path clones are calculated. The web application is also set up for checking clone coverage, display tiling path clones graphically, propose gapping clones' sequences, and dump the selected tiling path clones into the next step workflow system.

55

The Commercial Viability of EXCAVATOR: A Software Tool for Gene Expression Data Clustering

Robin Zimmer[1] (robzimmer@apocom.com), Morey Parang[1], Dong Xu[2], and Ying Xu[2]

[1]ApoCom Genomics and [2]Oak Ridge National Laboratory

ApoCom Genomics, in collaboration with Oak Ridge National Laboratory, is being funded under a DOE Phase I SBIR Grant (DE-FG02-02ER83365) to assess the commercial viability of a novel data clustering tool developed by Drs. Ying Xu, Victor Olman and Dong Xu (Xu, et.al., 2001). As we enter into an era of advanced expression studies and concomitant voluminous databases, there is a growing need to rapidly analyze and cluster data into common expression and functionality groupings. To date, the most prevalent approaches for gene and/or protein clustering have been hierarchical clustering (Eisen et.al., 1998), K-means clustering (Herwig et al., 1999), and clustering through Self-Organizing Maps (SOMs) (Tamayo et al.,1999). While these approaches have all clearly demonstrated their usefulness, they all have inherent weaknesses. First, none of these algorithms can, in general, rigorously guarantee to produce globally optimal clustering for any non-trivial objective function. Moreover K-means and SOMs heavily depend upon the 'regularity' of the geometric shape of cluster boundaries, and they generally do not work well when the clusters cannot be contained in some non-overlapping convex sets.

For cases where boundaries between clusters may not be clear, an objective function addressing more global properties of a cluster is needed. Three clustering algorithms, along with a minimum spanning tree (MST) representation, have been implemented within a computer program called EXpression data Clustering Analysis and VisualizATiOn Resource (EXCAVATOR™). Our research team has conducted a comparison between the EXCAVATOR™ clustering algorithm and the widely used K-means clustering algorithm using rat central nervous system (CNS) data.

Two criteria were employed for the comparison. The first was based on the jackknife approach to assess the predictive power of the clustering algorithm, and the second was based on the separability quality of clusters. All three of the EXCAVATOR™ algorithms (MST-hierarchical, MST-iterative, and MST-global optimal) outperformed the K-means algorithm relative to predictive power and separability quality.

In addition to comparative studies to assess the usefulness of EXCAVATOR™, the team has developed an advanced graphical user interface (GUI). The GUI has been designed to afford maximum flexibility incorporating the multi-clustering data visualization, as well as user driven comparison and editing capabilities. EXCAVATOR's™ data visualization component is based on a modular/flexible approach so as to extend its capability to other clustering/classification areas, such as phylogeny, sequence motif recognition, and protein family recognition.

References

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl Acad. Sci. USA, 95, 14 863-14 868.

Herwig, R., Poustka, A.J., Müller, C., Bull, C., Lehrach, H. and O'Brien, J. (1999) Large-scale clustering of cDNA-fingerprinting data. Genome Res., 9, 1093-1105.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl Acad. Sci. USA, 96, 2907-2912.

Xu, Y., Olman, V. and Xu, D. (2001) Clustering Gene Expression Data Using A Graph-Theoretic Appraoch: An Application of Minimum Spanning. Bioinformatics. Vol.18, no.2002.

Environmental Genomics

56

Community Genomics-Enabled Study of a Low Complexity, Geochemically-Simple Acid Mine Drainage Ecosystem

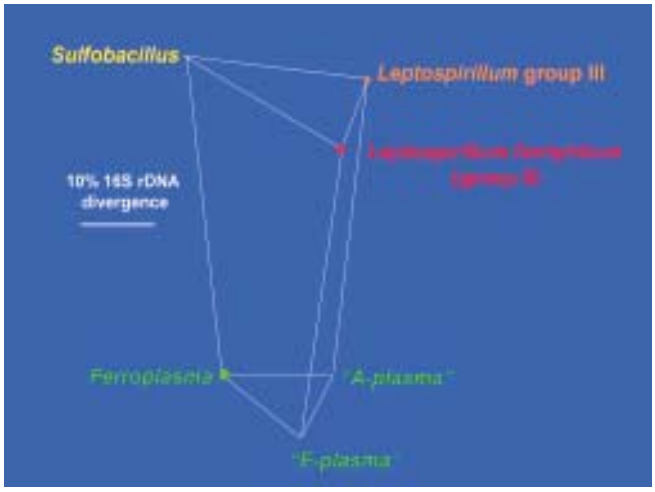
Gene W. Tyson, Philip Hugenholtz, and **Jillian F. Banfield** (jill@eps.berkeley.edu)

Department of Environmental Science, Policy and Management, and Earth and Planetary Sciences, University of California, Berkeley, CA; and the DOE Joint Genome Institute, Walnut Creek, CA

Subsurface acid mine drainage (AMD) ecosystems are ideal models for the genome-enabled study of microbial ecology and evolution because they are physically isolated from other ecosystems and are relatively geochemically and biologically simple. We are using culture-independent genome sequencing of an AMD community from the Richmond mine, Iron Mountain, CA, to evaluate the extent and character of lateral gene transfer (LGT) within the community and to resolve microbial community function at the molecular level.

Microbial communities exist in several distinct habitats within the Richmond mine, including biofilms (subaqueous slime streamers and subaerial slimes) and cells attached directly to pyrite granules. All communities investigated to date by 16S rDNA clone libraries comprise only a handful of phylogenetically distinct organisms, typically dominated by the iron-oxidizing genera *Leptospirillum* and *Ferroplasma*. A *Leptospirillum*-dominated biofilm community was chosen for detailed analysis. 16S rDNA clone libraries and fluorescence in situ hybridization (FISH) using group-specific oligonucleotide probes indicated that the community is comprised of only 6 prokaryotic populations (see Figure; the size of colored circles indicates 16S rDNA divergence within populations).

We analyzed 7 Mbp of community genome sequence data from a 3 Kb shotgun library of the biofilm to estimate the community genome size. This analysis used an implementation of the Lander-Waterman equation that took into consideration species abundances determined from the data and by FISH. Results indicate a community genome size of approximately 19 Mbp, suggesting that each population is dominated by a single genome type. The conclusion was robust, even when assembly criteria were varied to improbable extremes and large uncertainties in population structure were included. However, more sequence data are needed to statistically validate the finding. Furthermore, the analysis is insensitive to genome types that occur in low abundance. One isolated member of the AMD biofilm community, *Ferroplasma acidarmanus*, has been previously sequenced by the JGI and could be used as an internal calibration for the community genome size estimate. We found that approximately 350,000 bp of the 7 Mbp community genome sequence data had ≥98% sequence identity to the *F. acidarmanus* genome, i.e. an ca. 0.2X coverage of this 1.5 Mbp genome. Therefore, a 1X coverage of this genome would



require 35 Mbp of community genome data, which can be viewed as a crude estimate of the community genome size. These results suggest that reassembly of the dominant genomes of AMD community members will be tractable with a modest sequencing effort. The apparent population homogeneity may arise due to the specific characteristics of the AMD habitat or may be a widespread phenomenon in microbial ecosystems.

Similarity searches of the initial community genome sequence data revealed many genes consistent with the chemoautotrophic lifestyle of the community, including CO_2 fixation genes, nitrogen fixation genes and heme biosynthesis genes (heme is responsible for the pink color of the biofilm). None of these genes had close matches to genes from the *F. acidarmanus* genome and probably belong to *Leptospirillum* genomes. The most represented functional categories in the community genome data are DNA replication, recombination and repair (19%; due in part to a large number of transposases that belong to this group), amino acid transport and metabolism (12%) and energy production and conversion (11%). Organism-resolved metabolic pathway information will be used to develop methods to monitor microbial activity in the environment.

LGT is thought to play a crucial role in the ecology and evolution of prokaryotes. The extreme conditions (pH < 1.0, molar concentrations of iron sulfate and mM concentrations of arsenic, copper and zinc, and elevated temperatures of up to 50° C) largely isolate the AMD community from most potential gene donors. Naked DNA, phage and prokaryotes native to neutral pH habitats do not persist at pH <1.0, precluding influx of genes by transformation, transduction and conjugation, respectively. However, prophage have been recognized in the *Ferroplasma* genome sequence and acidophilic phage have been detected in the biofilm community. Phage may be important vectors for gene exchange. We have initiated a collaboration to sequence the phage community to assess their diversity and enhance our ability to detect prophage in the prokaryote community genome data.

Comparative genome analyses indicate that *F. acidarmanus* and the ancestor of two acidophilic *Thermoplasma* species belonging to the *Euryarchaeota* have traded many genes with phylogenetically remote acidophilic *Sulfolobus* species (*Crenarchaeota*). The putatively transferred sets of *Sulfolobus* genes in *Ferroplasma* and the *Thermoplasma* ancestor are distinct, suggesting independent LGT events between organisms living in the same, and adjacent habitats. In both cases, however, the majority of transferred genes are involved in metabolism, particularly energy production/conversion and amino acid transport/metabolism. The lack of genes transferred from the (sequenced) genomes of other prokaryotes is consistent with the hypothesis that extreme acidophiles have limited access to genes from organisms outside their ecotype. Interestingly, *Sulfolobus*, *Ferroplasma* and *Thermoplasma* are all bounded by a single tetraether-dominated membrane, which may facilitate conjugation. To date, no *Sulfolobus* species have been detected at Iron Mountain, suggesting two possibilities to explain the observed pattern of putatively transferred genes to *Ferroplasma* from *Sulfolobus*: 1) *Sulfolobus* is present at Iron Mountain but in regions currently inaccessible to sampling and/or 2) the transfers occurred prior to introduction of *Ferroplasma* into the current geological setting. Comparative analyses of the community genome data should improve the resolution of LGT in the community.

Ultimately, our goal is to develop an understanding of how acidophilic organisms evolved and function as communities to control acid mine drainage generation. The community genomics data are essential for this effort.

Technology Development

57

Developing a Lox-Based Recombinatorial Cloning System for Ligand Libraries

Robert Siegel[1], Nileena Velappan[2], Peter Pavlik[2], Leslie Chasteen[2], and **Andrew Bradbury**[2] (amb@lanl.gov)

[1]Pacific Northwest National Laboratory and [2]Los Alamos National Laboratory

The selection of binding ligands (e.g. single chain Fvs - scFvs) against protein targets can be done using a number of different systems, including phage, phagemid, bacterial or yeast display vectors. Genetic selection methods have also been developed based on yeast two hybrid and enzyme complementation systems. In general, selection vectors are not suitable for subsequent production. Furthermore, once scFvs have been selected, they can be usefully modified by cloning into other destination vectors (e.g. by adding dimerization domains, detection domains, eukaryotic expression in eukaryotic vectors etc.). However, this is relatively time consuming, and requires checking of each individual construct after cloning. An alternative to cloning involves the use of recombination signals to shuttle scFvs from one vector to another. These have the advantage that DNA restriction and purification can be avoided. Such systems have been commercialized in two general systems: Gateway™, uses lambda att based recombination signals, while Echo™ uses a single lox based system to integrate a source plasmid completely into a host plasmid.

We have examined the potential for using heterologous lox sites and cre recombinase for this purpose. Five apparently heterologous lox sites (wild type, 511, 2372, 5171 and fas) have been described. A GFP/lacZ based assay to determine which of these were able to recombine with each other was designed and implemented. Of the five, three (2372, 511 and wt) were identified which recombined with one another at levels less than 2%.

To use recombination as a cloning system, it is important to be able to select against host vectors which do not contain the insert of interest. Two toxic genes were examined for this purpose. The tetracycline gene confers sensitivity to nickel, while the sacB gene confers sensitivity to sucrose. We confirmed these sensitivities, although found that some antibiotic resistances interfere with survival of bacteria hosting non-tetracycline containing plasmids.

In preliminary experiments we have demonstrated that recombination from one plasmid to another, using 2272 and wild type lox sites and sacB or tetracycline, can occur in vivo at very high efficiency. This opens the possibility of using this system to easily transfer scFvs after selection to other plasmids. However, the utility of this system is not limited to scFvs - any DNA fragment (gene, open reading frame, promoter etc.) can easily be shuttled from one plasmid to another using these lox based signals.

58

Towards High-Throughput Selection of Binding Ligands

Milan Ovecka, Nileena Velappan, Leslie Chasteen, Peter Pavlik, and **Andrew Bradbury** (amb@lanl.gov)

Los Alamos National Laboratory

Phage display libraries represent a relatively easy way to generate binding ligands against a vast number of different targets. Although in principle, phage display selection should be amenable to automation, this has not yet been described and present selection protocols are far from high throughput. We have examined the selection process in a systems approach and attempted to automate each individual step. Selection is carried out in the microtiter format using 24 targets as the individual selection lot size. Output is plated onto large assay trays, and a program to pick colonies in specific orders corresponding to the selection arrangement is in the process of being developed for the Qbot picking robot. High density dot blots (400 clones in the footprint of 4 "96 well" wells) as a first round clone testing method is in the process of being developed, while ELISA as final confirmation has been completely automated in the 384 well format for the Tecan Genesis workstation.

59

Fluorobodies: Fluorescent Binding Ligands for Genomic Studies

Ahmet Zeytun, Geoff Waldo, and **Andrew Bradbury** (amb@lanl.gov)

Los Alamos National Laboratory

Antibodies are the most widely used binding ligands in research, and recent molecular diversity techniques (e.g. phage display) permit the generation of antibody fragments, comprising the binding domain alone, without the use of animals. As a result, much hope has been placed in the idea of creating genome-wide panels of antibodies selected with high throughput procedures. Such antibodies could be used in the high throughput genome wide study of gene products by immunofluorescence, immunoprecipitation and western blotting, as well as novel applications such as antibody chips and intracellular inhibition studies. While we are in the process of automating the selection of antibodies against different targets, and feel this is feasible, we have noted that antibody fragments suffer from a number of problems, foremost among these is the inability to detect binding without the use of secondary enzymatically, or fluorescently, labeled reagents. Other problems include relatively poor expression levels in bacteria and poor stability.

The use of GFP as a binding scaffold, rather than antibodies, would resolve many of these problems. However, due to destabilization of GFP folding upon the insertion of extraneous sequences, attempts to engineer standard GFP have been unsuccessful to date. We have overcome these problems with two essential modifications. The first involves the use of a novel form of GFP (superfolder GFP) which is far more stable than traditional GFP. The second modification involves the use of CDRs from antibodies as diversity elements, rather than random peptides encoded by oligonucleotides. We have created a small library of 5e6 fluorobody clones and displayed them on phage. From this small library we selected specific binders for a number of different targets. These fluorobodies bind their targets specifically as shown by band shift assays, immunofluorescence, ELISAs and crude microarrays. The affinities are similar to those expected from antibody libraries of similar sizes, and they can be expressed at very high levels (100mg/L) compared to antibodies. We have also shown that loss of fluorescence is associated with loss of binding function, permiting easy monitoring of their use. They hold tremendous potential in genomics, proteomics, diagnostics and drug screening.

60

Microbioreactor Arrays with Parametric Control for High-Throughput

Michel Maharbiz, William Holtz, Afshan Shaik, Roger Howe, and **Jay D. Keasling** (keasling@socrates.berkeley.edu)

Departments of Electical Engineering and Computer Sciences and of Chemical Engineering, University of California, Berkeley, CA 94720

We present a scalable array technology for parametric control of high-throughput cell cultivations. The technology makes use of commercial printed circuit board (PCB) technology, integrated circuit sensors, and an electrochemical gas generation system. We present results and for an array of eight 250 μl microbioreactors. Each bioreactor contains an independently addressable suite that provides closed-loop temperature control, generates feed gas electrochemically, and continuously monitors optical density. The PCB technology allows for the assembly of additional off-the-shelf components into the microbioreactor array; we demonstrate the use of a commercial ISFET chip to continuously monitor culture pH. The electrochemical dosing system provides a powerful paradigm for gas delivery to high-density arrays of microreactors. We show growth data for *Escherichia coli* cultured in the array with varying microaerobic conditions using electrochemically generated oxygen. Additionally, we present data on carbon dioxide generation and pH control.

61

Selective Genotyping of Individual Cells by Capillary Polymerase Chain Reaction

Hanlin Li and **Edward S. Yeung** (yeung@ameslab.gov)

Ames Laboratory, Iowa State University, Ames, IA 50011

On-line capillary polymerase chain reaction (PCR) coupled with laser-induced fluorescence detection was successfully demonstrated for individual human cells. A single 50-mm i.d. fused-silica capillary served both as the reaction vessel and for isolating single cells. SYBR Green I dye was added into the reaction mixture for dynamic fluorescence labeling. Because of the small i.d. of the capillary, PCR-amplified DNA fragments from single cells were localized in the capillary, providing discrete product zones with concentrations at readily detectable levels. With selective primer design, only cells containing the DNA of interest were amplified. By counting the number of peaks in the capillary via electromigration past a detection window, the number of targeted cell templates could be determined. Identification of the 295-bp fragment beta-actin gene from individual human lymphoblast cell was demonstrated. Independent on-column cell counting provided positive correlation between the starting cell templates and the final PCR products. This opens up the possibility of highly selective and sensitive disease diagnosis at an early stage, when only a few cells in the population are defective.

Ethical, Legal, and Social Issues

62

Solutions to the Anticommons in Genome Patenting: Recent Events

David J. Bjornstad[1] (dub@ornl.gov) and Lee A. Greer[2]

[1]Oak Ridge National Laboratory and [2]PNC Financial Services Group, Inc.

Since spring of 1998, when concerns over the implications of patenting gene fragments for access to the base genome sequence first surfaced, a number of changes have occurred that have as their goal overcoming inefficiencies due to what Heller and Eisenberg described as a "tragedy of the anticommons." The anticommons, too many property rights issued for too few goods, allegedly restricts access to the base genome through a combination of incentives, transactions costs, and institutional rigidities. Since 1998, the base genome has been completed and published, major private sector players have revised their business plans, the U.S. Patent Office has issued new guidance for evaluating gene fragment patent applications, new analysis of the anticommons has occurred and new data have been collected, the Justice Department has revised its Antitrust policy that potentially governs policy options in this area, and a number of scholars have checked in with new suggestions for reducing anticommons costs. This paper reviews these events and arguments, and offers a unique, if untested, solution to the anticommons in base genome patenting policy.

63

The UC Discovery Grant

David Gilbert (dgilbert@uclink.berkeley.edu)

University of California, Office of The President

The UC Discovery Grant, awarded by the Industry-University Cooperative Research Program (IUCRP), creates a 3-way partnership between UC, Industry, and the State of California, and help advance research and education while simultaneously strengthening the competitiveness of California businesses.

Launched by the State of California in 1996, the IUCRP, together with Industry and State contributions, invests up to $60 million a year in UC Discovery Grants to encourage research at UC campuses in collaboration with California companies. The program is unusual in its emphasis on early-stage investigations that promise to yield new products and technologies and boost California's economic productivity.

The UC Discovery Grant supports the following five fields: biotechnology; communications, networking and operating systems; digital media; electronics manufacturing and new materials; and information technology for life sciences. In addition, the program encourages and welcomes interdisciplinary research proposals across these five fields.

Researchers with Principal Investigator status at the ten UC campuses, the three National Laboratories, and the Agriculture Experiment Station are eligible to apply. Business Sponsors must have relevant R&D operations in California, or an R&D alliance with a firm in California. There are three competitive application rounds each year—fall, winter, and spring. For more information, visit: www.ucdiscoverygrant.org.

64

Design of a Survey of Licensing Practices of DNA-Based Patents

Bi Ade[1], Robert Cook-Deegan[2], Stephen McCormack[3], **Lori Pressman**[1] (lori@loripressman.com), and LeRoy Walters[1]

[1]Georgetown University, [2]Duke University, and [3]AlleCure

A Web-based survey on the licensing policies and practices of academic institutions regarding their DNA based patents has been designed, and is currently being tested by a major northeastern university, prior to being administered to two dozen other academic institutions with the highest numbers of DNA based patents. The survey starts with policy questions, and then provides an interface which enables respondents to map licensing information to the bibliographic data of their own set of DNA based patents. The survey will report the percentage of DNA based patents managed by the responding institutions which have been, at one time, licensed. It will also report the number of times these licensed patents have been licensed. More detailed questions will be asked on a dozen licenses for every institution, including dates of execution and termination of the license, degrees of exclusivity and fields of use of the licenses, times when certain income thresholds have been reached, and diligence provisions in the license agreements. These data are expected to yield information on the distribution of times between patent filing, patent issuance, license execution, and other license outcomes, such as product introduction. It will also be possible to associate diligence provisions with degrees of exclusivity and licensing outcomes, where such outcomes may range from termination for non performance to product introduction. No results will be reported on a patent by patent basis, nor will licensee names be identified. Licensing data will be aggregated according to variables such as degrees of exclusivity, or elapsed times between patent filings and license grants.

This is a pilot survey funded jointly by the NIH and DOE.

Author Index

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

U

V

W

X

Y

Z

Institution Index