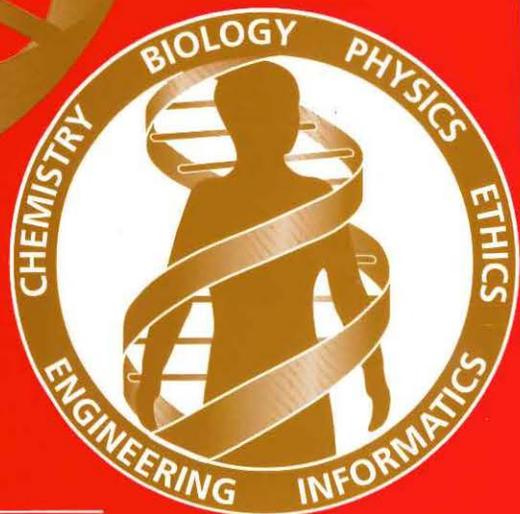


DOE



Human Genome Program

Contractor-Grantee Workshop VIII
Santa Fe, New Mexico
February 27-March 2, 2000



Human Genome Program
U.S. Department of Energy
Office of Biological and Environmental Research
SC-72 GTN
Germantown, MD 20874-1290
301/903-6488, Fax: 301/903-8521
E-mail: *genome@science.doe.gov*

A limited number of print copies are available. Contact:

Sheryl Martin
Human Genome Management Information System
Oak Ridge National Laboratory
1060 Commerce Park, MS 6480
Oak Ridge, TN 37830
865/576-6669, Fax: 865/574-9888
E-mail: *s22@ornl.gov*

An electronic version of this document will be available on February 27, 2000, at the Human Genome Project Information Web site under Publications (<http://www.ornl.gov/hgmis>).

Abstracts for this publication were submitted via the web.

DOE Human Genome Program Contractor-Grantee Workshop VIII

February 27–March 2, 2000
Santa Fe, New Mexico

Date Published: February 2000

Prepared for the
U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research
Washington, DC 20874-1290

Prepared by
Human Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830

Managed by
LOCKHEED MARTIN ENERGY RESEARCH CORP.
for the
U.S. DEPARTMENT OF ENERGY
UNDER CONTRACT DE-AC05-96OR22464

Contents¹

Introduction to Contractor-Grantee Workshop VIII	1
Sequencing	3
1. Sequence Analysis of Human Chromosome 19 Anne Olsen , Paul Predki, Ken Frankel, Laurie Gordon, Astrid Terry, Matt Nolan, Mark Wagner, Amy Brower, Andrea Aerts, Marnel Bondoc, Kristen Kadner, Manesh Shah, Richard Mural, Miriam Land, Denise Schmoyer, Sergey Petrov, Doug Hyatt, Morey Parang, Jay Snoddy, Ed Uberbacher, and the JGI Production Sequencing Team	3
2. Draft Sequencing Procedures for Chromosome 16 Sequencing Mark O. Mundt , David C. Bruce, Leslie Chasteen, Judith Cohn, Lynne Goodwin, Kristina Kommander, Chris Munk, Robert Sutherland, Norman Doggett, and Larry Deaven	3
3. Large-Scale Finishing of Human and Mouse Genomic Sequences Richard M. Myers , Jeremy Schmutz, Jane Grimwood, the Sequencing Group at Stanford Human Genome Center, and the Joint Genome Institute	4
4. A Tale of Three Loci Lee Rowen , Anup Madan, Shizhen Qin, and Lee Hood	4
5. Human Telomere Mapping and Sequencing Robert K. Moyzis , Deborah L. Grady, Han-Chang Chi, and Harold C. Riethman	5
6. Targeted cDNA Sequencing Kimberly Prichard, Susi Wachocki, Mira Dimitrijevic-Bussod, Mark Mundt, Judith Cohn, David Bruce, Cliff Han, Norman Doggett, Christa Prange, and Michael R. Altherr	6
7. Determining Quality of Oligonucleo-tides Synthesized in a High Throughput Process Linda S. Thompson , David C. Bruce, Norman A. Doggett, Mark O. Mundt, and Larry L. Deaven	6
8. Progress of Concatenation cDNA Sequencing at the BCM-Human Genome Sequencing Center Richard Gibbs	7
9. Full-Length cDNA Sequencing Using Differential Extension with Nucleotide Subsets (DENS) O. Chertkov, C. Naranjo, D. Zevin-Sonkin, H. Hovhanissyan, A. Ghochikyan, L. Lvovsky, A. Liberzon, M.C. Raja, and L.E. Ulanovsky	8
10. pZIP: A Versatile Vector for Sequencing by Nested Deletions John J. Dunn	8
11. pUC-SV: A New Double Adaptor Plasmid System for Sequencing Complex Genomes Jonathan L. Longmire , Nancy C. Brown, Larry L. Deaven, and Norman A. Doggett .	9
12. A Fluorescent Sequencing Vector for High-Throughput Clone Selection by Cell Sorting Juno Choe and Ger van den Eng	10
13. An Isothermal Amplification System for the Production of DNA Templates for DNA Sequencing Stanley Tabor and Charles Richardson	10

¹Contact authors are bolded.

<u>Poster Number</u>	<u>Page</u>
14. Universal Energy-Transfer Cassettes for Facile Construction of Energy-Transfer Fluorescent Labels Jin Xie, Lorenzo Berti, Richard A. Mathies, and Alexander N. Glazer	11
15. Fimer Chemistry for Sequencing off BAC and Genomic DNA Templates S. Kozyavkin, N. Polouchine, A. Malykh, O. Malykh, and A. Slesarev	11
16. Chemical Conversion of Boronated PCR Products into Bidirectional Sequencing Fragments Barbara Ramsay Shaw, Kenneth W. Porter, Ahmad Hasan, Kaizhang He, and Jack Summers	12
17. Human and Mouse BAC Libraries for Genome Sequencing, Mapping, and Functional Analysis Kazutoyo Osoegawa, Chung Li Shu, Aaron Mammoser, Joe Catanese, and Pieter J. De Jong	13
18. Human and Mouse BAC Ends Shaying Zhao, Mark D. Adams, Joel Malek, Lily Fu, Bola Akinretoy, Sofiya Shatsman, Maureen Levins, Stephany McGann, Keita Geer, Getahun Tsegaye, Margaret Krol, Peter Choi, Tamara Feldblyum, William Nierman, and Claire Fraser	14
19. Library Strategy for Genome Sequencing Projects William C. Nierman	15
Instrumentation	17
20. New Technologies for Genome Sequencing and Expression Analysis Wayne P. Rindone, John Aach, Martha Bulyk, George M. Church, Jason Hughes, Abby Mcguire, Pam Ralston, Martin Steffen, and Saeed Tavazoie	17
21. High-Performance DNA Sequencing and Analysis Richard A. Mathies	17
22. Radial Capillary Array Electrophoresis Microplate and Scanner for High-Performance DNA Sequencing and Analysis Yining Shi, Brian M. Paegel, James R. Scherer, Peter C. Simpson, David Wexler, Christine Skibola, Martyn T. Smith, and Richard A. Mathies	18
23. Turn Geometries for Minimizing Band Broadening in Microfabricated Capillary Electrophoresis Channels Brian M. Paegel, Lester D. Hutt, Peter C. Simpson, and Richard A. Mathies	19
24. Integrated Microfluidic DNA Amplification and Analysis Systems Eric T. Lagally, Daojing Wang, Charles Emrich, and Richard A. Mathies	19
25. High-Speed High-Throughput Mutation Detection Qiufeng Gao, Ho-Ming Pang, and Edward S. Yeung	20
26. Micro-Fabricated Devices for Concentrating DNA by Induced-Dipole Trapping Charles Asbury and Ger Van Den Engh	21
27. Fully Automated Multiplexed Capillary Systems for DNA Sample Analysis Qingbo Li, Thomas E. Kane, Changsheng Liu, Harry Zhao, Gary W. Loge, John Kernan, Songsan Zhou, Kevin Levan, Heidi Monroe, and David Fisk . . .	21
28. Development and Evaluation of a PCR-Based Sequencing Routine for Use on the ABI 3700 Capillary Machine Lynne Goodwin, Owatha Tatum, Olga Chertkov, Judith Cohn, and P. Scott White . . .	22

<u>Poster Number</u>	<u>Page</u>
29. Rapid and Accurate Detection of Human Functional SNPs Using a Base Stacking Microelectronic DNA Chip Glen Evans, David Canter, Purita Ramos, Ray Radtkey, Ron Sosnowski, Gene Tu, James O'Connell, and Michael Nerenberg	23
30. DNA Sequencing by Single Molecule Detection Peter M. Goodwin, Hong Cai, James H. Werner, James H. Jett, and Richard A. Keller	23
31. New Optical Methods for Sequencing Individual Molecules of DNA Jonas Korfach, Michael Levene, Stephen W. Turner, Mathieu Foquet, Harold G. Craighead, and Watt W. Webb	24
32. High Throughput Multiplexed mtDNA SNP Scoring Using Microsphere-Based Flow Cytometry P. Scott White, Alina Deshpande, Lance Green, Yolanda Valdez, David C. Torney, and John P. Nolan	24
33. Mass Spectrometric Analysis of Genetic Variations Lloyd M. Smith	25
34. Affinity Capture and Mass Spectrometry of Targeted Proteins in Mice Stephen J. Kennel, Gregory B. Hurst, Linda J. Foote, and Michelle V. Buchanan ...	25
35. Rapid Quantitative Measurements of Proteomes Richard D. Smith, Ljiljana Pasa Tolic, Mary S. Lipton, Pamela K. Jensen, Gordon A. Anderson, and Timothy D. Veenstra	26
36. DNA Characterization by Electrospray Ionization-FTICR Mass Spectrometry David S. Wunschel, Bingbing Feng, Ljiljana Pasa Tolic, Mary S. Lipton, and Richard D. Smith	26
37. DNA Sequencing via Electrospray and Ion/Ion Chemistry in an Electrodynamic Ion Trap Scott A. McLuckey, James L. Stephenson, Jr., and Gregory B. Hurst	27
38. DNA and Protein Analyses on Microfabricated Devices R. S. Foote, Y. Khandurina, I. M. Lazar, Y. Liu, T. McKnight, L. C. Waters, S. C. Jacobson, R. S. Ramsey, and J. M. Ramsey	28
39. Stable Isotope Assisted Mass Spectrometry Allows Accurate Determination of Nucleotide Compositions of PCR Products Xian Chen, Zhengdong Fei, Lloyd M. Smith, E. Morton Bradbury, and Vahid Majidi	28
40. Hybridization Detection Tom J. Whitaker and Kenneth F. Willey	29
41. Automation Using Packard Multiprobe Robots for Finishing Christine Munk, Judy Buckingham, Marie Krawczyk, Elizabeth Saunders, David Bruce, and Mark Mundt	29
42. Automated, Low Cost Isolation of Blood or Bacterial Genomic DNA Brian Bauman, Tuyen Nguyen, Zuxu Yao, Tony Zucca, Dan Langhoff, and William MacConnell	30
43. The Use of Electrode Arrays for the Synthesis of Biomolecular Affinity Probes Francis Rossi, Christopher Ashfield, Karl Maurer, and Donald Montgomery	30
44. Development of a High Throughput Peptide Nucleic Acid Synthesizer J. Shawn Roach, Simon Rayner, Lynn Mayfield, David R. Corey, and Harold "Skip" Garner	31

<u>Poster Number</u>	<u>Page</u>
45. MicroArray of Gel Immobilized Compounds on Chip V. Vasiliskov, A. Stomakhin, B. Strizhkov, S. Tillib, V. Mikhailovich, A. Sobolev, A. Kuhktin, and A. Mirzabekov	31
Mapping	33
46. Analysis of WUSTL's Human BAC Fingerprint Database R. Sutherland, M. Mundt, and N. Doggett	33
47. Human Chromosome 16 Mapping Update Cliff S. Han, Robert D. Sutherland, Phillip B. Jewett, Mary L. Campbell, Linda J. Meincke, Judy G. Tesmer, Mark O. Mundt, Larry L. Deaven, and Norman A. Doggett	33
48. Annotation and Analysis of the Draft Sequence of 16Q12 Jung-Rung Wu, Mark O. Mundt, Cliff S. Han, Kristina Kommander, Robert D. Sutherland, Lela Tatum, Norman A. Doggett, and Larry L. Deaven	34
49. Progress in Mapping the Mouse Genome Cliff S. Han, Linda J. Meincke, Larry L. Deaven, and Norman A. Doggett	34
50. Rapid Construction of Mouse Sequence-Ready Maps Using a Homology-Driven Approach Lisa Stubbs, Joomyeong Kim, Laurie Gordon, Hummy Badri, Mari Christensen, Matt Groza, Chi Ha, Sha Hammond, Michelle Vargas, and Eddy Wehri	35
51. Structural and Functional Analysis of a Conserved Imprinted Region of Human Chromosome 19q13.4 and Mouse Chromosome 7 Joomyeong Kim, Vladimir Noskov, Xiaochen Lu, Anne Bergmann, Tiffany Warth, Paul Richardson, Vladimir Larionov, Natasha Kouprina, and Lisa Stubbs ...	35
52. Mapping and Functional Analysis of the Mouse Genome D. K. Johnson, C. T. Culiati, M. L. Klebig, Y. You, D. R. Miller, L. B. Russell, E. J. Michaud, and E. M. Rinchik	36
53. Toward Completion of a Human Chromosome 5 BAC Map and a Mouse Syntenic BAC Map Steve Lowry, Ze Peng, Duncan Scott, Yiwen Zhu, Mei Wang, Roya Hosseini, Michele Bakis, Joel Martin, Ingrid Plajzer-Frick, Jeff Shreve, Le-Thu Nguyen, and Jan-Fang Cheng	37
54. Sequence-Ready Characterization of the Pericentromeric Region of 19p12: A Strategy for the Analysis of Complex Regions of the Human Genome Evan E. Eichler, Anthony P. Popkie, Laurie A. Gordon, and Anne S. Olsen	37
55. IMAGEne 3.0: Clustering All Sequences Obtained from I.M.A.G.E. Clones Peg Folta, Tom Kuczumarski, Tim Harsch, and Christa Prange	38
Bioinformatics	39
56. Software to Support BAC Mapping Cliff S. Han and Norman A. Doggett	39
57. Automated Optimization of Expert System for Base-Calling in DNA Sequencing Arthur W. Miller and Barry L. Karger	39
58. Is Q20 a Sufficient Measure of Quality to Use for DNA Sequencing Process Analysis? D. C. Bruce, M. D. Jones, J. E. Bryant, R. Lobb, J. R. Griffith, M. O. Mundt, N. A. Doggett, and L. L. Deaven	40

<u>Poster Number</u>	<u>Page</u>
59. Annotation of Draft Genomic Sequence Generated at the JGI Richard Mural, Miriam Land, Frank Larimer, Morey Parang, Manesh Shah, Doug Hyatt, Ed Uberbacher, P. Folta, T. Bobo, Zhengping Huang, and T. Slezak	40
60. Information Systems to Support Experimental and Computational Research into Complex Biological Systems and Functional Genomics: Several Pilot Projects Jay Snoddy, Denise Schmoyer, Kathe Fischer, Gwo-Lin Chen, Miriam Land, Sergey Petrov, Sheryl Martin, Ed Michaud, Bob Barry, Gene Rinchik, Peter Hoyt, Mitch Doktycz , and E. Uberbacher	41
61. Navigation, Visualization, and Query of Genomes: The Genome Channel and Beyond Morey Parang, Miriam Land, Denise Schmoyer, Jay Snoddy, Doug Hyatt, Richard Mural, and Ed Uberbacher	42
62. Continuation of the Genome Database Christopher J. Porter, C. Conover Talbot Jr., Jay Snoddy, Ed Uberbacher, and A. Jamie Cuticchia	43
63. Reconstruction and Annotation of Transcribed Sequences: The TIGR Gene Indices John Quackenbush, Ingeborg Holt, Feng Liang, Geo Pertea, Jonathan Upton, and Thomas S. Hansen	44
64. An Informatics Framework for Transcriptome Annotation Brian Brunk, Jonathan Crabtree, Mark Gibson, Chris Overton, Debra Pinney, Jonathan Schug, Chris Stoeckert, Jian Wang, Ihor Lemischka, Kateri Moore, and Robert Phillips	44
65. Protein Domain Dissection and Functional Identification Temple F. Smith, Sophia Zarakovich, and Hongxian He	45
66. Finding Remote Protein Homologs Kevin Karplus	45
67. Multi-Way Protein Folding Classification Using Support Vector Machines and Neural Networks C. H. Q. Ding and I. Dubchak	46
68. Comparative Analyses of Syntenic Regions using Pattern Filtering Jonathan E. Moore and James Lake	46
69. Discovery of Distant Regulatory Elements by Comparative Sequence- Based Approaches. Inna Dubchak, Chris Mayor, Lior Pachter, Gabriella Cretu, Edward M. Rubin, and Kelly A. Frazer	47
70. Identification of Novel Functional RNA Genes in Genomic DNA Sequences S.R. Holbrook, C. Mayor, and I. Dubchak	47
71. Automatic Discovery of Sub-Molecular Sequence Domains in Multi-Aligned Sequences: A Dynamic Programming Algorithm for Multiple Alignment Segmentation Eric Poe Xing, Ilya Muchnik, Denise Wolf, Inna Dubchak, Casimir Kulikowski, Manfred Zorn, and Sylvia Spengler	48
72. Classification of Multi-Aligned Sequence Using Monotone Linkage Clustering and Alignment Segmentation Eric Poe Xing, Ilya Muchnik, Manfred Zorn, and Sylvia Spengler	48
73. Extensions to the Arraydb Micro-Array LIMS Donn Davy, Daniel Pinkel, Donna Albertson, Gregory Hamilton, Joel Palmer, Donald Uber, Arthur Jones, Joe Gray, and Manfred Zorn	49

<u>Poster Number</u>	<u>Page</u>
74. Identifying Single Nucleotide Polymorphisms (SNPs) in Human Candidate Genes Deborah A. Nickerson, Scott L. Taylor, and Mark J. Rieder	49
75. Integrating Sequence and Biology: Developing an Informatics Infrastructure for Mouse/Human Comparative Genomics C. J. Bult, J. T. Eppig, J. A. Blake, J. E. Richardson, and J. A. Kadin	50
76. WIT2 — An Integrated System for Genetic Sequence Analysis and Metabolic Reconstruction Ross Overbeek, Gordon Pusch, Mark D’Souza, Evgeni Selkov Jr., Evgeni Selkov, and Natalia Maltsev	50
77. PUMA2 — An Environment for Comparative Analysis of Metabolic Subsystems and Automated Reconstruction of Metabolism of Microbial Consortia and Individual Organisms from Sequence Data Natalia Maltsev and Mark D’Souza	51
78. Progress Report on EMP Project Evgeni Selkov, Nadezhda Avseenko, Valentina Dronova, Galina Dyachenko, Aleksandr Elefterov, Milyausha Galimova, Nadezhda Fedotcheva, Maria Fomkina, Tatiana Kharybina, Irina Krestova, Aleksandr Kuzmin, Elena Mudrik, Nikolay Mudrik, Valentina Nenasheva, Valeri Nenashev, Evgeni Nikolaev, Aleksandr Osipov, Lyudmila Pronevich, Anna Rykunova, Aleksey Selkov, Evgeni Selkov, Jr., Vladimir Semerikov, Tatiana Sirota, Anatoly Sorokin, Oleg Stupar’, Vadim Ternovsky, and Olga Vasilenko	51
79. BCM Search Launcher - Providing Distributed, Enhanced Sequence Analysis M. P. McLeod, Z. Yang, and K. C. Worley	52
80. Data Submission Tool Manfred D. Zorn and David Demirjian	52
81. Working Examples of XML in the Management of Genomic Data J. D. Cohn and M. O. Mundt	53
82. The Genome Database — Integrating Maps with Sequence Christopher J. Porter, C. Conover Talbot Jr., and A. Jamie Cuticchia	53
83. A Visual Data-Flow Editor Capable of Integrating Data Analysis and Database Querying Dong-Guk Shin, Ravi Nori, Rich Landers, and Wally Grajewski	54
84. Annotating DNA with Protein Coding Domains Winston A. Hide, Robert Miller, Gary L. Sandine, and David C. Torney	55
85. Clustering and Visualizing Yeast Microarray Expression Data Using VxInsight™ George Davidson, Edwina Fuge, and Margaret Werner-Washburne	56
86. Comprehensive Microbial Genome Display and Analysis Frank Larimer, Doug Hyatt, Miriam Land, Richard Mural, Morey Parang, Manesh Shah, Jay Snoddy, and Ed Uberbacher	56
87. Infrastructure and Tools for High Throughput Computational Genome Analysis Doug Hyatt, Phil Locascio, Victor Olman, Manesh Shah, and Inna Vokler	57
88. Genome Information Warehouse: Information and Databases to Support Comprehensive Genome Analysis and Annotation Miriam Land, Denise Schmoyer, Morey Parang, Jay Snoddy, Sergey Petrov, Richard Mural, and Ed Uberbacher	58
89. BiSyCLES: Biological System for Cross-Linked Entries Search Michael Brudno, Igor Dralyuk, Sylvia Spengler, Manfred Zorn, and Inna Dubchak	59

<u>Poster Number</u>	<u>Page</u>
90. Updated ASDB: Database of Alternatively Spliced Genes I. Dralyuk, M. Brudno, M.S. Gelfand, S. Spengler, M. Zorn, and I. Dubchak	60
91. Splice Site Recognition Terry Speed and Simon Cawley	61
92. Refreshing Curated Data Warehouses Using XML Susan B. Davidson, Hartmut Liefke, and G. Christian Overton	61
93. Genome-Scale Protein Structure Prediction in <i>Prochlorococcus europae</i> Genome Ying Xu, Dong Xu, Oakley H. Crawford, J. Ralph Einstein, and Ed Uberbacher	61
94. The Ribosomal Database Project II: Providing an Evolutionary Framework James R. Cole, Bonnie L. Maidak, Timothy G. Lilburn, Charles T. Parker, Paul R. Saxman, Bing Li, George M. Garrity, Sakti Pramanik, Thomas M. Schmidt, and James M. Tiedje	62
Function and cDNA Resources	65
95. The I.M.A.G.E. Consortium: Progress Toward a Complete Set of Human Genes Christa Prange, Peg Folta, Tim Harsch, Genevieve Johnson, Tom Kuczmarksi, Bernadette Lato, Leanne Mila, David Nelson, and Anthony Carrano	65
96. Analysis of Uncharacterized Human cDNAs which Encode Large Proteins in Brain M. Oishi, T. Nagase, R. Kikuno, M. Hirosawa, and O. Ohara	65
97. Novel Approaches to Facilitate Gene Discovery and the Development of a Non-Redundant Arrayed Collection of Full-Length cDNAs Sergey Malchenko, Brian Berger, Vera Da Costa Soares, Maria De Fatima Bonaldo, and Marcelo Bento Soares	66
98. From EST to High Quality cDNA: The BDGP Pipeline for the Construction of <i>Drosophila</i> cDNA Resources Mark Stapleton, Damon Harvey, Peter Brokstein, and Gerald M. Rubin*	66
99. The RIKEN Mouse Full-Length cDNA Encyclopedia Piero Carninci, Kazuhiro Shibata, Masayoshi Itoh, Hideaki Konno, Jun Kawai, Yuko Shibata, Yuichi Sugahara, T. Endo, Y. Ozawa, Yoshifumi Fukunishi, Atsushi Yoshiki, M. Kusakabe, Masami Muramatsu, Yasushi Okazaki, and Yoshihide Hayashizaki	67
100. Tissue Gene Expression Profiling Using RIKEN Full-Length Mouse 20K cDNA Microarray Yasushi Okazaki, Rika Miki, Yosuke Mizuno, Yasuhiro Tomaru, Kouji Kadota, Piero Carninci, Kazuhiro Shibata, Masayoshi Itoh, Yasuhiro Ozawa, Jun Kawai, Hideaki Konno, Yoshifumi Fukunishi, Toshinori Kusumi, Hitoshi Goto, Hiroyuki Nitanda, Yohei Hamaguchi, Itaru Nishiduka, Masami Muramatsu, Atsushi Yoshiki, Moriaki Kusakabe, Joseph Derisi, Vishy Iyer, Michael Eisen, Patric O. Brown, and Yoshihide Hayashizaki	67
101. The Molecular Genetics of DNA Repair in <i>Drosophila</i> K. C. Burtis, R. S. Hawley, C. Boulton, K. Hollis, A. Laurencon, and D. Milliken	68
102. The Tennessee Mouse Genome Consortium D. K. Johnson, D. R. Miller, J. Snoddy, B. A. Berven, and E. M. Rinchik	68

*Invited speaker

<u>Poster Number</u>	<u>Page</u>
103. Designing Genetic Reagents to Facilitate the Mutagenesis and Functional Analysis of the Mouse Genome Edward J. Michaud, Qing G. von Arnim, Carmen M. Foster, Yun You, Dabney K. Johnson, and Eugene M. Rinchik	69
104. Mouse Genetics and Mutagenesis for Functional Genomics: Phenotype-Driven Regional Mutagenesis and Genomics at the Oak Ridge National Laboratory E. M. Rinchik, D. A. Carpenter, E. J. Michaud, Y. You, P. R. Hunsicker, L. B. Russell, D. R. Miller, M. L. Klegig, and D. K. Johnson	70
105. Defining Complex Genetic Pathways with Gene-Expression Microarrays M. J. Doktycz, B. H. Jones, C. T. Cuiat, P. R. Hoyt, B. W. Harker, R. E. Barry, D. D. Schmoyer, S. Petrov, E. M. Rinchik, K. L. Beattie, J. R. Snoddy, and E. J. Michaud	71
106. Genome-Wide Expression Analysis Prove that Distinct Sets of Genes Participate in Cardiac Hypertrophy and the Regression of Hypertrophy Carl Friddle, James Bristow, Teiichiro Koga, and Edward M. Rubin	71
107. Ribozyme Gene Vector Libraries Identify Putative Tumor Suppressor Genes Qi-Xiang Li, Eric Marcusson, Joan Robbins, Mark Leavitt, Flossie Wong-Staal, and Jack R. Barber	72
108. New Vectors for TAR Cloning and Retrofitting of Mammalian Genes Maxim Y. Koriabine, Gregory G. Solomon, Lois A. Annab, J. Carl Barrett, and Vladimir L. Larionov	72
109. Defining the Minimal Length of Sequence Homology Required for Selective Gene Isolation by TAR Cloning Vladimir Noskov, Maxim Koriabine, Greg Solomon, Natalay Kouprina, J. Carl Barrett, Lisa Stubbs, and Vladimir Larionov	73
110. Contamination of BAC Clones by <i>E. coli</i> IS186 Insertion Elements Owatha L. Tatum, Andrew W. Womack, Mark O. Mundt, and Norman A. Doggett	73
111. Developing General Methods to Select Phage Antibodies Against Gene Products Peter Pavlik, Robert Siegal, Daniele Sblattero, Vittorio Verzillo, Roberto Marzari, Jianlong Lou, Jim Marks, and Andrew Bradbury	74
112. Search and Identification of Proteins that Bind Specifically to the Satellite DNAs Ivan B. Lobov and Olga I. Podgornaya	75
113. Diversity in the Proteome: Homologous DNA Replicase Genes Use Alternatives of Transcriptional Slippage or Translational Frameshifting for Gene Expression Norma M. Wills, Bente Larsen, Chad Nelson, John F. Atkins, and Raymond F. Gesteland	75
114. The Transcriptional Program of Gametogenesis in Budding Yeast Ira Herskowitz*	76

*Invited speaker

Microbial Genome Program	77
115. The Comprehensive Microbial Resource Owen White , Jeremy Peterson, Jonathan A. Eisen, and Steven L. Salzberg	77
116. The <i>Pseudomonas putida</i> KT2440 Genome Sequencing Project Karen E. Nelson , Hoda Khouri, Erik Holtzapple, Jeff Buchoff, Michael Rizzo, Azita Moazzez, Kelly Moffat, Kevin Tran, Hean Koo, P. Chris Lee, Daniel Kosack, Bradley Slaven, Helmut Hilbert, Burkhard Tuemmler, and C.M. Fraser	77
117. The Genome of <i>Geobacter sulfurreducens</i> B. A. Methe , L. Banerjee, W. C. Nierman, O. Snoeyenbos-West, S. Sciuffo, and D. E. Lovley	78
118. The <i>Haloferax volcanii</i> Genome Project Rajendra J. Redkar, Joe J. Shaw, Gary G. Bolus, Mary Lee Ferguson, Troy A. Horn, and Vito G. Delvecchio	78
119. The <i>Caulobacter crescentus</i> Genome Sequencing Project Tamara Feldblyum , William C. Nierman, Nikhil Phadke, Peter Ulintz, and Janine Maddox	79
120. Unusual Features of Radioresistant Bacterium <i>Deinococcus radiodurans</i> Genome Revealed by Comparative-Genomic Analysis Kira S. Makarova , L. Aravind, Roman L. Tatusov, Eugene V. Koonin, and Michael J. Daly	80
121. Protein Expression in <i>Methanococcus jannaschii</i> and <i>Pyrococcus furiosus</i> C. S. Giometti , S. L. Tollaksen, H. Lim, J. Yates, J. Holden, A. Lal Menon, G. Schut, M. W. W. Adams, C. Reich, and G. Olsen	80
122. Detection of Non-Cultured Bacterial Divisions in Environmental Samples using 16S rRNA-Based Fluorescent in situ Hybridization Cheryl R. Kuske , Susan M. Barns, and Stephan Burde	81
123. Diversity of Metal Reducing Bacteria from Ecological, Physiological and Genomic Perspectives Jizhong Zhou , Guangshan Li, Alison Murray, Yul Roh, Heshu Huang, Ray Stapleton, Qiaoyun Qiu, John Heidelberg, Claire Fraser, Douglas Lies, Kenneth H. Nealson, and James M. Tiedje	81
124. Pangenomic Microbial Comparisons by Subtractive Hybridization Peter Agron, Lyndsay Radnedge, Evan Skowronski, Madison Macht, Jessica Wollard, Sylvia Chin, Aubree Hubbell, Marilyn Seymour, Christina Nocerino, and Gary Andersen	82
125. <i>Prochlorococcus</i> : The Smallest and Most Abundant Photosynthetic Microbe in the Oceans Sallie W. Chisholm* , Gabrielle Rocap, and Lisa Moore	83
126. Sequencing Microbial Genomes of Relevance to Global Climate Change J. E. Lamerdin , K. Burkhart-Schultz, A. Arellano, S. Stilwagen, A. Erler, A. Kobayashi, M. Shah, D. J. Arp, A. B. Hooper, S. W. Chisholm, G. Rocap, E. Branscomb, and F. Larimer	83

*Invited speaker

<u>Poster Number</u>	<u>Page</u>
Ethical, Legal, and Social Issues	85
127. Electronic Scholarly Publishing: Foundations of Classical Genetics Robert J. Robbins	85
128. <i>Genes, Environment, and Human Behavior: A Curriculum Module</i> Mark V. Bloom, Rodger W. Bybee, Michael J. Dougherty, and Joseph D. McInerney	85
129. High School Students as Partners in Sequencing the Human Genome Kristi Sanford, Maureen Munn, and Leroy Hood	86
130. The Science and Issues of Human DNA Polymorphisms: An ELSI Training Program for High School Biology Teachers David Micklos, Matt Christensen, and Scott Bronson	86
131. Using the Power of Informal Learning to Address Science Literacy: A Report from the Microbial Literacy Collaborative Cynthia A. Needham	87
132. Hispanic Role Model and Science Education Outreach Project - Human Genome Project Education and Outreach Component Clay Dillingham	88
133. Seeking Truth, Finding Justice: A PBS Documentary Special Noel Schwerin	89
134. SoundVision Science Literacy Project Barinetta Scott and Jude Thilman	89
135. Getting the Word Out on the Human Genome Project: A Course for Physicians Sara L. Tobin and Ann Boughton	90
136. Dilemmas in Commercializing Human Genome and Biotechnology Products: Developing a Case-Based Business Ethics Curriculum for Industry Barbara Koenig	90
137. The Responsibility of Oversight in Genetics Research: How to Enable Effective Human Subjects Review of Public and Privately Funded Research Programs Barbara Handelin	91
138. Confidentiality Concerns Raised by DNA-Based Tests in the Market-Driven Managed Care Setting Jeroo S. Kotval, Kathleen Dalton, and Patricia Salkin	91
139. An Economic Analysis of Intellectual Property Rights Issues Concerning the Human Genome Program David J. Bjornstad and Steven Steward	92
140. EINSHAC's Genetics Adjudication Resource Project Franklin M. Zweig	93
Infrastructure	95
141. The Human Genome Management Information System: Making Genome Project Science and Implications Accessible Betty K. Mansfield, Anne E. Adamson, Denise K. Casey, Sheryl A. Martin, Marissa D. Mills, John S. Wassom, Judy M. Wyrick, and Laura N. Yust ...	95
142. DOE Genome Program Coordination and Outreach Sylvia J. Spengler, Janice L. Mann, and Leonora I. Castro	96

143. DOE Alexander Hollaender Distinguished Postdoctoral Fellowships	
Linda Holmes and Wayne Stevenson	97
144. JASON Study on Data Mining and the Human Genome	
G. Joyce, H. Abarbanel, C. Callan, W. Dally, F. Dyson, T. Hwa, S. Koonin, H. Levine, O. Rothaus, R. Schwitters, C. Stubbs, and P. Weinberger	97
Appendix A: Author Index	99
Appendix B: National Laboratory Index	107

Introduction to Contractor-Grantee Workshop VIII

Welcome to the Eighth Contractor-Grantee Workshop sponsored by the Department of Energy (DOE) Genomics Program. This workshop provides a unique opportunity for DOE genome investigators to discuss and share the successes, problems, and challenges of their research as well as new material resources and software capabilities. The meeting also provides scientists and administrative staff with an overview of the program's progress and content, a chance to assess the impact of new technologies, and, perhaps most important, a forum for initiating new collaborations.

We hope you will take full advantage of the opportunities offered by this meeting and by the beautiful surroundings of the Santa Fe area. Last time we got together we were in Oakland, California, so that you could visit the DOE Joint Genome Institute's new Production Sequencing Facility in Walnut Creek. This facility was opened officially in the spring of 1999 by Secretary of Energy Bill Richardson.

The 144 abstracts in this booklet describe the most recent activities and accomplishments of grantees and contractors funded by DOE's long-running human and microbial genome programs, as well as ongoing efforts in model organism and functional genomics research. We also have included talks from invited guests who will discuss related genomics research efforts and opportunities for postgenomic biology enabled by genome research. All projects funded by the Biological and Environmental Research (BER) Program will be represented at poster sessions at the Santa Fe Convention Center on East Marcy Street one block north of the Plaza in the center of Santa Fe, so you will have an opportunity to meet with the researchers who make this program a success. New informatics resources also will be demonstrated at the poster sessions. I urge you to take full advantage of all the formal and informal opportunities for discussion and exchange of information that are available at this workshop.

The main challenge facing the DOE genome program today remains the imperative to sequence to "Bermuda Standards" human chromosomes 5, 16, and 19, which constitute DOE's commitment to the International Human Genome Project. Three years ago, DOE addressed this challenge by forming the Joint Genome Institute (JGI) under the direction of Elbert Branscomb. Using the complementary strengths of DOE's three largest genome programs and those at other laboratories and universities, the JGI has made efficient and effective use of diverse expertise and resources to establish a facility for high-throughput sequencing. By the date of this meeting, JGI is expected to have completed the draft sequencing of chromosomes 5, 16, and 19. This represents around 10% (or 250 million base pairs) of the euchromatin in the human genome. In addition, JGI has determined the DNA sequences of two entire microbial genomes, about 4 "finished" Mb. Currently the sequencing rate at the JGI's Production Sequencing Facility is averaging about 300 million bases of raw sequence per month, a large and remarkable increase over last year.

Although many challenges lie ahead, particularly in anticipating and preparing for the "postgenomic" world, we are more optimistic than ever about the success of this grand project and

its many contributions to science and society. Yet we cannot afford to be complacent, and the workshop speakers on ethical, legal, and social implications (ELSI) will remind and challenge all of us that our science has societal impacts that we cannot avoid. We cannot be aloof and disengaged from those interactions. In particular, ELSI must go forward with the active participation of all scientists in the program. ELSI implications are not just research topics for ELSI investigators but real-life issues that need to be considered in the context of genome research today.

There are other genomes to be sequenced besides the human, and the BER Microbial Genome Program continues to contribute complete sequences to public databases—11 complete microbes to date with 11 more in the pipeline. Each microbial sequence has its surprises and its exciting science. The entire 3-Mb sequence of *Deinococcus radiodurans*, the most radioresistant microbe yet known, was published in *Science* (November 19, 1999). Its astounding DNA-repair capacities represent longstanding and continuing high-priority DOE interests and the opportunity, perhaps through genetic engineering of toxin-degrading enzyme systems, to address DOE's mission of mixed-waste remediation. This single example underscores the opportunities that lie ahead to further exploit the interdisciplinary biological approaches that we view as important guiding principles for the science we support. It also reminds us that we must continue to take responsibility for using our science to better our world. The complete sequence of another microbe, *Thermotoga maritima*, published in the May 27, 1999, issue of *Nature*, suggested that the swapping of genetic elements by microbes, much as our children might trade Pokemon cards, is more common in evolution than had ever been suspected. Most recently, JGI has completed two genomes of microbes that play enormous roles in carbon-management processes on our planet. You will hear more about these interesting microbes at this workshop.

We look forward to a very interesting and productive meeting and offer our sincere thanks to all the organizers and to you, the scientists, whose vision and efforts have realized and continue to realize the promises of genome research.

Sincerely,



Ari Patrinos
Associate Director of Science for Biological and Environmental Research
Office of Biological and Environmental Research
U.S. Department of Energy
genome@science.doe.gov

Sequencing

1. Sequence Analysis of Human Chromosome 19

Anne Olsen¹, Paul Predki², Ken Frankel², Laurie Gordon¹, Astrid Terry¹, Matt Nolan¹, Mark Wagner¹, Amy Brower¹, Andrea Aerts², Marnel Bondoc², Kristen Kadner², Manesh Shah³, Richard Mural³, Miriam Land³, Denise Schmoyer³, Sergey Petrov³, Doug Hyatt³, Morey Parang³, Jay Snoddy³, Ed Uberbacher³, and the JGI Production Sequencing Team

¹Lawrence Livermore National Laboratory and
²Lawrence Berkeley National Laboratory, DOE Joint Genome Institute, Production Sequencing Facility, Walnut Creek, CA; ³Genome Annotation Consortium, Oak Ridge National Laboratory, Oak Ridge, TN
olsen2@llnl.gov

Chromosome 19 has an estimated size of ~65 Mb and is the most GC-rich human chromosome. The higher than expected proportion of genes and ESTs mapped to this chromosome suggests that it is exceptionally gene-rich, consistent with its high GC content. Sequencing of chromosome 19 will thus be especially rewarding in terms of gene discovery and elucidation of higher order gene organization. The sequence-ready BAC/cosmid map of chromosome 19 constructed at Lawrence Livermore National Laboratory currently consists of 17 ordered, restriction mapped BAC/cosmid contigs of average size 3.3 Mb spanning 56.3 Mb, or approximately 97% of the estimated 58 Mb comprising the p- and q-arms. A minimally overlapping set of clones (468 cosmids and 290 BACs) spanning the chromosome has been selected for sequencing. About 15 Mb of unique sequence has been finished and submitted to

Genbank. Draft sequence (minimum coverage 3X) has been generated for about 68% of the remaining territory with an average depth of 7.7X. Sequence contigs have been ordered and oriented for about 5.2 Mb of the draft sequence. Updated map and sequence data is available from the LLNL web site at <http://www-bio.llnl.gov/bbrp/genome/genome.html> and the JGI web site at <http://jgi.doe.gov/>. Sequence is being analyzed through the Genome Annotation Pipeline at Oak Ridge National Laboratory. The analysis of 15 Mb of finished genomic sequence yielded 719 gene models predicted by Genscan, and 766 gene models predicted by GRAIL-EXP. About two-thirds of the gene models predicted by GRAIL-EXP were aligned with one or more ESTs. 500 of the Genscan predicted proteins and 456 GRAIL-EXP predicted proteins had homologs with BLAST E-values <1.0e-5. Annotation summaries are available from the ORNL Genome Catalog and Genome Channel at <http://genome.ornl.gov>. Detailed analyses of specific chromosomal regions will be presented.

Supported by USDOE under Contracts W-7405-Eng-48 (LLNL), DE AC0376SF00098 (LBNL) and DE-AC05-96OR22464 (ORNL).

2. Draft Sequencing Procedures for Chromosome 16 Sequencing

Mark O. Mundt, David C. Bruce, Leslie Chasteen, Judith Cohn, Lynne Goodwin, Kristina Kommander, Chris Munk, Robert Sutherland, Norman Doggett, and Larry Deaven
Bioscience Division and DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM 87545
mom@telomere.lanl.gov

As the amount of human sequence that is publicly available increases, efficient use of tools to monitor sequencing progress becomes a more important issue. Our current strategy for monitoring chromosome 16 draft data produced by the JGI includes immediate collection of information on 1) sequence marker content using ePCR, 2) BAC end sequence overlap with BLAST, 3) *E. coli* contamination level, 4) short subclone level, 5) Q20 quality analysis, and 6) confirmation of suspected overlaps based on our maps. These data are measured after the first two plates of forward and reverse sequencing, and decisions are then made for the continuation and desired level of sequence coverage

Order and orientation of contigs becomes the main concern as the draft depth is increased. We present Java tools to assist in both detecting assembly and tracking/handling errors as well as to help order contigs when possible. These tools take advantage of the paired end relationships that are available from our double-end plasmid sequencing approach. Finally, we demonstrate the importance of proper BAC end orientation in choosing clones to extend sequence as well as in feeding information back to a more accurate mapping process.

3. Large-Scale Finishing of Human and Mouse Genomic Sequences

Richard M. Myers, Jeremy Schmutz, Jane Grimwood, the Sequencing Group at Stanford Human Genome Center, and the Joint Genome Institute

The Stanford Human Genome Center and Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305-5120 and the Joint Genome Institute, 2800 Mitchell Drive, B100, Walnut Creek, CA 94598
<http://www.jgi.doe.gov>
<http://www-shgc.stanford.edu>
myers@shgc.stanford.edu

We have begun a new collaboration with the the Joint Genome Institute to generate large amounts of finished human and mouse sequence from “draft” sequences produced by the JGI and its associated

laboratories. Our goal is to produce about 100 Mb of finished sequence each year, focusing first on finishing human chromosome 19 while also finishing clones from human chromosomes 5 and 16 and syntenic mouse sequences. Our criteria for considering a large-insert clone as finished is that it has an estimated base-pair error rate of less than one in 10,000 bp, and that the entire sequence is contiguous, with exception for small, difficult-to-fill gaps of known size in a small fraction of the clones.

We receive subclones and sequence traces from the JGI and reassemble the data, generally resulting in assemblies with 10-20 contigs per 100 kb for the 6X of shotgun sequence data produced for each clone. We then use a computationally-driven process that requires almost no human decision-making to choose subclones, directed sequencing reactions, and, for a portion of the reactions, oligonucleotide primers. After applying this automated stage, all or almost all of the gaps are filled and the clone is passed to a group of finishers, who reassemble the sequence data and design specialized sequencing reactions to fill remaining gaps and to bring up the quality of the sequence so that the entire clone meets our finished criteria. The final sequence is checked by our informatics group and then submitted to GenBank. We have finished more than 3 Mb of sequence with the JGI since this collaboration began, and expect to have an additional 8 Mb finished by the end of February. We hope to achieve an average of 8-10 Mb of finished sequence per month within the next quarter.

4. A Tale of Three Loci

Lee Rowen, Anup Madan, Shizhen Qin, and Lee Hood
Multimegabase Sequencing Center, University of Washington, Seattle, WA 98195 and Institute for Systems Biology, Seattle, WA
leerowen@u.washington.edu

One of the fascinating results of large-scale sequencing is the revelation of vastly different types of genomic landscapes. Based on our accumulation of finished sequence over long contiguous stretches of

the genome, we plan to present data analyses pertaining to three of these landscapes:

A) The human and mouse beta T cell receptor loci, which exemplify the rapid evolution of a multigene family. Here, genes are embedded in long repeats which are added to or deleted from the genome via unequal cross-over. In this regard, human and mouse have undergone somewhat different evolutionary paths.

B) The human and mouse major histocompatibility complex class III regions, which exemplify high gene density (> 15% coding sequence). Here, the orthologous relationship between human and mouse is highly conserved. This landscape raises interesting questions about gene regulation and why it is that genes with apparently unrelated functions might be so closely spaced.

C) The human neurexin III gene on chromosome 14, which exemplifies a large-intron gene that spans over a megabase. Neurexins are noteworthy for their large number of alternative splice forms and their differential expression in neurons, depending supposedly on the alternative splices.

The notion of genomic landscapes provides a framework for thinking beyond individual genes to the organization of the genome as a whole and how this organization bears on different types of function for individual genes and gene families.

5. Human Telomere Mapping and Sequencing

Robert K. Moyzis, Deborah L. Grady, Han-Chang Chi, and Harold C. Riethman
Department of Biological Chemistry, College of Medicine, University of California at Irvine, Irvine, CA 92697 and The Wistar Institute, Philadelphia, PA 19104
rmoyzis@uci.edu

The Human Genome Project has undergone a dramatic shift to the goal of obtaining a “working draft” sequence of human DNA by the end of this year. Such a framework sequence will catalyze gene discovery and functional analysis, and allow finished sequencing to be focused on regions of the highest biomedical priority. Over 80% of human DNA can be rapidly sequenced in the next few years by highly automated, high throughput sequencing centers. However, a significant fraction of the human genome will not be sequenced and/or assembled to completion by such approaches, as demonstrated by the recent sequence of human chromosome 22 (Dunham et. al., *Nature* 402, 489-495, 1999). These are regions that contain 1) a high percentage of repetitive DNA sequences; 2) internal tandem duplications, including multigene families; and/or 3) are unstable in all current sequencing vectors. Producing quality DNA sequence of these regions, which faithfully represents genomic DNA, will be a continuing challenge.

Telomeres, the ends of the linear DNA molecules in human chromosomes, exhibit both high levels of repetitive DNA composition and cloning instability. In addition, extensive heterogeneity exists in these regions between various individuals. Half-YAC clones are uniquely suited as starting material for the sequence analysis of human telomeric regions. The inability to clone the extreme end of human chromosomes in bacterial vectors, including BACs, is well known. Due to the lack of appropriate restriction sites in the terminal (TTAGGG)_n regions, as well as the necessary size selection involved in BAC library construction, the most terminal BAC clones will be 20-200Kb from the true DNA ends. By functional complementation in yeast, however, the true human telomeric end can be cloned. To date, 43 of the 46 unique human telomeres have been obtained as half-YACs.

Using RARE (RecA-Assisted Restriction Endonuclease) cleavage, 20 of these telomere half-YAC clones (representing the telomeres of human chromosomes 1p, 1q, 2p, 2q, 4p, 6q, 7p, 7q, 8p, 8q, 9p, 12q, 13q, 14q, 17p, 17q, 18p, 18q, 19p, and 21q)

have now been confirmed to represent the true telomere. An additional 11 clones (representing the telomeres of 3q, 4q, 9q, 10p, 10q, 11p, 11q, 15q, 16q, 19q, and 20p) are currently being confirmed by RARE cleavage analysis. Given the new goals of the Human Genome Project, we have initiated framework sequencing on these clones, as well as the most terminal BACs identified from our chromosome 5 mapping project (Peterson et.al., *Genome Res* 9, 1250-1267, 1999). These framework telomere sequences will provide a "cap" to the worldwide genome sequencing efforts. A combination of cosmid and plasmid end sequence analysis, combined with extensive restriction enzyme mapping of the original YAC, results in highly ordered framework sequences. To date, framework sequence of 1q, 5p, 9q, 11q, 17p, and 18p have been completed. Following framework sequencing, finished sequencing will be conducted in select regions, with priority given to areas with high biological interest and/or relevant to the JGI, i.e., chromosomes 5, 16, and 19. An important QC/QA aspect of our sequence analysis is the extensive confirmation of the sequence against genomic DNA by PCR-resequencing. Numerous polymorphisms in these regions, including SNPs, VNTRs, and rearrangements have been identified. Using pooled DNA PCR/sequencing, the population distribution of many of these polymorphisms can be determined rapidly.

6. Targeted cDNA Sequencing

Kimberly Prichard, Susi Wachocki, Mira Dimitrijevic-Bussod, Mark Mundt, Judith Cohn, David Bruce, Cliff Han, Norman Doggett, Christa Prange, and Michael R. Altherr
Los Alamos National Laboratory, Los Alamos, NM 87545
ALTHERR@LANL.GOV

The sequencing of cDNAs that are co-linear to genomic sequencing targets adds considerable value to the information generated from both efforts. Through the use of sequence analysis tools, comparisons of these distinct data sets reveal details of gene organization, splice sites and, because the sequences are derived from different sources, gene based single nucleotide polymorphisms (SNPs). We

intend to exploit gene predictions derived from genomic sequencing data to identify full-length cDNAs (from initiation codon to the poly adenylation site) for complete cDNA sequencing. We will use "overgo" probes to identify cDNAs corresponding to the gene predictions. We have chosen the strategy of cDNA insert concatenation as our sequencing method. To model this effort, we have embarked by sequencing the complete inserts of cDNAs from the IMAGE collection previously mapped to chromosomes 5, 16, and 19. Approximately, 1800 clones were identified for this effort. We used the Unigene database to identify cDNAs for which the sequence data was incomplete and to identify the largest predicted member of the clone set. These clones are undergoing concatenation cDNA sequencing. Subsequent analyses are being done to identify those coding sequences for which co-linear genomic sequence exists, to characterize their gene structure and to identify SNPs.

This work was supported by Los Alamos National Laboratory LDRD funds and by US DOE OBER under contract W-7405-Eng36.

7. Determining Quality of Oligonucleotides Synthesized in a High Throughput Process

Linda S. Thompson, David C. Bruce, Norman A. Doggett, Mark O. Mundt, and Larry L. Deaven
Bioscience Division and DOE Joint Genome Institute,
Los Alamos National Laboratory, Los Alamos, NM 87545
thompson@telomere.lanl.gov

LANL obtained from the University of Texas Southwest Medical Center a liquid chemical dispensing robot, LCDR or Mermade, to produce large numbers of oligos. The Mermade is capable of making two 96-well plates of oligos per day. The instrument protocol adheres to the standards of DNA synthesis using deprotection, evaporation, resuspension, and quantitation. We also utilize a Biomek 2000 to resuspend the samples, prepare standard dilutions for OD readings on a plate reader, and to set up the samples for gel electrophoresis. Quality control consists of running a representative

sample of the plate on BioRad 15% TBE-Urea Ready gels, 12 wells. If those samples are confirmed to be the required length without n-x species, i.e. oligos missing one or more bases, and if quantitation turns up no zero values, the plate can be given to the user. A quality control issue at this time is whether one needs to PAGE every oligo made on the Mermade. A recent random test of a plate of Mermade oligos vs. "factory made" oligos showed a success rate of 93% for both plates, indicating that 7% failed or were n-x. UTSW and LANL both report an average success rate of 95%, showing there's little or no difference between oligos made on the Mermade or those purchased through an oligo company.

LANL has also purchased a MALDI mass spectrometer. This instrument can be utilized to look at the DNA samples made by the Mermade and has been considered to take over the quality control aspect of DNA synthesis. The MALDI-MS could be used as a qualitative tool to screen all oligos for the desired length and the presence of n-x species, followed up with PAGE to identify the extent of the n-x in the suspect sequences.

8. Progress of Concatenation cDNA Sequencing at the BCM-Human Genome Sequencing Center

Richard Gibbs
Baylor College of Medicine, Houston, TX
agibbs@bcm.tmc.edu

Concatenation cDNA sequencing (CCS) has been used to complete sequencing of more than 750 clones from the human brain cDNA library (INIB) and 30 clones representing childhood leukemia. Together these represent a total length of 1.2 megabases of assembled sequence. An additional 390 clones are currently in the sequence pipeline. Statistics from 14 completed projects continue to show that CCS is as efficient as sequencing of single large DNA fragments, with an average of 17 reads and one custom primer to complete each kb of sequence.

Methodological improvements, including pooling of clones during growth and the use of Phred and Phrap for assembly, have further simplified CCS.

For 596 different clones from the INIB brain library, a similarity search was performed against the public cDNA database. Of these 58% were novel and the remaining 42% (251) had partial matches to known sequences or genes from human or other organisms. Of the latter, 159 clones displayed similarity matches to known proteins. A comparison against the Unigene cDNA dataset revealed that 61% of the cDNAs or submitted cDNAs represented novel contributions, 258 from among a set of 424.

Of 159 clones with partial protein matches, only 32 (20%) had a complete ORF (open reading frame). This indicates a low percentage of cDNA clones representing full length mRNAs from the libraries. To generate better libraries, a postdoctoral student made a trip to Japan where four cDNA libraries (one human infant brain, one mouse brain and two childhood leukemia) were constructed, using the CAP-trapping technology developed by Hayashizaki's group at the RIKEN Institute. Three libraries have been evaluated in detail, acquiring ESTs (expressed sequence Tags) from 192 clones of each library. The data show good quality through little contamination with vector (1.8-2.5%), ribosomal DNAs (1.3-1.8%), and low redundancy (3.1-4.2%). About half of the ESTs lacked matches with Unigene cDNA sequences. Some 65.0-66.6% of clones possessed the first ATG codon of the encoded protein, indicating very high quality of the libraries. Thus the three analyzed cDNA libraries are suitable for large-scale and full-length sequencing. About 8,000 ESTs have been generated from these cDNAs and potentially novel clones are being selected for subsequent full-length sequencing.

9. Full-Length cDNA Sequencing Using Differential Extension with Nucleotide Subsets (DENS)

O. Chertkov¹, C. Naranjo¹, D. Zevin-Sonkin³, H. Hovhanissyan³, A. Ghochikyan³, L. Lvovsky³, A. Liberzon³, M.C. Raja^{2,3}, and L.E. Ulanovsky^{1,2}

¹Los Alamos National Laboratory, Los Alamos, NM 87545; ²Argonne National Laboratory, Argonne, IL 60439; and ³Weizmann Institute of Science, Rehovot 76100, Isreal
levy@anl.gov

Upon moving to LANL, we are setting up a full-length cDNA sequencing facility using our technology termed Differential Extension with Nucleotide Subsets (DENS) which is essentially primer walking without primer synthesis (Raja et al., 1997, *NAR* 25, pp. 800-805). DENS works by converting a short primer (selected from a pre-synthesized library of 8-mers with 2 degenerate bases each) into a long one on the template at the intended site only. DENS starts with a limited initial extension of the primer (at 20 °C) in the presence of only 2 out 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 termination (sequencing) reaction at 60 °C then accepts the primer extended at the intended site, but not at alternative sites where the initial extension (if any) is generally much shorter.

DENS primer walking seems to be tailor-made for full-length cDNA sequencing, as the absence of the primer synthesis step facilitates closed-loop automation of primer walking with the benefit of unattended operation. Earlier, in a pilot experiment we used DENS for sequencing both strands of four cDNA clones containing inserts of 1.9, 2.3, 3.8 and 4.9 kb. The success rate of the DENS sequencing reactions was 72% yielding 27,864 base-calls. The median PHRED quality value was 40, corresponding to the error probability of approximately 10⁻⁴. The plotted distribution showed that base-calls with PHRED values less than 20 occurred only 1% of the time. The 8-mer primers for DENS sequencing were selected using our dedicated software.

10. pZIP: A Versatile Vector for Sequencing by Nested Deletions

John J. Dunn

Biology Department, Brookhaven National Laboratory, Upton, NY 11973
jdunn@bnl.gov

We have constructed a low-copy, amplifiable vector that should be particularly useful for cloning and sequencing full-length cDNAs and highly repeated DNAs. This pZIP vector is maintained in *Escherichia coli* at low copy number by the F replicon and can be amplified 300 fold from an IPTG-inducible phage P1 replicon (repL). A relatively small size of 4.4 kbp was achieved by removing the 2.5-kb sop (stability of plasmid genes) region of F, but the plasmid is stably maintained by selective growth in the presence of kanamycin. A multiple cloning region (MCR) is flanked by sites that allow the biochemical generation of unidirectional nested deletions crossing the cloned DNA. The resulting deletion clones can be ordered by size, and an ordered, overlapping set of sequences can be obtained by priming within the flanking vector sequence to produce the complete sequence of both strands. The correspondence of plasmid lengths with those predicted by the assembled sequence aids in and verifies the correctness of the assembly. The low copy number should allow the cloning of DNAs that might not be stable in higher copy vectors, and amplification provides ample DNA for generating the nested deletions.

Unidirectional nested deletions are produced by cutting the DNA specifically near one end of the cloned DNA to generate an end that is sensitive to digestion by *E. coli* exonuclease III (ExoIII) and an end that is resistant, or by specifically nicking the appropriate strand. The ends or nick are oriented so that ExoIII will digest one strand across the cloned DNA. The resulting single-strand gaps are converted to double-strand gaps by treatment with S1 nuclease, and the ends are repaired and ligated with T4 DNA polymerase and ligase. ExoIII digests quite synchronously, and treating pooled samples from several different ExoIII digestion times, followed by electroporation, produces a population of clones with a distribution of different deletion end points. ExoIII-

resistant ends are produced by intron-encoded endonucleases that cut at very rare sites to produce 4-base 3' overhangs. I-CeuI and I-SceI flank the MCR on one side and I-PspI on the other. ExoIII-sensitive ends can be generated by cutting with a restriction endonuclease at any of several different 8-base or other rare cleavage sites located between the sites cut by the intron-encoded nucleases and the cloning sites in the MCR. The *fd* origin of replication is also located on the I-PspI side of the MCR, oriented so that the specific nick by the gene 2 protein can be extended across the cloned DNA by ExoIII.

In collaboration with the Joint Genome Institute, we are evaluating the capability of the pZIP vector and the nested-deletion sequencing strategy to close gaps that have resisted closure by standard sequencing strategies in several different regions of human chromosome 19. We hope to demonstrate cloning in the low-copy pZIP of regions that are difficult to clone in standard sequencing vectors, and to determine accurate sequences of highly repeated regions by the nested-deletion strategy.

11. pUC-SV: A New Double Adaptor Plasmid System for Sequencing Complex Genomes

Jonathan L. Longmire, Nancy C. Brown, Larry L. Deaven, and Norman A. Doggett
Bioscience Division and DOE Joint Genome Institute,
Los Alamos National Laboratory, Los Alamos, NM
87545
longmire@telomere.lanl.gov

The sequencing of complex genomes requires shotgun cloning (or subcloning) of genomic DNA (or BACs) into vectors that carry smaller inserts and that can serve as templates in sequencing reactions. Such subcloning vectors typically include plasmids or M13. For sequencing purposes at Los Alamos, we have previously used blunt end ligation of inserts into the HincII site of pUC-18. In addition, we have also used the double adaptor approach described by

Andersson et al. ([1996] *Analytical Biochemistry* 236: 107-113) to subclone BAC fragments into pBluescript. Both of these approaches have distinct advantages and disadvantages. For example, blunt end subcloning is technically straight forward but can result in clones with multiple inserts and nonrecombinants even when vector ends are dephosphorylated. Double adaptor subcloning into pBluescript can reduce the frequency of nonrecombinants and clones with multiple inserts. However, the sequencing priming sites are located at a greater distance from the cloning site in Bluescript compared to pUC. Consequently, some of the sequence data that is generated in Bluescript clones is vector readthrough that has to be trimmed prior to assembly of the data.

In order to improve upon existing systems, we have developed a new cloning vector that allows double adaptor shotgun subcloning of large target molecules into pUC-18. The vector pUC-SV was constructed by cloning a 2 kb human DNA insert fragment into the XbaI and PstI sites of pUC-18. The insert serves as a "stuffer" and enables one to easily monitor for complete digestion when the plasmid is being processed to produce subcloning-ready vector. Vector adaptors are ligated to the SacI and SphI ends of the linearized vector. This produces 12 nt overhangs that are complimentary to adaptors that are ligated to the repaired ends of the fragmented target DNA. Nonrecombinants and clones with multiple inserts are eliminated because neither the vector adaptors nor the insert adaptors are self-complimentary. The pUC-SV vector yields cloning (subcloning) efficiencies greater than 105 colonies per microgram target DNA with zero nonrecombinant background. Highly representative subclone libraries can be made using as little as 10 ng of processed target DNA. In addition, the amount of DNA sequence data that is produced using pUC-SV is increased compared to Bluescript due to placement of the priming sites. In the adapted pUC-SV, the primer sites are located 26 nt and 28 nt away from the insert ends (compared to 53 and 64 nt in adapted Bluescript). Thus, 63 nt less data is lost to trimming for every subclone that is

processed. This increased data yield becomes very significant when several thousands of subclones are processed.

12. A Fluorescent Sequencing Vector for High-Throughput Clone Selection by Cell Sorting

Juno Choe and Ger van den Engh
Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195-7730
engh@biotech.washington.edu

With the advent of automated high-throughput DNA sequencers, clone selection and preparation is rapidly becoming a major bottleneck in genome sequencing. We are developing an integrated clone-selection / clone-preparation process that has the potential to dramatically increase the speed of sample production. The process makes use of a sequencing vector that contains fluorescent proteins so that insert-containing bacteria can be selected with a cell sorter. The cell sorter will deposit individual bacteria onto a carrier ribbon that moves the samples in a linear procession along processing stations.

We have constructed two vectors that can be used in this process. Most recently we developed a vector containing a tandem of Blue and Green Fluorescent Proteins separated by a cloning site. This vector indicates integration of a cloned insert through the shifting of ratios of two fluorescent proteins. In the native pBGFP, a fusion protein composed of Blue Fluorescent Protein (BFP2, Clontech) and Green Fluorescent Protein (GFPmut3.1, Clontech) genes is expressed at high levels. The BFP/GFP fusion protein can be excited by UV light in the 350–400 nm range. This causes excitation of the BFP, followed by GFP due to fluorescence resonance energy transfer (FRET). When an insert is successfully ligated into the cloning site in the linker region between BFP and GFP, there is a loss of function of the GFP portion of the protein. In this case, increased BFP fluorescence will be observed with loss of observable green fluorescence. We can quantitate the ratios of these two fluorescent proteins very accurately by flow cytometry. This provides the ability to rapidly sort individual bacteria with high BFP and low GFP

content at a rate of up to 10,000 per hour. Amplification of cloned inserts can be achieved by growing bacteria in culture medium and/or PCR amplification.

Under actual test conditions, ligation of fragments between 2.3–5.6 kb resulted in clear separations of two populations of *E. Coli* grown in liquid media: BFP/GFP expressing bacteria containing native pBGFP plasmid and BFP expressing bacteria containing pBGFP with cloned insert. Furthermore, bacteria were observed under UV fluorescence microscopy. Two clearly distinguishable phenotypes appearing either green or blue were observed.

13. An Isothermal Amplification System for the Production of DNA Templates for DNA Sequencing

Stanley Tabor and Charles Richardson
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115
stabor@hms.harvard.edu

We are developing DNA polymerases for use in DNA sequencing and amplification applications. We will describe a very efficient isothermal amplification system which provides an attractive alternative to conventional methods of generating plasmid and BAC templates for DNA sequencing. Amplification of from 1 pg to 1 µg of template DNA results in the synthesis of DNA to a final concentration of 0.5 µg/µl in 15 min at 37 °C (corresponding to an amplification of up to several million-fold). Amplification is nonspecific; all sequences present are amplified equally. The reaction requires no exogenous primers. This system is based on the replication apparatus of bacteriophage T7; the principle enzymes required are two forms of T7 DNA polymerase, the T7 helicase/primase, and single-stranded DNA binding protein. The products are linear double-stranded DNA fragments several thousand base pairs in length. When the products are used as templates for capillary-based fluorescent sequencing, the fluorescent signal produced is several fold higher than a comparable amount of supercoiled plasmid DNA, and results in 20% more base calls

that have a quality score greater than Phred 20. The attractive features of this system for large sequencing projects is its simplicity and the constant, reproducibly high yield of DNA that can be used directly in DNA sequencing reactions without further purification. This nonspecific amplification reaction could also be of use in immortalizing small, precious samples of genomic DNA required for genotype analysis. We are also using this technology to amplify single DNA molecules embedded in agarose. This enables one to construction and amplification DNA libraries in vitro without the need to transform bacterial cells. Finally, we will present an update of our work modifying DNA polymerases to increase their processivity and their use of nucleotide analogs for use in DNA sequencing.

14. Universal Energy-Transfer Cassettes for Facile Construction of Energy-Transfer Fluorescent Labels

Jin Xie¹, Lorenzo Berti², Richard A. Mathies², and Alexander N. Glazer¹

¹Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720 and

²Department of Chemistry, University of California, Berkeley, CA 94720

glazer@uclink4.berkeley.edu

Energy-Transfer fluorescent labels are advantageous in DNA sequencing applications because of their improved emission signal strength and spectral purity¹. To facilitate the production of ET labels it is useful to develop cassette labeling strategies. In one approach an ET cassette was synthesized as a part of the oligonucleotide primer using sugar-phosphate spacers². Here, we have designed and synthesized a universal cassette DMT-O-(CH₂)₆-S-S-(CH₂)₆T(Rox) SSSSSS-Fam (where DMT is dimethoxytrityl and S is 1',2'-dideoxyribose phosphate) for attachment to appropriate terminator or primer derivatives which can be readily prepared by automated methods. With 488 nm excitation, the emission of this cassette is 10-fold higher than that of

free Rox. The thiol group on the universal ET cassette is exposed by reduction, while an amino-derivative of the terminator (or primer) is substituted with a bifunctional NHS-ester-maleimide reagent. The conjugation of the universal ET cassette with maleimide-derivatized terminator or primer is almost quantitative in 2 hours at room temperature. When used in sequencing, cassette-labelled primers gave excellent results.

1. Xie, J., Hung, S.-C., Glazer, A. N. and Mathies, R. A. Energy Transfer Fluorescent Labels for DNA Sequencing and Analysis, in Topics in Fluorescence Spectroscopy, Volume 7: ANA Technology, in press (1999).

2. Ju, J., Glazer, A. N. and Mathies, R. A. Cassette Labeling for Facile Construction of Energy Transfer Fluorescent Primers, Nucleic Acids Research 24, 1144-1148 (1996).

15. Fimer Chemistry for Sequencing off BAC and Genomic DNA Templates

S. Kozyavkin, N. Polouchine, A. Malykh, O. Malykh, and A. Slesarev
Fidelity Systems, Inc., 7961 Cessna Avenue,
Gaithersburg, MD 20879-4117
<http://www.fidelitysystems.com>
fsi1@fidelitysystems.com

Robust sequencing off BAC and genomic templates presents a new challenge in technology development. The problems associated with the use of standard oligonucleotides as primers in genomic cycle sequencing protocols include insufficient specificity of primer annealing, non-specific amplification, low sensitivity and premature truncation at secondary structures in template DNA.

To overcome these problems we have developed a new method to generate combinatorial libraries of chemically modified oligonucleotides (fimers). The method is based on the use of our proprietary monomers containing MOX or SUC reactive

moieties. We assessed the effects of modifications on DNA melting, electrophoretic mobility and DNA-protein interaction for individual oligonucleotides and their small libraries. We have developed rapid procedure for modification, deblocking and purification of fimers in 96-well plate format. Different design strategies for fimers have been tested with ThermoFidelase-2A, -2B and -2C, deaza-dGTP and dGTP in various thermal cycling protocols. We found that fimer design eliminates many restrictions on choosing primer sequence. Our results demonstrate feasibility of suppressing non-specific PCR amplification and primer-dimer formation after 100-400 cycles, synergy of chemical and enzymatic tools to sequence through strong stop and long simple repeats and sequence directly off sub-microgram quantities of bacterial genomic templates. The implementation of fimers in high-throughput projects will be presented.

We have achieved contiguity and high total and local quality of base calls starting from 2x - 5x shotgun coverage in draft human BAC projects. The major conclusion is that workflow for finishing low-coverage projects differs significantly from that for full shotgun projects and has become manageable due to the increased power of sequencing chemistry. For BAC-end sequencing projects we have developed long fimers and ThermoFidelase-2E to accelerate kinetics of primer annealing to minute quantities of template DNA and 400x(1 min) sequencing protocol. We have increased detection sensitivity to 10 ng BAC and obtained high quality reads from 30 ng BAC. New protocol is compatible with the yield of BAC DNA from 1-ml cultures in 96 well plate format.

Enhanced reaction chemistry has allowed us to overcome major obstacles in bacterial genomic sequencing associated with high fluorescent background and low signal. We obtained high quality reads from as low as 100-300 ng genomic template. Applications of direct genomic sequencing to the discovery of novel genes and characterization of bacterial populations will be presented.

Supported in part by DOE and NIH (DE-FG02-98ER82557 and 2R44GM55485-02).

16. Chemical Conversion of Boronated PCR Products into Bidirectional Sequencing Fragments

Barbara Ramsay Shaw, Kenneth W. Porter, Ahmad Hasan, Kaizhang He, and Jack Summers
Department of Chemistry, Duke University, Durham, NC 27708-0346
brshaw@chem.duke.edu

We developed an alternate sequencing chemistry which avoids cycle sequencing, allows direct bidirectional genomic sequencing, and permits direct loading of PCR products onto the separating system. The method employs template-directed enzymatic, random incorporation of small amounts of boron-modified nucleotides (i.e. 2'-deoxynucleoside 5'-alpha-[P-borano]-triphosphates) during PCR amplification. The position of the modified nucleotide in each PCR product can be revealed in two ways, either enzymatically (as previously described¹) or chemically. Both approaches take advantage of differences in reactivity of the normal and boronated nucleotidic linkages to generate PCR sequencing fragments that terminate at the site of incorporation of the modified nucleotide. By employing labeled PCR primers, the original PCR products are able to be converted directly into bidirectional sequencing fragments.

In the enzymatic approach, the modification of a phosphate into a boranophosphate internucleotidic linkage prolongs its lifetime toward degradation by nucleases. The sequential hydrolysis by 3'-5' exonuclease III is thereby blocked by a boranophosphate, resulting in fragments that terminate in a nucleoside boranophosphate. However, normal and borano-phosphate linkages with a 3'-cytosine are more susceptible to exonuclease degradation than other purines and pyrimidines, which reduces band uniformity. A series of base-modified cytosine derivatives were therefore synthesized and tested for nuclease resistance. Substitution at the C-5 position of cytosine by alkyl groups (ethyl and methyl) markedly enhances the cytidine boranophosphate resistance towards exonuclease III (i.e., 5-ethyl-dC > 5-methyl-dC > dC \approx 5-bromo-dC > 5-iodo-dC). The best analog, 5-

ethyl-a-borano-dCTP, not only showed an increased resistance to exonuclease III compared to the a-borano-dCTP used previously in our method, but did so without affecting incorporation and resulted in more even banding patterns². Analysis with Basefinder software (M. Giddings) takes into account any mobility changes, permitting increased consistency and accuracy. The enzymatic approach may find use in applications where high resolution of longer fragments requires stronger signals at longer read lengths, because the distribution of fragments produced by nuclease digestion is skewed to long fragments.

In the chemical approach, we have examined several methods for generating sequencing fragments, as an alternative to exonuclease chew-back. First, we identified reagents that selectively cleave the backbone of the PCR product at deoxy boranophosphate linkages, while leaving the normal phosphodiester linkages intact. Second, we synthesized a new boranophosphate RNA dimer analogue³ and found conditions under which the ribo boranophosphate linkage is considerably more susceptible to cleavage than a deoxy or normal phosphodiester linkage. We then synthesized diastereomers of ribonucleoside 5'-(a-P-borano)triphosphates⁴ and showed that one isomer can be incorporated readily into RNA with T7 RNA polymerases, yielding boronated transcripts that are thousands of nucleotides long. We are now examining DNA polymerases that can incorporate the boronated RNA triphosphates into DNA. Also under investigation are agents that can result in colorimetric detection of boranophosphate. Direct sequencing of PCR products by cleavage of boranophosphates should simplify mono- and bidirectional sequencing and provide a simple, direct, and complementary method to cycle sequencing.

1. K.W. Porter, J. D. Briley, and B. R. Shaw, "One-Step PCR Sequencing with Boronated Nucleotides", *Nucleic Acids Research* 25, 1611-1617 (1997).

2. K. He, A. Hasan, Bozena Krzyzanowska and B. Ramsay Shaw, "Synthesis and Separation of Diastereomers of Ribonucleoside 5'(a-P-Borano)triphosphates". *Journal of Organic Chemistry* 63(17), 5769-5773 (1998).
3. K. He, D. S. Sergueev, Z. A. Sergueeva and B. Ramsay Shaw, "Synthesis of Diuridine 3',5'-Boranophosphate: H-Phosphonate Approach." *Tetrahedron Letters* 40, 4601-4604 (1999).
4. K. He, K. W. Porter, A. Hasan, J. D. Briley and B. R. Shaw, "Synthesis of 5-Substituted 2'-Deoxycytidine 5'-(a-P-Borano)triphosphates, their Incorporation into DNA and Effects on Exonuclease". *Nucleic Acids Research* 27, 1788-1794 (1999).

17. Human and Mouse BAC Libraries for Genome Sequencing, Mapping, and Functional Analysis

Kazutoyo Osoegawa, Chung Li Shu, Aaron Mammoser, Joe Catanese, and Pieter J. De Jong
 Department of Cancer Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263 and Children's Hospital Oakland Research Institute, Oakland, CA 94609
pieter@dejong.med.buffalo.edu

Our earlier 25-fold redundant human BAC library (RPCI-11; EcoRI fragments) has recently been expanded with an additional 7-fold genome redundancy from the same donor using MboI-digested DNA. Insert sizes averaged 173 and 195 kb for the early and late parts of the library, respectively. A 1.5-Mb BAC contig was extensively analyzed to test the human BAC library for clonal integrity and fidelity. The results indicate the absence of chimeric clones and 19 rearranged clones in the contig of 169 BACs. Three murine libraries (each 11-fold genome redundant) have previously been constructed various digest strategies, two strains (129S6/SvEvTac and C57BL/6J) and either PAC or BAC vectors. The BAC library for the C57BL/6J strain (designated RPCI-23) is most significant in

view of the large EcoRI-inserts (average 200 kb) and because it was selected as a preferential source for murine genome sequencing. To obtain additional representation for the C57BL/6J strain, we created an additional the C57BL/6J BAC library (RPCI-24) from male DNA partially digested with MboI (average insert sizes about 155 kb). To permit the cloning of sheared DNA, a new vector, pTARBAC6, was constructed with two BstXI sites for cloning. The BstXI sites have non-complementary ends to avoid vector self-ligation. Blunt-ended DNA fragment are ligated to a BstXI linker to create ends complementary with the vector. In pilot experiments, fragments were cloned from HincII partially-digested DNA and DNaseI partially-digested DNA resulting in average insert sizes around 100 and 60 kb, respectively. To maximize randomness of the BAC cloning process, we are optimizing the cloning of blunt-ended fragments obtained by shearing. Information on current libraries can be found at <http://bacpac.med.buffalo.edu>.

Supported by grants from the U.S. DOE (#DE-FGO3-94ER61883) and NIH (#1RO1RGO1 165).

18. Human and Mouse BAC Ends

Shaying Zhao, Mark D. Adams, Joel Malek, Lily Fu, Bola Akinretoy, Sofiya Shatsman, Maureen Levins, Stephany McGann, Keita Geer, Getahun Tsegaye, Margaret Krol, Peter Choi, Tamara Feldblyum, William Nierman, and Claire Fraser
The Institute for Genomic Research, Rockville, MD 20850
szhao@tigr.org

End sequences from Bacterial Artificial Chromosomes (BACs) provide highly specific sequence markers in large-scale sequencing projects. To date, we have generated >300,000 BAC end sequences (BESs) from >186,000 human BAC clones with the following properties. 1) Over 60% of the clones have BESs from both ends representing 5X coverage of the human genome by the paired-end clones. 2) The average read length is ~460 bp providing a total of 141 MB covering ~4.7% of the genome. 3) The average phred Q20 length is ~400 bp giving an identity of >99% to the human finished

sequences. 4) Over 90% of the BESs faithfully represent the original clones and over 85% of the paired-end clones have both ends tracked correctly. This high quality of data gives BAC end users a high confidence in 1) retrieving the right clones from the BAC libraries based on the BAC end sequence matches; and 2) building a minimum tiling path of sequence-ready clones across the genome and building genome assembly scaffolds. Our sequence analyses indicate that BESs from human BAC libraries developed at The California Institute of Technology (CalTech) and Roswell Park Cancer Institute (RPCI) have similar properties. The analyses have highlighted differences in insert size for different segments of the CalTech library. Problems with the fidelity of tracking of sequence data back to physical clones have been observed in some subsets of the overall BES dataset. The annotation results of BESs for the contents of available genomic sequences, sequence tagged sites (STSS), expressed sequence tags (ESTs), protein encoding regions and repeats indicate that this resource will be valuable in many areas of genome research. The URL for human BAC ends is http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_intro.html.

We have been funded to end sequence the mouse BACs from RPCI-23 library within the next year (http://www.tigr.org/tdb/bac_ends/mouse/bac_end_intro.html). To date, we have over 25,000 mouse BESs with a quality similar to our human ends. In addition, all end sequencing are being conducted on the ABI 3700 sequencers to eliminate the lane tracking errors experienced on the ABI 377 sequencers. We expect that our mouse ends will have 1) an average read length of 500 bp; 2) an average phred Q20 bases of 400; 3) over 90% of the clones having paired-ends; and 4) a clone tracking accuracy of 99%. The mouse resource will have an even higher quality than the current human ends.

19. Library Strategy for Genome Sequencing Projects

William C. Nierman

The Institute for Genomic Research, Rockville, MD
20850

wnierman@tigr.org

Microbial genome sequencing projects at TIGR have been conducted using two-ended clone sequencing primarily from a small 1.5 to 2 kb insert size plasmid library supplemented with sequence reads from both ends of several hundred 15 - 20 kb lambda clones. We have recently implemented a new shotgun library strategy which incorporates sequence reads from both a small 2 kb insert size library and a larger 10 kb insert size plasmid libraries. This strategy has resulted in a dramatic decline in the number of gaps at the end of the random phase of sequencing for which there is no clone coverage, greatly simplifying the process of closure of the genome. Data from several TIGR sequencing projects will be provided to document this conclusion.

BAC based projects for organisms such as the human and mouse are undertaken to minimize the assembly and closure problems of large repeat rich genomes. The BAC libraries supporting these projects were constructed using partial restriction digests to fragment the genomic DNA prior to ligation to the BAC vector. Due to the non-random distribution of restriction sites for any enzyme in genomic DNA libraries thus constructed always have over-representation and underrepresentation of some regions and no coverage of some small fraction of the genome. These regions of no coverage are revealed as gaps in the BAC contig maps produced by analysis of restriction fingerprints of the BAC clones.

In order to develop a resource for providing clone coverage across these gaps in the BAC contig maps for the human and mouse sequencing efforts, we are constructing BAC libraries from random sheared genomic DNA. The targeted insert sizes are 50 and 100 kb. The human libraries are being constructed

with donor DNA collected in strict accordance with appropriate informed consent at Celera Genomics (Hamilton Smith). The mouse libraries are being constructed from male C57BL/6J DNA provided by The Jackson Laboratory with appropriate animal committee review at both The Jackson Laboratory and at TIGR. All libraries are being constructed with very narrow insert size cuts to facilitate easy detection of consequential deletions by clone insert size determinations.

Instrumentation

20. New Technologies for Genome Sequencing and Expression Analysis

Wayne P. Rindone, John Aach, Martha Bulyk, George M. Church, Jason Hughes, Abby Mcguire, Pam Ralston, Martin Steffen, and Saeed Tavazoie
Department of Genetics, Harvard Medical School, Boston, MA 02115
wrindone@arep.med.harvard.edu

We pursue new, basic technologies for collecting low-cost, quantitative molecular data, useful for a variety of sequencing and/or functional genomics goals. We have developed a set of new methods for producing microarrays of Acrydite-immobilized PCR colonies and subsequent Fluorescent In Situ Sequencing (FISSEQ) on the slides. We have integrated mRNA expression data from SAGE, Chip, and microarray experiments into the ExpressDB database and developed a flexible Web-based search engine for exploring the data for clustering and DNA-motif analysis. The process also lead to insights into methodological improvements that would make such information more amenable to comparison among other laboratories.

Nucleic Acids Res 27(24):e34;pp. 1-6
Nature Genetics 22:281-5
<http://arep.med.harvard.edu/>

21. High-Performance DNA Sequencing and Analysis

Richard A. Mathies
Department of Chemistry, University of California, Berkeley, CA 94720
rich@zinc.cchem.berkeley.edu

Capillary array electrophoresis (CAE) systems¹ coupled with high sensitivity detection provided by

energy-transfer labeling reagents² is now the accepted standard for high-throughput DNA sequencing facilities. Further advances are focused on the development of capillary array systems capable of running more than 96 capillaries as illustrated by Scherer et al.³ and the development of microfabricated CAE systems that provide higher throughput as well as the important ability to integrate microfluidic chemistries. Toward this end we have recently shown that microfabricated CE channels only 7-cm long can produce >500 bp 4-color sequencing reads in under 30 minutes⁴.

Microfabrication also permits the production of very high density electrophoretic analysis devices that provide unprecedented analysis throughput. Radial microplates coupled with a novel rotary confocal scanning system have been developed that can rapidly analyze 96 genotyping samples in parallel in seconds on 5-cm long channels⁵. To integrate arrays of 10-15 cm long channels on a radial microplate, it is necessary to devise ways to fold channels without reducing resolution. We have determined that "pinched turns", where the channel width is reduced before the turn and widened after the turn, enable folded designs that minimize turn-induced broadening while maintaining facile matrix introduction⁶.

Sequencing results on radial microplates with 96 15-cm long channels fabricated on a 6"-diameter wafer will be presented. We have also developed microfabrication methods for the integration of nanoliter volume PCR sample preparation directly with microfabricated CE analysis systems⁷.

Genotyping results will be presented along with plans for integrated thermal cycling devices. The implementation of microfabricated separation systems with integrated chemistries will be the next paradigm shift in DNA sequencing and genomic analysis.

1. Kheterpal, I. and Mathies, R. A. Capillary Array Electrophoresis DNA Sequencing, *Analytical Chemistry*, 71, 31A-37A (1999).
2. Xie, J., Hung, S.-C., Glazer, A. N. and Mathies, R. A. Energy Transfer Fluorescent Labels for DNA Sequencing and Analysis, in *Topics in Fluorescence Spectroscopy*, Volume 7: ANA Technology, in press (1999).
3. Scherer, J. R., Kheterpal, I., Radhakrishnan, A., Ja, W. W. and Mathies, R. A. Ultra-High Throughput Rotary Capillary Array Electrophoresis Scanner for Fluorescent DNA Sequencing and Analysis, *Electrophoresis* 20, 1508-1517 (1999).
4. Liu, S., Shi, Y., Ja, W. W. and Mathies, R. A. Optimization of High-Speed DNA Sequencing on Microfabricated Capillary Electrophoresis Channels, *Anal. Chem.* 71, 566-573 (1999).
5. Shi, Y., Simpson, P., Scherer, J. R., Wexler, D., Skibola, C., Smith, M. T. and Mathies, R. A. Radial Capillary Array Electrophoresis Microplate and Scanner for High-Performance Nucleic Acid Analysis, *Anal. Chem.* 71, 5354-5361 (1999).
6. Paegel, B. M., Hutt, L. D., Simpson, P. C. and Mathies, R. A. Turn Geometries for Minimizing Band Broadening in Microfabricated Capillary Electrophoresis Channels, *Analytical Chemistry*, submitted.
7. Lagally, E., Simpson, P. C. and Mathies, R. A. Monolithic Integrated Microfluidic DNA Amplification and Capillary Electrophoresis System, *Sensors and Actuators B*, submitted (1999).

22. Radial Capillary Array Electrophoresis Microplate and Scanner for High-Performance DNA Sequencing and Analysis

Yining Shi¹, Brian M. Paegel¹, James R. Scherer¹, Peter C. Simpson¹, David Wexler¹, Christine Skibola², Martyn T. Smith², and Richard A. Mathies¹
¹Department of Chemistry and ²School of Public Health, University of California, Berkeley, CA 94720
 yining@zinc.cchem.berkeley.edu

The design, fabrication and operation of a radial capillary array electrophoresis (CAE) microplate and scanner for high-throughput DNA analysis are presented¹. The microplate consists of a central common anode reservoir coupled to 96 microfabricated separation channels connected to sample injectors on the perimeter of the wafer. Detection is accomplished by a laser-excited rotary confocal scanner with four-color detection. Loading of 96 samples in parallel is achieved using a pressurized capillary array system. High-quality separations of 96 pBR322 restriction digest samples are achieved in <120 s using a 4"-diameter microplate. The practical utility and multicolor detection capability of this system is demonstrated by analyzing 96 methylenetetrahydrofolate reductase (MTHFR) alleles in parallel using a non-covalent 2-color staining method. This work establishes the feasibility of high-performance genotyping with capillary array electrophoresis microplates.

To explore the capabilities of our radial CAE microplate and scanner for high-speed and high-throughput DNA sequencing, we have designed and fabricated a CAE microplate containing 96 folded 12-cm-long separation channels on a 6"-diameter wafer. While high-quality four-color sequencing separations can be achieved on 7-cm-long straight microchannels², integration of an array of such straight channels into a 6"-diameter wafer in the radial format is difficult due to the dimensional restrictions of the wafer. To address this issue, Paegel et al. introduced an optimized channel design which allows the fabrication of 96 folded separation channels on a 6"-diameter wafer. Each of the 96 folded 12-cm-long channels has two complementary tapered turns that minimize turn-induced band broadening during electrophoresis separations^{3,4}. Four-color sequencing separations and automatic base-calling analyses of 96 single stranded M13mp18 DNA sequencing samples with our 6"-diameter radial CAE microplate and scanner will be presented.

1. Y. Shi, P. C. Simpson, J. R. Scherer, D. Wexler, C. Skibola, M. T. Smith and R. A. Mathies. *Anal. Chem.* 1999, 71, 5354-5361.
2. S. Liu, Y. Shi, W. W. Ja and R. A. Mathies. *Anal. Chem.* 1999, 71, 566-573.
3. B. M. Paegel, L. D. Hutt, P. C. Simpson and R. A. Mathies. *Anal. Chem.* (submitted).

4. See abstract "Turn Geometries for Minimizing Band Broadening in Microfabricated Capillary Electrophoresis Channels," by B. M. Paegel, L. D. Hutt, P. C. Simpson, and R. A. Mathies.

23. Turn Geometries for Minimizing Band Broadening in Microfabricated Capillary Electrophoresis Channels

Brian M. Paegel, Lester D. Hutt, Peter C. Simpson, and Richard A. Mathies

Department of Chemistry, University of California, Berkeley, CA 94720

brian@zinc.cchem.berkeley.edu

Microfabricated capillary electrophoresis (CE) devices have dramatically increased the speed and performance of chemical and biochemical analyses¹. As larger numbers of microfabricated structures are placed on a wafer to form capillary array electrophoresis microplates², or as one attempts to fabricate longer channels to enhance resolution in sequencing applications³, it becomes necessary to fold channels. Folding CE channels gives rise to turn-induced band broadening due to dispersion forces in the turn. To circumvent this limitation, microfabricated channels were constructed with a variety of turn geometries for the purpose of minimizing turn-induced band broadening. Column efficiencies of channels with a variety of turn designs were determined by quantitating the resolution of separations of HaeIII digests of phiX174 bacteriophage DNA. Most advantageously, tapered turns were created by narrowing the channel width at the start of a turn to reduce the channel width, followed by widening the channel at the end of the turn. The radius of curvature of the turn, the length of the tapered region, and the degree of tapering were explored. These experiments were performed using our novel rotary scanner⁴, which permits the simultaneous interrogation of a separation at three or more points along a serpentine channel. Serpentine channels were monitored before the first turn, after one turn, and after two turns. Turns with a minimum radius of curvature (250 μm), a minimum length of

taper (55 μm), and a maximum tapering ratio (4:1 separation channel width to turn channel width) were found to provide the highest number of theoretical plates for the 271 and 281 base pair fragments of the phiX174 HaeIII ladder. The optimal turn configuration was then used to perform M13 DNA sequencing separations with an effective separation length of 15 cm. High-quality separations to 800 bp were observed in only 35 minutes. These extended length channel designs have been incorporated in high-throughput 96-channel microplates for DNA sequencing⁵.

1. Simpson, P. C., Woolley, A. T. and Mathies, R. A. *J. Biomedical Microdevices* 1998, 1, 7-26.
2. Simpson, P. C., Roach, D., Woolley, A. T., Thorsen, T., Johnston, R., Sensabaugh, G. F. and Mathies, R. A. *Proc. Natl. Acad. Sci. U.S.A.* 1998, 95, 2256-2261.
3. Liu, S. R., Shi, Y., Ja, W. W. and Mathies, R. A. *Anal. Chem.* 1999, 71, 566-573.
4. Shi, Y., Simpson, P. C., Scherer, J. R., Wexler, D., Skibola, C., Smith, M.T. and Mathies, R.A. *Anal. Chem.* 1999, 71, 5354-5361.
5. See abstract entitled "Radial Capillary Electrophoresis Microplate and Scanner for High-Performance DNA Sequencing and Analysis," Y. Shi, B. M. Paegel, J. R. Scherer, D. Wexler, C. Skibola, M. T. Smith and R. A. Mathies.

24. Integrated Microfluidic DNA Amplification and Analysis Systems

Eric T. Lagally¹, Daojing Wang², Charles Emrich², and Richard A. Mathies²

¹UCB/UCSF Joint Graduate Group in Bioengineering and ²Department of Chemistry, University of California, Berkeley, CA 94720
lagally@zinc.cchem.berkeley.edu

Microfabrication technology is an effective method for creating integrated devices for chemical and biochemical analysis¹⁻³. Our early work in the development of integrated devices included the

manufacture of a hybrid Si polymerase chain reaction (PCR) reactor mated with a glass capillary electrophoresis (CE) device⁴ and the development of a CE device with an integrated electrochemical detector⁵. In more recent work, we have developed a fully integrated DNA analysis system microfabricated in glass consisting of controlled fluid delivery using active and passive elements, PCR amplification, and direct coupling to a capillary electrophoretic separation⁶. Samples are introduced at a common sample bus and loaded precisely into a 280-nanoliter volume PCR reactor using valves and hydrophobic vents. The sample is cycled between three temperatures using a resistive heater mounted on the bottom of the chip, and the amplification products are then directly injected and separated on a capillary electrophoresis channel. The device takes only 33 seconds/cycle, representing a vast improvement over conventional thermal cycling systems, which can take up to 5 minutes/cycle. Amplicons from the M13/pUC19 plasmid have been produced from only 20 starting copies/ μL or 5 copies in the reactor. This amplification is among the most sensitive compared both to previous static systems, which require $\sim 6,000$ starting copies⁷, and to continuous-flow geometries which require as many as ~ 108 starting copies⁸. The high sensitivity of this device allows studies at the single molecule level.

We have also developed a microfluidic DNA capture chamber for sequencing sample clean-up and concentration. This device uses microfluidic elements to flow raw sequencing samples through a filter chamber filled with oligonucleotide-labeled capture beads. The extension products of interest are selectively captured on the beads and subsequently released using formamide and heat. The capture chamber is directly connected to a capillary electrophoresis channel for immediate sequencing. These results demonstrate a key link in the development of an integrated microfluidic system that performs complete genetic analyses at sub-microliter volumes.

1. Simpson, P. C., Roach, D., Woolley, A. T., Thorsen, T., Johnston, R., Sensabaugh, G. F. and Mathies, R. A. *Proc. Natl. Acad. Sci. U. S. A.* 1998, 95, 2256-2261.
2. Simpson, P. C., Woolley, A. T., Mathies, R. A. *Journal of Biomedical Microdevices* 1998, 1, 7-26.

3. Liu, S. R., Shi, Y., Ja, W. W. and Mathies, R. A. *Anal. Chem.* 1999, 71, 566-573.
4. Woolley, A. T., Hadley, D., Landre, P., deMello, A. J., Mathies, R. A., et al. *Anal. Chem.* 1996, 68, 4081-4086.
5. Woolley, A. T., Lao, K. Q., Glazer, A. N., Mathies, R. A. *Anal. Chem.* 1998, 70, 684-688.
6. Lagally, E., Simpson, P. C. and Mathies, R. A. *Sensors and Actuators B*, in press (2000).
7. Cheng, J., Shoffner, M. A., Hvichia, G. E., Kricka, L. J. and Wilding, P. *Nucleic Acids Res.* 1996, 24, 380-385.
8. Kopp, M. U., de Mello, A. J. and Manz, A. *Science* 1998, 280, 1046-1048.

25. High-Speed High-Throughput Mutation Detection

Qiufeng Gao, Ho-Ming Pang, and Edward S. Yeung
Ames Laboratory, Iowa State University, Ames, IA 50011
yeung@ameslab.gov

Single-nucleotide polymorphism (SNP) detection has been the focus of much attention recently. Although many methods have been reported, low-cost, high-throughput and high-detection-rate methods are still in demand. We present a fast and reliable mutation detection scheme based on temperature-gradient capillary electrophoresis. A large temperature gradient ($10\text{ }^\circ\text{C}$) was applied with a precision of $0.02\text{ }^\circ\text{C}$ and a temperature ramp of $0.7\text{ }^\circ\text{C}/\text{min}$. Multiple unlabeled samples from PCR reaction were injected and analyzed. Ethidium bromide was used as the intercalating dye for laser-induced fluorescence detection. The mutations were identified by comparing the electrophoretic patterns of the heteroduplex with that of a homoduplex reference without prior knowledge of the DNA sequence. Mutations in all five test samples were successfully detected with high confidence. This scheme is demonstrated in 96-capillary array electrophoresis for screening single-point polymorphism in large numbers of samples.

26. Micro-Fabricated Devices for Concentrating DNA by Induced-Dipole Trapping

Charles Asbury and Ger Van Den Engh
Department of Molecular Biotechnology, University
of Washington, Seattle WA 98195-7730
engh@biotech.washington.edu

DNA molecules placed in a divergent electrical field experience an attractive force towards regions of higher field strength. This force is a result of a charge-dipole that is induced along the molecule's axis. The dipole interacts with the field gradient. Because induced dipoles always oppose the field, the attractive force is independent of the fields polarity. Migration due to dipole forces can be observed with both AC and DC fields. In contrast, electrophoretic forces, which are due to the native charge of DNA molecules, always move the molecules towards the positive electrode. Homogeneous oscillating fields with a 50% duty cycle do not cause a net displacement of DNA. In oscillating fields with steep gradients the molecules move towards the field's origin.

We are developing small chambers for manipulating DNA that allow independent application of both electrophoretic and induced-dipole forces. The devices consist of thin metal layers on a quartz substrate. By combining the two types of forces, cohorts of DNA molecules can be concentrated and moved with high precision. Dipole traps concentrate DNA out of a dilute solution. Electrophoretic forces can then be employed to move DNA cohorts between traps.

We are seeking the optimal conditions for dipole trapping of DNA. The ionic composition of the medium and the frequency, strength, and gradient, of the field are important. We current use a salt concentration below 10 mM and use fields oscillating between 30-100 Hz. By comparing the rate of Brownian movement and diffusion the trapping forces can be quantitated. We are using this information to develop devices in which DNA can be

separated by size without the use of a sieving medium. We will present a gold-on-quartz device that consists of a capillary lined with dipole traps. Such capillaries can be combined with other modules to perform complex operation of small cohorts of DNA molecules.

27. Fully Automated Multiplexed Capillary Systems for DNA Sample Analysis

Qingbo Li, Thomas E. Kane, Changsheng Liu,
Harry Zhao, Gary W. Loge, John Kernan, Songsan
Zhou, Kevin Levan, Heidi Monroe, and David Fisk
SpectruMedix Corporation, 2124 Old Gatesburg
Road, State College, PA 16803
qbli@spectrumedix.com

SpectruMedix has developed a commercial 96-capillary electrophoresis instrument for DNA analysis. All operation steps are automated, including capillary conditioning, gel filling, sample introduction, electrophoresis, and data acquisition. It can perform seven consecutive runs without human intervention. Simple yet highly efficient optical design renders an extremely robust detection system that shows excellent stability. The instrument uses a CCD detector to simultaneously record fluorescence signal from all 96 capillaries with on-column laser excitation. Multi-wavelength detection is implemented with a miniaturized spectrometer utilizing a transmission grating. A replaceable linear-polymer matrix provides high-performance separation for DNA sequencing and genotyping. Currently, the instrument is capable of routinely separating 500+ bases with 98% basecalling accuracy in a 2-hr run. By using a gel matrix that is optimized for longer read, up to 770 bp separation with 98% basecalling accuracy has been achieved in a 3-hr run including capillary conditioning, gel filling, sample introduction, and electrophoresis. The instrument includes an electronic unit that allows monitoring the electrophoresis current for each of the 96 capillaries. Further, a recently developed algorithm allows automated color deconvolution

matrix file construction, avoiding the need for a calibration run. This process further enhances the robustness of the instrument. A comparison of the SCE9600 performance to an ABI377 using the same sample has been performed.

A prototype 384-capillary array electrophoresis instrument has been developed for higher throughput analysis. The 384-capillary instrument design is based on the SCE9600 platform, so the 96-capillary instrument can be readily upgraded to obtain 4X higher throughput. The injection end of the 384-capillary array is configured in 16x24 format so that it is compatible with commercial 384-well microtiter tray technology. The instrument is capable of performing one genotyping or sequencing run within 2 hours. In fully automated mode, the instrument will analyze 4,608 DNA samples in a 24-hour day.

The instrument includes an electronic unit that allows monitoring the electrophoresis current for each of the 96 capillaries. This has proved to be a valuable technique for protocol diagnosis and development. Automated basecalling software analyzes a set of 96-capillary data within minutes. Further, a recently developed algorithm allows automated color deconvolution matrix file construction, avoiding the need for a calibration run. This process further enhances the robustness of the instrument since instrumental drift, if there is any, is automatically corrected by the algorithm. A comparison of the sequencing performance of a SCE9600 to the performance of an ABI377 using the same sample has been performed.

A prototype 384-capillary array electrophoresis instrument has been developed for higher throughput analysis. The 384-capillary instrument design is based on the SCE9600 platform, so the 96-capillary instrument can be readily upgraded to obtain 4X higher throughput. The injection end of the 384-capillary array is configured in 16x24 format so that it is compatible with commercial 384-well microtiter tray technology. The instrument is capable of performing one genotyping or sequencing run within 2 hours. In fully automated mode, the instrument will analyze 4,608 DNA samples in a 24-hour day.

28. Development and Evaluation of a PCR-Based Sequencing Routine for Use on the ABI 3700 Capillary Machine

Lynne Goodwin, Owatha Tatum, Olga Chertkov, Judith Cohn, and P. Scott White
Bioscience Division and DOE Joint Genome Institute,
Los Alamos National Laboratory, Los Alamos, NM
87545
swhite@telomere.lanl.gov

The rapid throughput of the new generation of capillary electrophoresis instruments for automated DNA sequencing presents unique problems for high throughput applications. Increased demands for sequencing substrate of high quality place strains on template preparation routines and equipment, and alternative strategies that scale well and cost less are needed. We have implemented a PCR-based strategy for creating sequencing template that requires only a few, easily automated steps, and costs a fraction of commercial robotic plasmid prep protocols. The template production process begins with overnight culture of subclones, followed by pin-stamp inoculation of PCR plates (96 or 384-well), and finishes with an enzymatic treatment of the PCR products to provide suitable sequencing template. Forward and reverse sequencing reactions are then performed directly on these templates using vector primers, followed by a single isopropanol precipitation, and resuspension in 1/10 TE. The 1/10 TE resuspension buffer allows samples to remain on the capillary instrument at ambient temperature for extended periods of time, which is necessary to obtain essentially hands-off automated sequencing of 6 to 8 96-well plates in a 24 hr period.

Concerns about the quality of sequence derived from PCR templates generally focus on homopolymer and simple repeat stretches that are difficult to accurately reproduce using Taq DNA polymerase. In addition, the number of failed reactions and the read length of PCR template-generated sequence is thought to be less than sequence generated from highly purified plasmid DNA. If the costs are favorable and the data of sufficient quality, then such a template production strategy is attractive; otherwise, the higher up front cost of plasmid template preparation is money well spent, as long as throughput is not limited. We have

examined the quality of data obtained from an ABI 3700 using this process, and directly compared subclone sequence generated from plasmid templates on ABI 377 instruments. The results will be presented, as will a discussion of the implementation of this process.

29. Rapid and Accurate Detection of Human Functional SNPs Using a Base Stacking Microelectronic DNA Chip

Glen Evans¹, David Canter², Purita Ramos¹, Ray Radtkey², Ron Sosnowski², Gene Tu², James O'Connell², and Michael Nerenberg²

¹Department of Internal Medicine, University of Texas Southwestern Medical Center at Dallas, TX and ²Department of Molecular Biology, Nanogen, Inc., San Diego, CA
gaevans@home.com

Large scale genomic sequencing is revealing thousands of useful functional and non-functional single nucleotide polymorphisms (SNPs). Technology for the accurate, rapid and expandable assessment of large numbers of human SNPs in parallel is needed for research and medical applications. We describe a novel technology for SNP assessment on microelectronic silicon-based DNA chips that utilizes short fluorescently-labeled oligonucleotide reporters and targets of amplified source DNA. This assay takes advantage of base stacking energies in the design of probes and has the ability of accommodate different sized amplicons in parallel. This assay has been utilized to assay functional SNPs in parallel controlled by electronic fields induced on the chip surface and two-color fluorescence detection. A panel of model functional SNPs has been developed to evaluate the utilize of this method. These markers include polymorphic sites in genes for HH (hemochromatosis), Factor V, EHI (epoxide hydrolase 1), EPHX2 (epoxide hydrolase 2) CYP19 (cytochrome P450), DTD (diastrophic dysplasia sulfate transporter), GSTA1 (alpha glutathione S transferase), GSTA12 (microsomal glutathione S transferase), NAT1 (N-acetyl transferase 1), NAT2

(N-acetyl transferase 2), ColA1 (type IV collagen), ApoCIII (apolipoprotein cIII), MGC24 (PNA-binding glycoprotein), PPP2R1B (lung cancer susceptibility polymorphism) and others. SNP detection from amplified genomic target DNA can be carried out on 100 SNPs on a single microelectronic chip with accuracy exceeding that of DNA sequencing. We have utilized this system for the systematic genotyping of more than 200 individuals for 15 SNPs, also determined by DNA sequencing, with virtually 100% accuracy. This system is imminently suited to point-of-care genetic diagnosis, forensic applications, medical diagnostics as well as large scale human genotyping for pharmacogenomics applications.

30. DNA Sequencing by Single Molecule Detection

Peter M. Goodwin, Hong Cai, James H. Werner, James H. Jett, and Richard A. Keller
Bioscience Division, Los Alamos National Laboratory, M888, Los Alamos, NM 87545 USA
pmg@lanl.gov

We are developing a method, based upon single fluorescent molecule detection, to sequence individual DNA strands. The method consists of: (1) polymerase incorporation of fluorophore-labeled nucleotides into a strand of DNA complementary to the target sequence; (2) anchoring a single fragment of fluorescently-labeled DNA, in flow, upstream of the detection volume of an ultrasensitive fluorescence flow cytometer; (3) exonuclease digestion of the free end of the anchored DNA strand to sequentially release single, fluorophore-labeled nucleotides into the flow stream; and (4) detection and identification of the individual, released fluorophore-labeled nucleotides in the order of exonuclease cleavage. We have made considerable progress towards a demonstration of single molecule DNA sequencing. Up to three of the nucleotide types, labeled with fluorophores, have been incorporated into strands of DNA 2-7 kilobases in length. Multiple strands of fluorescently-labeled DNA have been attached to

microspheres and individual microspheres have been anchored in flow upstream of the detection volume. Exonuclease cleavage of fluorescently-labeled DNA on individual microspheres anchored in flow has been observed. We have detected and identified single, tetramethylrhodamine-labeled dUMPs and Rhodamine-6G-labeled dCMPs enzymatically released from DNA strands containing both types of labeled nucleotides. The two fluorescent species were identified by correlated measurements of single molecule fluorescence burst intensity and intra-burst fluorescence lifetime. We present preliminary data demonstrating the detection of single, labeled nucleotides released by the processive exonuclease digestion of a single DNA strand.

This work was supported by the US Department of Energy, Office of Biological and Environmental Research.

31. New Optical Methods for Sequencing Individual Molecules of DNA

Jonas Korfach, Michael Levene, Stephen W. Turner, Mathieu Foquet, Harold G. Craighead, and Watt W. Webb
Applied & Engineering Physics, Clark Hall, Cornell University, Ithaca, NY 14853
jk109@cornell.edu

A new method for determining the base pair sequence of a single molecule of DNA by following the dynamical stepwise activity of DNA polymerase synthesizing the complementary strand of a given template strand is under development. The technical challenges consist in the development of suitable enzymatic systems and in the recognition of individual sequential base additions. Replacing spatial resolution of bases in the DNA by temporal resolution of sequential nucleotide additions is made possible by using near-field and multiphoton laser optics for chromophore processing, and time-resolved photon counting for detection. Confinement of the excitation volume far below the diffraction limit by nanostructured devices permits an increase of substrate concentrations by about three orders of magnitude above the nanomolar range (required for

the enzymatic systems under study), but still allowing sequential single molecule recordings and analysis.

The approach should enable the creation of a very fast sequencing protocol with long read lengths, and potentially highly parallel, integrated systems with large throughput. Each development step towards the sequencing goal appears fertile for the generation and improvement of analytic research systems capable of following biochemical and molecular biological processes (e.g., enzymatic activities) at the single molecule level. The optical tools will enable a characterization of these processes previously unattainable by conventional biochemical analysis. Specifically in respect to the sequencing proposal, this amounts to new knowledge of the photophysical and dynamical behavior of single DNA molecules, the generation and use of new fluorescent labels that can be incorporated into DNA in high densities, and the study of enzymes acting on DNA at the level of individual bases.

32. High Throughput Multiplexed mtDNA SNP Scoring Using Microsphere-Based Flow Cytometry

P. Scott White, Alina Deshpande, Lance Green, Yolanda Valdez, David C. Torney, and John P. Nolan
Bioscience Division and DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM 87545
white_paul_scott@lanl.gov

We have developed a flow cytometry-based minisequencing platform capable of extremely high-throughput, low cost assays. These are no-wash assays analyzed at less than 1 minute/sample, with superior sensitivities. Furthermore, with commercially available multiplexed microspheres we can score dozens to hundreds of SNPs simultaneously. Multiplexing, coupled with high throughput rates, makes it possible to score several million SNPs/day at costs that are a fraction of competing technologies. In addition, these assays can easily be integrated into conventional liquid handling automation systems, and require no unique instrumentation for setup and analysis.

Multiplexing is enhanced by universal capture tags consisting of carefully designed, unique DNA tails incorporated into each minisequencing primer. These are complementary to address tags attached to discrete populations of microspheres in multiplexed sets. This enables simultaneous minisequencing of many SNPs in solution, followed by capture onto the appropriate microspheres for multiplexed analysis by flow cytometry. High signal-to-noise ratios, ease of setup, flexibility in format and scale, and low cost of these assays make them versatile and valuable tools for studies at a wide range of scales where SNP scoring is needed.

A typing method that could rapidly and inexpensively score 100 or more SNPs located in the mitochondrial genome would be extremely valuable in population and evolutionary biological studies, as well as a powerful forensics tool. We present results from multiplexed analyses of mtDNA and HLA SNPs, performed on a few large PCR amplicons, each containing numerous SNPs that have been scored simultaneously.

33. Mass Spectrometric Analysis of Genetic Variations

Lloyd M. Smith

Department of Chemistry, University of Wisconsin-Madison, Madison, WI
smith@chem.wisc.edu

In the last decade two powerful new tools for the mass spectrometric analysis of biomolecules have been developed, Matrix-Assisted Laser Desorption Mass Spectrometry (MALDI-MS), and Electrospray Ionization Mass Spectrometry (ESI-MS). The power of these methods lies in their ability to produce and mass analyze intact gas phase ions from very large molecules such as proteins and nucleic acids. The speed, accuracy, and sensitivity of the technologies make them well-suited to address a number of problems in genetic analysis, including the analysis of DNA sequence, genetic variations, and gene expression. Results in these areas will be presented,

including recent work in which single nucleotide polymorphisms (SNPs) in genomic DNA may be analyzed without need for a prior PCR amplification step.

34. Affinity Capture and Mass Spectrometry of Targeted Proteins in Mice

Stephen J. Kennel¹, Gregory B. Hurst², Linda J. Foote¹, and Michelle V. Buchanan^{1,2}

¹Life Sciences Division and ²Chemical and Analytical Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6124
buchananmv@ornl.gov

We are developing new mass spectrometry-based methods for large-scale screening of targeted proteins in mice, taking advantage of the ability of mass spectrometry to detect compounds sensitively and to identify both normal and modified proteins unambiguously. In this initial study, we are working with the Mammalian Genetics and Development Section at ORNL to develop a method for large-scale screening of cytokine levels in mouse serum samples as a means to detect subtle abnormalities leading to chronic inflammatory diseases. Modified levels of cytokines in serum are indicative of inflammation, a condition associated with a wide variety of disease states. The new methodology involves capture of targeted cytokines from mouse serum onto antibody-derivatized aminopolystyrene beads, washing, elution, and analysis by matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS). The selectivity of the affinity separation is complemented by the m/z measurement capability of mass spectrometry, providing speed and specificity advantages over conventional ELISA techniques. Tumor necrosis factor α (TNF- α), a 17 kDa proinflammatory cytokine that has a relatively high serum concentration in acute reactions and seems to be an early effector in inflammatory cytokine cascades, has been used as a model analyte (Hurst, G.B.; Kennel, S.J.; Foote L.J.; Buchanan, M.V. *Anal. Chem.* 1999, 71, 4727-4733). In parallel with

MALDI-MS, experiments using ¹²⁵I-labeled TNF- α and gamma detection allow independent optimization of the affinity capture, and indicate that the capture methodology is viable from <100 pg/mL to >50 ng/mL. MALDI-MS currently allows reliable detection down to 1 ng TNF- α in an initial 100 μ L sample volume (mouse limited), and we are working to improve this figure. To demonstrate selectivity, TNF- α spiked into mouse serum can be concentrated onto the beads and detected by MALDI-MS with little interference from the many other components present in serum. Control experiments indicate that non-specific binding is minor. Preliminary MALDI-MS results on other cytokines (IL-6, IL-1 β , IL-2, and IFN- γ) indicate that the MALDI matrix conditions must be carefully optimized for each cytokine to allow sensitive detection.

Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by Lockheed Martin Energy Research Corp. for the U. S. Department of Energy under Contract No. DE-AC05-96OR22464.

35. Rapid Quantitative Measurements of Proteomes

Richard D. Smith, Ljiljana Pasa Tolic, Mary S. Lipton, Pamela K. Jensen, Gordon A. Anderson, and Timothy D. Veenstra
Pacific Northwest National Laboratory, Richland, WA 99352
rd_smith@pnl.gov

The patterns of gene expression, protein post-translational modifications, covalent and non-covalent associations, and how these may be affected by changes in the environment, cannot be accurately predicted from DNA sequences. In addition, direct protein measurements now constitute the most effective method for determining open reading frames for small proteins. Therefore, proteome characterization is increasingly viewed as a necessary complement to complete sequencing of the genome. Approaches for proteome characterization are increasingly based upon mass spectrometric analysis of in-gel digested electrophoretically separated proteins, allowing relatively rapid protein identification compared to conventional approaches. However, this technique remains constrained by the

speed of the 2-D gel separations, the sensitivity needed for protein visualization, the speed and sensitivity of subsequent mass spectrometric analyses for identification, and the limitations of spot visualization for quantitation.

Our objective is to circumvent the limitations of this approach by directly characterizing the cell's polypeptide constituents by combining fast separations and the mass accuracy and sensitivity obtainable with Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. Several approaches are presently being pursued; one based upon the analysis of intact proteins and the second upon global approaches for protein digestion and accurate peptide mass analysis (i.e. the use of "accurate mass tags"). A key attraction of FTICR is the enhanced facility for protein identification based upon the use of genome sequence data. Alternative versions of proteomes using stable isotope labeling are applied for the purposes of accurate quantitation. We describe the status of our efforts towards the development of a high throughput proteomics capability.

We thank the Office of Biological and Environmental Research, U. S. Department of Energy, for support of this research under contract DE-AC06-76RLO 1830.

36. DNA Characterization by Electrospray Ionization-FTICR Mass Spectrometry

David S. Wunschel, Bingbing Feng, Ljiljana Pasa Tolic, Mary S. Lipton, and Richard D. Smith
Pacific Northwest National Laboratory, Richland, WA 99352
rd_smith@pnl.gov

Mass spectrometry offers the potential for high speed DNA sequencing and ultra-sensitive characterization. Ongoing work in the laboratory is exploring approaches based upon electrospray ionization (ESI) and/or Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. These efforts have included advanced methods for the characterization of polymerase chain reaction (PCR) products¹, enzymatically produced oligonucleotide mixtures, modified DNA and the development of methods for

the analysis of DNA large fragments. High mass accuracy measurements for PCR products allowing a single base substitutions to be detected at >250 bp level with de novo identification of an unreported base substitution. This capability also allows the identification of small differences in mass such as those arising from methylation². Study of DNA damage/modifications in their sequence context will likely have to occur from within multi-component mixtures. The capability for this has been demonstrated using a multi-component reaction where a base pair deletion was identified with the putative identification of inter-operon variability within a single bacterial strain. These efforts are also being extended to exploit the non-destructive nature of FTICR for recovery (i.e., "soft-landing") of mass-selected modified DNA segments, following high resolution FTICR analysis and separation (i.e., high resolution sorting), for subsequent cloning or PCR³. This capability allows for the direct selection and analysis of individual components from within mixtures. Alternatively, DNA species that cannot be identified through traditional sequencing methodologies, those containing base modifications, can be isolated and the nature and position of the modification identified. Most importantly this provides an approach for identification of low abundance modifications where few if any alternatives for their detection exist.

1. D. S. Wunschel, D. C. Muddiman and R. D. Smith, *Advances in Mass Spectrometry, Volume 14*, E.J. Karjalainen, A.E. Hesso, J.E. Jalonen, U.P. Karjalainen, Eds., Elsevier Science Publishers B.V., Amsterdam, 377-406 (1998).
2. D. S. Wunschel, L. Pasa Tolic, B. Feng and R. D. Smith, *J. Amer. Soc. Mass Spectrom.*, in press.
3. B. Feng, D. S. Wunschel, C. D. Masselon, L. Pasa-Tolic and R. D. Smith, *J. Amer. Chem Soc.*, 121, 8961-8962 (1999).

37. DNA Sequencing via Electrospray and Ion/Ion Chemistry in an Electrodynamic Ion Trap

Scott A. McLuckey¹, James L. Stephenson, Jr.², and Gregory B. Hurst²

¹Department of Chemistry, Purdue University, West Lafayette, IN 47907-1393 and ²Chemical and Analytical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831
mcluckey@purdue.edu

We are pursuing a methodology for high speed DNA sequencing based on electrospray ionization mass spectrometry employing gas-phase ion/ion chemistry in a quadrupole ion trap. DNA sequencing via mass spectrometry has been pursued by a number of groups in recent years due to its promise for the obviation of time-consuming electrophoresis-based separations required with established sequencing strategies. By far, most effort has been directed toward matrix-assisted laser desorption ionization (MALDI) combined with time-of-flight mass spectrometry. While a MALDI-based approach may yet fulfill its promise, limitations encountered in ionizing relatively large DNA oligomers have proved to be difficult to overcome. In contrast, ionization of large DNA oligomers is not a limitation for electrospray ionization. However, electrospray-based approaches for high speed DNA sequencing have not been extensively pursued due to spectral congestion associated with the multiple charging phenomenon that is characteristic of electrospray. The formation of multiple charge states from a single oligomer severely limits the mixture complexity amenable to direct analysis via electrospray. For this reason, electrospray usually follows a separation method, such as liquid chromatography or capillary electrophoresis, when applied to mixtures. We have recently shown that gas-phase ion/ion chemistry involving oppositely charged ions within a quadrupole ion trap greatly expands the mixture analysis capability of electrospray. In this work, the idea is to subject Sanger mixtures to electrospray and ion/ion chemistry as a core element in a strategy for

high speed DNA sequencing. This talk describes the methodology and progress to date.

38. DNA and Protein Analyses on Microfabricated Devices

R. S. Foote, Y. Khandurina, I. M. Lazar, Y. Liu, T. McKnight, L. C. Waters, S. C. Jacobson, R. S. Ramsey, and J. M. Ramsey
Oak Ridge National Laboratory, P.O. Box 2008,
Oak Ridge, TN 37831-6142
footers@ornl.gov

Microfabricated, microfluidic devices are being developed for both nucleic acid and protein analyses. An integrated system for rapid PCR-based analysis on a microchip has been demonstrated. The system couples a compact thermal cycling assembly based on dual Peltier thermoelectric elements with a microchip gel electrophoresis platform. This configuration allows fast (~ 1 min/cycle) and efficient DNA amplification on-chip followed by electrophoretic sizing and detection on the same chip. On-chip DNA concentration has been incorporated into the system to further reduce analysis time by decreasing the number of thermal cycles required. The concentration-injection scheme enables detection of PCR products after performing as few as 10 thermal cycles with a total analysis time under 20 min, with a starting template copy number of fewer than 15 molecules per injection volume.

Electrophoretic separations of proteins have been carried out on microchips with on-chip, post-column labeling for detection by laser-induced fluorescence. The post-column labeling format avoids peak broadening and loss of resolution due to heterogeneous product formation in prelabeling reactions. Two-dimensional separations of tryptic peptides were demonstrated on microchips that combine micellar electrokinetic chromatography (MEKC) and high-speed capillary electrophoresis (CE). Effluent from the first dimension is sampled onto the second dimension every few seconds and the entire analysis is completed within 10 minutes. Structures that incorporate an electrospray element have also been devised and sub-attomole sensitivity demonstrated for peptide samples on a time-of-flight (TOF) mass analyzer. Proteolytic digestions with

trypsin can be performed directly on the chip and the peptide fragments analyzed by TOFMS for protein identification. Tryptic peptides could be generated in less than 10 min for analysis of femtomol or subfemtomol amounts of protein.

39. Stable Isotope Assisted Mass Spectrometry Allows Accurate Determination of Nucleotide Compositions of PCR Products

Xian Chen¹, Zhengdong Fei², Lloyd M. Smith², E. Morton Bradbury^{3,4}, and Vahid Majidi¹
¹CST-9, Chemical Science and Technology Division and ³B-3, MS M888, Biological Division, Los Alamos National Laboratory, Los Alamos, NM 87544; ²Department of Chemistry, University of Wisconsin-Madison, 1101 University Avenue, Madison, WI 53706-1396; and ⁴Department of Biological Chemistry, School of Medicine, University of California at Davis, Davis, CA 95616
xchen@telomere.lanl.gov

In parallel with the large-scale sequencing effort, the human genome project will need the next generation tools for accurate and efficient analyses of the enormous pool of DNA sequences. Such analyses are required for; (a) validation of DNA sequences; (b) comparison of a parent (known) sequence with a related (unknown) sequence, and (c) characterization of sequence polymorphisms in various genes especially those associated with genetically inherited human diseases. Here, we report a novel method that combines stable isotope ¹³C/¹⁵N-labeling of PCR products of the target sequences with analysis of the mass shifts by mass spectrometry (MS). The mass-shift due to the labeling of a single type of nucleotide (i.e., A, T, G, or C) will reveal the number of that type of nucleotide in a given DNA fragment. Using this technique, we have accurately determined nucleotide compositions of DNA fragments. The method has also been applied to score a known single nucleotide polymorphism. The comparisons of nucleotide compositions determined by our method among homologous sequences are useful in sequence validation, sequence comparison, and characterizations of sequence polymorphisms.

40. Hybridization Detection

Tom J. Whitaker and Kenneth F. Willey
Atom Sciences, Inc., Oak Ridge, TN 37830
whitaker@atom-sci.com

Two projects aimed at developing new techniques that measure DNA hybridization to oligonucleotide (ODN) probes on DNA chips will be described. Although these techniques have similar goals, they differ widely in cost, complexity, sensitivity, and application. One method uses laser-based mass spectrometry analysis of stable isotopes of Sn atoms sputtered from Sn-labeled DNA target molecules. A 10-kV ion beam is used to sputter the sample from the surface and a wavelength-tunable laser efficiently and selectively ionizes the neutral Sn atoms for time-of-flight mass spectral analysis. The extreme sensitivity of this laser ionization technique has previously enabled high spatial resolution measurements of trace impurities in a variety of substrates. The technique also facilitates quantitative measurements by avoiding the effects of variation in secondary ion yield, a problem that plagues SIMS (secondary ion mass spectrometry) analyses. We are currently working with samples supplied by Affymetrix, Inc. to explore application of the technique for quality control of in-situ formation of ODN probes and DNA hybridization.

The second technique uses an inexpensive electronic detection method to determine if hybridization has occurred at a specific probe site. The low cost and small dimensions of this device make it ideal for point-of-care applications. The method takes advantage of the fact that an ODN probe can form an insulating self-assembled monolayer between a gold surface (on which the dielectric is attached) and a conducting liquid. Because of the extremely thin dielectric, the resulting capacitor has a very high specific capacitance. Calculations indicate that the change in effective dielectric thickness caused by the binding of a fraction of a monolayer of DNA to the ODN probe layer will produce a significant, and easily measurable, change in capacitance. Results of initial experiments aimed at verifying this concept

will be provided and future experiments to enhance the capacitance change will be described.

41. Automation Using Packard Multiprobe Robots for Finishing

Christine Munk, Judy Buckingham, Marie Krawczyk, Elizabeth Saunders, David Bruce, and Mark Mundt
Bioscience Division and DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM 87545
cmunk@lanl.gov

We have recently adopted an automated approach for cherry-picking of subclone DNA to streamline our finishing process. The input for this process is a list of subclones for finishing reactions that is generated using the Phrap .ace file. A script sorts the resequencing reactions by single strand gap size and calculates source and destination for each subclone DNA. The script outputs two text files, one of which is formatted to be imported into an MP Table sample transfer program and contains source and destination locations for the DNA transfer. The second output file is used to make the sample sheet, which is imported into the data collection software on the ABI 377.

For cherry-picking subclones, we initially tried a Packard Multiprobe I equipped with disposable tips to transfer DNA from deepwell cluster tubes. We missed ~20% of the DNAs because the edge near the top of the tips got hung up on the top of the DNA tube, preventing the tip from reaching the DNA at the bottom of the tube. To address this problem, we moved to a fixed-tip robot, and the DNA prep procedure was changed to collect DNA into microtiter plates. The small diameter of the fixed tips allows them to reach the bottom of both the deepwell and microtiter plates. In our current process, we use a fixed-tip Multiprobe I robot with MP Table software to transfer subclone DNAs from round-bottom microtiter plates and to rearray these into cycleplates for sequence reaction setup. We have not experienced

problems with DNA cross-contamination using a simple water rinse of the tips between samples. The cost of disposable tips has been eliminated. Use of microtiter plates instead of deepwells has greatly reduced the amount of storage space needed.

42. Automated, Low Cost Isolation of Blood or Bacterial Genomic DNA

Brian Bauman, Tuyen Nguyen, Zuxu Yao, Tony Zucca, Dan Langhoff, and William MacConnell
MacConnell Research Corporation, San Diego, CA 92121
macres@macconnell.com

The isolation of the genomic DNA from blood, bacteria, and virus is a necessary starting point for molecular diagnosis of infection, genetic disease, inherited traits and identity determination, and other research applications. The ability to rapidly and reproducibly isolate DNA from blood and other bodily samples will be continually required to identify, characterize and treat factors involved in human disease and disorders. This process needs to be automated by means that are affordable to clinical or research labs. In Phase II we are further developing prototypes for a fully automatic high-throughput blood or bacteria genomic DNA isolation instrument and disposable processing cassettes. This instrument uses a derivative of electrophoretic separation technology that we developed for purification of plasmid DNA. Proof-of-concept data has been obtained during Phase I. The instrument gives high yields of pure DNA over a wide range of sample concentrations.

The separation technology makes use of programmed electrophoresis of the lysed sample which is placed in between boundaries of agarose separation material. The process can be accomplished using a microprocessor-controlled programmable power supply in conjunction with a multi-sample disposable processing cassette, electrophoresis rig and fluid handling components.

Thus far, our prototype instrument and cassettes allows automatic purification of human blood or bacterial DNA in as little as 24 minutes. The resulting DNA is pure enough for use in restriction

digests, and could be used as a template for PCR, PCR sequencing and RFLP analysis. In addition, we found that the process successfully purifies genomic DNA from a wide range of sample cell numbers, such as 10⁴ to 10⁹ bacterial cells or 0.2 to 500 microliters of human blood. Interestingly, the DNA yield from the above trials was greater than 75% of the theoretical amount of genomic DNA present in these samples.

In Phase II we are completing trials to: (1) optimize the electrophoretic purification technique, (2) determine the full range of sample volumes and cell numbers that can be processed, (3) test the activity of purified genomic DNA in a variety of molecular biology procedures, (4) test for cross-contamination between samples purified in the same run, (5) construct a 24 well and 96 well prototype instruments that automate the overall method, (6) write control software and (7) contract for the construction of an injection mold for the production of the processing cassettes. The method is being expanded to process DNA from a variety of sample types including viruses.

The products developed in this work will be commercialized by MacConnell Research in the form of instruments and supplies sold for genomic DNA isolation.

This research is being supported by DOE SBIR Phase II grant number DE-FG03-98ER82612.

43. The Use of Electrode Arrays for the Synthesis of Biomolecular Affinity Probes

Francis Rossi, Christopher Ashfield, Karl Maurer, and Donald Montgomery
CombiMatrix Corporation, 887 Mitten Road, Suite 200, Burlingame, CA 94010
frossi@combimatrix.com

We have developed an active semiconductor chip composed of over 1000 individually addressable electrodes that is used to synthesize microarrays of biomolecular affinity probes. Arrays are prepared by coating the semiconductor chip with a porous polymer support in which synthesis occurs. The

underlying electrodes are used to electrochemically generate reagents from inert precursors. By switching on individual electrodes, or patterns of electrodes, reactions can be conducted at defined locations of the chip.

The technology has been successfully used to prepare DNA oligonucleotide probe arrays from commercially available reagents. To synthesize an array, electrodes are biased as anodes at defined locations. This generates acid, which removes the DMT protecting group from the 5'-hydroxyl group of nascent oligonucleotides. Extension of the deprotected hydroxyl group using standard DNA phosphoramidite reagents adds the next base of the oligonucleotide.

We are now extending this technology to the synthesis of peptide probe arrays for the identification and analysis of gene products. Peptides were prepared by first immobilizing an Fmoc-protected amino linker to the porous polymer support. The amino groups were deprotected at the desired locations with an electrochemically generated base. The resulting free amines were reacted with an activated Fmoc-protected amino ester using conventional techniques to give peptides.

44. Development of a High Throughput Peptide Nucleic Acid Synthesizer

J. Shawn Roach¹, Simon Rayner¹, Lynn Mayfield², David R. Corey², and Harold "Skip" Garner¹

¹Center for Biomedical Inventions, Department of Internal Medicine and ²Howard Hughes Medical Institute, Department of Pharmacology and Department of Biochemistry, University of Texas Southwestern Medical Center at Dallas, Dallas, TX
roach@ryburn.swmed.edu

Peptide Nucleic Acids (PNAs) are synthetic analogs of DNA in which the phosphodiester backbone has been replaced with 2-aminoethyl glycine linkages, but maintaining the four natural nucleobases. A PNA strand will bind to DNA with the same sequence

complementarity as standard DNA/DNA base pairing, but PNA/DNA binding occurs more rapidly and more tightly than DNA/DNA binding. Much research has gone into the potential applications of PNAs as antisense and diagnostic agents.¹ However, a major obstacle in PNA research becoming more widespread has been the high cost of the PNAs. A high throughput PNA synthesizer will afford an economy of scale and reduce the synthetic cost of PNAs. We report the development of a high throughput PNA synthesizer capable of producing up to 192 different PNAs in one synthesis run. The synthesizer is based on high throughput DNA synthesis technology developed at the University of Texas Southwestern Medical Center at Dallas to support the Human Genome Project.² The synthesizer consists of an XY table, a series of valves plumbed to an injection head for reagent delivery, two vacuum chucks for reagent removal and a computer that controls the synthesis procedure. Synthesis is conducted in 96-well fritted filter plates, using standard solid phase Fmoc-PNA synthesis chemistry. The quality of the PNAs produced from the synthesizer is assessed using RP-HPLC and MALDI MS.

1. Nielsen, P.E. *Curr. Op. Biotech.* 1999, 10:71-75.
2. Rayner, S., Brignac, S., Bumeister, R., Belosludtsev, Y., Ward, T., Grant, O., O'Brien, K., Evans, G. and Garner, H. *Genome Res.* 1998, 8, 741-747.

45. MicroArray of Gel Immobilized Compounds on Chip

V. Vasiliskov, A. Stomakhin, B. Strizhkov, S. Tillib, V. Mikhailovich, A. Sobolev, A. Kuhktin, and A. Mirzabekov
Joint Human Genome Program: Biochip Technology Center, Argonne National Laboratory, Argonne, IL 60439 and Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 117984

Micro Array of Gel Immobilized Compounds on a Chip (MAGICChip™) are produced by immobilizing

oligonucleotides¹, DNA, enzymes, antibodies, and other proteins² on a photopolymerized micromatrix of polyacrylamide gel pads. Recently, a photocopolymerization technique was introduced for more rapid and inexpensive manufacturing of the microchips containing gel pads from 10'10'5 mm to 100'100'20 mm and larger in size³.

MAGIChips are efficient for carrying out direct hybridization tests, as well as for oligonucleotide ligation, single-base extension⁴, and PCR amplification of DNA. The fluorescence microscope has been devised for quantitative and real-time monitoring of hybridization, measuring the thermodynamic parameters of DNA duplexes, and measuring kinetics of enzymatic reaction on MAGIChips. On-chip MALDI-TOF mass spectrometry was successfully tested for polymorphism analysis of DNA⁵. These technologies are demonstrated for identification of microorganisms, detection of their genes, and screening for mutations. Simple equipment and procedures have been developed for monitoring the hybridization of amplified DNA with oligonucleotide microchips and on-chip amplification. This allows us to carry out fast and inexpensive screening of *Mycobacterium tuberculosis* drug resistant mutations.

Work supported by the Department of Energy, Office of Health and Environmental Research under Contract No. W-31-109-ENG-38; Cooperative Research and Development Agreement No. 970192 between Argonne National Laboratory, Motorola, and Packard Instruments; Defense Advanced Research Project Agency under Interagency Agreement No. AO-E428; and the Russian Foundation of Fundamental Research under Grant 96-04-49858.

References

1. Yershov, G., Barsky, V., Belgovskiy, A., Kirillov, Eu., Kreindlin, E., Ivanov, I., Parinov, S., Guschin, D., Drobishev, Dubiley, S. & Mirzabekov, A. (1996) *Proc. Natl. Acad. Sci. USA* 93, 4913-4918.
2. Arenkov, P., Kuhktin, A., Gemmell, A., Chupeeva, V., & Mirzabekov, A. (2000) *Anal. Biochem.* (in press).
3. A. Vasiliskov, V., Timofeev E., Surzhikov S., Drobyshev, A., Shick, V., & Mirzabekov, A. *BioTechnique* 27, 592-606
4. Stomakhin, A, Vasiliskov, V., Timofeev, E., Schulga, D., Cotter, R., & Mirzabekov, A. (2000) *Nucleic Acids Res.* (in press).

Mapping

46. Analysis of WUSTL's Human BAC Fingerprint Database

R. Sutherland, M. Mundt, and N. Doggett
Bioscience Division and DOE Joint Genome Institute,
Los Alamos National Laboratory, Los Alamos, NM
87545
rds@lanl.gov

We have used the LANL human chromosome 16 BAC contig data to evaluate the Washington University's Genome Sequencing Center Human BAC Fingerprint Database.

WU has fingerprinted 162,272 RPCI-11 BACs and assembled them into 12,549 contigs.

LANL has identified 4085 BACs using 1106 overgos and STSs from 16 q-arm. The 16 q-arm is 45 Mb and covers 1.35% of the human genome.

For this exercise, only BACs from sections 1 and 2 of the RPCI-11 library were considered. For these sections, there are 125,979 WU's BACs within contigs and 3530 LANL mapped BACs.

The first results are a straight set-to-set comparison of the two data sets to see which WU contigs can be linked to 16 q-arm map. BACs occurring in the LANL set were used to query WU contigs. The results were as follows: 10,882 BACs from 657 contigs were identified from the WU data, 2034 BACs were in common with the LANL data. 57% of the LANL BACs could be found in a WU contig but only 18.7% percent of WU BACs in these contigs were found in the LANL set.

For the second analysis we discounted all WU contigs that contained only a single LANL mapped

BAC. The results were as follows: 3,618 BACs from 245 contigs were identified from the WU data, 1622 BACs were in common with the LANL data. 46% of the LANL BACs could be found in a WU contig while 44.8% percent of WU BACs in these contigs were found in the LANL set.

For the third analysis we limited the WU set further. Only BACs that are contained in contigs that range from 2-40 members were considered; this is a 1 sigma distribution. The results were as follows: 2,766 BACs from 236 contigs were identified from the WU data, 1491 BACs were in common with the LANL data. 42% of the LANL BACs could be found in a WU contig while 53.9% percent of WU BACs in these contigs were found in the LANL set.

We believe that the LANL BAC map provides >90% coverage of the 16 q-arm and that we identified the great majority of 16 q-arm BACs from sections 1 and 2 of the RPCI-11 library. Thus, the percentages above suggest to us that there is a significant level of false overlaps in the WU BAC contigs.

47. Human Chromosome 16 Mapping Update

Cliff S. Han, Robert D. Sutherland, Phillip B. Jewett, Mary L. Campbell, Linda J. Meincke, Judy G. Tesmer, Mark O. Mundt, Larry L. Deaven, and Norman A. Doggett
Bioscience Division and DOE Joint Genome Institute,
Los Alamos National Laboratory, Los Alamos, NM
87545
chan@telomere.lanl.gov

We have used sequence-based markers from an integrated YAC STS-content/somatic cell hybrid

breakpoint physical map and radiation hybrid maps of human chromosome 16 to construct a new sequence-ready BAC map of this chromosome. The integrated physical map was previously generated in our laboratory and contains 1150 STSs, providing a marker on average every 78 Kb on the euchromatic arms of chromosome 16. The other two maps utilized for this effort were the radiation hybrid maps of chromosome 16 from Whitehead Institute and Stanford University. To create large sequenceable targets of this chromosome we used a systematic approach to screen high density BAC filters with probes generated from overlapping oligonucleotides (overgos). We first identified all available sequences in the three maps. These include sequences from genes, ESTs, STSs, and cosmid end sequences. We then used BLAST to identify 36 bp unique fragments of DNA for overgo probes. A total of 906 overgos were selected from the long arm of chromosome 16. After a total of 212 hybridizations we have constructed an initial probe-content BAC map of chromosome 16q consisting of 828 overgo markers and 3363 BACs providing greater than 85% coverage of the long arm of this chromosome. Gaps in the map are being closed with the following methods: 1) PCR screening the RPCI-11 library with the BAC end sequence-derived STSs. 2) BAC end sequence database searches with draft sequences of BAC clones near the gaps. 3) Screening the RPCI-11 library with overgos generated from the BAC end sequences near the gaps. To date, 400 PCR screening and 5 pooled overgo hybridization have been completed for the gap closing effort, extending the coverage of the BAC map to over 90%.

Supported by the US DOE, OBER under contract W-7405-ENG-36.

48. Annotation and Analysis of the Draft Sequence of 16Q12

Jung-Rung Wu, Mark O. Mundt, Cliff S. Han, Kristina Kommander, Robert D. Sutherland, Lela Tatum, Norman A. Doggett, and Larry L. Deaven
Bioscience Division and DOE Joint Genome Institute,
Los Alamos National Laboratory, Los Alamos, NM 87545
wu@telomere.lanl.gov

We have completed a map and chosen a tiling set of more than 40 spanning BACs, most often from the RPCI-11 library, for a region that covers more than 5 Mb of human 16q12, encompassing a locus for inflammatory bowel disease (IBD1). These BACs have been selected by a combination of overgo hybridization, restriction map assembly by both fingerprinting and sequence prediction, and BAC end sequence searches. Sequencing has been completed to levels ranging from light shotgun to Bermuda finished bases. We will present annotation and statistical results of our analysis of the draft sequence we have achieved thus far and show how these results compare with ~ 5 Mb of sequence from 16p13.3.

49. Progress in Mapping the Mouse Genome

Cliff S. Han, Linda J. Meincke, Larry L. Deaven, and Norman A. Doggett
Bioscience Division and DOE Joint Genome Institute,
Los Alamos National Laboratory, Los Alamos, NM 87545
chan@telomere.lanl.gov

The mouse genome is the second major target for sequencing by the JGI. The initial focus is on regions of biological interest and regions of synteny to human chromosome 5, 16, and 19. The mapping group at LANL is now focusing on mouse genome targets syntenic to human chromosome 16 for BAC map construction. The probes for this effort are derived from STSs and cDNA sequences.

1) STS mapping: We utilize the sequences from STSs located on mouse chromosomes syntenic to human chromosome 16. These STSs come from various map sources. To date, 960 overgos generated from STSs have been screened against a 5X portion of the RPCI-23 mouse library and 2403 BACs identified. Overgos from 833 STSs were located to at least one BAC.

2) cDNA mapping: We use two approaches to find cDNAs in the region syntenic to human chromosome 16: 1) BLAST against mouse unigene database with unigene sequences from human chromosome 16 that are masked with repeatmasker. 2) BLAST against mouse unigene database with genomic sequences of human chromosome 16 that are masked with

repeatmasker. A total 600 cDNA sequences were found after the two BLAST searches. Overgos from 96 of the cDNAs have been screened against the RPCI-23 library. 570 BACs were hit by 94 overgos. Average hit per overgo probe is 12. The first eighty six mouse BAC clones have been sent to the sequence queue.

Supported by the US DOE, OBER under contract W-7405-ENG-36.

50. Rapid Construction of Mouse Sequence-Ready Maps Using a Homology-Driven Approach

Lisa Stubbs, Joomyeong Kim, Laurie Gordon, Hummy Badri, Mari Christensen, Matt Groza, Chi Ha, Sha Hammond, Michelle Vargas, and Eddy Wehri

DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598 and Genome Division, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore CA 94550
 stubbs5@llnl.gov

We have developed a rapid and efficient "homology-driven" strategy for assembling mouse BAC clone contigs for comparative sequencing, and are using this approach to generate large contigs spanning all mouse regions related to gene-containing segments of human chromosome 19. The strategy uses overgo probes designed from matches detected between human genomic DNA sequence and mouse ESTs or other cDNA fragments in pooled hybridization against gridded mouse BACs. The overgoes are chosen with 50-100 kb spacing and hybridized in pools corresponding to position of the homologous sequence in the human chromosome. The human map is used as a model for assembly of corresponding mouse contigs, and contig assembly, clone integrity, and overlap are verified by restriction fingerprinting, completed at a depth that permits the creation of a detailed restriction map of the mouse contig. This strategy has been used to assemble maps of more than 15 Mb of mouse DNA as of the date of this

submission (12/99), and we expect to complete maps of all chromosome 19-related regions within 2-3 months. The maps we have generated provide an important source of clones for directed comparative sequencing, and reagents for basic studies of genome evolution and for analysis of mouse mutations.

51. Structural and Functional Analysis of a Conserved Imprinted Region of Human Chromosome 19q13.4 and Mouse Chromosome 7

Joomyeong Kim^{1,2}, Vladimir Noskov³, Xiaochen Lu^{1,2}, Anne Bergmann^{1,2}, Tiffany Warth², Paul Richardson¹, Vladimir Larionov³, Natasha Kouprina³, and Lisa Stubbs^{1,2}

¹DOE Joint Genome Institute, 2800 Mitchell Avenue, Walnut Creek, CA; ²Genome Division, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA; and ³National Institute of Environmental Health Sciences, Laboratory of Molecular Genetics, Research Triangle Park, NC
 stubbs5@llnl.gov

Mouse genetics studies have long ago predicted that a genomically imprinted domain would be found near the centromere of mouse chromosome 7, a region with known syntenic homology to human chromosome 19q. Animals that inherit only maternal alleles of this region die as neonates, suggesting the presence of a gene or genes, expressed exclusively from the paternal chromosome, which is required for normal development. In earlier studies, we mapped a known paternally-expressed gene, PEG3, to human 19q13.4, and we reasoned that other imprinted genes would be found nearby in a clustered imprinted domain. We have used the emerging human sequence and data derived from related mouse clone contigs to identify new genes near PEG3, and have demonstrated that these novel genes are also imprinted in mice. These new genes are expressed highly in embryos; one represents a strong candidate for the paternally expressed, neonatal lethal factor predicted by mouse genetics. Our studies demonstrate that, despite many basic similarities, human and

mouse regions surrounding PEG3 have undergone significant changes in gene content and organization. We will discuss the structure, expression and evolution of genes in this imprinted region, discuss the potential functions of each gene, and speculate on the possible implications of these findings for the genetics of chromosome 19-linked disorders in humans.

52. Mapping and Functional Analysis of the Mouse Genome

D. K. Johnson,¹ C. T. Cuiat,¹ M. L. Klebig,² Y. You,¹ D. R. Miller,¹ L. B. Russell,¹ E. J. Michaud,¹ and E. M. Rinchik^{1,2}

¹Mammalian Genetics and Development Section, Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8077 and

²Department of Biochemistry, Molecular, and Cellular Biology, University of Tennessee, Knoxville, TN 37996

johnsondk@ornl.gov

As part of a functional-genomics strategy to determine how altered genes and proteins impact complex biological systems in mammals, the Mammalian Genetics Program at ORNL is characterizing a number of regions of the mouse genome on both the physical and functional levels, using mouse mutations as tools. This project forms a logical partnership with our regional mutagenesis program, which is designed to detect, maintain, and partially characterize new chemically induced mutations in ~8-10% of the mouse genome by utilizing new genetic tools and broad-based phenotype screening. The integrated efforts of these projects will advance the post-genome sequencing mission of annotating human DNA sequence with whole-organism functional information from the mouse model system.

Our goal is to acquire the DNA sequence of each region, to develop a validated transcription/expression map, and to ascribe whole-organism functional information to each coding sequence through analysis of heritable gene mutations. Our chosen genome regions and subregions will be physically delimited by identifiable DNA landmarks (typically chromosomal rearrangements); hence, we

can easily co-map mutant phenotypes with coding units to establish unambiguous sequence/function relationships by superimposing mutation maps onto transcription maps. The target regions include the 5- to 6-cM pink-eyed dilution (p) region in mouse Chromosome (Chr) 7 (human Chrs 11p, 15p, and 15q homologies); the 14 cM between p and the albino (Tyr; c) region (human Chr 15q); the 6- to 11-cM Tyr region (human Chrs 6p, 11p, 11q, and 15q); all of Chr 15 (human Chrs 5p, 8q, 12q, 22q), concentrating initially on the distal half; and mid-Chr 10 (human Chrs 6q, 10q, 12q, 21q, and 22q). With available molecular and embryonic stem (ES)-cell techniques, the growing emphasis on regional-mutagenesis strategies and the development of mouse reagents with which to carry out those strategies, we and others can extend this same discovery approach to any genome region.

Complete DNA sequence for these regions will be obtained by collaboration with the Joint Genome Institute or by mining of public databases created by the NIH mouse sequencing efforts. After ascertainment of potential transcription units from EST mapping and from computational analysis of raw DNA sequence by ORNL's Computational Biosciences Section, predicted transcription units will be verified by RNA analyses (Northern, RT-PCR, RNase protection, and/or microarray procedures). The ultimate correlation of dense mutation maps with the transcription/expression maps has begun by identifying candidate mutant genes bearing ENU mutations, using densely mutagenized regions within the p- and Tyr regions as initial models with which to develop efficient mutation-scanning techniques. Phenotype gaps can also be filled with knockout/gene-trap mutations for genes discovered in DNA sequence analysis but not represented as ENU mutations. All new DNA sequence information, expression information, and mutations will be advertised to interested partners via the WWW.

[Research sponsored by the Office of Biological and Environmental Research, USDOE, under contract DE-AC05-96OR22464 with Lockheed Martin Energy Research, Inc.]

53. Toward Completion of a Human Chromosome 5 BAC Map and a Mouse Syntenic BAC Map

Steve Lowry, Ze Peng, Duncan Scott, Yiwen Zhu, Mei Wang, Roya Hosseini, Michele Bakis, Joel Martin, Ingrid Plajzer-Frick, Jeff Shreve, Le-Thu Nguyen, and Jan-Fang Cheng
Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
jcheng@mhgc.lbl.gov

Physical mapping of BACs on human chromosome 5 is in a final stage. The current map consists of 7,618 BAC clones anchored to the chromosome by 2,954 STSs. The distribution of STSs is not even across the chromosome. Approximately 50% of the STSs were derived from 1/3 of the chromosome at the end of the q arm where the average size of contigs is greater than 1 Mb. Most BACs were isolated as single colonies. Restriction maps and FISH maps were constructed for all contigs and are available on the web (<http://www-hgc.lbl.gov/human-maps.html>). The maps are updated regularly.

To date, 1,702 BACs have been selected for sequencing. These BACs contain a total restriction fragment length of 163.5 Mb or approximately 90% of the euchromatin portion of the 190 Mb chromosome. In an independent experiment, we tested the degree of coverage provided by our map by probing the mapped clones with 796 new STSs (580 ESTs, 216 randomly derived). We found 88% of the STSs were contained by the mapped clones. Both restriction map length and STS analysis indicate that the selected BAC tiling path covers approximately 88-90% of the chromosome.

Sequence already generated for our mapped clones is enabling us to expand contigs by detecting overlaps between BACs that were undetected by restriction fragment analysis and STS content mapping. BAC end sequences in the TIGR database also enable extension of contigs.

The clone map and sequence information for human chromosome 5 are being used to isolate syntenic mouse BACs. Several large contigs have been built using mouse ESTs that were identified by high sequence similarity to human chromosome 5 sequences. To streamline the identification of syntenic mouse ESTs, we have generated a web interface to facilitate (1) blasting of GenBank databases with batches of masked sequences, (2) parsing of output based on length and level of similarity, and (3) reduction of background matches by identifying successive exons in the genomic sequence.

54. Sequence-Ready Characterization of the Pericentromeric Region of 19p12: A Strategy for the Analysis of Complex Regions of the Human Genome

Evan E. Eichler¹, Anthony P. Popkie¹, Laurie A. Gordon², and Anne S. Olsen²

¹Case Western Reserve University, Cleveland, OH, 44106 and ²DOE Joint Genome Institute, Walnut Creek, CA 94598
eee@po.cwru.edu

The pericentromeric region of 19p12 represents one of the most poorly mapped and sequenced regions of chromosome 19. This is due, in large part, to the virtual absence of unique sequence identifiers within this region. The proximal portion of 19p12 possesses sequence attributes consistent with both euchromatic and heterochromatic DNA including a large cluster of ZNF (zinc-finger) genes, an overabundance of human endogenous retroviral elements and an atypical higher-order (~10-30 kb) beta-satellite repeat structure. Analysis of ~425 kb of seed sequence from 19p12 has revealed that less than 15% of the region consists of bonafide unique sequence. This unusual organization has hampered the development of sets of large contiguous clones in this region, resulted in relatively poor clonal coverage (<60%) and has greatly limited the selection of suitable templates for

sequencing. To complete mapping and sequencing in this region, we have designed a strategy that takes advantage of the known biological properties of 19p12 repetitive sequences. Our approach has been to distinguish between "generic" and 19p12-specific repeat elements; develop assays to rapidly identify 19p12 clones from different genomic libraries and to confirm the position of these clones at the level of sequence-overlap. Both high-resolution FISH techniques (extended chromatin analysis) and restriction fragment overlap are being implemented as complementary tools to confirm the integrity of the map and to identify potential sites of heteromorphism in the region. To date, a total of 403 19p12 BAC clones have been identified from RPCI-11 and CIT-D libraries. A subset of these have been used to extend clonal coverage into 12 different gap regions of 19p12; four of which are now tentatively closed. The data generated will be used in the selection of the most parsimonious tiling path of BAC clones to be sequenced as part of the JGI effort on chromosome 19 and should serve as a model for the sequence characterization of other difficult regions of the human genome. The complex organization of this region will be discussed in the context of its unusual biology.

55. IMAGEne 3.0: Clustering All Sequences Obtained from I.M.A.G.E. Clones

Peg Folta, Tom Kuczmarski, Tim Harsch, and Christa Prange
Lawrence Livermore National Laboratory,
Livermore, CA 94550
pfolta@llnl.gov

To date over 1.9 million sequences have been submitted to GenBank from the 2.9 million available I.M.A.G.E.¹ clones. This number will increase sharply due to the new Mammalian Gene Collection² project. To maximize the value of this information, the IMAGEne³ product has been extended to group the human sequences into clusters that represent both known genes and "candidate genes". For known genes, clustering eliminates redundancy by providing the best representative clone for a gene. For clusters not associated with a known gene, the results provide evidence of a possible gene discovery.

IMAGEne was first released to the public in 4/98 to provide the user community with known gene clusters of I.M.A.G.E. clones. Since then the product has undergone significant enhancements, including use of NCBI's RefSeq to base the known gene set, indication of sequence verified clones, repeat masking, enhanced error checking, and faster response times. Version 3.0 is the largest enhancement, which extends the functionality by forming clusters on clones not associated with known genes.

Clusters are formed by sequence similarity, clone membership, and internal I.M.A.G.E. project knowledge. The user can query the resulting cluster database and view the cluster members, ranked primarily by size, in a user-friendly Java-based display. Currently I.M.A.G.E. has clone representatives for 93% of the known genes. It defines 61,083 multi-member candidate gene clusters and over 236,000 singletons. By the conference date, IMAGEne 3.0 will be publicly available at <http://bbrp.llnl.gov/imagene/bin/search>.

1. Lennon, G., et al (1996) The I.M.A.G.E. Consortium: An Integrated Molecular Analysis of Genomes and Their Expression. *Genomics*, 33,151-152.
2. Strausberg, R.L., et. al. The Mammalian Gene Collection, *Science* 1999 Oct 15;286(5439): 455-7
3. Cariaso, M., et. al. IMAGEne I: Clustering and Ranking of I.M.A.G.E. cDNA Clones Corresponding to Known Genes, *Bioinformatics*, in-press

This work was performed by LLNL under the auspices of U.S. DOE, Contract No. W-7405-Eng-48.

Bioinformatics

56. Software to Support BAC Mapping

Cliff S. Han and Norman A. Doggett
Bioscience Division and DOE Joint Genome
Institute, Los Alamos National Laboratory, Los
Alamos, NM 87545
chan@telomere.lanl.gov

The LANL production mapping team has a major responsibility for supplying mapped BAC clones to the JGI's Production Sequencing Facility. Our approach involves the use of overgo probes to screen BAC filters and construct probe-content BAC maps. To facilitate this work we have developed and implemented several software programs. They are 1) Overgo selection program, 2) Automated contig assembly, and 3) Contig graphic draw.

Overgo selection program: The program is written to select overgos from a sequence. The program combines repeat screening and secondary structure screening with a nr Genbank search. Overgo sequences are checked for hairpins, selfdimers, heterodimers, and against RepBase and nr GenBank.

Automated contig assembly: The program assembling contigs automatically according to the input data. It can accept previously ordered backbone STS information as a constraint.

Contig graphic draw: The program is designed to draw a contig graphic map from probe(STS)-content map data that are presented in map order. The output is a Postscript III file. The file can be printed by dragging onto the printer window or translated into a portable document format with Acrobat Distiller and viewed with Acrobat Reader. The input for the program is generated with the Automated contig

assembly program or by a macro which can translate Excel data into the format for this program.

Supported by the US DOE, OBER under contract W-7405-ENG-36.

57. Automated Optimization of Expert System for Base-Calling in DNA Sequencing

Arthur W. Miller and Barry L. Karger
Barnett Institute, Northeastern University, 360
Huntington Ave., Boston, MA 02115
miller@ccs.neu.edu

A recurring issue in automated DNA sequencing is that base-calling lags behind improvements to instrumentation and sequencing chemistry. This is because base-callers require retraining, or because the preprocessing of the data prior to base-calling must be changed. We have previously presented an expert system for long-read base-calling, capable of read lengths up to 1300 bases in sequencing by capillary electrophoresis on optimized separation matrices (A. W. Miller and B. L. Karger, *DOE Human Genome Program Contractor-Grantee Workshop VII*, 1999). The expert system supplies probabilistic confidences on base-calls, with statistics computed for several different types of miscall. Here we present tools for the automated retraining and optimization of this base-caller, including preprocessing and confidences, by nonprogrammers. Training takes into account template effects, low signal, and other factors observed in production sequencing. Results are shown for large amounts of data from both ABI 3700 and MegaBACE 1000 sequencers. In addition to software, other recent

developments in long-read sequencing by capillary electrophoresis will also be presented.

This work is being supported by DOE grant DE-FG02-90ER 60985.

58. Is Q20 a Sufficient Measure of Quality to Use for DNA Sequencing Process Analysis?

D.C. Bruce, M.D. Jones, J.E. Bryant¹, R. Lobb¹, J.R. Griffith¹, M.O. Mundt, N.A. Doggett, and L.L. Deaven

Bioscience Division and DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM 87545 and ¹University of New Mexico, Department of Biochemistry and Molecular Biology, Albuquerque, NM 87131
dbuce@lanl.gov

We examined whether Q20 is a sufficient measure of sequence quality to predict the impact of a process change on the quality of data generated in the shotgun phase. Q20 is a threshold metric derived from DNA sequence trace data by the base-calling program Phred, and is commonly used to report the read length of DNA sequence data. This practice follows from the rubric that 1) the number of Q20 base pairs is a necessary and sufficient quality metric and 2) long Q20 reads improve sequence assembly and hence simplifies finishing. In addition, Q20 is used to perform cost/benefit analysis (\$/Q20 base) bearing on sequencing process changes. As a test bed to model a process change, we analyzed data produced from sequencing runs produced with three different sequencing gel formulations. In addition to Q20 statistics, we determined the number of correctly aligned bases and the probability of error at each base for the assembled data. We present data that shows Q20 does not fully predict the benefit or harm associated with process alterations.

59. Annotation of Draft Genomic Sequence Generated at the JGI

Richard Mural¹, Miriam Land¹, Frank Larimer¹, Morey Parang¹, Manesh Shah¹, Doug Hyatt¹, Ed Uberbacher¹, P. Folta², T. Bobo², Zhengping Huang², and T. Slezak²

¹Computational Biosciences and Toxicology and Risk Analysis Sections, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Oak Ridge, TN 37831 and ²Human Genome Center, Lawrence Livermore National Laboratory, Livermore, CA

The JGI is a major player in the effort to complete a 90% draft of the sequence of the human genome within the next few months. Draft sequence poses special problems for the annotation process, however it is clear that a 3 to 5X coverage of genomic DNA can yield large amounts of biologically meaningful data if the appropriate analysis methods can be applied. There are a number of features that can be located and annotated in draft-sequence which are useful for further analysis, these include: STS's, BAC ends (STCs), and EST's. These features can be annotated by standard similarity methods given sufficient computational resources. Using various gene identification programs, particularly those that incorporate similarity data such as Grail-Exp, which can use both EST and complete cDNA data, provide another level to the analysis of draft data. These analyses allow not only gene identification but it can also provide some ordering information for contigs that make up the clone being analyzed. Also recall that essentially all of the genes that can be found in finished sequence can be identified in draft sequence at about 3X coverage.

To help add biologically valuable information to the draft sequence being generated at the JGI/PSF a configurable analysis pipeline has been developed to provide analysis of draft data. Draft data produced at the JGI/PSF is analyzed and the analysis results are parsed into the JGI database. The initial annotation of draft sequence is a catalog of the clone contents (STS's, STC's, genes models predicted by Grail-Exp and Genscan, as well as Blast searches of their translations against the NR protein database) which are provided in a tabular form which is accessible from the JGI web page. Further analysis of this information will help to define relationships among

draft clones and will allow ordering, within and between clones.

To date we have analyzed over 1500 draft clones from human chromosomes 5, 16 and 19. The results of these analyses can be viewed at:
www.jgi.doe.gov/Data/JGI_finished.html.

(Research sponsored by the Office of Biological and Environmental Research, USDOE under contract number DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.)

60. Information Systems to Support Experimental and Computational Research into Complex Biological Systems and Functional Genomics: Several Pilot Projects

Jay Snoddy¹, Denise Schmoyer⁴, Kathe Fischer⁵, Gwo-Lin Chen⁵, Miriam Land³, Sergey Petrov¹, Sheryl Martin⁴, Ed Michaud³, Bob Barry⁷, Gene Rinchik³, Peter Hoyt³, Mitch Doktycz², and E. Uberbacher¹

¹Computational Biosciences Section, Life Sciences Division; ²Biochemistry and Biophysics Section, Life Sciences Division; ³Mammalian Genetics and Development Section, Life Sciences Division; ⁴Toxicology and Risk Analysis, Life Sciences Division; ⁵Computational Physics Division; ⁶Computer Science and Mathematics Division; and ⁷Robotics and Process Systems Division; Oak Ridge National Laboratory, P. O. Box 2009, Oak Ridge, TN 37831; ⁵Department of Biochemistry, Cellular, and Molecular Biology, University of Tennessee, Knoxville, TN 37996

In order to promote a research capability that can help understand complex biological systems, the ability to acquire, manage, and interpret the complex information of biology is a prerequisite. To study biological systems, computerized systems must function to automate the routine operations that are needed in large-scale, data-driven research projects. Secondly, information systems must permit the

biologist to analyze data from both data-driven and hypothesis-driven research; this analytical support needs to connect genome-scale data or other data-driven approaches with more focused, smaller-scale hypothesis-driven research into more complex biosystems. These analytical connections need to be made, in part, by generating inferences and supporting the decision-making of biologists.

Chip-based technologies for mRNA expression analysis are a large-scale, data-driven approach that can supply experimental information that is useful in exploring tissue-specific systems and pathways. In addition, advances in genomics, mutagenesis/phenotype screening, and other areas facilitate a higher-throughput mouse biology research for insights into complex traits and systems. For example, a pilot project was recently initiated to begin developing information systems that can help ORNL and Tennessee Mouse Genome Consortium (TMGC) and other collaborators to acquire insights into complex biological systems. This will result in a Complex Biosystems Information Warehouse (CBIW) that will be developed in ORACLE 8i and is closely associated with the Genome Information Warehouse (see related abstract of Petrov et al.).

Users will enter data and acquire data from specific modules that are application-specific. Four related bioinformation modules are currently planned that will be supported by this data warehouse. These are:

- Mouse Tracking and Phenotype System (MuTrack)
- Genosensor Information Management System (GIMS)
- Gene and Protein Catalog
- Comparative Genomics Inferencing System (CompariSys)

These systems are all interrelated, will need to share some data, and will need to work together.

The first proposed information module, MuTrack, must acquire information about specific mice, tissues, and especially molecular samples and track them as they are processed through mouse phenotype screens and other experiments. This system, once out of pilot

stage, must track the distribution of mice, track mouse tissue samples, and catalog observations about the phenotypes of these mice. Part of this problem that needs to be solved for the TMGC is moving mutagenized mice and samples around from ORNL to UT Memphis, Vanderbilt University, and other sites and returning phenotype screen data to a central, shared information system. This system also needs to connect to the GIMS chip expression system, especially in sharing information about mouse RNA samples sent by the mouse biologists to the chip lab for expression analysis. The chip lab must also return some data to be integrated with other observations about specific mice or mice strains.

An electronic notebook is being developed to provide some of this needed functionality and will be demonstrated at the meeting. The general electronic notebook approach should be able to allow a reasonable compromise among the power of an information system (e.g. ability to query the data) and the required flexibility in different kinds of lab data that can be stored.

GIMS, the second information module, will acquire, automate, and interpret data produced by the Genosensor chips and other similar microarrays (see related abstract of Doktycz et al.) This information system will address one of the major current bottlenecks to this technology—the data handling and, especially, the computational data interpretation to find patterns in this expression data. We are using commercially available software modules for some of this component—at least initially—but other operational logic and analytical reasoning will need to be developed to glue together these different components and provide for both operational and analytical support.

Gene and Protein Catalog is a user interface to new data about the structure and system functions of genes and proteins. This user interface is being designed and developed so that it can take data from both the Genome Information Warehouse and the Complex Biosystems Information Warehouse. It should provide access to relevant new data discovered by our experimental collaborators, any expert-curated information, predicted gene and protein models from genome annotation, and any cross-links to the underlying archival data from community databases. A pilot project, for example is

testing the addition of single nucleotide polymorphisms (SNPs) that are in or next to GenScan and Grail-EXP-predicted genes.

The last information module, CompariSys, is proposed to follow on after the other systems are further developed. This system should help create classify and cross-link homologous genes and proteins. This will assist the user to extrapolate from genes and systems in the mouse, for example, to genes and systems in the human. It will use existing methods of sequence similarity, conservation of synteny, protein classification, and possibly other developing methods like large-scale phylogenetic gene tree generation, to help navigate and create links among the gene and system data found in MuTrack, Gene and Protein Catalog, and GIMS. This should allow a user or another computer system to automatically move from data about one gene to data about homologous genes, proteins, and systems. This will provide a comparative approach that is critical to understand and navigate the biological data about genes, proteins, and the pathways or systems that involve those genes and proteins.

(Research sponsored by the Office of Biological and Environmental Research, USDOE under contract number DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.)

61. Navigation, Visualization, and Query of Genomes: The Genome Channel and Beyond

Morey Parang, Miriam Land, Denise Schmoyer, Jay Snoddy, Doug Hyatt, Richard Mural, and Ed Uberbacher

Computational Biosciences and Toxicology and Risk Analysis Sections, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831
<http://compbio.ornl.gov/>

The Genome Channel Browser is a Java based viewer capable of representing a wide variety of genomic-sequence annotation and links to a large number of related information and data resources. It relies on a number of underlying data resources, analysis tools, and data-retrieval agents to provide an up-to-date view of genomic sequences as well as computational

and experimental annotation. The current version of the Genome Channel Browser (v2.0) provides a diverse set of functional features. New features in this version of the Genome Channel include: additional features such as tRNA and BAC ends, additional organisms including microbes, genetic and radiation hybrid maps, extended and detailed listing of features and generation of summary reports, text-based searches and query of underlying data, BLAST searches against individual or combined assembled sequences and products, and pattern searches against genomes that return genome location and context of related sequences.

In addition to Java-based browsing, the Genome Catalog, an HTML-based interface to the Genome Channel is under development. Genome, chromosome, contig, and clone summary reports, gene and protein lists, homologies, and other features are available for browsing and querying through this interface. We are researching the feasibility of providing interfaces to additional types of analysis results, such as protein threading and structural classification that might provide clues to the functions of predicted genes. Other features being studied for future implementation and visualization include gene expression data, polymorphisms and mutations.

(Research sponsored by the Office of Biological and Environmental Research, USDOE under contract number DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.)

62. Continuation of the Genome Database

Christopher J. Porter, C. Conover Talbot Jr., Jay Snoddy, Ed Uberbacher, and A. Jamie Cuticchia
The Johns Hopkins School of Medicine, Baltimore, MD
cporter@gdb.org

Shortly after the last DOE Contractor and Grantee Workshop, in November 1997, DOE announced the

termination of funding for the Genome Database effective June 30th 1998. Consequently, this has been a year of transition for GDB.

In the months following the announcement, work continued on version 6.4 of GDB, which was released in March. GDB 6.4 introduced a simplified query form for regional queries, enhancements to the display of integrated map information, and multiple modifications to improve the manner in which results are displayed and increase the speed with which they are returned. A new version of Mapview features a more intuitive user interface, and allows markers selectively to be hidden. Plans to overhaul the handling of polymorphisms were withdrawn, but display of allele size and frequency information were integrated.

Despite the project's announced termination, we received four requests and set up three new international nodes in Taiwan, Belgium and Canada. At their annual meeting, representatives from the GDB nodes offered their support for continuation of the project. Previously, at their meeting at HGM'98 in Turin, members of HUGO's HGMC expressed a strong desire that the GDB project continue.

Subsequent to meetings with representatives from NCBI, OMIM and the HUGO Nomenclature Committee to explore the disposition of essential GDB activities, it became evident that the database could not be brought to a proper close in the six months allotted. Consequently we received a six month extension for the database shutdown and plans to migrate the database to Oak Ridge National Laboratory, there to be maintained as a static resource.

In late October, however, staff at the new GDB node in Toronto located a potential source of private funds for GDB continuation. Diligent work over the following two months has resulted in a rescue program for the database. The primary, editable, copy of GDB will move to the Bioinformatics Centre at the Hospital for Sick Children in Toronto, from whence it will be replicated to all international nodes. Editing and curation of the database will continue

and we will strengthen our relationship with HUGO. We are working with HUGO to supplement the 'classic' HUGO editors, who oversee genes and mapping, with editors from the sequencing community who will help to integrate physical maps and, ultimately, sequence information.

Plans are being made with the Sequence Annotation Consortium at ORNL to work on integrating GDB data into the Sequence Annotation project. Possibilities of other collaborations are also being investigated as resources allow. ORNL has received a copy of GDB, which will now be updated, to serve as primary U.S. node of the database.

63. Reconstruction and Annotation of Transcribed Sequences: The TIGR Gene Indices

John Quackenbush, Ingeborg Holt, Feng Liang, Geo Pertea, Jonathan Upton, and Thomas S. Hansen
The Institute for Genomic Research, Rockville, MD 20850
johnq@tigr.org

A goal of the Human Genome Project is identification of the complete set of human genes and the role played by these genes in development and disease. The sequencing of Expressed Sequence Tags (ESTs) has provided a first glimpse of the collection of transcribed sequences in humans and other organisms, but significant additional information can be obtained by a thorough analysis of the EST data. TIGR's analysis of the world's collection of EST sequence data, captured in our Gene Indices (<http://www.tigr.org/tdb/tdb.html#gi>), provides assembled consensus sequences that are of high confidence and represent our best estimate of the collection of transcribed sequences underlying the ESTs. In addition to the Human Gene Index (HGI; <http://www.tigr.org/tdb/hgi/hgi.html>), we maintain Gene Indices for a variety of other species, including mouse, rat, *Drosophila*, zebrafish, rice, tomato, and *Arabidopsis*. Collectively, the Gene Indices represent a unique resource for the comparative analysis of mammalian genes and may provide insight into gene function, regulation, and evolution.

We have recently expanded the TIGR Gene Index project to include quarterly releases, expanded annotation, integration with mapping and genomic sequence data, and more robust search capabilities. In addition, we are developing a database of mammalian orthologues based on comparison of the human, mouse, and rat TC sequences and a web-based presentation to allow the data to be effectively explored. This database will provide direct links between the human, mouse, and rat assemblies and represent the most extensive catalog of eukaryotic orthologues available, providing a valuable resource for gene identification, elucidation of functional domains, and analysis of gene and genome evolution.

64. An Informatics Framework for Transcriptome Annotation

Brian Brunk¹, Jonathan Crabtree¹, Mark Gibson¹, Chris Overton¹, Debra Pinney¹, Jonathan Schug¹, Chris Stoeckert¹, Jian Wang¹, Ihor Lemischka², Kateri Moore², and Robert Phillips²
¹Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104-6021 and
²Princeton University, Princeton, NJ
coverton@pcbi.upenn.edu

It is now feasible to define the transcriptional state of a eukaryotic cell with reasonable precision by combining multiple gene expression technologies, e.g., EST analysis with microarrays. However, few of the 10,000 – 20,000 different transcripts expressed in a cell are well characterized in terms of function and cell role. In a collaborative effort, we have begun the identification and characterization of the transcripts produced in the mouse hematopoietic stem cell. The Princeton group has enriched for the stem cell from fetal mouse liver by sorting for cells positive for the markers AA4.1, Sca-1 and c-Kit and low in Lin. A normalized, subtracted (against a stromal cell cDNA library) cDNA library was generated from these cells. A similar strategy was adopted in the construction of a stromal cell library. ESTs were generated from both libraries and analyzed through an automated computational annotation pipeline followed by expert manual annotation. Currently approximately 4000 stem cell and 3000 stromal cell ESTs have been carefully annotated leading to a well-defined "molecular

phenotype” of each cell type and opening the way for follow-up analyses of novel genes of interest. Based on this prototype annotation process, we have developed an integrated informatics framework for the systematic annotation of cell-specific transcriptomes. The system combines data management and visualization facilities with automated and manual data analysis components accessible through a Java servlet-based architecture. Using the K2 technology for accessing distributed databases, it integrates computationally annotated mouse and human genomes (GAIA system), computationally annotated mouse and human transcriptomes built from dbEST ESTs and known mRNAs (DOTS), and protein sequences in SwissProt. The K2 facility also provides access to a number of other remote databases and analysis services. Computational annotation steps include: clustering and assemble of ESTs/mRNAs to form consensus transcribed sequences (TSs); gene finding by similarity to TSs; similarity of TSs across species and in proteins; and assignment of cell roles/functions to TSs using computational and manual analyses. Manual annotation steps include: assessment of quality of consensus sequence to identify artifacts; refinement of cell role/function assignment; and characterization of alternative splicing. Results of the characterization of the stem and stromal cell molecular phenotypes will be presented.

65. Protein Domain Dissection and Functional Identification

Temple F. Smith, Sophia Zarakovich, and Hongxian He
BioMolecular Engineering Research Center, College of Engineering, Boston University, 36 Cummington Street, Boston, MA 02215
tsmith@darwin.bu.edu

Using various multialignment and conserved pattern tools (e.g., psiBLAST, BLOCKS, pfam, pimaII, etc.) protein domains as “evolutionary modules” can generally be identified. Using a set of 20 completely sequenced microbial genomes (including yeast), we

have generated over 1300 profiles representing diagnostic sequence domains. The majority either cover the entire length of the proteins matching the profile or identify a sequence region clearly identifiable in multiple distinct domain contexts. The relationship between such sequence domains and structural domains will be discussed with examples. The problems involved in associating these domains to a given biochemical function and/or the cellular role played by that function will also be addressed.

66. Finding Remote Protein Homologs

Kevin Karplus
University of California, Santa Cruz, Baskin School of Engineering, Santa Cruz, CA 95064
karplus@cse.ucsc.edu

Since Spring 1996, the bioinformatics group at UCSC has been working on ways to find and align homologs of proteins, even when the sequences of the proteins are quite diverged. Our main approach has been to use hidden Markov models with Dirichlet mixture regularizers for both the search and the alignment. The method uses only sequence information, not structural information, and so can be applied even to proteins whose structure is still unknown.

The main tests for the method are fold-recognition and alignment tests—searches and alignments are made for proteins whose structure is known (but not used in the search or alignment), and the results are compared with the results of structural alignment. In a test against other sequence-based search and alignment methods (including PSI-BLAST and ISS), our SAM-T98 method found more true homologs (based on SCOP) than other methods at any level of accepted errors.

A common use of remote homologs is to predict the structure of the protein. We have participated in both the CASP2 and CASP3 experiments for blind prediction of protein structure. In both, we were in the top six groups (invited to the special issue of

Proteins) for fold recognition and alignment. In CASP3, our alignments of the comparative-homology targets were consistently among the best (approximately top 3), even though we made no use of structural information.

For CASP3, we also tested a secondary-structure predictor using a neural net and the SAM-T98 multiple alignments used by our fold-recognition method. This predictor was the second best of the 31 groups participating, and we have since improved it.

We have installed an automatic server on the Web to take a sequence (or seed alignment) and produce the multiple alignment of similar sequences in NCBI's non-redundant protein database, search results for proteins with structures in PDB, and secondary-structure predictions: <http://www.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html>

For more information about our projects, see <http://www.cse.ucsc.edu/research/compbio>

67. Multi-Way Protein Folding Classification Using Support Vector Machines and Neural Networks

C.H.Q. Ding and I. Dubchak
National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
ildubchak@lbl.gov

In bioinformatics research, classification methods for multiple classes recognition employed so far are mostly based on the one-vs-others approach. We investigated two advanced approaches, the unique one-vs-others approach and the all-vs-all approach with increased classification accuracy.

We analyzed the traditional sensitivity and selectivity measures for multi-class classification from a new perspective of contingency table in categorical analysis, and provided some insights. These true positive and false positive based measures are combined and generalized to a new unique accuracy measure which characterize more accurately the performance of a recognition system. This measure can be applied consistently and uniformly to all

multi-class classification approaches thus facilitating inter-comparisons of different classification methods.

We used the state-of-art Support Vector Machine (SVM) together with an earlier neural network (NN) two-class classifiers. SVM gives higher accuracy and runs much faster than NN.

Of the six different physico-chemical based parameter sets extracted from protein sequences, we found that the amino acid composition based parameter set is the most effective for the discriminative methods. The secondary structure based parameters are also quite effective. These are followed by parameter sets extracted from hydrophobicity, polarity, van der Waals volume, and polarizability properties.

68. Comparative Analyses of Syntenic Regions using Pattern Filtering

Jonathan E. Moore and James Lake
Molecular Biology Institute and MCD Biology, University of California, Los Angeles, CA 90095

Comparative computational analyses of syntenic DNA sequences among the bilateral animals hold great potential for the identification of genomic features, such as protein coding regions, gene boundaries, introns, and genetic regulatory elements. A computational method, developed in our lab, called pattern filtering optimally separates the signals conserved between sequences from the noise caused by the stochastic process of nucleotide substitution. We are currently developing related methods having more statistical power for the identification of protein coding regions and their boundaries. Preliminary analyses, utilizing only pattern filtering methods, of mammalian mitochondrial DNAs and of the human chromosome 12p13 locus and its syntenic region in mouse show the considerably promise of this method. We plan to continue our progress toward rapid and effective analysis through pattern filtering of genomic features in the syntenic regions of mouse, human, and other species.

69. Discovery of Distant Regulatory Elements by Comparative Sequence-Based Approaches.

Inna Dubchak¹, Chris Mayor¹, Lior Pachter², Gabriella Cretu¹, Edward M. Rubin¹, and Kelly A. Frazer¹

¹Genome Sciences Department, Lawrence Berkeley National Laboratory, 1 Cyclotron Road MS 84-171, Berkeley, CA. 94720 and ²Mathematics Department, University of California, Berkeley, CA 94720
ildubchak@lbl.gov

Distant regulatory elements, such as enhancers, silencers, and insulators, are experimentally difficult to identify. Exploiting the fact that these elements tend to be highly conserved among mammals we are using comparative sequence-based approaches to discover them. To find conserved non-coding sequences with physical attributes of distant regulatory elements we compared ~ 1 Mb of orthologous human (5q31 interleukin cluster region) and mouse (chromosome 11) sequences. Ninety non-coding sequences (≥ 100 bp and $\geq 70\%$ identity) were identified - analysis of 15 found that ~ 70% were conserved across mammals but unique in the human genome. Although this study discovered numerous conserved non-coding sequences with features of distant regulatory elements only one of the two enhancers previously identified in the human 5q31 region was detected.

To improve the ability of comparative sequence analysis to identify distant regulatory elements we have developed a new method which globally aligns the sequences being compared and plots the percent identity of a moving average point (MAP). The advantage of MAP analysis over the previous method used is that it can detect conserved non-coding sequences with small insertions/deletions and is capable of three-way species comparisons. Comparison of ~ 200 kb of orthologous human (5q31), mouse (chromosome 11), and dog (chromosome 4) sequences using MAP analysis found all the known conserved non-coding sequences (≥ 100 bp and $\geq 70\%$ identity) in region as well as

additional non-coding elements, including the enhancer previously undetected by comparative analysis. The overall pattern of non-coding sequence conservation in the orthologous human, dog and mouse genomic DNA is strikingly similar suggesting the majority of elements identified based on conservation are likely to have biologic function. Experimental characterization of the largest non-coding element identified in these studies determined it to be a potent regulatory element of three genes, IL-4, IL-13 and IL-5, spread over 120 kb.

70. Identification of Novel Functional RNA Genes in Genomic DNA Sequences

S.R. Holbrook, C. Mayor, and I. Dubchak
Physical Biosciences Division and National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
ildubchak@lbl.gov

Finding the location of functional RNA genes in genomic sequences is much more difficult than the assignment of ORFs as potential protein coding genes. To date, the only method of identifying functional RNA genes is by homology.

Our initial approach to locating novel RNA genes was based on the premise that all stable, functional RNAs share common structural elements and that sequences corresponding to these elements occur preferentially in RNA genes. These elements include tetraloops, uridine turns, tetraloop receptors, adenosine platforms, and a high percentage of double helical base pairing. We have also used the free energy of folding as a structural parameter representing double helicity in RNA sequences. Since the frequency of occurrence of RNA structural elements can not be expected to identify non-RNA sequences in a positive manner, we identified additional sequence preferences based on global sequence descriptors (previously applied to protein fold prediction) to discriminate RNA genes from non-

RNA genes. These descriptors include composition, distribution, and transition parameters.

A total of 610 examples of *E. coli* sequence windows (305 from RNA genes, 305 from non-assigned regions) were used to calculate the descriptors and train neural networks. In order to optimize prediction, we used a voting procedure in which predictions were accepted only when predicted by both types of networks. The accuracy of RNA gene prediction using different combinations of global and structural parameters was estimated by the cross-validation test. Similarly we trained neural network to recognize RNA genes in other species.

Using trained neural networks we have predicted putative RNA genes in complete genomes of *E. coli*, *M. genitalium*, *M. pneumoniae*, and *P. horikoshii*. The weights from the trained neural networks are now used in a public web server to allow users to make predictions using their sequences. We will be enlarging the number of organisms present in the database of our server, including other bacteria, lower eukaryotes such as yeast and ultimately human.

71. Automatic Discovery of Sub-Molecular Sequence Domains in Multi-Aligned Sequences: A Dynamic Programming Algorithm for Multiple Alignment Segmentation

Eric Poe Xing, Ilya Muchnik¹, Denise Wolf, Inna Dubchak, Casimir Kulikowski¹, Manfred Zorn, and Sylvia Spengler
Center for Bioinformatics and Computational Genomics, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 and ¹DIMACS, Rutgers University, Piscataway, NJ 08855
EPXing@lbl.gov

Automatic identification of sub-structures in multiple sequence alignment is of great importance for effective and objective structural/functional domain annotation, phylogenetic treeing and many other types of molecular analyses. We present a segmentation algorithm that optimally partitions a given multi-alignment into a set of potentially biologically sensible segments based on the statistical

profile of sequence compositions of the multi-alignment, such as gap frequency and character heterogeneity, through dynamic programming and progressive optimization. Using this algorithm, a large multi-alignment of eukaryotic 16S rRNA was analyzed. Three types of sequence patterns: shared conserved domain; shared variable motif; and rare signature sequence, were identified automatically in a very short time compared to manual annotation, and the result was consistent with the patterns identified through independent phylogenetic approaches. This algorithm potentially facilitates the automation of sequence-based sub-molecular structural and evolutionary analyses through statistical modeling and high performance computation.

72. Classification of Multi-Aligned Sequence Using Monotone Linkage Clustering and Alignment Segmentation

Eric Poe Xing, Ilya Muchnik¹, Manfred Zorn, and Sylvia Spengler
Center for Bioinformatics and Computational Genomics, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 and ¹DIMACS, Rutgers University, Piscataway, NJ 08855
EPXing@lbl.gov

Optimal clustering of a set of sequence based on arbitrary set function is often of exponential complexity. In this paper, a low order polynomial procedure, which is based on the quasi-concavity of a special type of objective functions, was developed to cluster the multi-aligned sequences based on each of the segments resulted from the aforementioned segmentation process. It clusters sequences according to their degree of similarity to a pre-specified reference pattern (i.e. a consensus sequence or a particular organismal sequence of choice). A combination of such clustering from multiple segments results in a fairly fine-grained classification of all the sequences in the alignment, with a general pattern that is reminiscent of the branching order in a corresponding phylogenetic tree, but with additional information regarding the assumption of modular evolution. This algorithm can be applied to a broad spectrum of molecular sequence analysis purposes such as phylogenetic subtree construction or

recognition, tree updating and labeling, and can serve as a framework to organize sequence data in an efficient and easily searchable manner.

73. Extensions to the Arraydb Micro-Array LIMS

Donn Davy¹, **Daniel Pinkel**², **Donna Albertson**², **Gregory Hamilton**², **Joel Palmer**², **Donald Uber**¹, **Arthur Jones**¹, **Joe Gray**², and **Manfred Zorn**¹

¹Lawrence Berkeley National Laboratory, Berkeley, CA 94720 and ²University of California, San Francisco Cancer Center, San Francisco, CA dfdavy@lbl.gov

We have extended our Arraydb Laboratory Information System (LIMS) collaboration with UCSF Cancer Center to allow separate laboratories' users to separately track progress as micro-array slides are printed and used. The system prevents users from seeing each others' data, except where specified, while tracking clones, DNA, as it is prepared and plated, microtiter plates, print-run specifications, slide-printing runs and slides printed, experiments on and images of the slides, and image-analyses. Both slide printer and image-analysis software write directly to the database. Users have the option of downloading selected tabular reports to Excel spreadsheets. Enhancements underway will also allow upload into the database from spreadsheets.

The system is implemented on an Oracle 8 database and served on the Web by a NetDynamics application server, providing a highly scaleable, flexible, and responsive solution. It is accessible from java-compatible web browsers, and provides a fine-grained control over security and accessibility.

74. Identifying Single Nucleotide Polymorphisms (SNPs) in Human Candidate Genes

Deborah A. Nickerson, **Scott L. Taylor**, and **Mark J. Rieder**

Department of Molecular Biotechnology, Box 357730, University of Washington, Seattle, WA 98195

debnick@u.washington.edu

Single nucleotide substitutions and unique base insertions and deletions are the most common form of polymorphism and disease-causing mutation in the human genome. Based on the natural frequency of these variants, they are likely to be the underlying cause of most phenotypic differences in humans. Because of their functional importance, a number of methods have been developed to identify single nucleotide polymorphisms (SNPs). Among these, direct sequence analysis has many advantages because it provides complete information about the location and nature of any variants in a single pass, is automatable, widely available, and simple to apply. To further automate the detection of SNPs by direct sequence analysis we have developed PolyPhred which together with Phred, Phrap, and Consed identifies nucleotide substitutions within a target sequence. Over the past year, we have developed several approaches to increase the accuracy and selectivity of PolyPhred as well as a tool known as PolyPhred2db that simplifies the development of databases of SNPs using information obtained by PolyPhred. The application of these tools in analyzing the diversity of human candidate genes will be described. Our results suggest that the levels and patterns of sequence variation found in human genes could pose challenges in identifying the sites, or combination of sites, that influence variation in the risk of disease within and among human populations.

75. Integrating Sequence and Biology: Developing an Informatics Infrastructure for Mouse/Human Comparative Genomics

C.J. Bult, J.T. Eppig, J.A. Blake, J.E. Richardson, and J.A. Kadin

The Jackson Laboratory, Bar Harbor, ME 04609
cjb@informatics.jax.org

Sequence similarity provides a powerful mechanism for predicting orthogonal relationships between mouse and human genes. However, it is the extension of sequence level correspondence to the detailed knowledge about the genes and the relationships of genes to phenotype that makes the comparative genomics approach such a powerful one for understanding biological processes. As the capacity to collect large data sets of complex biological information grows, integration of data from diverse sources about the same genomic feature from diverse sources will be key to developing new insights into human biology using mouse as a model organism.

Although a number of highly automated sequence annotation pipelines have been developed to support large-scale genomic sequence projects, relatively little attention has been paid to developing the infrastructure needed to support the integration of genomic sequence data with related biological information and data. The Mouse Genome Sequence (MGS) database is being developed to provide access annotated mouse genome sequence that has been integrated with existing biological knowledge about the laboratory mouse (e.g., phenotype, expression data, gene homology that are represented in other databases (see the Mouse Genome Database and Gene Expression Database at <http://www.informatics.jax.org>). MGS represents a critical component of the informatics infrastructure that is needed to support comparative, computational, and functional genomics.

76. WIT2 — An Integrated System for Genetic Sequence Analysis and Metabolic Reconstruction

Ross Overbeek^{1,2}, Gordon Pusch^{1,2}, Mark D'Souza¹, Evgeni Selkov Jr.^{1,2}, Evgeni Selkov^{1,2}, and Natalia Maltsev¹

¹Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL and
²Integrated Genomics Inc., Chicago, IL
maltsev@mcs.anl.gov

The WIT2 system (<http://wit.mcs.anl.gov/WIT2>) was designed and implemented to support genetic sequence and comparative analysis of sequenced genomes as well as metabolic reconstructions from the sequence data. It now contains data from 38 distinct genomes. WIT2 provides access to thoroughly annotated genomes within a framework of metabolic reconstructions, connected to the sequence data; protein alignments and phylogenetic trees; as well as data on gene clusters, potential operons and functional domains. We believe that the parallel analysis of a large number of phylogenetically diverse genomes simultaneously can add a great deal to our understanding of the higher level functional subsystems and physiology of the organisms. The unique features of WIT2 include: 1. WIT2 is based on the unique EMP/MPW collection of the enzymes and metabolic pathways developed by E. Selkov et al., which contains extensive information on enzymology and metabolism of different organisms. 2. WIT2 allows researchers to perform interactive genetic sequence analysis within a framework of metabolic reconstructions and to maintain user models of the organism's functionality. 3. WIT2 provides access to a set of Web-based and original batch tools that offer extensible query access against the data. 4. WIT2 supports both shared and nonshared annotation of features and the maintenance of multiple models of the metabolism for each organism. 5. WIT2 supports metabolic reconstructions from Expressed Sequence Tags (EST) data.

77. PUMA2 — An Environment for Comparative Analysis of Metabolic Subsystems and Automated Reconstruction of Metabolism of Microbial Consortia and Individual Organisms from Sequence Data

Natalia Maltsev and Mark D'Souza
Mathematics and Computer Science Division,
Argonne National Laboratory, Argonne, IL
maltsev@mcs.anl.gov

We have developed a working prototype of interactive environment PUMA2 which intends to accomplish the following goals:

1. Allow comparative analysis of the metabolic subsystems in different organisms
2. Provide a framework for the automated reconstruction of the metabolism of microbial consortia and individual species
3. Provide a framework for representation of the expression data.

Analysis of data in PUMA2 is based on an original approach for representation of metabolism as a network of interconnected modules connected to sequence data. The results of such analyses will be presented in graphical form based on hierarchical representation of the functional subsystems and annotated with sequence data and literature information.

78. Progress Report on EMP Project

Evgeni Selkov, Nadezhda Avseenko, Valentina Dronova, Galina Dyachenko, Aleksandr Elefterov, Milyausha Galimova, Nadezhda Fedotcheva, Maria Fomkina, Tatiana Kharybina, Irina Krestova, Aleksandr Kuzmin, Elena Mudrik, Nikolay Mudrik, Valentina Nenasheva, Valeri Nenashev, Evgeni Nikolaev, Aleksandr Osipov, Lyudmila Pronevich, Anna Rykunova, Aleksey Selkov, Evgeni Selkov, Jr., Vladimir Semerikov, Tatiana Sirota, Anatoly Sorokin, Oleg Stupar', Vadim Ternovsky, and Olga Vasilenko

EMP Project Inc, Russian Subsidiary, Institutskaya 4, suite 121, 142292 Pushchino, Moscow Region, Russia 142292 Pushchino, Moscow Region, Russia

In the first quarter of this FY, the main focus of the EMP Project was on the following:

A high database annotation rate of about 800 records; there were 2,423 and 6,850 records encoded during the last three and twelve months, respectively (See Tab. 1), totaling 29,991 records by the end of 1999.

Table 1. EMP Updating Rate For the Last Twelve Months

Month	Records Encoded
Jan-99	147
Feb-99	395
Mar-99	372
Apr-99	461
May-99	484
Jun-99	615
Jul-99	795
Aug-99	599
Sep-99	559
Oct-99	827
Nov-99	899
Dec-99	697
Total	6850

Updating the EMP content with new records for novel enzymes classified by the Supplements 5 and 6 to the Enzyme Nomenclature.

Developing a new EMP format to simplify the encoding and retrieving the information with a special effort of extending the format toward signal transduction pathways and phenotype.

Organizing and training groups of annotators in some specific information domains, e.g. Enzyme Kinetics, Signal Transduction, Metabolic Pathways, Phenotype, etc.

Organizing a Software Development Team of 8 professional developers. Designing a new web site interface for EMP (<http://www.empproject.com>) to be installed with next two months.

Opening a new office in Pushchino of 120 m2 with 19 working stations; improving hardware, network

connection, and technical for annotators working at home.

In the coming months, we plan to develop a new EMP format documentation for annotators and users. This will slow down the annotation rate. Still, it will remain to be at the planned level of not less than 500 records a month.

With a very active support of GlaxoWellcome, we began a dialogue with the Swiss-Prot staff to unify the nomenclature and software development and to coordinate the information processing and encoding for Swiss-Prot and EMP. A trilateral workshop to discuss this cooperation between Swiss-Prot, EMP Project, and GlaxoWellcome is to be held in Geneva on March 27-29.

79. BCM Search Launcher - Providing Distributed, Enhanced Sequence Analysis

M. P. McLeod, Z. Yang, and K. C. Worley
Department of Molecular and Human Genetics,
Baylor College of Medicine, Houston, TX 77030
mmcleod@bcm.tmc.edu

We provide web access to a variety of enhanced sequence analysis search tools via the BCM Search Launcher. The BCM Search Launcher (<http://www.hgsc.bcm.tmc.edu/SearchLauncher/>) is an enhanced, integrated, and easy-to-use interface that organizes sequence analysis servers on the WWW by function, and provides a single point of entry for related searches. This organization makes it easier for individual researchers to access a wide variety of sequence analysis tools. The Search Launcher extends the functionality of other web-based services by adding hypertext links to additional information that can be extremely helpful when analyzing database search results.

The BCM Search Launcher Batch Client provides access to all of the searches available from the Search Launcher web pages in a convenient interface. The Batch Client application automatically 1) reads sequences from one or more input files, 2) runs a specified search in the background for each sequence, and 3) stores each of the search output files as

individual documents directly on a user's system. The HTML formatted result files can be browsed at any later date, or retrieved sequences can be used directly in further sequence analysis. For users who wish to perform a particular search on a number of sequences at a time, the batch client provides complete access to the Search Launcher with the convenience of batch submission and background operation, greatly simplifying and expediting the search process.

BEAUTY, our Blast Enhanced Alignment Utility makes it much easier to identify weak, but functionally significant matches in BLAST protein database searches. BEAUTY is available for DNA queries (BEAUTY-X) and for gapped alignment searches. Up-to-date versions of the Annotated Domains database present annotation information. The latest version of this database includes domain information from DOMO and Prodom in addition to BLOCKS, PRINTS, Pfam, Entrez sequence records, and Prosite.

Recent enhancements to the BCM Search Launcher include the addition of searches for human genomic sequences, additional domain information with BEAUTY, and an updated help system designed to assist researchers with little or no experience in computational biology.

Our collaboration with the Genome Annotation Consortium (<http://compbio.ornl.gov/channel>) provides BEAUTY search results for all of the predicted protein sequences found in the human genomic sequences produced by the large scale sequencing centers.

Support provided by the DOE (DE-FG03-95ER62097/A000).

80. Data Submission Tool

Manfred D. Zorn and David Demirjian
Lawrence Berkeley National Laboratory, Berkeley,
CA 94720
DGDemirjian@lbl.gov

SubmitData provides researchers with a simple and intuitive solution for annotated data submissions to public databases.

Current feature list:

- Incorporates XML as the data exchange format
- User defined database definitions using XML documents
- Smart GUI interfaces
- Complete data validation using data type syntax and rule processing
- Point & Click error correction on invalid data elements
- Simple and Complex Batch Submission support
- Create persistent reusable Batch Submission Templates
- Batch Submission process can merge data from external files into selected elements using Batch Submission Template Variables
- User definable data export format
- XML GUI Document Editor / Parser / Validation
- Complete Help Pages
- Uses current Java Technologies: Java2, JavaMail, JavaHelp, Java Project X

81. Working Examples of XML in the Management of Genomic Data

J. D. Cohn and M. O. Mundt

Bioscience Division and DOE Joint Genome Institute,
Los Alamos National Laboratory, Los Alamos, NM
87545

jcohn@lanl.gov

XML is fast becoming the universal format for structured data exchange and documents on the Web. Standards for XML were developed by the World Wide Web Consortium (W3C) and have been adopted for a wide range of applications from e-commerce to mathematics and chemistry. Unlike HTML, XML was designed to be extended and offers a much richer base to build upon (including capabilities for using binary as well as ASCII data).

Until now, exchange of genomic data has been limited primarily to FastA files and a few proprietary or application-specific formats. XML seems to offer an ideal means of enhancing our capability of exchanging data within the genomic community.

Using a growing array of XML parsers and other development tools, XML formatted data can be utilized by software applications written in a variety of different languages across multiple hardware platforms. Major database systems (e.g. Oracle) are beginning to offer XML output for SQL database queries. Further, the W3C is working on a standard for an XML query language for searching XML documents directly.

Recently we have taken the first steps in making use of XML in our distributed sequencing informatics system. Among the applications for XML which we will describe are: 1) automated loading and analysis of sample files from multiple sources (production sequencing, finishing, cDNA, outside laboratories, etc) using naming convention documents; 2) distributed data management; 3) user preference files; and 4) sequence annotation. All of these have been accomplished using the XML Parser for Java from Datachannel. Examples of XML code as well as descriptions of the applications will be presented.

82. The Genome Database — Integrating Maps with Sequence

Christopher J. Porter, C. Conover Talbot Jr., and A. Jamie Cuticchia

The Johns Hopkins University, Baltimore, MD and
The Hospital for Sick Children, Toronto, ON,
Canada

jamie@bioinfo.sickkids.on.ca

As reported at the 1999 Contractor-Grantee Workshop, the Genome Database (GDB) is now hosted by the Bioinformatics Supercomputing Centre (BiSC) of the Hospital for Sick Children, Toronto. The database was transferred in May 1999, and has since been moved to an SP supercomputer, donated to the GDB project by IBM.

GDB introduced a number of new tools during 1999. We have used NCBI's electronic PCR (e-PCR) software in a tool that retrieves GDB Amplimers predicted to amplify from a sequence. GDB-BLAST

uses the BLAST server on BiSC's Origin supercomputer, and displays GDB objects linked to the sequences retrieved. Additionally, output from BiSC's public high speed BLAST server (<http://bioinfo.sickkids.on.ca/>) was modified to display links to GDB. BiSC's supercomputers made feasible the use of e-PCR to create a database of potential amplification sites for GDB Amplimers in all public human sequence.

These resources are serving to improve GDB's mapping of Amplimers and Genes. Work progresses to integrate the extensive body of variation and SNP data from GDB into sequence-level maps.

The recent release of a full sequence of chromosome 22 has served as a proving ground for the integration of GDB data with the complete human sequence. BiSC's sequence analysis tools were used to map GDB objects onto the sequence. These results are displayed as a interactive graphical map, and are being integrated into GDB's comprehensive map. These approaches show how GDB will integrate classical mapping information with the rapidly emerging genomic sequence.

Collaborations with sequencing centers and the Genome Annotation Consortium are continuing to load clone tiling paths into GDB, and to create bidirectional links between GDB records and the annotated sequence.

International interest in GDB continues - 1999 saw the establishment of a new GDB node in Beijing, China, and work is underway to create a node in Bangalore, India.

83. A Visual Data-Flow Editor Capable of Integrating Data Analysis and Database Querying

Dong-Guk Shin¹, Ravi Nori², Rich Landers², and Wally Grajewski²

¹Computer Science and Engineering, University of Connecticut, Storrs, CT 06269-3155 and

²CyberConnect EZ, LLC, Storrs, CT 06268
shin@enr.uconn.edu

Determining mapping sequence variations or polymorphism between homologous genomic regions requires access to genomic data available from different sources and use of many data analysis and visualization programs. It is imperative that software be developed to enable genome scientists to automate tedious and repetitive data handling, database querying and analysis tasks. Our approach is to develop a data-flow editing environment in which genome scientists with minimal computer training can easily describe data analysis tasks. The scientists' use of the software tool involves organizing and coordinating individual tasks of data retrieval from different data sources, combined with data analysis tasks to derive answers to biologically significant questions.

Phase I aimed at developing prototype software which demonstrates the feasibility of a full-scale development of a data-flow editing environment in which interactions between data access and data analysis can be freely described by genome scientists with minimal computer training. The feasibility study is based on a working scenario of determining homology relationships between some known DNA sequences from one species and unknown sequences from a taxonomically-related species.

Software of this kind is expected to be immediately usable by molecular biology and the pharmaceutical industry both of which are becoming more computationally intensive. Since data-flow management problems are not unique to computational biology, the software developed is expected to be useful in many other data and computationally intensive areas, e.g., physics, chemistry, engineering and finance.

The proposed software will enable scientists to automate the repetitive analysis tasks involving an enormous amount of DNA sequence data that must be analyzed to understand its implications to biological and environmental processes. Without the software tool, the difficulties involved in conducting these large scale data analysis projects could be insurmountable due to the magnitude of data available and the variety of analysis techniques involved.

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

84. Annotating DNA with Protein Coding Domains

Winston A. Hide¹, Robert Miller¹, Gary L. Sandine², and David C. Torney²

¹South African National Bioinformatics Institute, University of the Western Cape, South Africa and

²Los Alamos National Laboratory, Los Alamos, NM
dct@ipmatil.lanl.gov

DNA genomic sequence is now becoming readily available for the human and fly genomes. Reliably finding genes and annotating gene information remains, however, at a premium. Coding domains within gene sequences are detected both by gene prediction programs, locating exons based on predictive models, and by similarity to known expressed sequences.

Predictive gene detection methods have yet to be sufficiently sensitive to be able to accurately predict all exons of a given gene. In addition, once predicted, only sequence comparison provides reliable corroboration. Once located, exonic DNA sequences need to be correctly translated into their corresponding proteins. The proteins may then be compared with known protein sequences corresponding to known structures as determined by direct or modelled homology.

Our annotation approach employs the novel paradigm of direct annotation of DNA based upon the secondary-structure properties of its translate (e.g. helix, sheet, and turn). To accomplish this, we have developed Bayesian classification methods for biological sequences. These methods use examples for 'training'. We have used several secondary-structure classes of polypeptides from the CATH database (Orengo et al 1997). (The latter is a valuable resource because it has strictly hierarchically classed secondary structures and presents homologous superfamilies found in genome sequences).

We have successfully completed an analysis of such classes. The integrals of the Bayes'-rule formulas are

approximated by finding the global maximum of the integrand the product of probabilities of the sequences in the sample. This has been a challenge for numerical analysis, but, a constrained quadratic programming program yielded near-optimal points. For example, for peptides of length four, taken from alpha-helix sequences, each point consists of 300 parameters characteristic of the sequences in the sample. These parameters yield posterior likelihoods for all peptides of length four to belong to the class.

The DNA sequences for these polypeptides may also be used for 'training'. Our direct approach, however, has been to submit genomic sequence to exon prediction engines such as 'Genome Annotator Pipeline <http://compbio.ornl.gov/tools/pipeline/index.shtml>', and, in addition, to generate a large number of processed expressed sequence tag fragments for reduced redundancy and consensus generation using clustering. Proteins predicted from both the exons and the clustered consensus sequences are submitted for analysis using our statistical methods.

The results are presented via a web system which reveals likely structural domains within exons and coding expressed sequence. We have implemented a web tool which accepts raw DNA sequence and generate predicted coding regions from expressed sequences or accept predicted exonic information and process for statistical states of structural class and display predicted states for each of the structurally encoded parameters.

Our next steps will be to

- Determine sensitivity and selectivity of the statistic with respect to current secondary structural prediction (tools that rely models/empirical derivatives)
- Analyze 100,000 EST consensus sequences, producing peptides and predicting their domains.
- Determine the efficacy of employing an implementation of our methods for synergistic support of gene finding tools
- Map gene prediction outputs onto structurally predicted states to determine jointly predicted exons.

- Annotate jointly predicted exons onto known gene and protein structures
- Determine the efficacy of combining our methods with other methods for finding genes
- Implement these methods in other important contexts, such as functional promoter class characterization and annotation.

85. Clustering and Visualizing Yeast Microarray Expression Data Using VxInsight™

George Davidson¹, Edwina Fuge², and Margaret Werner-Washburne²

¹Sandia National Laboratories, Albuquerque, NM and ²Biology Department, University of New Mexico, Albuquerque, NM
maggieww@unm.edu

The database visualization tool VxInsight™ (<http://www.cs.sandia.gov/projects/VxInsight.html>) was used to cluster and visualize two sets of data from *Saccharomyces cerevisiae*. These included Spellman's cell-cycle data (<http://genomics.stanford.edu/yeast/cellcycle.html>) and data obtained in our laboratory examining gene expression during exit from stationary phase. Microarray hybridization data were ordinated using correlation, after Eisen, and displayed in VxInsight™. Chromatin genes expressed in S-phase are shown to group closely together in this 2D visualization environment, as expected. This visualization paradigm can be used to identify uncharacterized genes that are closely grouped with well studied genes, for example genes associated with stationary phase. VxInsight™ automatically assembles HTML pages with expression plots and with the associated links into the Proteome database. Further development of this capability will enable faster and more user-friendly interpretation of huge volumes of microarray data.

86. Comprehensive Microbial Genome Display and Analysis

Frank Larimer, Doug Hyatt, Miriam Land, Richard Mural, Morey Parang, Manesh Shah, Jay Snoddy, and Ed Uberbacher

Computational Biosciences and Toxicology and Risk Analysis Sections, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831
<http://compbio.ornl.gov>

We are now representing all completed microbial genomes in the Genome Channel and the Genome Catalog, providing comprehensive sequence-based views of genomes from a full genome display, to the nucleotide sequence level. We have developed a tool for comparative multiple genome analysis that provides automated, regularly updated, comprehensive annotation of microbial genomes using consistent methodology for gene calling and feature recognition. The visual genome browser currently represents ca. 51,000 Microbial GRAIL gene models as well as providing over 45,000 GenBank gene models. Precomputed BEAUTY searches are provided for all gene models, with links to original source material as well as links to additional search engines. Comprehensive representation of microbial genomes will require deeper annotation of structural features, including operon and regulon organization, promoter and ribosome binding site recognition, repressor and activator binding site calling, transcription terminators, and other functional elements. Sensor development is in progress to provide access to these features. Linkage and integration of the gene/protein/function catalog to phylogenetic, structural, and metabolic relationships are being developed.

A draft analysis pipeline has been constructed to provide annotation for the microbial sequencing projects being carried out at the Joint Genome Institute. The pipeline is being applied to annotating the *Nitrosomonas europaea* and *Prochlorococcus marinus* genomes currently being sequenced. Multiple gene callers (currently Generation, Glimmer and Critica) are used to construct a candidate gene model set. The conceptual translations of these gene models are used to generate similarity search results and protein family relationships; from these results a metabolic framework is constructed and functional roles are assigned. Simple repeats, complex repeats, tRNA genes and other structural RNA genes are also identified. Annotation summaries are made available through the JGI Microbial Sequencing web site; in addition, draft results are being integrated into the

interactive display schemes of the Genome Channel/Catalog.

(Research sponsored by the Office of Biological and Environmental Research, USDOE under contract number DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.)

87. Infrastructure and Tools for High Throughput Computational Genome Analysis

Doug Hyatt, Phil Locascio, Victor Olman, Manesh Shah, and Inna Vokler
Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831
<http://compbio.ornl.gov/>

The Computational Biosciences Section at Oak Ridge National Laboratory provides Computational Genome Analysis resources to the DOE Joint Genome Institute, other major Genome Centers, and to the international biology community. In addition, these resources are also used internally to support the analysis of sequences in the ORNL Genome Channel and Genome Catalog systems. With the imminent publishing of the draft human genome sequence by the spring of 2000, challenges in the computational analysis of biological data are now critical. We have constructed a computational infrastructure to meet these new demands for processing sequence and other biological data for genome centers and for the biological community at large. Utilizing OBER's timely investment in a high performance resource at ORNL, we have developed the **Genomic Integrated Supercomputing Toolkit (GIST)** to address this critical throughput challenge and to provide advanced capabilities for the **Genome Analysis Toolkit (GAT)**. Both systems are described below.

Genome Analysis Toolkit (GAT)

The Genome Analysis Toolkit incorporates a wide variety of analysis tools: exon and gene prediction

tools, other kinds of feature recognition systems and database homology search systems. The exon and gene recognition systems include Grail, GrailExp and Genscan; and microbial gene prediction systems, Generation and Glimmer. Additionally, Grail suite of tools, consisting of CpG islands, PolyA sites, Simple and Complex Repeats, and BAC End analysis tools, have also been incorporated. Also included are NCBI STS E-PCR, RepeatMasker and TRNAScan-SE systems. Database homology systems include NCBI BLAST and Beauty post-processing. Supported organisms include human, mouse, *arabidopsis*, *drosophila*, and most sequenced microbial organisms.

Access to these resources is provided by the GAT client server system. Genome Analysis Toolkit is structured as a layered system. The innermost layer is the tool layer, which comprises the binary executables for the individual tools and associated configuration and data files required by these tools. The binaries are compiled for all supported hardware platforms and operating systems. The service layer, implemented in Perl, provides a platform-independent mode of tool execution. When a service script is invoked by the server, it determines the platform on which it is running and calls the appropriate tool binary. Rigorous error checking has been added at this layer to guarantee that errors in tool execution will be caught and reported to the server.

Access to individual services is provided through a master-slave server layer. The master server receives all analysis requests from clients and distributes them among the heterogeneous pool of slave machines to best utilize the available compute resources and to achieve optimal throughput. Compute-intensive analysis tasks like BLAST searches are directed to the GIST server, running on ORNL's IBM RS/6000 SP infrastructure, described below.

A generic, platform-independent command line (client) interface, written in Perl can be used to submit individual analysis requests to the server. A specialized batch processing tool, `ornl_pipeline`, has been developed to facilitate specification of customized analysis pipelines. `ornl_pipeline`, on

invocation, reads a user specified configuration file, consisting of a set of analysis directives. A single directive can consist of a logical chain of analysis to be performed on the given sequence. The pipeline then interacts with the server, submitting the specified requests along with associated input data, and collecting the server responses. The output of one analysis is typically fed as input to the next analysis in the chain, in a pipelined fashion. All results are then suitably organized and reported.

GIST (Genomic Integrated Supercomputing Toolkit)

The initial tools included in **GIST** are a framework of high performance biological application servers that include massively parallel BLAST codes (versions of BLASTN, BLASTP, and BLASTX), which are at the heart of analyses processes such as gene modeling with GRAIL-EXP. We are currently in the process of adding gene modeling tools (e.g., GRAIL-EXP) and plan multiple sequence alignment, protein classification, protein threading, and phylogeny reconstruction (for both gene trees and species trees).

The **GIST** resources are utilized by the **GAT** server in a transparent fashion, permitting the gradual introduction of new algorithms and tools, without jeopardizing existing operations. Due to the logical decoupling of the query infrastructure, we have been able to produce an infrastructure with both excellent scaling abilities and many fault-tolerant characteristics. In testing the ability to run multiple instances of tools requiring BLAST we have demonstrated that the removal of any dependent services does not cause loss of data. Instead, where processing power is removed, we observe a graceful degradation of services as long as there is some instantiation of service available, and options to permit "never fail" operation, to cope with network failure and long running operations. **GIST**'s logical structure can be thought of as having three overall components: client, administrator, and server. All components share a common infrastructure consisting of a naming service and query agent, with an administrator having policy control over agent behavior, and namespace profile.

The tools and servers are transparent to the user but able to manage the large amounts of processing and

data produced in the various stages of enriching experimental biological information with computational analysis. The goal of **GIST** is not only to provide one-stop shopping to a genome sequence-data framework and interoperable tools but also to run the codes in the toolkit on platforms where the kinds of questions users can ask are not greatly affected by hardware limitations.

Located at Oak Ridge National Laboratory within both the Center for Computational Sciences and the Computational Biosciences section, the computational infrastructure consists of the centerpiece IBM SP3, some SGI SMP machines, a DEC Alpha Workstation cluster, and a trial Linux PC cluster. We are rapidly approaching beta-stage deployment testing; after testing performance and stability, we hope to deploy the framework at NERSC, other high-performance computing sites, and other collaborators.

(Research sponsored by the Office of Biological and Environmental Research, USDOE under contract number DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.)

88. Genome Information Warehouse: Information and Databases to Support Comprehensive Genome Analysis and Annotation

Miriam Land, Denise Schmoyer, Morey Parang, Jay Snoddy, Sergey Petrov, Richard Mural, and Ed Uberbacher

Computational Biosciences and Toxicology and Risk Analysis Sections, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830
<http://compbio.ornl.gov/>

Genome Information Warehouse (GIW) supports the ORNL-based genome annotation and analysis effort by integrating experimental data and computational predictions within a single framework. This is a heterogeneous collection of different databases and data stores. The primary purpose of this data warehouse is to provide the data management for user interfaces and other analytical functions for genome information and genome sequence annotation. Some current user interfaces supported

by this data warehouse include Genome Channel, Genome Catalog, U.S. node of Genome DataBase, and a SRS mirror of community databases. The information found in GIW includes comprehensive annotation for human and mouse genomic sequences and completed microbial genomes. While the genomic sequences, themselves, are available from NCBI, EBI or DDBJ, the genome features, especially predicted genes and proteins, that can be inferred from each sequence are not being annotated at a rate that matches the rate of sequencing. As the world's knowledge-base about gene, proteins, and their interrelationships continues to grow, new insights can be gained by analyzing and reanalyzing all existing data with a consistent, managed process. One function of the GIW is to provide automated operation support for a consistent annotation process that uses the Analysis Pipeline and its analysis tools to acquire this very useful information.

GIW makes the assumption that the computed and annotated links are not going to be permanent. Since the underlying databases and knowledge change, results are likely to change. For example, the archival data sets like the nonredundant database (NR) at NCBI continue to grow and change, so that a new Blast analysis of a specific gene model can identify additional proteins with good similarities. As the knowledge-base about genes grows, gene modeling methods continue to be refined and improved which provide the impetus for recalculating the gene predictions. Libraries of BAC ends and repetitive sequences continue to grow and can provide new analysis insights by reexamining established sequences. The GIW supports the rerunning of annotation in order to provide researchers with good information and insight that was not available at the time a sequence was first published.

A significant challenge of GIW is to reanalyze existing sequences in a timely fashion while maintaining currency of the underlying archival data from legacy databases. Many of these critical, underlying archival databases do not have a very robust update mechanism; for example, new and modified sequences from NCBI must be recognized and processed and should not be confused with any

previous versions of sequences or contigs. Changes to underlying databases may occur during an analysis cycle. To maintain consistency over all sequences, we need to create analysis versions or epochs that use a consistent archival dataset. Another challenge is to present rapidly evolving information to the user in a way that provides some consistency in navigation and retrieval of data. One major challenge is to continue to develop flexible data structures in biology that can adapt to the evolving understanding of how biological entities relate to each other and new desired user functions.

The GIW primarily uses Oracle 8i to store and manage new experimental and computational data that is created at ORNL. Archival data from other legacy databases in GIW is stored and managed with SRS, flat files, GDB (Sybase-backed), XML files, and others; this archival data must be stored and updated to facilitate the value-added computational cross-linking and annotation.

A few of the completed user interfaces to these GIW databases can be accessed through <http://genome.ornl.gov/>.

(Research sponsored by the Office of Biological and Environmental Research, USDOE under contract number DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.)

89. BiSyCLES: Biological System for Cross-Linked Entries Search

Michael Brudno, Igor Dralyuk, Sylvia Spengler, Manfred Zorn, and Inna Dubchak
Center for Bioinformatics and Computational Genomics, National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
ildubchak@lbl.gov

We have developed a prototype of an Object Oriented search system, which would allow researchers in biological sciences and medicine to combine the

information found in diverse databases. The information a researcher needs in most cases is not to be found in any single source, but is divided among several. Further, these sources are often cross-linked, either by the curators of the databases or by the contributing authors. The vast amount of information available to the researcher make manual searches extremely time-consuming, so biologists require the assistance of bioinformatics specialists to process and retrieve the relevant information. The number of projects addressing similar concerns, of which the TAMBIS and the K2 are just two examples, underscores the importance of this problem.

BiSyCLES possesses features, which should be both of immediate use to biomedical researchers and of interest to the bioinformatics community: it is easy-to-use, flexible, and extendable. We built an intuitive user interface, accessible through the World Wide Web (<http://cbcg.lbl.gov/bisycles/>). BiSyCLES allows user-defined queries to be executed across all of the recognized databases. Simple query syntax, similar to AltaVista™, makes our program easy to learn and use. Furthermore, the set of databases is easily extendable because of our use of inheritance and other object-oriented techniques. This prototype works with the two databases most often used in biological research, Genbank and Medline, and it will be further extended in the near future to include others.

90. Updated ASDB: Database of Alternatively Spliced Genes

I. Dralyuk, M. Brudno, M.S. Gelfand¹, S. Spengler, M. Zorn, and I. Dubchak
National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA and ¹State Scientific Center for Biotechnology NIIGenetika, Moscow, 113545, Russia
ildubchak@lbl.gov

Version 2.1 of ASDB consists of two divisions, ASDB(proteins), which contains 1922 amino acid sequences, and ASDB(nucleotides) with 2486 genomic sequences. ASDB(nucleotides) was developed in 1999, while ASDB (proteins) was updated with the latest data from SwissProt and

improved clustering procedures. The database can be assessed at the URL <http://cbcg.neresc.gov/asdb>. SwissProt uses two formats for description of alternative splicing. Thus the protein sequences were selected from SwissProt using full text search for the words “alternative splicing” and “varsplic”.

In order to group proteins that could arise by alternative splicing of the same gene, we developed the clustering procedure. Two proteins were linked if they had a common fragment of at least 20 amino acids, and clusters were initially defined as maximum connected groups of linked proteins. Each cluster was represented by multiple alignment of its members.

It turned out that some clusters were chimeric, in the sense that they contained members of multigene families, but not alternatively spliced variants of one gene. Therefore the multiple alignments were subject to additional analysis aimed at detection of chimeric clusters.

This processing covers the cases when alternatively spliced variants are described in separate SwissProt entries. The other kinds of ASDB records, originating from the SwissProt entries with the “varsplic” field in the feature table, usually provide the information on the variable fragments of the several proteins which result from the alternative splicing of a single gene. Thus ASDB(proteins) entries are marked with different symbols to allow for easy differentiation among the three types: those proteins which are part of the ASDB clusters and the corresponding multialignments, those which have the information on different variants in the associated SwissProt entries, and those for which the information on the variants is not available at the present time. ASDB contains internal links between entries and/or clusters, as well as external links to Medline, GenBank and SwissProt entries.

The ASDB(nucleotides) division was generated by collecting all GenBank entries containing the words “alternative splicing” and further selection of those entries that contain complete gene sequences (all CDS fields are complete, i.e. they do not have continuation signs).

91. Splice Site Recognition

Terry Speed and Simon Cawley
University of California at Berkeley, Berkeley, CA
94720-3860

With the increasing abundance of completely sequenced genomes the automation of genome annotation has become an important research goal. We focus on the classification of splice sites in eukaryotic genes, an integral sub-task in most successful genefinding programs. In particular we focus on probabilistic models for splice sites, since they can be readily incorporated into probabilistic genefinders without having to worry about how to weight the evidence of splice site classifiers. We make use of variable length Markov chains (also known as context models). VLMCs can capture long-range dependencies in splice sites without having the usual problem of exponential increase in the number of parameters encountered with regular Markov models. We compare these VLMCs with existing splice site recognition methods, both as a stand-alone problem and within PfParser, a hidden Markov model genefinding program for *Plasmodium falciparum* (a Malaria parasite).

92. Refreshing Curated Data Warehouses Using XML

Susan B. Davidson, Hartmut Liefke, and G. Christian Overton
Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104-6389
susan@cis.upenn.edu

The process of building a new database relevant to some field of study in biology involves transforming, integrating and cleansing multiple external data sources, as well as adding new material and annotations. Such databases are commonly called curated warehouses (or materialized views) due to the fact that they are derived from other databases with value added. Building them entails two primary problems:

- 1) specifying and implementing the transformation and integration from the underlying source databases to the view database.
- 2) automating the refresh process.

Previously, we have reported on the development of the Kleisli system for implementing data transformation and integration (the first problem). In this abstract, we focus how XML can be used to solve the second.

XML is a “self-describing” or semi-structured data format that is increasingly being used for data exchange. More recently, XML query languages and storage techniques have been proposed which enable its use in data-warehousing; we study the problem of using XML to detect and propagate updates. Note that determining how the underlying data sources have changed is a complicated problem, due to the fact that biomedical databases propagate their updates in one of three ways:

- a) Producing periodic new versions;
- b) Timestamping data entries; and
- c) Keeping a list of additions and corrections; each element of the list is a complete entry.

We have developed efficient “diff” techniques for comparing old versions of entries with updated versions of entries which produce the minimal updates in XML. Using these minimal updates, we show that the curated warehouse can be incrementally updated rather than recomputed from scratch for a large class of warehouse definitions.

93. Genome-Scale Protein Structure Prediction in *Prochlorococcus europae* Genome

Ying Xu, Dong Xu, Oakley H. Crawford, J. Ralph Einstein, and Ed Uberbacher
Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6480
xyn@ornl.gov

The goal of this pilot project is to assign the maximum amount of structural information to proteins, computationally identified from genes, of the *Prochlorococcus europae* genome, using a combination of a number of existing methods. Proteins are first classified into four categories: (1) proteins having high level (> 40%) of sequence similarity with their homologs in PDB, as identified by BLAST searches; (2) proteins having medium level (25-40%) of sequence similarities with their homologs in PDB, as detected by PSI-BLAST and (super-)family-specific profiles like HMM models; (3) proteins having low level (< 25%) of sequence similarity with their homologs in PDB, as detected by threading methods; and (4) proteins having no homologs in PDB, as determined by threading and statistical analysis. For each protein of the first class, our prediction system applies MODELLER and SWISS-MODEL to generate a few all-atom structure models. Structure models are generated similarly for proteins of the second class after some refinement on the BLAST-generated alignment based on information extracted from HMM models, active site/motif search results, residue-residue contact patterns, etc. The initial alignments of proteins of the third class are generated by threading methods, including our own program PROSPECT, and refinements are done in a similar fashion. Loop regions are first modeled using mini-threading methods; all-atom models are then generated using MODELLER, SWISS-MODEL, and CNS, based on the threading alignments and modeled loop regions. A combined method of threading and statistical analysis is used to determine if a protein has a new structural fold. Instead of attempting to generate full 3D structures for proteins of class 4, our prediction system searches for possible active sites and predicts structural motifs using the local threading option of PROSPECT. For each prediction, the system assigns a confidence value of the prediction based on our performance analysis on a benchmark data set. Preliminary prediction results will be presented in this presentation.

(Research sponsored by the Office of Biological and Environmental Research, USDOE under contract number DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.)

94. The Ribosomal Database Project II: Providing an Evolutionary Framework

James R. Cole, Bonnie L. Maidak, Timothy G. Lilburn, Charles T. Parker, Paul R. Saxman, Bing Li, George M. Garrity, Sakti Pramanik, Thomas M. Schmidt, and James M. Tiedje
Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824
colej@msu.edu

The Ribosomal Database Project - II (RDP-II) provides rRNA related data and tools important for researchers from a number of fields. These RDP-II products are widely used in molecular phylogeny and evolutionary biology, microbial ecology, bacterial identification, characterizing microbial populations, and in understanding the diversity of life. As a value-added database, RDP-II offers aligned and annotated rRNA sequence data, analysis services, and phylogenetic inferences derived from these data to the research community. These services are available through the RDP-II web site (<http://www.cme.msu.edu/RDP/>).

Release 7.1 (September 1999) contained more than 10,000 aligned and annotated small subunit (SSU) rRNA sequences. A special focus of this release was the identification and annotation of sequences from type material. Over 3,000 type sequences representing 636 distinct prokaryotic genera were included in release 7.1. These type sequences provide a mechanism for users to place new sequences in a taxonomic as well as phylogenetic framework. This release also included the introduction of an interactive assistant to help with the planning and analysis of T-RFLP experiments (TAP T-RFLP).

We are now preparing release 8, scheduled for March 2000. For this release we are enhancing the alignment to match a new set of alignment guidelines to help us provide an alignment with more consistent treatment of secondary structure regions. This release will contain over 20,000 aligned prokaryotic SSU rRNA sequences, including the vast majority of prokaryotic SSU sequences available through GenBank release 114 (October 15th 1999). Initially, release 8 will be made available without manual curation of annotation information. We hope the RDP advisory

panel we are in the process of establishing will help us set new annotation standards that better serve our users with available curation resources. Release 8 will also mark a turning point for RDP. It will be the first release since 1994 where the delay between sequences becoming available through GenBank and being released in aligned format by RDP has actually decreased. We expect both the time to release and frequency of releases to continue to improve through 2000.

Function and cDNA Resources

95. The I.M.A.G.E. Consortium: Progress Toward a Complete Set of Human Genes

Christa Prange, Peg Folta, Tim Harsch, Genevieve Johnson, Tom Kuczmarksi, Bernadette Lato, Leeanne Mila, David Nelson, and Anthony Carrano
Biology and Biotechnology Research Program,
Lawrence Livermore National Laboratory,
Livermore, CA 94550
prange1@llnl.gov

The I.M.A.G.E. Consortium is the largest publicly available collection of cDNAs, containing approximately three million clones. cDNAs are currently derived from five different species, with an emphasis on sampling from both normal and abnormal human and mouse tissue types at a variety of developmental stages. As a collaborative effort between the National Cancer Institute (NCI) and various academic groups, the Cancer Genome Anatomy Project employs EST sequence data to characterize normal, pre-cancerous, and cancerous cell types. We have also recently begun arraying cDNAs from full-length enriched libraries as part of the Mammalian Gene Collection, a collaborative effort between the I.M.A.G.E. Consortium, the National Institutes of Health, the National Center for Biotechnology Information, and many academic groups.

Another goal of the I.M.A.G.E. Consortium is to provide web-based software to aid in the analysis of clones derived from I.M.A.G.E. libraries. Re-arrayed sets of clones representing specific target genes will

be chosen based on this clustering analysis, and made available for use by the community.

Further information about the I.M.A.G.E. Consortium is available by email (info@image.llnl.gov) or through the WWW (<http://www-bio.llnl.gov/bbrp/image/image.html>).

This work was performed under the auspices of the U.S. Department of Energy, Office of Health and Environmental Research (OHER) by Lawrence Livermore National Laboratory under contract no. W-7405-Eng-48.

96. Analysis of Uncharacterized Human cDNAs which Encode Large Proteins in Brain

M. Oishi, T. Nagase, R. Kikuno, M. Hiroswawa, and O. Ohara
Department of Human Gene Research, Kazusa DNA
Research Institute, Kisarazu, Chiba, Japan

The aim of Kazusa cDNA project, which was initiated five years ago, is to accumulate sequence and other information of unidentified full-length human cDNA clones which encode large proteins in brain. Our ultimate goal is to characterize brain functions on molecular basis and to identify genes responsible for serious neurodisorders such as schizophrenia and bipolar disorders. From human brain cDNA libraries, which are quite versatile in their contents compared to those of other tissues, we have been focussing on large cDNA clones (with more than 4 kb in size), mainly because (1) genes known to be responsible for diseases tend to be large

in size and (2) large cDNAs, for some reasons, have been left out in world-wide efforts of cDNA characterization. After initial screening for the clones to make large proteins in vitro, selected clones are subjected to entire sequencing, expression pattern analysis among major tissues and brain sub-tissues and determination of their chromosomal location. To date, characterization of more than 1200 cDNA clones has been completed and the information is accessible through a data base for Human Unidentified Genes Encoding (the HUGE protein data base; <http://www.kazusa.or.jp/huge>).

97. Novel Approaches to Facilitate Gene Discovery and the Development of a Non-Redundant Arrayed Collection of Full-Length cDNAs

Sergey Malchenko¹, Brian Berger¹, Vera Da Costa Soares¹, Maria De Fatima Bonaldo¹, and Marcelo Bento Soares^{1,2}

Departments of ¹Pediatrics and ²Physiology and Biophysics, The University of Iowa, Iowa City, IA 52242

bento-soares@uiowa.edu

(A) Gene Discovery. Serial subtraction of normalized cDNA libraries has proven powerful to expedite gene discovery in large-scale EST programs. This strategy enabled us to generate large non-redundant collections of rat (41,000), mouse brain (22,000), and human (15,700) cDNAs within a two-year time frame, with minimal sequencing effort. This process, however, can only facilitate the identification of mRNAs that are represented in starting libraries. Since the number of primary recombinants needed to guarantee representation of rare mRNAs exceeds those typically attained in standard libraries, it is anticipated that a fraction of such transcripts will not be represented. Furthermore, mRNAs whose expression is limited to a small number of cells within a tissue may also not be appropriately represented in a bulk tissue library regardless of their level of expression. To address this problem, we developed a method aimed at the cloning of mRNAs that are either under- or not-represented in standard normalized libraries. We have applied this procedure to construct a mouse hippocampus cDNA library significantly enriched for rare mRNAs. Enrichment

was documented by analysis of over 1,000 ESTs as well as by Southern hybridization of library DNA with a number of cDNA probes that were under-represented in the non-normalized or normalized mouse hippocampus libraries.

(B) Development of non-redundant collections of full-length cDNAs. Full-length-enriched libraries have been and continue to be constructed and made available to the Mammalian Gene Collection program, a trans-NIH initiative to generate high accuracy sequence of large numbers of full-length cDNAs. However, given that enrichments are typically of the order of 50%, some screening strategy is necessary for en masse selection of the full-length clones in these libraries. We are developing novel methods and strategies to address this problem with the goal of generating comprehensive non-redundant collections of arrayed full-length cDNAs.

98. From EST to High Quality cDNA: The BDGP Pipeline for the Construction of *Drosophila* cDNA Resources

Mark Stapleton¹, Damon Harvey, Peter Brokstein, and Gerald M. Rubin

Berkeley *Drosophila* Genome Project-University of California, Berkeley, CA and ¹Lawrence Berkeley National Laboratory, Berkeley, CA
staple@bdgp.lbl.gov

The *Drosophila* Genome Center's future goals are centered around functional genomics. By taking advantage of the *Drosophila* genomic sequence, we intend to develop tools and technologies for answering biological questions in a high-throughput environment. Our first step in this direction is to create a publicly available unigene set of *Drosophila* cDNAs and sequence them to high quality.

We have finished the second stage of creating a set of *Drosophila* cDNAs. The first stage consisted of sequencing greater than 80,000 5' ESTs and was finished March 19, 1999. For the second stage, these ESTs were then clustered based on their 5' ends to reduce redundancy, which resulted in a set of 12,198 clusters. The clone extending most 5' in each cluster

has been selected and rearranged. We have sequenced the 5' and 3' ends of these clones to verify their identity and to further collapse the set on the basis of their 3' identities. We also determined the length of the cDNA insert in each clone so that optimal sequencing strategies can be applied to specific size ranges. Finally, we have performed pilot experiments for full-length sequencing utilizing transposon-based methods that resulted in the completion of 283 high quality cDNAs.

99. The RIKEN Mouse Full-Length cDNA Encyclopedia

Piero Carninci, Kazuhiro Shibata, Masayoshi Itoh, Hideaki Konno, Jun Kawai, Yuko Shibata, Yuichi Sugahara, T. Endo, Y. Ozawa, Yoshifumi Fukunishi, Atsushi Yoshiki, M. Kisakabe, Masami Muramatsu, Yasushi Okazaki, and Yoshihide Hayashizaki
Genome Science Laboratory, RIKEN, Tsukuba Life Science Center, 3-1-1 Koyadai, Japan
carninci@rtc.riken.go.jp

We report the ongoing efforts to prepare the mouse full-length cDNA Encyclopedia. We have, to the date of the submission of this abstract, constructed more than 150 full-length cDNA libraries. Libraries were prepared with the CAP trapper full-length cDNA selection method coupled to the trehalose-thermoactivated reverse transcriptase, in order to clone long full-length cDNAs. Additionally, in most libraries, cDNAs were normalized and subtracted to isolate rarely expressed full-length cDNAs. To date, we have clustered 100,907 3' end sequences containing at least half of all mouse full-length cDNAs in 690,130 successful sequencing reactions from 5'-end validated libraries. We have also established technology and mRNA resources to collect the majority of remaining mouse expressed sequences.

We will discuss about library preparation and tissue selection strategy, quality of cDNA libraries in terms of complexity and full-length cDNAs presence and coverage of mouse genes by our clones.

100. Tissue Gene Expression Profiling Using RIKEN Full-Length Mouse 20K cDNA Microarray

Yasushi Okazaki^{1,2}, Rika Miki^{1,2,3}, Yosuke Mizuno^{1,2,3}, Yasuhiro Tomaru¹, Kouji Kadota^{1,2,4}, Piero Carninci¹, Kazuhiro Shibata^{1,2}, Masayoshi Itoh^{1,2}, Yasuhiro Ozawa¹, Jun Kawai^{1,2}, Hideaki Konno^{1,2}, Yoshifumi Fukunishi^{1,2}, Toshinori Kusumi¹, Hitoshi Goto^{1,5}, Hiroyuki Nitanda^{1,5}, Yohei Hamaguchi^{1,6}, Itaru Nishiduka^{1,6}, Masami Muramatsu^{1,2}, Atsushi Yoshiki⁷, Moriaki Kusakabe⁷, Joseph Derisi⁸, Vishy Iyer⁹, Michael Eisen⁹, Patric O. Brown⁹, and Yoshihide Hayashizaki^{1,2,3}

¹Laboratory for Genome Exploration Research Project, Genomic Sciences Center (GSC) and Genome Science Laboratory, Tsukuba Life Science Center, The Institute of Physical and Chemical Research (RIKEN), Koyadai, Japan; ²CREST, Japan Science and Technology Corporation (JST); ³Tsukuba University; ⁴University of Tokyo; ⁵Tohoku University, Sendai, Japan; ⁶Yokohama City University; ⁷Experimental Animal Research Division, Tsukuba Life Science Center, The Institute of Physical and Chemical Research (RIKEN), Koyadai, Japan; ⁸University of California; and ⁹Stanford University, Stanford, CA
yoshihide@rtc.riken.go.jp

The target of the Genome Science Laboratory of RIKEN is to clone and sequence the largest number possible of full-length mouse cDNAs and then to sequence these cDNAs in two phases. The first phase is to classify the cDNAs and the second is to complete full-length sequencing and functional annotations. We have developed two original methods to construct full-length cDNAs efficiently: "cap-trapper" which preferentially recognizes the Cap site of mRNA and the "trehalose-thermoactivated reverse transcriptase (RT)" which allows the RT reaction at higher (60 °C) temperature. We have constructed over 80 libraries from embryonic tissues of different developmental stages and adult tissues in order to ensure the greatest possible coverage of the expressed mRNA.

More than 200,000 successful sequencing passes have been performed with the use of two in house developed tools; a high-throughput plasmid preparation system and the RISA 384 capillary sequencer. Most of the sequences were performed from 3' end in order to select individual cDNAs. We have selected more than 65,000 different cDNAs.

Using these sets of RIKEN full-length cDNA, we have established Gene Expression Microarrays containing 20 K set of RIKEN full-length cDNA unique mouse genes (<http://genome.rtc.riken.go.jp>). These set have been used to profile expression patterns of various adult and embryonic tissues. Target DNAs were PCR amplified and printed on the Poly-L-lysine coated slide glasses. Target DNAs were blocked by excess amount of Cot1DNA. Probes were labeled by two-color fluorescent dye using random primer and reverse transcriptase. Normalization has been achieved using a global normalization method. We have also developed a program to filter the noise. The experiment was done twice and the reproducible results were extracted and clustered. We will present a large set of database, which show the spatial and temporal expression patterns of mice. These mouse full-length 20 K cDNA microarrays are widely applicable to analyze the global expression profiling of normal and diseased status of the mice.

101. The Molecular Genetics of DNA Repair in *Drosophila*

K.C. Burtis, R.S. Hawley, C. Boulton, K. Hollis, A. Laurencon, and D. Milliken
Section of Molecular and Cellular Biology,
University of California at Davis, Davis, CA 95616
shawley@netcom.com

Screening for new repair-deficient mutants:

To date we have screened approximately 12,100 of 12,500 available lines of EMS induced mutations on the 2nd and 3rd chromosomes from the Zuker collection. Thus far we have both identified and confirmed by retest approximately 60 lines that display significant sensitivity to one or more mutagens. We are currently assigning the newly-isolated mutants to complementation groups as well

as testing them for allelism with existing 2nd and 3rd chromosome *mus* mutations.

Characterization of existing mutagen-sensitive mutations:

We are in the process of carefully mapping the existing collection of mutagen-sensitive mutations. In several cases we have refined the map positions down to small genetic intervals and are testing P element insertions in the region for their ability to complement these mutations. In most cases however we are still in the process of positioning the mutant to within a numbered unit on the polytene chromosome. We have also continued our molecular and genetic characterization of two *Drosophila* ATM homologs. For one of these genes, mei-41, we have completed a synthetic lethal screen and begun a genetic fine structure analysis of existing mutant alleles.

Microarrays:

Glass slide microarrays have been produced that include over 10,000 *Drosophila* cDNAs. The cDNAs are derived from the Berkeley *Drosophila* Genome Project Unigene set, as well as from a set of testes-specific cDNAs generated by Dr. Brian Oliver at the NIDDK. We will report the results of our initial array experiments examining changes in gene expression resulting from exposure of *Drosophila* to various doses of ionizing radiation.

Genomics:

A comprehensive summary of *Drosophila* homologs of approximately 90 known DNA repair genes will be presented. This summary is based on analysis of the complete sequence of the *Drosophila* genome developed by Celera Genomics in collaboration with the Berkeley *Drosophila* Genome Project. Also integrated into this analysis is our current data regarding the association of these sequences with extant *Drosophila* mutagen-sensitive mutations.

102. The Tennessee Mouse Genome Consortium

D. K. Johnson¹, D. R. Miller¹, J. Snoddy¹, B. A. Berven¹, and E. M. Rinchik^{1,2}

¹Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8077 and ²Department of Biochemistry, Cellular, and

Molecular Biology, The University of Tennessee,
Knoxville, TN 37996
k29@ornl.gov

In order to maximize our capabilities for screening mice for a wide variety of mutant phenotypes, we have joined with institutions across the state of Tennessee to form the Tennessee Mouse Genome Consortium (TMGC). Our goal is to combine and exploit the clinical and academic expertise resident at Oak Ridge National Laboratory, the University of Tennessee, Vanderbilt University, the University of Memphis, St. Jude's Children's Hospital, and Meharry Medical College to induce and analyze genetic mutations that alter development, behavior, biochemistry, and morphology in mice. Each institution has confirmed its commitment to the goals of the Consortium by signing a Memorandum of Cooperation, by agreeing to a set of scientific, administrative, and veterinary principles governing institutional interactions, and by providing start-up funding for investigators to develop analytical methods and tools that can contribute to TMGC research projects.

In addition to the broad-based screening supported by state-wide expertise in many fields, the factors that distinguish the TMGC are ORNL's unique history in mouse genetics/mutagenesis and in the design of genetic screens, as well as ORNL's strength in bioinformatics and computational biology. Our genetics strategy is designed to produce multiple mice that may express new recessive mutations in a visually-identifiable "test class", which permits multi-site screening, screening for innately variable phenotypes, and screening in an aged, test-class colony.

Pilot screens for mutations in the central nervous system have currently identified fifteen potential mutants in about 450 pedigrees screened from mutagenesis experiments targeting two regions of mouse chromosome 7 (see abstract by Rinchik, et al.). The infrastructure provided by the TMGC provides a basis for the pooling of our expertise in joint proposals to federal and non-federal sponsors for long-range support for this unified, large-scale

effort to develop mouse models as community resources for human genetics research.

103. Designing Genetic Reagents to Facilitate the Mutagenesis and Functional Analysis of the Mouse Genome

Edward J. Michaud^{1,2}, Qing G. von Arnim^{1,2},
Carmen M. Foster^{1,4}, Yun You^{1,2}, Dabney K.
Johnson^{1,2}, and Eugene M. Rinchik^{1,2,3}

¹Mammalian Genetics and Development Section, Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8077;

²University of Tennessee - Oak Ridge National Laboratory Graduate Program in Genome Science and Technology; and ³Department of Biochemistry, Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996-0845, and

⁴Department of Pathology, College of Veterinary Medicine, University of Tennessee, Knoxville, TN 37901-1071

michaudej@bio.ornl.gov

Analysis of the molecular, cellular, and organismal consequences of induced and spontaneous mutations in mouse genes provides insight into the roles that genes play in human biological systems and disease. The complete DNA sequences of the human and mouse genomes will soon be available, and strategies are being developed to annotate the physical maps with gene-function maps. For many years, ORNL has used a phenotype-driven chromosome-region mutagenesis strategy in the mouse to map gene function in pre-selected segments of the genome. Currently, we are applying this strategy to approximately 8% of the mouse genome (see abstract by Rinchik *et al.*). In collaboration with the Joint Genome Institute (JGI), we are also conducting molecular, genomic, transcriptional, and DNA-sequence analyses of our mutagenized regions in order to integrate the genetic mutation maps with the transcript maps (see abstract by Johnson *et al.*).

This project forms the third component of the ORNL mutagenesis program; designing genetic reagents to facilitate regional-mutagenesis and functional-genomics analyses in additional portions of the genome. Our regional-mutagenesis strategy is based on having visibly marked (altered coat color, for example) chromosomal deletions or inversions in order to perform the mutagenesis and gene-function mapping in the most cost-effective, high-throughput, user-friendly, and error-free manner. However, the genetic reagents that facilitate these regional-mutagenesis screens are currently available for a limited portion of the mouse genome. We are employing embryonic stem-cell strategies to design marked chromosomal alterations in large, gene-rich regions of mouse chromosomes that are in synteny conservation with portions of the human genome being mapped and sequenced by the JGI. These reagents will facilitate additional mutagenesis screens in the mouse for the purpose of annotating human DNA sequence information with the whole-organism biological functions of genes. Our initial focus is on the proximal 23 cM of mouse Chromosome 7 (human 19q homology) and a 16 cM region of proximal mouse Chromosome 11 (human 5q homology). Efforts are also under way to complement the chromosome-region mutagenesis program by developing an integrated, systems-biological approach to analyzing complex multigenic traits in mice (see abstracts by Doktycz *et al.*, and Snoddy *et al.*).

104. Mouse Genetics and Mutagenesis for Functional Genomics: Phenotype-Driven Regional Mutagenesis and Genomics at the Oak Ridge National Laboratory

E. M. Rinchik^{1,2}, D. A. Carpenter¹, E. J. Michaud¹, Y. You¹, P. R. Hunsicker¹, L. B. Russell¹, D. R. Miller¹, M. L. Klegig², and D. K. Johnson¹

¹Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8077 and ²Department of Biochemistry, Cellular, and Molecular Biology, University of Tennessee, Knoxville, TN 37996
rinchikem@ornl.gov

A major goal of the mouse-genetics program at ORNL is to apply our experience in chemical germ-cell mutagenesis, mutation recovery and propagation, and broad-based phenotype screening, for creating a large, user-friendly mouse-mutation resource that can be used by the wider biological community for functional annotation of human DNA sequence. Our current overall program expands previous work that molecularly characterized regions of mouse Chromosome (Chr) 7 while also recovering N-ethyl-N-nitrosourea (ENU)-induced, recessive single-gene mutations. For example, in one screen of a ~5-cM Chr-7 region (human 11p and 15q homologies), simple phenotype-screening criteria had ascertained 19 new mutations in 1218 gametes, and, recently, broadly based phenotype-screening has yielded seven additional heritable mutations (including two subtle behavioral ones) in ~450 additional gametes, with another 13 subtle variants undergoing heritability testing. All mutations are being placed, by a simple set of genetic complementation crosses with overlapping deletions, into the rich DNA-sequence and expression map evolving for this region.

Mutations within two additional regions [mid-Chr 7 (human 15q homology), and mid-to-distal Chr 15 (human 8q, 22q, and 12q homologies)] are being recovered using dominantly and recessively marked inversion chromosomes in three-generation screens, which allows easy detection and low-cost maintenance of chromosomally “pre-mapped” deleterious recessive mutations without any molecular genotyping. In parallel, deletions are being developed in embryonic stem cells for use as finer-mapping and gene-identification reagents. Our experimental design also provides for the generation of multiple mutant test-class mice of a singular genotype for comprehensive multi-site phenotype screening (e.g., across the Tennessee Mouse Genome Consortium) and for establishment of aging colonies to be screened for later-onset recessive phenotypes. It also provides a facile means for placing any mutation on a number of inbred genetic backgrounds to analyze modifier effects in genetic-network analyses. We estimate that approximately 8-10% of the genome will be covered by our screens in the near term, with even wider coverage possible as additional genetic reagents are created.

105. Defining Complex Genetic Pathways with Gene-Expression Microarrays

M. J. Doktycz¹, B. H. Jones², C. T. Cuiat², P. R. Hoyt¹, B. W. Harker¹, R. E. Barry⁴, D. D. Schmoyer³, S. Petrov³, E. M. Rinchik^{2,5}, K. L. Beattie¹, J. R. Snoddy³, and E. J. Michaud²

¹Biochemistry and Biophysics Section, ²Mammalian Genetics and Development Section, and ³Computational Biosciences Section, Life Sciences Division, and ⁴Robotics and Process Systems Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831 and ⁵Department of Biochemistry, Cellular, and Molecular Biology, University of Tennessee, Knoxville, TN 37996
okz@ornl.gov

A primary goal of functional genomics is to understand the molecular mechanisms underlying complex interactions among genetically controlled biochemical pathways and the effects of environmental exposures and aging. The complete DNA sequences of the human and mouse genomes will soon be available, including the sequences of the estimated 100,000 genes present in each of these mammals. Even now there are over 893,000 mouse expressed sequence tags (ESTs) present in databases. The availability of these EST reagents, combined with recent advances in analytical technologies and bioinformatics tools are making a dramatic impact on our comprehension of complex genetic pathways. We are exploiting these EST reagents for determining the components of genetic pathways in a single organ system, the skin. Gene-expression profiles are being determined for anonymous skin ESTs, as well as ESTs from genes with known roles in skin development, differentiation, apoptosis, DNA repair, cancer, pigmentation, and skin and hair morphology. Gene expression is being examined during normal growth and differentiation processes, and compared to expression patterns elicited in response to genetic mutations or environmental exposures. To this end, we are combining three areas of expertise at ORNL (i.e., mouse molecular genetics, analytical technologies and instrumentation, and bioinformatics)

to develop an integrated-systems approach for defining gene function in genetic networks. Custom instruments, combining reagent-jets with precision movement stages, have been developed for the high throughput production of high-density microarrays. Automated procedures have been developed using commercial liquid handling systems for the preparation of tissue-specific cDNA probes, and for the parallel processing of 96 cell or tissue samples into fluorescently-labeled cDNA targets for hybridization to microarrays. Integration of these various instruments, tissue samples, cDNA clones, microarrays, and expression data will be accomplished with the aid of several inter-operating bioinformatics tools. Three bioinformation-system modules are being developed: (1) to track mice, tissues, and molecular samples; (2) to analyze the results of gene-expression arrays; and (3) to perform biologically meaningful reduction of data (e.g., by cluster analysis) and linking of the expression results to other databases containing structural and functional information (see abstract by Snoddy *et al.*). An important goal of this project is to make these data available to the scientific community through the web. These efforts complement the chromosome-region mutagenesis program at ORNL (see abstracts by Johnson *et al.*, Michaud *et al.*, and Rinchik *et al.*) by developing an integrated, systems-biological approach to analyzing complex multigenic traits in mice.

106. Genome-Wide Expression Analysis Prove that Distinct Sets of Genes Participate in Cardiac Hypertrophy and the Regression of Hypertrophy

Carl Friddle, James Bristow, Teiichiro Koga, and Edward M. Rubin
Department of Genome Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
EMRubin@lbl.gov

Cardiac hypertrophy is a significant risk factor for cardiac failure, affecting 15% of the adult population and 50% of those with hypertension. Single gene disorders account for a small fraction of these cases. Prior studies have identified a limited set of genes that play important roles in the onset progression and regression of cardiac hypertrophy. These studies have primarily focussed on genes known to function in the heart. In the present study, using pharmacological models of hypertrophy in mice, expression profiling was performed with fragments of more than 3,000 genes to characterize and contrast expression changes during induction and regression of hypertrophy. Administration of angiotensin II and isoproterenol by osmotic minipump produced increases in heart weight (15% and 40% respectively) that returned to pre-induction size following drug withdrawal. From multiple expression analyses of left ventricular RNA isolated at daily time-points during cardiac hypertrophy and regression, we identified sets of genes whose expression was altered at specific stages of this process. While confirming the participation of 25 genes and pathways known to be altered by hypertrophy, a larger set of 30 genes was identified whose expression had not previously been associated with cardiac hypertrophy or regression. Of the 55 genes that showed reproducible changes during the time course of induction and regression, 32 genes were altered only during induction and 8 were altered only during regression. This study identified both known and novel genes whose expression is affected at different stages of cardiac hypertrophy and regression and demonstrates that cardiac remodeling during regression utilizes a set of genes that are distinct from those used during induction of hypertrophy.

107. Ribozyme Gene Vector Libraries Identify Putative Tumor Suppressor Genes

Qi-Xiang Li, Eric Marcusson, Joan Robbins, Mark Leavitt, Flossie Wong-Staal, and Jack R. Barber
Immusol, Inc. San Diego, CA 92121 and University of California, San Diego, CA
barber@immusol.com

We have developed a method for gene identification, based on analysis of cellular function, that allows the

specific, directed identification and cloning of many genes. We have created viral vectors containing a highly complex library of Rz genes that can be stably and efficiently introduced into mammalian cells. By utilizing methodologies that enable selection of cells that have undergone a phenotypic change as a result of a specific Rz, we can isolate Rzs that inactivate genes associated with that phenotype. Once specific Rzs are selected and verified, their target recognition binding sequences can be used as tags to identify and clone the corresponding target genes.

We have used this approach to identify three novel tumor suppressor genes. We will present evidence that validates the role of these genes in the process of malignant transformation. Furthermore, we have used RNA expression profiling to begin the functional dissection of the pathways that these genes are involved in.

108. New Vectors for TAR Cloning and Retrofitting of Mammalian Genes

Maxim Y. Koriabine, Gregory G. Solomon, Lois A. Annab, J. Carl Barrett, and Vladimir L. Larionov
Laboratory of Molecular Genetics and Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709
koriabi1@niehs.nih.gov

The recent development of TAR (Transformation-Associated Recombination) cloning strategy for the selective isolation of specific regions and genes from complex genomes greatly advanced YAC cloning technology. Over the last two years the new technique was successfully applied for isolation of different genes and specific regions from human and mouse genomes. In this study we describe construction of a second generation of TAR cloning vectors, pVC604 (HIS3-CEN6-pBR), pVC604-A (HIS3-CEN6-pBR-Alu) and pVC604-B (HIS3-CEN6-pBR-B1), for gene isolation from human and mouse genomes. New vectors greatly simplify replacement of targeting sequences and subsequent physical analysis of the cloned material. In order to help to mobilize the DNA inserts in YACs for a variety of studies, we also have designed a set of vectors that retrofit YACs with different mammalian selectable markers and permit

their transferring into *E. coli* cells as circular YAC/BACs. The following vectors were constructed: BRV1-N [BAC-URA3-Neomycin phosphotransferase (Neo)], BRV2-H [BAC-URA3-Hygromycin phosphotransferase (Hyg)], BRV3-B [BAC-URA3-Blasticidin S deaminase (BSD)], BRV4-G [BAC-URA3-xanthine-guanine phosphoribosyl transferase (gpt)] and BRV5-C [BAC-URA3-Cytidine deaminase (codA)]. In this study we have shown that using these vectors YACs up to ~700 kb can be efficiently converted to YAC/BACs with mammalian selectable markers by *in vivo* recombination in yeast. We also show evidence that circular YAC/BACs of up to 300 kb can be subsequently transferred into *E. coli* cells by electroporation for further DNA isolation. The YAC retrofitting method is simple, and opens possibility to use the YACs generated by TAR cloning for structural and functional studies.

109. Defining the Minimal Length of Sequence Homology Required for Selective Gene Isolation by TAR Cloning

Vladimir Noskov, Maxim Koriabine, Greg Solomon, Natalay Kouprina, J. Carl Barrett, Lisa Stubbs¹, and Vladimir Larionov

Laboratory of Molecular Genetics and Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709 and ¹Human Genome Center, Lawrence Livermore National Laboratory, Livermore, CA 95616
noskov@niehs.nih.gov

Using the recently developed TAR cloning technique, it is possible to directly isolate specific chromosomal regions and genes from complex genomes as linear or circular YACs. Over the last two years the new technique has been successfully applied for isolation of different genes and specific regions of human and mouse genomes. In this study we investigated the minimal length of sequence homology required for gene isolation by TAR cloning using the Tg.AC

transgene as a model. The Tg.AC transgene unit consists of a zeta-globin promoter fused to the v-Ha-ras structural gene with a terminal simian virus 40 (SV40) polyadenylation signal sequence. We constructed a set of radial TAR cloning vectors containing the B1 repeat and different size SV40-specific hooks (from 800 bp to 20 bp). With a vector containing a 800 bp hook, cloning of Tg.AC transgene sequences from mouse genome was highly specific: one among fifty yeast transformants obtained contained a YAC with Tg.AC transgene. The same yield of positive clones was observed when length of homology was reduced to 60 bp. Therefore the minimal length of a unique sequence required for gene isolation is only 2 times larger than the minimal size of homology required for spontaneous mitotic recombination in yeast. This observation greatly facilitates selection of hooks for isolation of specific regions as well as construction of TAR vectors because the hooks can be synthesized as oligonucleotides instead of being isolated as genomic fragments.

110. Contamination of BAC Clones by *E. coli* IS186 Insertion Elements

Owatha L. Tatum, Andrew W. Womack, Mark O. Mundt, and Norman A. Doggett
Bioscience Division and DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM 87545
doggett@lanl.gov

The *E. coli* insertion element IS186 is a 1343 bp transposable element which is present at three to four copies in the *E. coli* genome. The transposon is flanked by a 23 bp inverted repeat and has been shown to insert preferentially into GC-rich targets. We have discovered 8 BAC and P1 clones from several widely used human genomic libraries which have a single copy of this insertion element included in the finished Genbank submission. These clones were sequenced at six different sequencing centers and in no case was the insertion element annotated as being derived from *E. coli*. The average G+C content

of a 100 bp window on either side of the insertion site in all clones is very high (75.8%) and appears to be within CpG islands. In two of the 8 cases the insertion site is flanked by G+C-rich SVA repeat elements (a retroviral LTR class of repeat). Earlier studies of IS186 insertions into plasmids have shown that target duplications of 8 to 12 bp occur at the insertion site. We looked for evidence of target duplication in a BAC clone containing this insertion by comparison with the finished sequence of a cosmid clone which overlapped the insertion site. This proved that the insertion of IS186 caused a 10 base pair duplication of human sequence surrounded the insertion site. Ten base pair duplications were surrounding the insertion site were found in the finished sequence of all other clones. In order to determine whether IS186 insertions occurred during library construction or propagation we performed PCR and sequencing experiments on several isolates of each clone. We found that RPCI-11 BAC clones sent directly from the Roswell Park resource did not contain insertion elements providing strong evidence that IS186 insertions are most likely occur during subsequent propagation of BACs. We estimate the frequency of this insertion in finished clones to be approximately 1 in 1000 but the actual frequency could be much higher if this element has been removed from some finished sequences prior to submission.

Supported by the US DOE.

111. Developing General Methods to Select Phage Antibodies Against Gene Products

Peter Pavlik¹, Robert Siegal¹, Daniele Sblattero², Vittorio Verzillo^{1,2}, Roberto Marzari³, Jianlong Lou², Jim Marks⁴, and Andrew Bradbury^{1,2}

¹Los Alamos National Laboratory, Los Alamos, NM 87545; ²SISSA, Trieste, Italy; ³University of Trieste, Trieste, Italy; and ⁴University of California San Francisco, San Francisco, CA
amb@lanl.gov

Phage display offers the possibility of selecting polypeptides (and the genes which encode them) from libraries of 1e10 or more different polypeptides on the basis of their abilities to bind target proteins and

subdomains. This diversity far surpasses the estimated number of total genes in the human genome. The application of this technology to the Human Genome Project will powerfully accomplish a central goal: the derivation of ligands that recognize protein products of all human genes, such ligands being either antibodies, or protein fragments.

Where the recognition ligands derived from this relatively new technology are antibody binding regions (single chain Fv) they can be employed in the same way as traditional antibodies. As such, they can play essential roles in assigning gene function, including the characterization of spatiotemporal patterns of protein expression and the elucidation of protein-protein interactions. Where the recognition ligands are protein fragments, they can be considered to be potential protein-interaction partners for the immobilized polypeptide and so a starting point for further biochemical studies.

This project has concentrated on trying to find a general way to isolate antibodies against gene products, preferably starting from gene sequence and using peptides to avoid the need for cloning and expression, although high throughput methods to select against recombinant products have also been developed.

Selection of antibodies against recombinant proteins has been reduced to the microtitre format. A comparison of the antibodies selected using this protocol with the standard selection procedure shows that the antibodies selected are on the whole different, although there is some overlap. Should gene products be available this is a very efficient way to select antibodies in a high throughput format.

In addition, selection on peptide surrogates of gene products has also been attempted. 192 scanning peptides corresponding to overlapping parts of four different proteins have been synthesised on microtitre pins and used to select phage antibodies. Some of the selected antibodies are able to recognise the full length protein. An analysis of the peptides which select antibodies recognising the full length protein has allowed us to develop an algorithm to predict which peptides are more likely to select useful proteins.

112. Search and Identification of Proteins that Bind Specifically to the Satellite DNAs

Ivan B. Lobov and Olga I. Podgornaya
Institute of Cytology RAS, 4 Tikhoretskii Ave.,
194064 St. Petersburg, Russia
ivan_lobov@hotmail.com

In the nucleus, chromosomes and individual chromosome domains are arranged by a non-random fashion. This organization is cell type-specific and undergoes rearrangements under the conditions that alter gene expression. Tandemly organized transcriptionally silent non-coding sequences, satellite DNAs (satDNAs), are localized in gene-poor heterochromatic chromosome regions. In interphase nucleus, these regions have a tendency to fuse, forming large chromocenters in a cell type-specific manner. Euchromatic regions can also associate with chromocenters that causes gene silencing as a result of transcription repression effect of heterochromatin.

Heterochromatin properties are mediated by proteins specifically associated with satDNAs that, however, are poorly characterized. To furthering our understanding of the role that satDNAs play in genome, we undertook search for proteins that bind specifically satDNAs of mouse and human. Chromosomal DNA anchored to the nuclear matrix (NM) or scaffold at the specific sites called M/SARs (for Matrix or Scaffold Attachment Regions) and via large blocks of satDNAs. We used electrophoretic mobility shift assay to reveal NM DNA-binding protein specific for the mouse major satDNA. We have developed a reliable approach for mild non-denaturing extraction of NM proteins that are generally insoluble under physiological conditions. The main DNA-binding protein revealed in these experiments was identified as a mouse homologue of SAF-A, an M/SAR-binding protein. We have also found that in interphase nuclei SAF-A predominantly decorates and covers heterochromatic areas.

Using Southwestern assay we have also identified four abundant DNA-binding proteins (p150, p120,

p83 and p66) in nuclei and NM preparations. These proteins bound specifically to mouse major satDNA and fragment of alphoid satDNA from locus alpha21-II of human chromosome 21. p120 and p66 were identified as SAF-A and lamin B correspondingly. p150 and p83 are apparently identical with SAF-B and ARBP, well-characterized M/SAR-binding proteins. Using an electrophoretic assay and computer modeling of DNA structure we have found that proteins prefer intrinsically bent DNA fragments over the straight ones. Thus, despite the lack of sequence homology, different satDNAs share structural features that might serve as a recognition signal for DNA-binding proteins of the NM.

Our data raise the possibility that different M/SAR-binding proteins can bind specifically to certain subsets of satDNAs of different species. The ability of NM proteins to recognize both M/SARs and satDNAs might serve as a general mechanism of gene association with heterochromatin.

113. Diversity in the Proteome: Homologous DNA Replicase Genes Use Alternatives of Transcriptional Slippage or Translational Frameshifting for Gene Expression

Norma M. Wills, Bente Larsen, Chad Nelson, John F. Atkins, and Raymond F. Gesteland
Department of Human Genetics, University of Utah,
15 N. 2030 East Room 7410, Salt Lake City, UT
84112-5330
nwills@genetics.utah.edu

A newly discovered contributor to the complexity of the proteome stems from generation of multiple RNAs from a single gene by transcriptional slippage. In the *Thermus thermophilus* dnaX gene, transcriptional slippage on a run of nine T residues results in a mixture of mRNAs differing in the number of A residues. Standard translation of a subpopulation of mRNAs yields the full-length tau protein while another subpopulation produces the

shortened gamma protein. Transcriptional slippage was implicated by determining the masses of PCR products spanning the run of A residues using mass spectrometry. With genomic DNA as the PCR template, the predominant signals correspond to molecules containing nine A/Ts. The pattern is strikingly different using reverse-transcribed mRNA as template. There are multiple signals corresponding to molecules containing 8-18 A/Ts showing heterogeneity in the mRNA population transcribed from the single *dnaX* gene.

This method of *dnaX* gene expression in *Thermus thermophilus* differs markedly from *dnaX* expression in *E. coli* where two analogous proteins are produced from a single *dnaX* gene by ribosomal frameshifting. Standard translation of the homogeneous mRNA population produces the full-length tau protein. Approximately 50% of the time, ribosomes shift to the -1 reading frame at a specific sequence, A AAA AAG, stimulated by signals in the mRNA and produce the shortened gamma protein. It is surprising that two rather similar *dnaX* sequences lead to very different modes of expression in the two organisms. The global importance of these and other alternative mechanisms of gene expression will be revealed by proteome analysis now underway.

114. The Transcriptional Program of Gametogenesis in Budding Yeast

Ira Herskowitz
Department of Biochemistry and Biophysics,
University of California, San Francisco, San
Francisco, CA 94143-0448
ira@cgl.ucsf.edu

Gametogenesis in yeast is the process whereby diploid cells of the a/alpha cell type undergo meiosis and form spores. Sporulation is initiated only when two conditions are met: cells are of the appropriate cell type (a/alpha), and cells receive the appropriate environmental stimulus (nutritional starvation). Under these conditions, a developmental program is initiated in which the following events occur: chromosomes are duplicated; then the chromosomes align and recombine with each other. After successful alignment and recombination, the duplicated sister chromatids are separated from each other (the first

meiotic division). Next, the sister chromatids are separated from each other (the second meiotic division). Finally, the separated sets of chromosomes (a haploid set) are wrapped up in spores. The end result of sporulation is production of four haploid spores encased in a single sac.

Microbial Genome Program

115. The Comprehensive Microbial Resource

Owen White, Jeremy Peterson, Jonathan A. Eisen, and Steven L. Salzberg
The Institute for Genomic Research, Rockville, MD 20850
owwhite@tigr.org

One of the challenges presented by large-scale genome sequencing efforts is the effective display of information in a format that is accessible to the laboratory scientist. Conventional databases offer the scientist the means to search for a particular gene, sequence, or organism, but do little in the way of displaying the vast amounts of curated information that are becoming available. TIGR has developed methods to effectively “slice” the vast amounts of data in the sequencing databases in a wide variety of ways, allowing the user to formulate queries that search for specific genes as well as to investigate broader topics, such as genes that might serve as vaccine and drug targets.

The Comprehensive Microbial Resource (CMR) is a facility for annotation of TIGR genome sequencing projects, and a web presentation of all of the fully sequenced microbial genomes, the curation from the original sequencing centers, and further curation from TIGR (for those genomes sequenced outside TIGR). The web presentation of the CMR includes the comprehensive collection of bacterial genome sequences, curated information, and related informatics methodologies. The scientist can view genes within a genome and can also link across to related genes in other genomes. The effect is to be able to construct queries that include sequence

searches, isoelectric point, GC-content, GC-skew, functional role assignments, growth conditions, environment and other questions, and isolate the genes of interest. The database contains extensive curated data as well as pre-run homology searches to facilitate data mining. The interface allows the display of the results in numerous formats that will help the user ask more accurate questions. This resource should be of value to the scientific community to design experiments and spur further research. Resources of this type are an essential tool to make sense of bacterial genome information as the number of completed genomes continues to grow.

116. The *Pseudomonas putida* KT2440 Genome Sequencing Project

Karen E. Nelson, Hoda Khouri, Erik Holtzapfle, Jeff Buchoff, Michael Rizzo, Azita Moazzez, Kelly Moffat, Kevin Tran, Hean Koo, P. Chris Lee, Daniel Kosack, Bradley Slaven, Helmut Hilbert, Burkhard Tuemmler, and C.M. Fraser
The Institute for Genomic Research, Rockville, MD 20850
kenelson@tigr.org

Pseudomonas putida is a ubiquitous soil bacterium that has significant potential for bioremediation of numerous compounds. To determine the complete genome sequence of strain KT2440, a genome sequencing project was initiated in January 1999 as a collaboration between The Institute for Genomic Research in Rockville MD, and a German Consortium (see <http://www.tigr.org/tdb/mdb/mdb.html>), with primary funding from the Department of Energy. The 6.1 Mbp genome is being

sequenced by the random shotgun method, and multiple sized large insert libraries (10 kb, 40 kb) are acting as a scaffold for the genome. At the end of random sequencing, there were a total of 392 sequencing gaps, many of which have been closed by editing off the ends of assemblies, or by sequencing clones that span the respective gaps. The high GC content of the genome is highlighted in long stretches of G's and C's encountered in both sequencing and physical gaps, and through which we have had problems sequencing by traditional methods. Sequencing of short reads that point into gaps, dye primer chemistry, transposon mutagenesis on selected spanning clones, as well as ET chemistry were used for most of the difficult areas. Multiplex PCR and micro-library construction, have also assisted in resolving and ordering the RNA operons. Grouper and autoprimers (TIGR softwares) were improved or modified to deal with the size of the genome. Ultimately, the *P. putida* genome sequence will identify the real potential of this organism in various biotechnological areas including the production of natural compounds, and remediation of polluted habitats.

117. The Genome of *Geobacter sulfurreducens*

B.A. Methe¹, L. Banerjee², W.C. Nierman², O. Snoeyenbos-West¹, S. Sciuffo¹, and D.E. Lovley¹

¹University of Massachusetts, Amherst, MA 01003 and ²The Institute for Genome Research, Rockville, MD 20850

bmethe@microbio.umass.edu

The complete genome sequence of *Geobacter sulfurreducens* is currently being determined to better understand its genetic potential. *G. sulfurreducens* is an important member of a family (*Geobacteraceae*) of delta Proteobacteria capable of oxidizing organic compounds including aromatic hydrocarbons to carbon dioxide with Fe(III) or other metals and metalloids including U(VI), Tc(VII), Co(III), Cr(IV), Au(III), Hg(II), As(V) and Se(VI) serving as the terminal electron acceptor. It is the dominant group of iron reducing microorganisms recovered from a wide variety of aquifer and subsurface environments when both molecular and traditional culturing techniques are used. *Geobacter* plays a critical role in

the biogeochemical cycling of carbon, iron and other metals. Its genetics and physiology are a subject of intense study in part due to the importance that these processes can play in the remediation of contaminated anaerobic subsurface environments. The determination of the *G. sulfurreducens* genome is being accomplished using a random shotgun cloning approach to provide at least six-fold coverage of a 1mb genome followed by closure of remaining physical or sequence gaps. TIGR Assembler software and other computer programs developed by The Institute for Genome Research are used to assemble the genome and aid in gap closing, finishing and annotation. Searches of sequences and contigs from the early random phase of sequencing using the BLAST algorithm and database have produced high scores with low expect values indicating significant homologies to proteins contained in the database. These include enzymes considered important to basic housekeeping functions such as tRNA synthases and amino acid synthesis as well as those essential to other metabolic processes known to occur in *G. sulfurreducens* including nitrogen fixation. A number of sequences have produced no significant alignments indicating the likelihood of genes encoding for novel functions. Of further significance has been the extension of N-terminal sequences previously obtained from cytochromes known to be important in dissimilatory iron reduction. Thus, the genome will provide information crucial to the further understanding of this important metabolic process.

118. The *Haloferax volcanii* Genome Project

Rajendra J. Redkar, Joe J. Shaw, Gary G. Bolus, Mary Lee Ferguson, Troy A. Horn, and Vito G. Delvecchio
Institute of Molecular Biology and Medicine,
University of Scranton, Corner of Monroe and Ridge
Row, Scranton, PA 18510
Vimbm@aol.com

Haloferax volcanii is a salt-loving archaeon belonging to the family *Halobacteriaceae*. Halophilic archaea exhibit obligate halophilism and require 2-5 M salt concentration for viability. At lower concentrations of salt (about 1 M), cells become distorted, leading to cell lysis and death. The

cytoplasm of halophiles contains very high internal concentrations of K⁺ and Na⁺, and is iso-osmotic with the environment. These bacteria have evolved metabolic and synthetic machinery that functions at high concentrations of salts, concentrations that are typically lethal for other organisms. Future research on halophilic archaea will provide better insight into early evolution of microorganisms and fundamental knowledge of biochemical and genetic events in organisms living in extreme environments.

H. volcanii cells are disk shaped and show involuted forms in the presence of NaCl. *H. volcanii* is a chemoorganotroph requiring complex nutrient medium and 1.5-2.5 M NaCl for growth. Cultures may be grown in the laboratory at 37°C with gentle shaking, however better growth is achieved at 42°C. In the laboratory, *H. volcanii* produces a characteristic pink pigment. Overall, the organism can be easily cultivated and maintained in the laboratory. Moreover, auxotrophic mutants are available, the cells are easily transformable, and genetic manipulation systems such as shuttle vectors and expression vectors are available for functional studies.

The genome size of the *H. volcanii* is estimated to be ~4.2 Mb. About 90% of the genome has 65% (G + C) content while remaining genome has 55% (G + C) content. The genome is composed of a chromosome (2,920 kb) and 4 plasmids, viz. pHV1 (86 kb), pHV2 (6.4 kb), pHV3 (442 kb) and pHV4 (690 kb). Plasmid pHV3 was selected for the first phase of sequencing operation. Sixteen overlapping cosmids were supplied by our collaborator Dr. Robert Charlebois, University of Ottawa, Ottawa, Canada, and used to make individual shotgun libraries. The average size of inserts in these randomly fragmented libraries is 2-3 kb. The sequencing reactions were performed on both ends of the clones using Big Dye Terminator chemistry to achieve 6-8X coverage. The data has been collated into several large assemblies using different software packages and the gaps in the assemblies are being located for closure. The annotation of data has started simultaneously to identify genes on pHV3. A progress report and the

future plans on *H. volcanii* sequencing project will be presented.

119. The *Caulobacter crescentus* Genome Sequencing Project

Tamara Feldblyum¹, William C. Nierman¹, Nikhil Phadke², Peter Ulintz², and Janine Maddox²
¹The Institute for Genomic Research, Rockville, MD and ²Department of Biology, University of Michigan
tamaraf@tigr.org

Caulobacter crescentus is a member of the alpha subclass of the proteobacteria which also include *Rickettsia*, *Rhizobium*, *Agrobacterium* and *Brucella* species. It is the most prevalent non-pathogenic bacterium in nutrient-poor fresh water streams and is also found in marine environments. It is one of the organisms responsible for sewage treatment. Caulobacters are being modified for use as a bioremediation agent for removing heavy metals from wastewater.

Caulobacter crescentus has been extensively studied because it exhibits a well-defined developmental pattern that is independent of environmental stress. The free-swimming morphologically distinct swarmer cell progresses to an anchored stalked cell, the only cell type capable of genome replication and cell division. Cell division of the stalked cell splits out a swarmer daughter cell.

C. crescentus has a genome size of 4 Mb, with G+C content of about 66.5%. Tremendous power for genome assembly was brought to this project through the use of a 2 and 10 kb insert size 2 plasmid library strategy. In sequencing this organism at TIGR, 65,588 random sequence reads from both ends of plasmid clones were used to assemble the genome into only five groups comprising essentially all of the genome sequence. A preliminary review of *C. crescentus* ORFs revealed by the sequence is provided.

120. Unusual Features of Radioresistant Bacterium *Deinococcus radiodurans* Genome Revealed by Comparative-Genomic Analysis

Kira S. Makarova^{1,2}, L. Aravind², Roman L. Tatusov², Eugene V. Koonin², and Michael J. Daly¹
¹Uniformed Services University of the Health Sciences, Bethesda, MD 20814-479 and ²The National Center for Biotechnology Information, The National Institutes of Health, Bethesda, MD 20814
makarova@ncbi.nlm.nih.gov

In-depth analysis of *Deinococcus radiodurans* genome reveals some unusual features, which may be relevant to its extreme radioresistance and desiccation resistance. Comparison of the *Deinococcus* gene products to the collection of Clusters of Orthologous Groups of proteins (COGs) allowed us to identify not only a set of genes which are shared with all or most bacteria, but some surprisingly missing genes, including those for several enzymes involved in repair and recombination. Using this information, we tentatively reconstructed the metabolic pathways, repair and recombination systems and stress response mechanisms of *D. radiodurans*. The comparative analysis helped in identifying phylogenetic affiliations of *Deinococcus* and sets of genes with unusual phylogenetic patterns, especially those that are shared with thermophilic archaea and bacteria, indicating a possible thermophilic ancestor for *Deinococcus*. We described several protein families that are specifically expanded in the genome of *D. radiodurans*, namely possible nuclease inhibitors, specific transcriptional regulators and desiccation-related proteins, which could contribute to the radioresistance and desiccation resistance of this bacterium. Some additional unique multidomain proteins, which could be involved in novel repair or stress-response-related mechanisms were detected. Investigation of short repeats in *Deinococcus* resulted in the identification of their mosaic nature and suggested that they could contribute to the recombinational proficiency of this organism.

121. Protein Expression in *Methanococcus jannaschii* and *Pyrococcus furiosus*

C.S. Giometti¹, S.L. Tollaksen¹, H. Lim², J. Yates², J. Holden³, A. Lal Menon³, G. Schut³, M.W.W. Adams³, C. Reich⁴, and G. Olsen⁴
¹Argonne National Laboratory, Argonne, IL; ²University of Washington, Seattle, WA; ³University of Georgia, Athens, GA; and ⁴University of Illinois, Urbana, IL
csgiometti@anl.gov

Complete genome sequences are now available for both *Methanococcus jannaschii* and *Pyrococcus furiosus*. The open reading frame (ORF) sequences from these completed genomes can be used to predict the proteins synthesized, but laboratory methods are needed to verify those predictions. Two-dimensional gel electrophoresis (2DE) coupled with mass spectrometry of peptides isolated from the gels is being used to determine the constitutive expression of proteins from these two *Archaea* and to explore the regulation of expression of non-constitutive proteins. The most abundant proteins (i.e., those easily detectable by staining with Coomassie Blue R250) from cells grown in minimal nutrient media have been isolated and analyzed. Using a combination of matrix-assisted laser desorption ionization (MALDI) and tandem mass spectrometry, 100 proteins expressed by *M. jannaschii* and 50 proteins expressed by *P. furiosus* have been related to specific ORFs in the respective genome sequences. The molecular weights and isoelectric points determined by the positions of proteins in the 2DE patterns are compared with the ORF-predicted molecular weights and isoelectric points for each microbe. Numerous instances of multiple proteins with different molecular weights or isoelectric points being associated with the same ORF have been observed. Possible reasons for such multiplicity include the incomplete unfolding of these highly stable proteins prior to electrophoresis, the non-dissociation of subunits, post-translational modifications such as phosphorylation (multiple proteins with the same identity but different isoelectric points) or peptide cleavage (multiple proteins with the same identity but different molecular weights). Preliminary experiments to change the protein expression of these organisms by altering growth conditions have revealed

significant quantitative changes in a small number of the proteins visible in 2DE patterns. Correlation of proteins expressed with specific ORFs is now focused on those proteins showing quantitative changes in expression and on less abundant proteins. The observed protein abundances and changes in abundance from these proteomic studies could be useful for validation of predictions of protein expression based on ORFs.

This work is supported under Contract No. W-31-109-ENG-38 with the U.S. Department of Energy.

122. Detection of Non-Cultured Bacterial Divisions in Environmental Samples using 16S rRNA-Based Fluorescent in situ Hybridization

Cheryl R. Kuske, Susan M. Barns, and Stephan Burde
Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
Kuske@lanl.gov

Microbial genome sequencing projects have focused primarily on species that can be easily cultured. However, readily cultured bacteria are only a small fraction of the total bacterial diversity present in the environment. Diverse bacteria representing novel divisions have been identified in many natural environments using 16S rDNA sequence analysis. Microbial processes in these environments are of critical importance to the biosphere and the non-cultured bacteria residing there are a valuable resource for novel genomic information. We have identified novel bacterial divisions from 16S ribosomal RNA gene libraries generated from DNA of a volcanic cinder field and an arid sandstone soil. Using RFLP and sequence analysis, we have analyzed 800 bacterial rDNA sequences obtained from the two arid environments. The majority of sequences were members of recently identified bacterial divisions that have no, or very few, cultivated members (Kuske et al. 1997. *AEM* 63:3614-3621, Hugenholtz et al. 1998 *J.Bact.*

180:366-376). Using PCR primers specific for two of these divisions, Acidobacterium and OP11, and their subgroups, we have detected both divisions in local hot/warm spring microbial mats and sediments. Analysis of cell abundance of members of these groups is under investigation using fluorescently labeled rRNA probes and fluorescence microscopy. We plan to collect bacterial cells directly from the environmental samples using flow cytometry and cell sorting. The pooled DNA of non-cultured bacteria will be a valuable resource of genetic material for comparative analyses of conserved and novel gene families, and for targeted genome sequencing.

123. Diversity of Metal Reducing Bacteria from Ecological, Physiological and Genomic Perspectives

Jizhong Zhou^{1,2}, Guangshan Li¹, Alison Murray², Yul Roh¹, Heshu Huang¹, Ray Stapleton¹, Qiaoyun Qiu², John Heidelberg³, Claire Fraser³, Douglas Lies⁴, Kenneth H. Nealson⁴, and James M. Tiedje²
¹Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831; ²Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824; ³The Institute for Genomic Research, Rockville, MD 20850; and ⁴Department of Geology and Planetary Sciences, Jet Propulsion Laboratory and California Institute of Technology, Pasadena, CA 91109
zhouj@ornl.gov

Microbial metal reduction plays an important role in biogeochemical cycling of carbon and nitrogen as well as in bioremediation of metals, radionuclides and organic contaminants. To further investigate their diversity, metal-reducing bacteria were isolated from a variety of extreme environments including the deep terrestrial subsurface, Siberia and Alaska permafrost soils, continental margin marine sediments and Hawaii deep-sea water. Thermophilic isolates from terrestrial subsurface formations that had been geologically and hydrologically isolated for about 200 million years were able to use glucose, pyruvate, lactate, acetate and hydrogen as electron donors, and

were able to reduce iron, manganese, chromium and uranium, as well as produce magnetite at 50-75 °C. The psychrotrophic isolates were able to use iron, manganese, and cobalt as electron acceptors, and were able to produce magnetite at 0 °C. A few isolates were able to reduce cobalt at - 4 °C and produce siderite using CO₂. Phylogenetic analyses indicated that the thermophilic iron-reducing bacteria were closely related to *Thermoanaerobacter ethanolicus* whereas the psychrotrophic iron-reducing bacteria were related to the members of the *Shewanella* genus. Although the psychrotrophic metal-reducing bacteria were able to use nitrate as electron acceptor, physiological studies and the comparisons of whole genome sequences from *Shewanella oneidensis* MR-1 (formerly *S. putrefaciens* MR-1) indicated that MR-1, and possibly the other psychrotrophic bacteria isolated appear to not be dissimilatory denitrifiers. In addition, whole genome sequence comparison indicated that MR-1 is more closely related to *Vibrio cholerae* O1 than to *Escherichia coli* K12. Finally, a partial microarray containing about 200 genes involved in energy metabolism and regulation were constructed and used to monitor gene expression patterns under anaerobic conditions. Substantial differences in gene expression patterns were observed under aerobic and anaerobic conditions. One interesting observation is that the genes (*mtrA*, *B*, and *C*) involved in metal reduction were highly expressed under both aerobic conditions and anaerobic metal reducing and denitrifying conditions, suggesting that the expression of these genes is not specific to metal reduction. The iron reduction rates in the deletion mutants of *mtrB* generated by newly developed suicide vectors were much slower than in the wild type strain. The partial microarrays were also used to assess the genome diversity among different metal-reducing bacteria. The results indicated that *S. oneidensis* DLM7 is more closely related to MR-1 than *Shewanella* sp. W3-6-1. Housing keeping genes and the genes involved in metal reduction appear to be highly conserved between MR-1 and W3-6-1 although the overall genomic diversity is low.

124. Pangenomic Microbial Comparisons by Subtractive Hybridization

Peter Agron, Lyndsay Radnedge, Evan Skowronski, Madison Macht, Jessica Wollard, Sylvia Chin, Aubree Hubbell, Marilyn Seymour, Christina Nocerino, and Gary Andersen
Biology and Biotechnology Research Program,
Lawrence Livermore National Laboratory,
Livermore, CA 94550
andersen2@llnl.gov

Sequencing of whole genomes is reshaping microbiology. However, as more sequence information is generated, there will be increased sequence redundancy between closely related species or strains. In the course of time, the amount of new sequence information obtained by whole genome sequencing with current technology will become increasingly less cost efficient. We are exploring the use of suppression subtractive hybridization (SSH) of total DNA as a means of focusing sequencing efforts on unique regions when a reference strain of known sequence is compared to a different isolate of the same species or genus. To rigorously examine this approach, two sequenced strains of *Helicobacter pylori* (J99 and 26695) were used as a model system, as this allows rapid determination and mapping of difference products based on sequencing alone. Using the high-throughput SSH methods, difference products can be rapidly cloned, sequenced, and then mapped by comparing the data to the *H. pylori* genome database. To increase the likelihood of amplifying difference products from any given region, several restriction enzymes were used in separate SSH experiments. So far we have obtained data from 2,123 clones that reveal 427 (20%) unique sequences. Control subtractions with an *Escherichia coli* strain containing the transposon Tn5 against its isogenic parent showed a 270-fold enrichment for Tn5 sequences, demonstrating that SSH is highly effective. Current efforts are focused on: 1) mapping of the difference products onto the relevant genome using the cross_match algorithm and Percent Identity Plots, 2) assessing coverage of the difference regions by the subtracted clones, 3) assessing the redundancy of this coverage and 4) determining the reproducibility of SSH. We will present data that

address the overall efficacy of the use of subtractive hybridization for pangenomic microbial comparisons.

125. *Prochlorococcus*: The Smallest and Most Abundant Photosynthetic Microbe in the Oceans

Sallie W. Chisholm, Gabrielle Rocap, and Lisa Moore

Departments of Civil and Environmental Engineering and Biology; Massachusetts Institute of Technology; 15 Vassar St. 48-425; Cambridge, MA 02139

chisholm@mit.edu

<http://web.mit.edu/chisholm/www/>

Prochlorococcus is a unicellular cyanobacterium that dominates the temperate and tropical oceans. It lacks phycobilisomes that are characteristic of cyanobacteria, and contains chlorophyll b as its major accessory pigment. This enables it to absorb blue light efficiently at the low-light intensities and blue wavelengths characteristic of the deep euphotic zone. It contributes 30-80% of the total photosynthesis in the oligotrophic oceans, and thus plays a significant role in the global carbon cycle and the Earth's climate. Description of the complete genome of this microbe will greatly advance our understanding of the regulation of these globally important processes.

To this end, colleagues at the DOE Joint Genome Institute (J. Lamerdin et al.) have been working on the complete genome sequence of *Prochlorococcus marinus* (MED4) (http://bbrp.llnl.gov/jgi/microbial/prochlorococcus_homepage.html). The work has progressed rapidly, and the sequence is almost complete. *Prochlorococcus* is an ideal candidate for complete genome sequencing because (1) it is the smallest known phototroph with a relatively small genome (1.8 Mb), (2) it is widespread and abundant and is easily identified and enumerated *in situ* using flow cytometry, (3) its unique photosynthetic pigment (divinyl chlorophyll) makes its contribution to total photosynthetic biomass in the oceans easily assessed,

and (4) we have an extensive culture collection of isolates from different oceans and environments.

Moreover, we have recently demonstrated that at least two ecotypes of *Prochlorococcus* coexist in the oceans that are distinguished by their photophysiology and molecular phylogeny. One is capable of growth at irradiances where the other is not. Ultimately, a comparison of the complete genomes of these two ecotypes would provide valuable insights into the regulation of this type of microdiversity in marine microbial systems. In addition, the use of microarray technology for the analysis of gene expression patterns will give us unprecedented insights into how these microbes cope with the dilute environment of the oligotrophic oceans.

126. Sequencing Microbial Genomes of Relevance to Global Climate Change

J. E. Lamerdin¹, K. Burkhardt-Schultz¹, A. Arellano¹, S. Stilwagen¹, A. Erler¹, A. Kobayashi¹, M. Shah⁴, D. J. Arp², A. B. Hooper³, S. W. Chisholm⁵, G. Rocap⁶, E. Branscomb⁷, and F. Larimer⁴

¹Joint Genome Institute, Lawrence Livermore National Laboratory, Livermore, CA., ²Botany and Plant Pathology Department, Oregon State University, Corvallis, OR, ³Department of Biochemistry, University of Minnesota, St. Paul, MN, ⁴Oak Ridge National Laboratory, Oak Ridge, TN, ⁵Departments of Civil and Environmental Engineering and Biology, Massachusetts Institute of Technology, Cambridge, MA, ⁶Massachusetts Institute of Technology/Woods Hole Oceanographic Institution Joint Program, Cambridge, MA, ⁷Joint Genome Institute, Production Sequencing Facility, Walnut Creek, CA
lamerdin1@llnl.gov

The Joint Genome Institute (JGI) has established a new program to sequence the genomes of microorganisms that may significantly impact global climate. This effort is focused initially on five

microorganisms: *Nitrosomonas europaea*, *Rhodospseudomonas palustris*, *Nostoc punctiforme*, and two marine cyanobacteria, *Prochlorococcus marinus* and *Synechococcus*. The common theme shared by these microbes is that all are autotrophic, fairly numerous within their respective ecosystems, and contribute materially to carbon cycling or biomass production (with the exception of *N. europaea*). By systematic analysis of each genome, we hope to identify specialized nutrient uptake systems, pathways that contribute to or regulate nitrogen fixation, carbon cycling and photosynthesis. With this knowledge, it may be possible to maximize the carbon recycling capabilities of these organisms.

We have completed the initial data generation phase for *N. europaea* and *P. marinus*, which yielded >95% of the genomic sequence for each microbe. (Progress towards completion can be monitored through our web site: <http://bbrp.llnl.gov/jgi/microbial/>). A similar level of coverage is anticipated for *R. palustris* by mid-March. Finishing is underway on the first two organisms, and we expect closure by Spring of 2000. The level of coverage achieved by the 'shotgun' phase is readily amenable to generating a rough inventory of the types of genes present in each organism. Preliminary analyses have been performed on *N. europaea* and *P. marinus* and the results are available on our web site. The resulting gene 'catalogues' provide the scientific user community access to the contents of unfinished sequence data in a consumable format, without the need for protracted data manipulations on their part. At this meeting, we will present some of the surprising findings that are emerging from the analyses of these two genomes.

This work was performed under the auspices of the U.S. DOE by LLNL under contract no. W-7405-ENG-48.

Ethical, Legal, and Social Issues

127. Electronic Scholarly Publishing: Foundations of Classical Genetics

Robert J. Robbins

Fred Hutchinson Cancer Research Center, Seattle, WA 98109

rrobbins@fhcrc.org

The HGP has significant ethical, legal, and social implications for all citizens. Public interest in the project is growing. In addition to its importance in the training of professional geneticists, the HGP is of special relevance for undergraduate training in basic biology, and even for high-school and other K-12 education. In a world soon to experience a flood of information and technology from genomic research, a basic understanding of genetic principles may become part of the expected knowledge base of the educated citizen.

Understanding HGP research, however, requires some familiarity with basic genetics. Access to materials that support an understanding of classical genetics can be difficult for those outside a university environment. We have created an informational and educational resource at which material related to the foundations of classical genetics is being republished in readily available, typeset-quality electronic form. We also publish additional material, such as pedagogical materials, items of general interest, biographical and autobiographical memoirs, and historical or analytical treatments.

Materials at our site are of interest to individual users, but they are especially valuable for teachers and other educators in the preparation of their course materials. Several textbook publishers are providing

links to our site at their value-adding textbook support sites. Many junior college and secondary school sites are also now referencing our site.

In the past year, we have emphasized software development to improve the efficiency with which we can publish works at our site and to improve the functionality of our site for users. We have also gained access to some critical copyrighted materials by establishing relationships with key publishers, such as the Genetics Society of America and the Cold Spring Harbor Laboratory Press.

128. *Genes, Environment, and Human Behavior: A Curriculum Module*

Mark V. Bloom¹, Rodger W. Bybee¹, Michael J. Dougherty², and Joseph D. McInerney³

¹Biological Sciences Curriculum Study (BSCS), Colorado Springs, CO 80918-3842; ²Hampden-Sydney College, Hampden-Sydney, VA 23943; and ³Foundation for Genetic Education and Counseling, Johns Hopkins University School of Medicine, Baltimore, MD
mbloom@bscs.org

Genes, Environment, and Human Behavior is a curriculum module designed for high school biology classes. It specifically addresses the roles of nature and nurture in human behavior. Developed by BSCS, a nonprofit curriculum development group, the module uses an inquiry-based approach that also provides well-structured and civil classroom analyses of ELSI-related issues. The curriculum is the fourth genome module produced by BSCS and is provided to teachers at no cost.

Students using the module should be familiar with Mendelian genetics, the chromosome theory of inheritance (including genetic linkage and recombination), the chemical nature of the gene (including the structure of DNA), and the central dogma, which states that genetic information resides in DNA, passes through an RNA intermediate, and is ultimately expressed as protein.

Designed for five periods of classroom instruction, the module consists of five student activities and includes extensive teacher background material. The student activities are organized into a conceptual whole that introduces students to the meaning of human behavior, to types of variation in populations, to the study of behavioral genetics, and then to the methodology for isolating genes that influence behavior. The first activity involves distinguishing between simple Mendelian traits and polygenic traits. The second activity builds on the first, and helps students understand multifactorial traits. The third activity introduces the use of twin studies in behavioral genetics, while the fourth activity demonstrates the use of association studies in the quest to locate genes influencing behavior. The concluding activity is a role-playing exercise that lets students use what they have learned to assess the wisdom of using genetic knowledge to formulate social policy.

129. High School Students as Partners in Sequencing the Human Genome

Kristi Sanford, Maureen Munn, and Leroy Hood
Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195-2145
ksanford@u.washington.edu

Through the generous support of the Department of Energy since March 1994, we have developed a unique program that encourages high school students to think constructively about the scientific and ethical issues of genomic research by enabling them to participate in both. High school biology teachers from around Washington State and throughout the USA attend a one-week summer institute at the University of Washington to learn about the process and applications of DNA sequencing. Throughout the

school year, local teachers are provided with the equipment and supplies necessary to carry out the sequencing experiment. Distant teachers are provided with DNA templates and primers, but they need to obtain the equipment and supplies independently.

Students are currently using manual or automated approaches to sequence the b2 subunit of the nicotinic acetylcholine receptor, as part of an authentic research project focused on understanding the genetic basis of nicotine addiction. Some of the activities, such as a DNA assembly and BLAST search, are accessed through the project web site. Students also learn to apply an ethical decision making process to resolve a complex topic based on presymptomatic genetic testing. Through discussions with scientist mentors who assist during classroom experiments, students learn about many career options in science. The program is used in general and advanced biology, as well as vocational biotechnology classrooms.

We have trained 74 teachers through one-week summer institutes and 160 teachers through one-day workshops. Since 1993, a total of 8400 students have participated in the DNA sequencing experiment as part of this program. We have found that teachers are much more likely to integrate this program into their classroom activities if they are eligible to borrow the HSHGP equipment kits. For this reason, we will channel our dissemination efforts through partnerships with several outreach programs throughout the nation.

This work is sponsored by the US Department of Energy under grant No. DE-FG03-98ER62547/AM01.

130. The Science and Issues of Human DNA Polymorphisms: An ELSI Training Program for High School Biology Teachers

David Micklos, Matt Christensen, and Scott Bronson
DNA Learning Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
micklos@cshl.org

We have implemented a nationwide training program to introduce high school biology teachers to the key

uses and societal implications of human DNA polymorphisms. The 2.5-day program targets an audience of highly professional faculty who have already implemented hands-on labs in molecular genetics, many of whom offer laboratory electives in biotechnology. The 10 workshops conducted to date have involved a total of 231 high school faculty at workshops held in eight states — 15% over the projected registration of 200 participants.

Program participants learn simplified lab techniques for amplifying two types of chromosomal polymorphisms: an Alu insertion and a VNTR. These polymorphisms illustrate the use of DNA variations in disease diagnosis, forensic biology, and identity testing — and provide a starting point for discussing the uses and potential abuses of genetic technology. Participants submit their Alu insertion data to the Student Allele Database at the DNALC's WWW site (vector.cshl.org/sad/). This elegant and easy-to-use interface allows students to test Hardy-Weinberg equilibrium, compare world populations, and test theories of human evolution.

We have also introduced a Sequencing Service to generate control region sequence from mt DNA samples submitted from biology classes around the country. Workshop participants amplified the mt control region from DNA prepared from their hair roots or cheek cells, and the amplicons were returned to Cold Spring Harbor Laboratory for cycle sequencing. The completed sequences were then posted at the DNA Sequence Server (vector.cshl.org/sequences/), a powerful database application that allows students to analyze their own mt DNA sequences — including similarity searches and multiple sequence alignments. We have replicated the process with student mt samples submitted by mail, and our database currently contains 850 teacher and student mt sequences, which can be reached at the DNA Sequence Server site.

131. Using the Power of Informal Learning to Address Science Literacy: A Report from the Microbial Literacy Collaborative

Cynthia A. Needham

American Society for Microbiology, Washington, D.C. and ICAN Productions, Ltd., Stowe, VT 05672
caneedham@aol.com

The Microbial Literacy Collaborative (MLC), a partnership of organizations committed to advancing scientific literacy, opened the door to the microbial world for millions of people over this past year.

“Intimate Strangers: Unseen Life on Earth,” the four hour science documentary that provided the centerpiece for the initiative, was broadcast simultaneously by over 90% of the stations within the PBS system during November 1999. The series was viewed by an average of 1.6 million households each week and received favorable reviews from both trade and scientific press.

“Meet the Microbes,” a collection of 17 hands-on activities that complements the major themes of the television series were used widely throughout the US. The activities are appropriate for both informal and formal learning environments. They support open-ended experimental design, help to address elements of National Science Education Standards, and can be downloaded from the MLC website www.microbeworld.org.

Two National Youth Leadership Summits, week-long experiences designed to introduce youth leaders and their adult sponsors to the microbial world and prepare them to implement the hands-on activities in their local community programs, were held. Participants for each summit were drawn from science museums, youth clubs, and school science clubs from 12 different regions in the U.S. A third summit is planned for summer 2000 in Portland, Oregon.

“Unseen Life on Earth: An Introduction to Microbiology,” a 12 part video series designed for use in both undergraduate and pre-college classrooms as well as for distance learning, was received with great enthusiasm by the teaching community. Each 30 minute film focuses on a different aspect of the microbial world. The video series addresses the curriculum standards endorsed by the ASM and is accompanied by teacher and student guides.

“Intimate Strangers: Unseen Life on Earth,” the companion book to the documentary, combines vivid, descriptive images from the television series with original artwork to tell the compelling story of the world of microbes and their role in the Earth’s ecosystems. Targeted to members of the scientifically interested public, the book puts the vitally important role of the microbial world into stories and terms familiar to the reader. The book is available through ASM Press.

www.microbeworld.org is the MLC’s award winning website. The site contains information about the microbial world for all, educational resources for teachers and students, and links to other sites with microbial information. www.microbeworld.org was chosen as an USA Today Hot Site and was featured in NetScape’s What’s Cool. The site received an A-plus review from Education World and was picked and the Top Site Award from Education Planet.

132. Hispanic Role Model and Science Education Outreach Project - Human Genome Project Education and Outreach Component

Clay Dillingham

Institute Of Genetics Education (IGE), Self Reliance Foundation (SRF) and the Hispanic Radio Network (HRN), Santa Fe, NM 87505
CDillingham@GenEd.org

This year completes the DOE sponsored project. The project continued to write, produce and broadcast radio programs, in Spanish, about the scientific, medical, and ELSI consequences of the HGP.

The project more than doubled the number of radio programs to 50 shows this year. These programs were divided among three popular radio shows:

1. “Fuente de Salud” (“Fountain of Health”) 30 programs
 - “Fuente de Salud” reaches an estimated 48.9% of the Hispanic population.
2. “Planeta Azul” (“Blue Planet”) 10 programs
 - “Planeta Azule” reaches an estimated 86.4% of the Hispanic population.
3. “Saber es Poder” (“To Know is to be Able”) 10 programs
 - “Saber es Poder” reaches an estimated 46.2% of the Hispanic population.

Topics covered in the episodes focus on:

- Introducing the Human Genome Project’s science, genetic medicine and ELSI issues. Particular emphasis is given to issues that Hispanics find important and interesting.
- Utilizing Hispanic individuals involved in aspects of Genome Research and its implications.
- The economic implications.
- Bio Industry’s involvement in the genetic revolution.
- Encouraging Hispanics students to pursue science and biotechnology as a career.

In addition, this year the project added two new dimensions to the project.

1. “Mundo 2000”

The project produced three shows for the acclaimed radio talk show “Mundo 2000.” These were one hour long shows featuring Spanish-speaking experts in genetics or related fields discussing the issues.

“Mundo 2000” is currently carried by 17 affiliates and reaches an estimated 20.9% of the US Hispanic population. However, it is also broadcast throughout Latin America, with a additional estimated audience of 15.3 million.

2. “La Columna Vertebral”

The project began providing information via the syndicated newspaper column, “La Columna Vertebral.” “La Columna Vertibral” is syndicated in 82 Spanish language newspapers, with a combined circulation of about 2.5 million readers.

133. Seeking Truth, Finding Justice: A PBS Documentary Special

Noel Schwerin

Backbone Media, San Francisco, CA 94131
schwerin@backbonemedia.org

TRUTH & JUSTICE will be a documentary special for national broadcast on PBS. Produced by the nonprofit corporation Backbone Media in association with Oregon Public Broadcasting, *TRUTH & JUSTICE* will stimulate the public to think critically about the real strengths and important limits of science — particularly genetic science — in both framing and resolving social conflict. It will demonstrate how new technologies — particularly genetic technologies — create unexpected, unprecedented legal conflicts which challenge fundamental legal, ethical and social principles.

In the style of *A Question of Genes*, the PI's award-winning, DOE-funded PBS program, each hour will closely observe two or three pairs of people — judges and scientists, lay people and lawyers — as they grapple with science and technology in a handful of actual legal cases. Through the interactions of these central “characters,” the program will explore the critical interplay of science and the courts in deciding conflicts raised by genetic, reproductive and life-creating technologies. By profiling people at the center of these conflicts, it will use compelling, accessible human drama as its vehicle.

134. SoundVision Science Literacy Project

Barinetta Scott and Jude Thilman

SoundVision Productions; 2991 Shattuck Ave.,
Ste. 304; Berkeley, CA 94705; 510/486-1185,
Fax: 510/486-1287

bariscot@aol.com, jthilman@aol.com
<http://www.dnfiles.org>

SoundVision Productions held its first week-long intensive science literacy training workshop for

public radio reporters and producers. The pilot workshop was held at KQED in San Francisco in October, 1999. Ten reporters/producers came from around the country to attend 15 sessions led by scientists from fields ranging from genetics to physics and statistics, as well as exceptionally creative radio producers. The training built on its award-winning public radio documentary series, *The DNA Files*, which consisted of nine one-hour documentaries broadcast in late 1998 and early 1999. As a result, much of the session content centered on genetics and the Human Genome Project.

This workshop targeted public radio reporters and producers who have specific needs that differ from those of print journalists and television reporters. The workshop was designed around three tracks 1) introduce basic science concepts that form the backdrop for the majority of science and technology stories prominent in the media today as well as provide the tools with which to evaluate developments within the science arena; 2) offer training in science journalism which requires the use of unique research tools for reporting science in a clear and comprehensible, as well as accurate way; and 3) explore techniques for honing the craft of radio writing and production specifically for science stories, which can be especially complex.

An evaluation of the effectiveness of this pilot workshop was conducted two months after its conclusion. Evaluators found that participants both retained a good deal of the information and found it useful in their work. SoundVision's plans for this pilot to be the first in a series of workshops that will bring together scientists and science educators with public radio news and information producers to create a venue for information sharing. The goal of this initiative is to improve the overall news coverage of science issues currently offered to the public radio audience.

*Supported by ELSI grant DE-FG03-99ER62782 from the Office of Health and Environmental Research of the U.S. Department of Energy.

135. Getting the Word Out on the Human Genome Project: A Course for Physicians

Sara L. Tobin¹ and Ann Boughton²

¹Program for Genomics, Ethics, and Society, Stanford University Center for Biomedical Ethics, Palo Alto, CA 94304 and ²Thumbnail Graphics tobinsl@leland.stanford.edu

Progressive identification of new genes and implications for medical treatment of genetic diseases appear almost daily in the scientific and medical literature, as well as in public media reports. However, most physicians completed their medical training prior to the application of recombinant DNA technology to medical diagnosis and treatment. Such individuals do not understand the promise or the limitations of the current explosion in knowledge of the human genome. This project is designed to fill two important functions: first, to provide physicians with a solid foundation in molecular medical genetics, including the impact, implications, and potential of this field for the treatment of human disease; second, to utilize physicians as informed community resources who can educate both their patients and community groups about the new genetics.

We are completing the development of a flexible, user-friendly, interactive multimedia CD-ROM designed for continuing education of physicians in applications of molecular medical genetics. Following the completion and evaluation of a prototype, we have focused on content creation and upgrades to the multimedia tool. The courseware will provide training in four areas: (1) Genetics, including DNA as a molecular blueprint and patterns of inheritance; (2) Recombinant techniques, stressing cloning and analytical tools and techniques applied to medical case studies; (3) Current and future clinical applications, encompassing technical advances and disease diagnosis and prognosis; and (4) Societal implications, focusing on issues such as confidentiality, discrimination, and impact on the family.

The multimedia format permits the use of animation, video, and audio, in addition to graphic illustrations and photographs. Novel features are utilized to tailor the CD to the needs of the user, and continuing medical education credits will be available through

Stanford. The CD will function as a “hybrid” product, capable of seamless interaction with Internet resources and updated content.

136. Dilemmas in Commercializing Human Genome and Biotechnology Products: Developing a Case-Based Business Ethics Curriculum for Industry

Barbara Koenig

Center for Biomedical Ethics; Stanford University School of Medicine; 701 Welch Rd. #1105; Stanford, CA 94305-5015
650/725-6103, Fax; -6131
bkoenig@leland.stanford.edu

This project will be conducted jointly by the Center for Biomedical Ethics and the Graduate School of Business of Stanford University. The general aim of the project is to research and develop instructional material on business ethics decision making for those involved in commercializing biotechnology products. There are four specific aims. They are to (1) research and identify the ethical and social issues that are raised when biotechnology and genomic research is commercialized, (2) develop an analytical business ethics decision making model or process that can be used by pharmaceutical and biotechnology corporations when their managers face these ethical and social issues, (3) develop comprehensive case studies in business ethics based on past pharmaceutical and biotechnology corporate behavior, and (4) apply the decision making process to these case studies as examples of how corporate managers can incorporate ethical reflection, debate, and analysis into business practices.

This project falls within the Department of Energy’s interest in the preparation and dissemination of educational materials that will enhance understanding of the ethical, legal, and social aspects of the Human Genome Project. The educational materials developed will foster corporate decision making that enhances responsible use of genomic and biotechnology information and products from research to postmarketing phases of development.

The results of this research project will be directed primarily to pharmaceutical and biotechnology corporate executives, managers, board members, and attorneys. The detailed case studies developed through collaboration between the Center of Biomedical Ethics and the Graduate School of Business will be utilized within Stanford's Executive Education Program. This unique program attracts leaders from the international business community, providing a singular opportunity to educate decision makers in the biotechnology and pharmaceutical industries. The case studies developed also will be available to augment ethics curricula of graduate schools of business. Case material will be distributed via the World Wide Web and eventually through a conventional textbook format. In this way, the case material will be accessible to anyone interested in the ethical and social consequences of commercializing human genome research.

137. The Responsibility of Oversight in Genetics Research: How to Enable Effective Human Subjects Review of Public and Privately Funded Research Programs

Barbara Handelin

PRIM&R, 132 Boylston Street, 4th Floor, Boston, MA 02116

BHandelin@compuserve.com

There are 3 specific aims 1) to collect relevant materials such as informed consent guidelines, protocol review guidelines, review board development guidelines, etc. which have been developed by individual IRBs, biopharmaceutical companies or policy boards; and 2) to implement focus groups discussions which will seek to articulate the specific stumbling blocks to overcome in genomics protocol review as well as suggestions for tools to remove these blocks; and 3) to develop and seek review of tools to help IRBs and biopharmaceutical companies enhance the overall protection of human subjects in genetics protocols.

Aim 1: Materials have been collected from a fairly wide variety of academic institution IRBs as well as from biopharmaceutical companies which provide a good foundation for further development of materials. These are under review.

Aim 2: Five focus groups and individual interviews with IRB administrators and IRB members have been completed. In addition, one focus group and multiple individual interviews with biopharmaceutical company representatives have also been completed.

Aim 3: We are currently working to develop guidelines and tools for IRBs and industry users. These documents will be the subject of a small review panel workshop where we will seek input and modifications from select users.

We have presented some of our findings at a recent national IRB conference sponsored by PRIM&R (December 5-7, 1999; Boston, MA). Dr. Handelin gave a plenary talk and lead a workshop at this conference of 1300 IRB audience members. Feedback from these presentations supports our conclusions about the needs of IRBs and about the inherent conflicts between industry sponsors and IRBs in the contemplation of genetics protocols.

The work products of this grant are being developed currently, in the last phase of the project. We are developing guidelines and tools for IRBs and industry users. These documents will be the subject of a small review panel workshop where we will seek input and modifications from select users.

138. Confidentiality Concerns Raised by DNA-Based Tests in the Market-Driven Managed Care Setting

Jeroo S. Kotval, Kathleen Dalton, and Patricia Salkin

The University at Albany, State University of New York, Albany, NY

JSK03@health.state.ny.us

The United States has embarked on a unique experiment in the financing and delivery of health care through for-profit, market-driven Managed Care Organizations (MCOs). A relevant feature of these institutions is that they raise financial capital through stock offerings and are therefore beholden to investors for solvency, and accountable to them for their practices. Between 1981 and 1997 for-profit MCO enrollment grew from 12% of all MCO enrollees to 62% of all MCO enrollees. As market pressures intensify, increasing numbers of health plans are shifting to a for-profit status. Cost-controls are more vigorously pursued by such institutions because their survival depends on keeping their stockholders satisfied. Genetic tests that provide information about future illnesses and their possible cost would be of great interest to such institutions. As the population ages and more persons seek insurance in the individual markets, underwriting of insurance will become an increasing concern.

Our study of data utilization in MCOs indicates the following: (1) MCOs have access to a full array of confidential medical information. (2) These data are indeed used for administrative and cost-containment purposes. (3) Current confidentiality practices are inadequate to protect existing medical data. (4) Not-for profit MCOs are engaged in the same practices as for-profits due to competition within the market. (5) As the industry consolidates, large data warehouses collect and analyze medical information from individual medical records for the purpose of better health care management but also for the purpose of controlling costs. (6) Segmentation in the market results in MCOs closing those businesses that they find unprofitable such as services to special needs populations, Medicaid and Medicare. (7) Existing laws are insufficient to protect individuals from uses of genetic information for purposes other than the provision of health care. (8) Policy options — such as the right to bring action against MCOs — do not give the average person control over discrimination by MCOs based on their genetic status.

139. An Economic Analysis of Intellectual Property Rights Issues Concerning the Human Genome Program

David J. Bjornstad¹ and Steven Steward²

¹Oak Ridge National Laboratory, Oak Ridge, TN and

²University of Tennessee
dub@ornl.gov

Information produced by the human genome program and by private investors is given asset status through the patent process. It is then traded among agents who assemble it along with other factors of production to produce products for final consumption. This project seeks to examine the efficiency of markets in effecting this assemblage and the role of patent policy and other public policies in changing this efficiency. It does this by constructing economic models and studying them using the methods of experimental economics.

Efforts during the first year have sought to develop a description of this overall process and to create a logic for studying its components. We view value being added to the stock of intellectual information through a series of phases. The first phase is the development of the base genome itself. The second phase is the identification of the functionality of the genome. The third phase is the identification of the significance of individual variability from the base genome. The fourth phase is the development of mitigating responses to variability. The final step is recovery of value by sales for final consumption. Patent policy establishes property rights across this continuum. Thus, the debate over the patentability of an EST may be framed according to the manner in which information on the base genome must be combined with information on functionality to define patent utility. Recovery of value through final consumption requires acquiring property rights across the four phases. Hence, the demand for final consumption is passed backward through the phases as a derived demand. In other words, the information about the base genome acquires value in combination with function, variability and mitigating response, as well as investment alternatives. Agents bid for patent rights not for their own sake but because of the combinatorial value.

We model this activity through a market in which agents make choices between investing in R&D to acquire property rights or by purchasing established property rights. We view agents as having resource endowments that may be “money,” property rights, or some combination thereof. They interact over a series of time periods during which they can choose to invest in R&D or to buy or sell property rights. They interact through a market environment that has a number of fixed properties and several variable properties. Variable properties are used to implement alternative government policies. Government develops patent policies to guide the establishment of intellectual assets and R&D policies to guide its contribution to the stock of knowledge. For example, government may set the standard for the number of pieces of information necessary to acquire a patent, with different requirements leading to different behavior by agents. Government may affect the potential value of R&D investments by agents through its own R&D investment strategies. Other policy parameters include the degree to which information, like R&D success, is made public, by the “technologies” through which R&D investment may take place, and by the role of uncertainty, for example, as an attribute of information quality. Uncertainty can enter the market in a number of additional ways as well, each of which complicates the analysis.

A key aspect of the market is the requirement that information from the different stages be assembled in specific combinations. This is similar to an airplane trip requiring rights to takeoff and land at specific times at specific airports or the assemblage of an urban site requiring the lease or purchase of a specific configuration of properties. For example, if ESTs are patented, several EST may to be required to form a gene. This, in turn, must be combined with information from other stages. For any given family of final consumption products, total value to be distributed among the agents is limited by the demand for the final consumption products. These distributions will be used to define market efficiency and departures from it. We are interested in how market forces, technical attributes and government policies lead to different distributions of rent.

We are now completing our study of the individual parts of this system and are preparing parameters for portions of the market that can be studied using experimental economics.

140. EINSHAC's Genetics Adjudication Resource Project

Franklin M. Zweig

The Einstein Institute for Science, Health & the Courts, 2 Wisconsin Circle, Suite 700, Chevy Chase, Maryland 20815

www.einshac.org

einshac@intr.net

Beginning in March, 1997, the Genetics Adjudication Resource Project (GARP) has operated from the center of the judicial system to provide science education in ELSI-related genetics to the nation's state and federal courts. 14 regional conference have been conducted involving 1,200 judges in three day meetings to apply the impact of the genomic era to the administration of justice in a changing legal environment. Approximately 250 science advisors from the National Laboratory system and academic science and health centers have anchored the series. An evaluation of the Western States Conference on Genetics in the Courtroom was recently published. See F. Zweig and D. Cowdrey, “Educating Judges for Adjudication of New Life Technologies,” 83,3 *Judicature* 157 (November-December, 1999).

The Project's aim is to involve 2003 judges by the Year 2003. In addition to basic genetics education conferences, advanced subjects and policy-oriented meetings have been conducted and are scheduled for the future. Biomarkers, neuro/behavioral genetics, and biological property conferences have been conducted. Planned for the future are conferences on the Adjudication of Claims Related to Bioremediation and Environmental Health & Safety Activities; Adjudication of Biomedical Claims of Causation and Treatment in Cases Involving Violent Conduct Disorders; and Adjudication of Claims Over

Genetically Modified Food and Agricultural Products.

Behavioral Genetics, will be available in Summer, 2000.

The leadership conference series concerns itself with the direction of the law and gaps in policy. The GARP's Policy Courts Conference on Genetics Issues is scheduled for the Summer of 2001. An International Judicial Conference on ELSI/Genetics Issues will be planned in The Hague in April, 2001 and conducted a year later if funding permits. A bi-lateral Canadian/American Working Conversation on Biotechnology Advance, Biological Property and Ethics is scheduled for Banff, Canada, June 23-25, 2000.

New adjudication issues require new tools for the courts. Accordingly, the Project is testing remote access of neutral, independent experts for the courts; instructional witnesses for juries; and procedures for science neutrals in court-annexed alternative dispute resolution proceedings. The help of the Lawrence Berkeley, Lawrence Livermore and Oak Ridge National Laboratories has been indispensable in the adjudication tools assessment program. The added assistance of the National Institute of Environmental Health Sciences has proved supportive as well. NIEHS and DOE have executed an interagency agreement to fuel the GARP's advanced conferences, leadership conferences, and court-related tools programs.

Cases are key to judicial perspectives about issues created by the explosive growth of life technologies. To assure the availability of durable handbooks for judges in the United States and elsewhere, a Case Review Workshop is scheduled April 20 - 23, 2000 in conjunction with the Medical University of South Carolina. Judges from the United States and three global regions have been invited to critique five cases presenting complex fact patterns hinging on scientific evidence; adjudication issues for courts; state of the science summaries; and state of the law summaries. While EINSHAC prescribes no outcomes and recommends no judicial orders, we have become known for the provision of a full spectrum of considerations in aid of effective judicial case management. The Case Review Workshop systematizes that quest and is resulting in several judicial handbooks built around the case law. The first, The Judges' Handbook on Neuro and

Infrastructure

141. The Human Genome Management Information System: Making Genome Project Science and Implications Accessible

Betty K. Mansfield, Anne E. Adamson, Denise K. Casey, Sheryl A. Martin, Marissa D. Mills, John S. Wassom, Judy M. Wyrick, and Laura N. Yust
Life Sciences Division; Oak Ridge National Laboratory; 1060 Commerce Park; Oak Ridge, TN 37830
bkq@ornl.gov
www.ornl.gov/hgmis

The large, multidisciplinary Human Genome Project (HGP)—the effort to find all human genes and characterize a reference human genome—promises to revolutionize the future so profoundly that this century has been dubbed the “biology century.” Applications of information and technologies derived from the HGP era of the late 20th century will affect almost everyone. Entirely new approaches to biological research and the practice of medicine and agriculture will be implemented. For an unprecedented understanding of the inner workings of whole biological systems, genetic data will provide the foundation upon which research from many biological subdisciplines will be layered. Because the multidisciplinary research model fostered by the HGP will continue, integrative approaches to understanding complex biological systems are likely to replace the reductionist strategies now prevalent.

Commercialization of numerous applications in genomic science is fueling the burgeoning life sciences economic sector. Legislation and litigation increasingly will be concerned with genetics and the intellectual property issues pertaining to genetic information. Educators, the media, students, and the

public need a good understanding of the “new genetics” and its implications to communicate, teach, and help people make related career and personal decisions. Democratizing access to genetic science information should help maximize HGP benefits while protecting against misuse of the data.

Since 1989 the Human Genome Management Information System (HGMIS) has been producing and distributing text on the HGP. This text includes the technical newsletter *Human Genome News*, progress reports, fact sheets, invited articles in peer-reviewed publications, and the DOE *Primer on Molecular Genetics*. Using knowledge-management experience gained in this work and in presentations, exhibitions, and judicial meetings, HGMIS initiated and is continually developing and expanding a suite of Web sites for a variety of audiences. Because genomics and the life sciences are becoming so pervasive in all sectors of society, HGMIS seeks to make information accessible to nontechnical audiences as well as to scientists, social scientists, and medical and legal practitioners who need an understanding of genetics to enhance their work and allow them to communicate across disciplines. *Human Genome News*, for example, contains a unique compilation of resources not found in any discipline-specific publication.

The *Human Genome Project Information (HGPI)* suite of Web sites, initiated in 1994, now supports some 170,000 unique user sessions and about 500,000 text-file transfers each month. The sites contain more than 2600 files, over 2200 of which are text files. About 3000 other sites link to *HGPI* and its individual pages. Although *HGPI* Web pages are presented from the HGP perspective, they extend well beyond the project’s primary goals into the technological and societal ramifications of genomic research. *HGPI* Web pages are primary resources for

nearly all major news outlets carrying stories on genomics including CNN, MSNBC, and Yahoo.

HGPI Web pages are updated daily, new pages are added several times a year, and the entire site receives a major overhaul at least annually. HGMIS incorporates constant feedback from *HGPI* users into the strategy for the sites' ongoing development. Main priorities are to meet user needs for accurate, understandable, and easy-to-locate information relevant to the HGP, its downstream science, and its societal implications.

A sampling of Web pages in the *HGPI* suite (www.ornl.gov/hgmis):

- Applications of Genome Science to biological research and for societal benefit
- HGP Goals and Progress, including ticker box on home page
- Educator and Student pages
- Images, Posters, and Electronic Presentations
- Medicine and the New Genetics and Genetic Disease Pronto! pages
- Ethical, Legal, and Social Issues including ELSI retrospective and products
- Research in Progress
- Meeting Calendars and Reports
- Chromosome Launchpad
- Publications including the November-December 1999 "Genes and Justice" issue of *Judicature*
- CERN Virtual Library Genetics
- Frequently Asked Questions and Fact Sheets on such topics as Gene Testing, SNPs, Functional Genomics, Cloning, and DOE Involvement in Genomics Research

Enhancements anticipated over the next few months include:

- Improved navigation incorporating a comprehensive, textbook-style index;
- Page for potential investors in genomics and biomedical sectors of the life sciences industry;
- "Gee Whiz" factoids on genetics;
- Fact sheet on genetically modified foods;
- Updated or new Web page suites for Medicine and the New Genetics, Legal Issues, Media, Education, Research, and Virtual Library Genetics

Because the *HGPI* sites are increasing in size and usage, HGMIS is implementing a database-facilitated

method to update and maintain links to external pages more efficiently and to place Web-transmitted *Human Genome News* print subscription requests directly in the mailing list database.

HGMIS has distributed more than 210,000 HGP documents, and there are 13,000 print subscribers to *Human Genome News*. Each month HGMIS staff processes about 100 new print subscriptions and 150 information requests received via e-mail, mail, and telephone.

Web Sites

General: www.ornl.gov/hgmis

Medicine and the New Genetics:

www.ornl.gov/hgmis/resource/medicine.html

Research: www.ornl.gov/hgmis/research.html

Constructive comments are appreciated.

This work is sponsored by the Office of Biological and Environmental Research, U.S. Department of Energy, under contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.

142. DOE Genome Program Coordination and Outreach

Sylvia J. Spengler, Janice L. Mann, and Leonora I. Castro

Lawrence Berkeley National Laboratory, Berkeley, CA 94720

SJSpengler@lbl.gov

The DOE Genome Program of the Office of Biological and Environmental Research (OBER) has developed a number of tools for management of the Program. Among these was the Human Genome Coordinating Committee (HGCC), established in 1988. In 1996, the HGCC was expanded to a broader vision of the role of genomic technologies in OBER programs, and the name was changed to reflect this broadening. The HGCC became the Biotechnology Forum. The Forum is chaired by the Associate Director, OBER.

Specifically, DOE is committed to encouraging the development of the next generation of scientists and engineers. The OBER has identified magnet schools, such as those devoted to science and technology, as a

target of opportunity for the initial development and testing of a new program involving DOE program managers, national laboratory scientists, and schools around the country. A broad program includes mentorship in student research projects, summer enrichment programs for teachers at the national laboratories, lecture series by scientists, student summer internships and a technical expertise, as needed. We used Thomas Jefferson High School for Science and Technology (TJHSST) in Alexandria, VA as our first school partner.

In addition, the coordinating group has been deeply involved in the development of the Office of Science Community College Intern Program.

143. DOE Alexander Hollaender Distinguished Postdoctoral Fellowships

Linda Holmes and Wayne Stevenson
Science and Engineering Education Programs; Oak Ridge Institute for Science and Education; Oak Ridge, TN 37831-0117
HOLMESL@ORAU.GOV

The Department of Energy Alexander Hollaender Distinguished Postdoctoral Fellowships were initiated in FY 1986 by the DOE Office of Biological and Environmental Research (OBER) to support research in the life, biomedical and environmental sciences. Fellowships of up to two years are tenable at any DOE, university or private laboratory, if the proposed advisor at that laboratory receives at least \$150,000 per year in support from OBER with support continuing throughout the anticipated tenure of the fellow. Fellows receive stipends of \$37,500 the first year and \$40,500 the second. Eligible applicants must be U.S. citizens or permanent resident aliens and must have received their doctoral or medical degrees within two years of the earliest possible starting date, which is May 1 of the appointment year.

The Oak Ridge Institute for Science and Education (ORISE), administrator of the fellowships for DOE, prepares and distributes program literature to

universities and laboratories across the country, accepts applications, convenes a panel to make award recommendations, and issues stipend checks to fellows. The review panel identifies finalists from which DOE chooses the award winners. Up to five awards are made in even numbered years and up to ten in alternate years. The deadline for applications is January 15. For more information or an application packet, contact Kay Ball at Science and Engineering Education Programs, ORISE, MS 36, P.O. Box 117, Oak Ridge, TN 37831-0117; (865) 576-9975; Fax (865) 241-5220.

144. JASON Study on Data Mining and the Human Genome

G. Joyce, H. Abarbanel, C. Callan, W. Dally, F. Dyson, T. Hwa, S. Koonin, H. Levine, O. Rothaus, R. Schwitters, C. Stubbs, and P. Weinberger
JASON Program Office, McLean, VA
gjoyce@scripps.edu

The JASON organization conducted a DOE-sponsored study on bioinformatics and the human genome project. The study sought to explore the problems that must be faced in bioinformatics and to identify information technologies that could help to overcome these problems. While the current influx of data greatly exceeds what biologists have experienced in the past, other scientific disciplines and the commercial sector have been handling much larger datasets for many years. Powerful datamining techniques have been developed in other fields that, with appropriate modification, could be applied to the biological sciences.

Clearly there is a need for more bioinformaticists, as well as computer scientists and engineers who are willing to become involved in bioinformatics research. An ample talent pool already exists from which to recruit these individuals. The DOE can facilitate cross-fertilization between biologists and the non-biological datamining community by sponsoring joint workshops, offering research fellowships to computer scientists who are interested in biological applications, providing access to the

unclassified resources of the Advanced Strategic Computing Initiative, and taking advantage of the commercial sector's willingness to make datamining tools freely available to the academic community. Greater emphasis must be placed on closing the loop between algorithmic analysis and experimental validation. This will require close cooperation between computer scientists and biologists. The DOE should support the development of experimental methods for validating bioinformatics algorithms and the establishment of statistical tests that can be used to assess the robustness of these algorithms. The DOE should take responsibility for ensuring the provenance of the primary data from the major sequencing centers and making that data freely available in a generic database format with minimal annotation.

Appendices

Appendix A: Author Index

Contact authors are in bold.

A

Aach, John	17
Abarbanel, H.	97
Adams, M.W.W.	80
Adams, Mark D.	14
Adamson, Anne E.	95
Aerts, Andrea	3
Agron, Peter	82
Akinretoy, Bola	14
Albertson, Donna	49
Altherr, Michael R.	6
Andersen, Gary	82
Anderson, Gordon A.	26
Annab, Lois A.	72
Aravind, L.	80
Arellano, A.	83
Arp, D. J.	83
Asbury, Charles	21
Ashfield, Christopher	30
Atkins, John F.	75
Avseenko, Nadezhda	51

B

Badri, Hummy	35
Bakis, Michele	37
Banerjee, L.	78
Barber, Jack R.	72
Barns, Susan M.	81
Barrett, J. Carl	72, 73
Barry, Bob	41
Barry, R. E.	71
Bauman, Brian	30
Beattie, K. L.	71
Berger, Brian	66
Bergmann, Anne	35
Berti, Lorenzo	11
Berven, B. A.	68
Bjornstad, David J.	92
Blake, J.A.	50
Bloom, Mark V.	85
Bobo, T.	40
Bolus, Gary G.	78
Bonaldo, Maria De Fatima	66
Bondoc, Marnel	3
Boughton, Ann	90
Boulton, C.	68
Bradbury, Andrew	74
Bradbury, E. Morton	28
Branscomb, E.	83
Bristow, James	71
Brokstein, Peter	66
Bronson, Scott	86
Brower, Amy	3
Brown, Nancy C.	9
Brown, Patric O.	67
Bruce, David C.	3, 6, 29, 40
Brudno, M.	60
Brudno, Michael	59
Brunk, Brian	44
Bryant, J.E.	40
Buchanan, Michelle V.	25
Buchoff, Jeff	77
Buckingham, Judy	29
Bult, C.J.	50
Bulyk, Martha	17
Burde, Stephan	81
Burkhart-Schultz, K.	83
Burtis, K.C.	68
Bybee, Rodger W.	85

C

Cai, Hong	23
Callan, C.	97
Campbell, Mary L.	33
Canter, David	23
Carninci, Piero	67
Carpenter, D. A.	70
Carrano, Anthony	65
Casey, Denise K.	95
Castro, Leonora I.	96
Catanese, Joe	13
Cawley, Simon	61
Chasteen, Leslie	3
Chen, Gwo-Lin	41
Chen, Xian	28
Cheng, Jan-Fang	37
Chertkov, Olga	8, 22
Chi, Han-Chang	5
Chin, Sylvia	82
Chisholm, Sallie W.	83
Choe, Juno	10
Choi, Peter	14
Christensen, Mari	35
Christensen, Matt	86
Church, George M.	17
Cohn, Judith	3, 6, 22, 53
Cole, James R.	62
Corey, David R.	31
Crabtree, Jonathan	44
Craighead, Harold G.	24
Crawford, Oakley H.	61
Cretu, Gabriella	47
Culiat, C. T.	36, 71
Cuticchia, A. Jamie	43, 53

D

D'Souza, Mark	50
Dally, W.	97
Dalton, Kathleen	91
Daly, Michael J.	80
Davidson, George	56
Davidson, Susan B.	61
Davy, Donn	49
De Jong, Pieter J.	13
Deaven, Larry L.	3, 6, 9, 33, 34, 40
Delvecchio, Vito G.	78
Demirjian, David	52
Derisi, Joseph	67
Deshpande, Alina	24

Dillingham, Clay	88
Dimitrijevic-Bussod, Mira	6
Ding, C.H.Q.	46
Doggett, Norman A.	3, 6, 9, 33, 34, 39, 40, 73
Doktycz, Mitch	41, 71
Dougherty, Michael J.	85
Dralyuk, Igor	59, 60
Dronova, Valentina	51
Dubchak, Inna	46-48, 59, 60
Dunn, John J.	8
Dyachenko, Galina	51
Dyson, F.	97

E

Eichler, Evan E.	37
Einstein, J. Ralph	61
Eisen, Jonathan A.	77
Eisen, Michael	67
Elefterov, Aleksandr	51
Emrich, Charles	19
Endo, T.	67
Eppig, J.T.	50
Erler, A.	83
Evans, Glen	23

F

Fedotcheva, Nadezhda	51
Fei, Zhengdong	28
Feldblyum, Tamara	14, 79
Feng, Bingbing	26
Ferguson, Mary Lee	78
Fischer, Kathe	41
Fisk, David	21
Folta, Peg	38, 40, 65
Fomkina, Maria	51
Foote, Linda J.	25
Foote, R. S.	28
Foquet, Mathieu	24
Foster, Carmen M.	69
Frankel, Ken	3
Fraser, Claire	14, 77, 81
Frazer, Kelly A.	47
Friddle, Carl	71
Fu, Lily	14
Fuge, Edwina	56
Fukunishi, Yoshifumi	67

G

Galimova, Milyausha	51
Gao, Qiufeng	20
Garner, Harold "Skip"	31
Garrity, George M.	62
Geer, Keita	14
Gelfand, M.S.	60
Gesteland, Raymond F.	75
Ghochikyan, A.	8
Gibbs, Richard	7
Gibson, Mark	44
Giometti, C.S.	80
Glazer, Alexander N.	11
Goodwin, Lynne	3, 22
Goodwin, Peter M.	23
Gordon, Laurie A.	3, 35, 37
Goto, Hitoshi	67
Grady, Deborah L.	5
Grajewski, Wally	54
Gray, Joe	49
Green, Lance	24
Griffith, J.R.	40
Grimwood, Jane	4
Groza, Matt	35

H

Ha, Chi	35
Hamaguchi, Yohei	67
Hamilton, Gregory	49
Hammond, Sha	35
Han, Cliff S.	6, 33, 34, 39
Handelin, Barbara	91
Hansen, Thomas S.	44
Harker, B. W.	71
Harsch, Tim	38, 65
Harvey, Damon	66
Hasan, Ahmad	12
Hawley, R.S.	68
Hayashizaki, Yoshihide	67
He, Hongxian	45
He, Kaizhang	12
Heidelberg, John	81
Herskowitz, Ira	76
Hide, Winston A.	55
Hilbert, Helmut	77
Hirosawa, M.	65

Holbrook, S.R.	47
Holden, J.	80
Hollis, K.	68
Holmes, Linda	97
Holt, Ingeborg	44
Holtzapple, Erik	77
Hood, Leroy	4, 86
Hooper, A. B.	83
Horn, Troy A.	78
Hosseini, Roya	37
Hovhanissyan, H.	8
Hoyt, Peter	41, 71
Huang, Heshu	81
Huang, Zhengping	40
Hubbell, Aubree	82
Hughes, Jason	17
Hunsicker, P. R.	70
Hurst, Gregory B.	25, 27
Hutt, Lester D.	19
Hwa, T.	97
Hyatt, Doug	3, 40, 42, 56, 57

I

Itoh, Masayoshi	67
Iyer, Vishy	67

J

Jacobson, S. C.	28
Jensen, Pamela K.	26
Jett, James H.	23
Jewett, Phillip B.	33
Johnson, Dabney K.	36, 68-70
Johnson, Genevieve	65
Jones, Arthur	49
Jones, B. H.	71
Jones, M.D.	40
Joyce, G.	97

K

Kadin, J.A.	50
Kadner, Kristen	3
Kadota, Kouji	67
Kane, Thomas E.	21
Karger, Barry L.	39
Karplus, Kevin	45

Kawai, Jun	67
Keller, Richard A.	23
Kennel, Stephen J.	25
Kernan, John	21
Khandurina, Y.	28
Kharybina, Tatiana	51
Khoury, Hoda	77
Kikuno, R.	65
Kim, Joomyeong	35
Kisakabe, M.	67
Klebig, M. L.	36
Klegig, M. L.	70
Kobayashi, A.	83
Koenig, Barbara	90
Koga, Teichiro	71
Kommander, Kristina	3, 34
Konno, Hideaki	67
Koo, Hean	77
Koonin, Eugene V.	80
Koonin, S.	97
Koriabine, Maxim Y.	72, 73
Korlach, Jonas	24
Kosack, Daniel	77
Kotval, Jerroo S.	91
Kouprina, Natalay	73
Kouprina, Natasha	35
Kozyavkin, S.	11
Krawczyk, Marie	29
Krestova, Irina	51
Krol, Margaret	14
Kuczmarksi, Tom	38, 65
Kuhktin, A.	31
Kulikowski, Casimir	48
Kusakabe, Moriaki	67
Kuske, Cheryl R.	81
Kusumi, Toshinori	67
Kuzmin, Aleksandr	51

L

Lagally, Eric T.	19
Lake, James	46
Lamerdin, J. E.	83
Land, Miriam	3, 40-42, 56, 58
Landers, Rich	54
Langhoff, Dan	30
Larimer, Frank	40, 56, 83
Larionov, Vladimir L.	35, 72, 73
Larsen, Bente	75
Lato, Bernadette	65
Laurencon, A.	68

Lazar, I. M.	28
Leavitt, Mark	72
Lee, P. Chris	77
Lemischka, Ihor	44
Levan, Kevin	21
Levene, Michael	24
Levine, H.	97
Levins, Maureen	14
Li, Bing	62
Li, Guangshan	81
Li, Qi-Xiang	72
Li, Qingbo	21
Liang, Feng	44
Liberzon, A.	8
Liefke, Hartmut	61
Lies, Douglas	81
Lilburn, Timothy G.	62
Lim, H.	80
Lipton, Mary S.	26
Liu, Changsheng	21
Liu, Y.	28
Lobb, R.	40
Lobov, Ivan B.	75
Locascio, Phil	57
Loge, Gary W.	21
Longmire, Jonathan L.	9
Lou, Jianlong	74
Lovley, D.E.	78
Lowry, Steve	37
Lu, Xiaochen	35
Lvovsky, L.	8

M

MacConnell, William	30
Macht, Madison	82
Madan, Anup	4
Maddox, Janine	79
Maidak, Bonnie L.	62
Majidi, Vahid	28
Makarova, Kira S.	80
Malchenko, Sergey	66
Malek, Joel	14
Maltsev, Natalia	50, 51
Malykh, A.	11
Malykh, O.	11
Mammoser, Aaron	13
Mann, Janice L.	96
Mansfield, Betty K.	95
Marcusson, Eric	72
Marks, Jim	74

Martin, Joel	37
Martin, Sheryl A.	41, 95
Marzari, Roberto	74
Mathies, Richard A.	11, 17-19
Maurer, Karl	30
Mayfield, Lynn	31
Mayor, Chris	47
McGann, Stephany	14
Mcguire, Abby	17
McInerney, Joseph D.	85
McKnight, T.	28
McLeod, M. P.	52
McLuckey, Scott A.	27
Meinke, Linda J.	33, 34
Menon, A. Lal	80
Methe, B.A.	78
Michaud, Edward J.	36, 41, 69-71
Micklos, David	86
Mikhailovich, V.	31
Miki, Rika	67
Mila, Leeanne	65
Miller, Arthur W.	39
Miller, D. R.	36, 68, 70
Miller, Robert	55
Milliken, D.	68
Mills, Marissa D.	95
Mirzabekov, A.	31
Mizuno, Yosuke	67
Moazzez, Azita	77
Moffat, Kelly	77
Monroe, Heidi	21
Montgomery, Donald	30
Moore, Jonathan E.	46
Moore, Kateri	44
Moore, Lisa	83
Moyzis, Robert K.	5
Muchnik, Ilya	48
Mudrik, Elena	51
Mudrik, Nikolay	51
Mundt, Mark O.	3, 6, 29, 33, 34, 40, 53, 73
Munk, Chris	3
Munk, Christine	29
Munn, Maureen	86
Mural, Richard	3, 40, 42, 56, 58
Muramatsu, Masami	67
Murray, Alison	81
Myers, Richard M.	4

N

Nagase, T.	65
Naranjo, C.	8
Nealson, Kenneth H.	81
Needham, Cynthia A.	87
Nelson, Chad	75
Nelson, David	65
Nelson, Karen E.	77
Nenashev, Valeri	51
Nenasheva, Valentina	51
Nerenberg, Michael	23
Nguyen, Le-Thu	37
Nguyen, Tuyen	30
Nickerson, Deborah A.	49
Nierman, William C.	14, 15, 78, 79
Nikolaev, Evgeni	51
Nishiduka, Itaru	67
Nitanda, Hiroyuki	67
Nocerino, Christina	82
Nolan, John P.	24
Nolan, Matt	3
Nori, Ravi	54
Noskov, Vladimir	35, 73

O

O'Connell, James	23
Ohara, O.	65
Oishi, M.	65
Okazaki, Yasushi	67
Olman, Victor	57
Olsen, Anne S.	3, 37
Olsen, G.	80
Osipov, Aleksandr	51
Osoegawa, Kazutoyo	13
Overbeek, Ross	50
Overton, G. Christian	44, 61
Ozawa, Y.	67
Ozawa, Yasuhiro	67

P

Pachter, Lior	47
Paegel, Brian M.	18, 19
Palmer, Joel	49
Pang, Ho-Ming	20
Parang, Morey	3, 40, 42, 56, 58

Parker, Charles T.	62
Pavlik, Peter	74
Peng, Ze	37
Pertea, Geo	44
Peterson, Jeremy	77
Petrov, Sergey	3, 41, 58, 71
Phadke, Nikhil	79
Phillips, Robert	44
Pinkel, Daniel	49
Pinney, Debra	44
Plajzer-Frick, Ingrid	37
Podgornaya, Olga I.	75
Polouchine, N.	11
Popkie, Anthony P.	37
Porter, Christopher J.	43, 53
Porter, Kenneth W.	12
Pramanik, Sakti	62
Prange, Christa	6, 38, 65
Predki, Paul	3
Prichard, Kimberly	6
Pronevich, Lyudmila	51
Pusch, Gordon	50

Q

Qin, Shizhen	4
Qiu, Qiaoyun	81
Quackenbush, John	44

R

Radnedge, Lyndsay	82
Radtkey, Ray	23
Raja, M.C.	8
Ralston, Pam	17
Ramos, Purita	23
Ramsey, J. M.	28
Ramsey, R. S.	28
Rayner, Simon	31
Redkar, Rajendra J.	78
Reich, C.	80
Richardson, Charles	10
Richardson, J.E.	50
Richardson, Paul	35
Rieder, Mark J.	49
Riethman, Harold C.	5
Rinchik, Eugene M.	36, 41, 68-70, 71
Rindone, Wayne P.	17
Rizzo, Michael	77
Roach, J. Shawn	31
Robbins, Joan	72

Robbins, Robert J.	85
Rocap, Gabrielle	83
Roh, Yul	81
Rossi, Francis	30
Rothaus, O.	97
Rowen, Lee	4
Rubin, Edward M.	47, 71
Rubin, Gerald M.	66
Russell, L. B.	36, 70
Rykunova, Anna	51

S

Salkin, Patricia	91
Salzberg, Steven L.	77
Sandine, Gary L.	55
Sanford, Kristi	86
Saunders, Elizabeth	29
Saxman, Paul R.	62
Sblattero, Daniele	74
Scherer, James R.	18
Schmidt, Thomas M.	62
Schmoyer, Denise	3, 41, 42, 58, 71
Schmutz, Jeremy	4
Schug, Jonathan	44
Schut, G.	80
Schwerin, Noel	89
Schwitters, R.	97
Sciufo, S.	78
Scott, Barinetta	89
Scott, Duncan	37
Selkov, Aleksey	51
Selkov, Evgeni	50, 51
Selkov, Jr., Evgeni	50, 51
Semerikov, Vladimir	51
Seymour, Marilyn	82
Shah, Manesh	3, 40, 56, 57, 83
Shatsman, Sofiya	14
Shaw, Barbara Ramsay	12
Shaw, Joe J.	78
Shi, Yining	18
Shibata, Kazuhiro	67
Shibata, Yuko	67
Shin, Dong-Guk	54
Shreve, Jeff	37
Shu, Chung Li	13
Siegal, Robert	74
Simpson, Peter C.	18, 19
Sirota, Tatiana	51
Skibola, Christine	18
Skowronski, Evan	82

Slaven, Bradley	77
Slesarev, A.	11
Slezak, T.	40
Smith, Lloyd M.	25, 28
Smith, Martyn T.	18
Smith, Richard D.	26
Smith, Temple F.	45
Snoddy, Jay	3, 41-43, 56, 58, 68, 71
Snoeyenbos-West, O.	78
Soares, Marcelo Bento	66
Soares, Vera Da Costa	66
Sobolev, A.	31
Solomon, Gregory G.	72, 73
Sorokin, Anatoly	51
Sosnowski, Ron	23
Speed, Terry	61
Spengler, Sylvia J.	48, 59, 60, 96
Stapleton, Mark	66
Stapleton, Ray	81
Steffen, Martin	17
Stephenson, Jr., James L.	27
Stevenson, Wayne	97
Steward, Steven	92
Stilwagen, S.	83
Stoeckert, Chris	44
Stomakhin, A.	31
Strizhkov, B.	31
Stubbs, C.	97
Stubbs, Lisa	35, 73
Stupar', Oleg	51
Sugahara, Yuichi	67
Summers, Jack	12
Sutherland, Robert D.	3, 33, 34

T

Tabor, Stanley	10
Talbot Jr., C. Conover	43, 53
Tatum, Lela	34
Tatum, Owatha L.	22, 73
Tatusov, Roman L.	80
Tavazoie, Saeed	17
Taylor, Scott L.	49
Ternovsky, Vadim	51
Terry, Astrid	3
Tesmer, Judy G.	33
Thilman, Jude	89
Thompson, Linda S.	6

Tiedje, James M.	62, 81
Tillib, S.	31
Tobin, Sara L.	90
Tolic, Ljiljana Pasa	26
Tollaksen, S.L.	80
Tomaru, Yasuhiro	67
Torney, David C.	24, 55
Tran, Kevin	77
Tsegaye, Getahun	14
Tu, Gene	23
Tuemmler, Burkhard	77
Turner, Stephen W.	24

U

Uber, Donald	49
Uberbacher, Ed	3, 40-42, 43, 56, 58, 61
Ulanovsky, L.E.	8
Ulintz, Peter	79
Upton, Jonathan	44

V

Valdez, Yolanda	24
van den Engh, Ger	10, 21
Vargas, Michelle	35
Vasilenko, Olga	51
Vasiliskov, V.	31
Veenstra, Timothy D.	26
Verzillo, Vittorio	74
Vokler, Inna	57
von Arnim, Qing G.	69

W

Wachocki, Susi	6
Wagner, Mark	3
Wang, Daojing	19
Wang, Jian	44
Wang, Mei	37
Warth, Tiffany	35
Wassom, John S.	95
Waters, L. C.	28
Webb, Watt W.	24
Wehri, Eddy	35
Weinberger, P.	97
Werner, James H.	23
Werner-Washburne, Margaret	56

Wexler, David	18
Whitaker, Tom J.	29
White, Owen	77
White, P. Scott	22, 24
Willey, Kenneth F.	29
Wills, Norma M.	75
Wolf, Denise	48
Wollard, Jessica	82
Womack, Andrew W.	73
Wong-Staal, Flossie	72
Worley, K. C.	52
Wu, Jung-Rung	34
Wunschel, David S.	26
Wyrick, Judy M.	95

X

Xie, Jin	11
Xing, Eric Poe	48
Xu, Dong	61
Xu, Ying	61

Y

Yang, Z.	52
Yao, Zuxu	30
Yates, J.	80
Yeung, Edward S.	20
Yoshiki, Atsushi	67
You, Yun	36, 69, 70
Yust, Laura N.	95

Z

Zarakhovich, Sophia	45
Zevin-Sonkin, D.	8
Zhao, Harry	21
Zhao, Shaying	14
Zhou, Jizhong	81
Zhou, Songsan	21
Zhu, Yiwen	37
Zorn, M.	60
Zorn, Manfred D.	48, 49, 52, 59
Zucca, Tony	30
Zweig, Franklin M.	93

Appendix B: National Laboratory Index

U.S. Department of Energy Laboratories

Human Genome Program work at DOE laboratories is described on the following pages.

Ames Laboratory	20
Argonne National Laboratory	8, 31, 50, 51, 80
Brookhaven National Laboratory	8
Joint Genome Institute	3, 4, 6, 9, 22, 24, 29, 33- 35, 37, 39, 40, 53, 73, 83, 108
Lawrence Berkeley National Laboratory	3, 37, 46-49, 52, 59, 60, 66, 71, 96
Lawrence Livermore National Laboratory	3, 35, 38, 40, 65, 73, 82, 83
Los Alamos National Laboratory	3, 6, 8, 9, 22-24, 28, 29, 33, 34, 39, 40, 53, 55, 73, 74, 81
Oak Ridge National Laboratory	3, 25, 27, 28, 36, 40-42, 56-58, 61, 68-71, 81, 83, 92, 95
Pacific Northwest National Laboratory	26
Sandia National Laboratories	56

