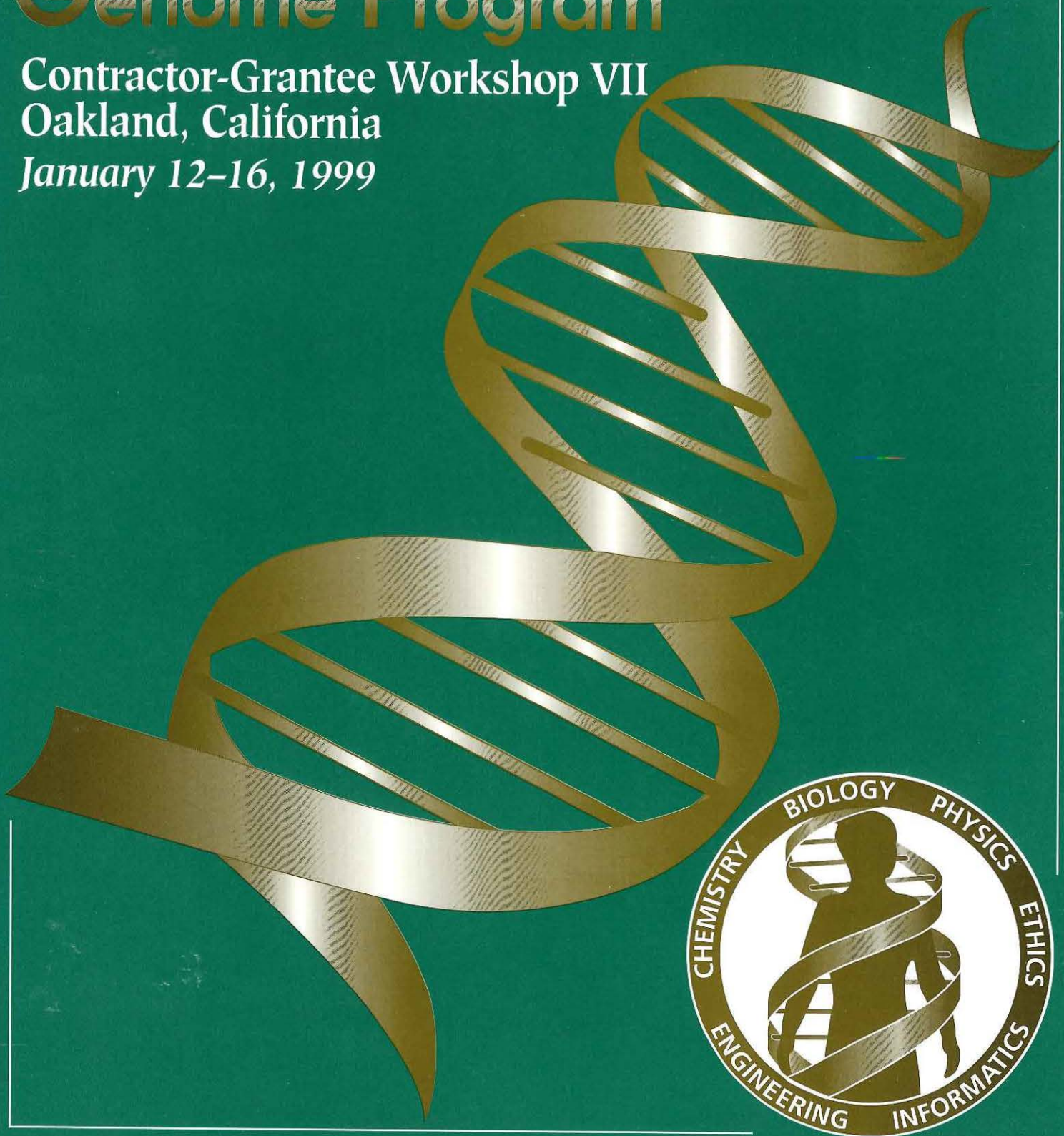
CONF-990104

DOE Human Genome Program

Contractor-Grantee Workshop VII

Oakland, California

January 12–16, 1999



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@oer.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
423/576-6669, Fax: 423/574-9888
E-mail: *s22@ornl.gov*

An electronic version of this document will be available on January 12, 1999 at the Human Genome Project Information Web site under Publications (*http://www.ornl.gov/hgmis*).

Abstracts for this publication were submitted via the web.

CONF-990104

DOE Human Genome Program Contractor-Grantee Workshop VII

January 12-16, 1999
Oakland, California

Date Published: December 1998

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

Prepared by
Human Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830-6480

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

Contents

Poster
Number Page

Poster
Number Page

Poster
Number Page

Mapping

Poster
Number Page

Informatics

Poster
Number Page

Poster
Number Page

Functional Genomics

Poster
Number Page

Microbial Genome Program

Poster
Number Page

Ethical, Legal, and Social Issues

Introduction to Contractor-Grantee Workshop VII

Welcome to the Seventh Contractor-Grantee Workshop sponsored by the Department of Energy (DOE) Human Genome Program (HGP). This workshop provides a unique opportunity for HGP 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 advantage of opportunities offered by this meeting and by the beautiful surroundings of the San Francisco-Oakland Bay area. We also hope you will visit the DOE Joint Genome Institute's new Production Sequencing Facility (PSF) in Walnut Creek. This facility is scheduled to be opened officially in the spring by the Secretary of Energy Bill Richardson.

The 193 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 early efforts in model organism and functional genomics research. In addition, we have included talks from invited guests who will discuss related efforts in other species and opportunities for postgenomic biological investigations enabled by genome research. All projects funded by the Office of Biological and Environmental Research (OBER) will be represented at poster sessions at the Oakland Marriott Hotel, so you will have the opportunity to meet with researchers. New informatics resources also will be demonstrated during the poster sessions, and I urge you to take full advantage of them.

The main challenge facing the genome program today remains high-throughput sequencing. Two years ago, DOE addressed this challenge by forming the Joint Genome Institute (JGI), under the direction of Elbert Branscomb. JGI employs the complementary strengths of DOE's three largest genome programs and those at other laboratories and universities to make more efficient and effective use of diverse expertise and resources.

In the past year, we gave JGI two very challenging and difficult tasks:

- Sequence a total of 20 Mb of DNA to "Bermuda standards," and
- Successfully occupy the PSF building in Walnut Creek.

JGI has met both of those ambitious goals. As of October 23, 1998, it had submitted to GenBank a total of 21 Mb, all with Phred/Phrap values of 40 or greater. This represents a tenfold increase over the amount of DNA sequenced by all three DOE genome centers in the previous year; currently, the sequencing rate averages 3 Mb per month. And, as noted, PSF is open, halfway between the Lawrence Berkeley and Lawrence Livermore national laboratories. PSF's goal this year is 40 Mb, a doubling of last year's yield.

Also last October, the U.S. Human Genome Project completed and published in *Science* its third 5-year plan, developed jointly with the NIH National Human Genome Research Institute. This plan was achieved with input from the broad genome scientific community. 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, either, and the workshop speakers on the ethical, legal, and social implications will remind and challenge all of us that our science has societal impacts and we cannot be aloof and disengaged from those interactions.

There are other genomes to sequence besides the human, and the OBER Microbial Genome Program continues to contribute complete sequences to public databases. Each microbial sequence has its surprises and its exciting science. One of the more exciting sequences completed in 1998 was the entire 3-Mb sequence of *Deinococcus radiodurans*, the most radio-resistant microbe yet known; 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 achievement also underscores the opportunity to exploit the interdisciplinary biological approaches that we view as important guiding principles for the science we support. We must continue to take responsibility for using our science to better our world.

We anticipate 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
Office of Biological and Environmental Research
U.S. Department of Energy
genome@oer.doe.gov

Sequencing

1. Uncovering the Riches of Human Chromosome 19 Through Genomic Sequencing

Jane E. Lamerdin, Karolyn Burkhart-Schultz, Linda Danganan, Laurie Gordon, Stephanie Stilwagen, Glenda Quan, Hoan Phan, Nelson Velasco, Andre Arellano, Brent Kronmiller, Long Do, Astrid Terry, Warren Regala, Vijay Viswanathan, Jennifer Dias, Amy Brower, Tim Andriese, Pat Poundstone, Julie Avila, Jackie Coefield, Susan Lucas, Tina Attix, Stephenie Liu, Robert Bruce, Evan Skowronski, Rick Colyaco, Arthur Kobayashi, David Ow, Matt Nolan, Anthony V. Carrano, Anne. S. Olsen, and Paula McCready
Joint Genome Institute, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550
lamerdin1@llnl.gov

Genomic sequencing of human chromosome 19 is well underway. Roughly 20% of the euchromatin of chromosome 19 is now available as finished genomic sequence in GenBank, with completion of most of the chromosome anticipated by 2001. Utilizing a high resolution physical map constructed largely in bacterial-based clones, we have seeded our current sequencing queue with many large (*1Mb) contigs from well-mapped regions, with representative contigs from almost every cytogenetic band on chromosome 19. As of Oct 30, we have finished over 11 Mb of genomic sequence, with roughly 10.5 Mb submitted to GenBank. Preliminary analyses of our data lend credence to the expectation that this GC-rich chromosome will be an excellent target for gene discovery through genomic sequencing. In this regard, several GC-rich regions (average GC content in excess of 58%) that have been sequenced on chromosome 19 exhibit a high gene density (on average, 1 gene per 20-25 kb) relative to the rest of the genome, and encode genes with compact genomic structure. Other regions with a slightly lower GC content (average GC= 50%) possess fewer genes which span larger genomic distances, e.g. the ryanodine receptor (RYR) region in 19q13.1.

One interesting feature of chromosome 19 is the large number of clustered gene families distributed throughout the length of the euchromatin. These include the pregnancy-specific glycoprotein family (PSG), multiple zinc finger families (ZNF), olfactory receptors (OLFR) and cytochrome P-450s (CYP). In order to understand their evolution and subsequent functional diversification, several of these clusters are current sequencing targets. Not surprisingly, the ages of these clusters differ significantly, with the PSG family having duplicated fairly recently in evolutionary time, while the OLFR and ZNF clusters appear much older, with many of their members possessing orthologs in mice and rats. One common feature of the genomic structure of these disparate families is the prevalence of specific repeat families, which may have contributed to the evolution and expansion of these regions. We are undertaking a more detailed comparison of the genomic content of these gene family regions on chromosome 19, as well as their orthologous counterparts in mouse. These comparisons will no doubt expand our recognition of the fluidity of the mammalian genome.

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

2. Genomic Sequencing of 3 Mb of Human Chromosome 16p13.3 Containing 4 Disease Genes

M.O. Mundt, D.O. Ricke, D.C. Bruce, A.C. Munk, D.L. Robinson, M.D. Jones, J.M. Buckingham, L.A. Chasteen, E.H. Saunders, L.S. Thompson, L.A. Goodwin, A.L. Williams, J.L. Longmire, P.S. White, L.L. Deaven, and **N.A. Doggett**
Joint Genome Institute, Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545
doggett@lanl.gov

We have nearly completed genomic sequencing of a 3.0 Mb cosmid/P1 contig of the human chromosome region in 16p13.3 extending from the tuberous sclerosis disease (TSC2) locus to the CREB binding protein (CREBBP) locus [responsible for Rubinstein-Taybi Syndrome and implicated in acute myeloid leukemias associated with translocations t(8;16)(p11;p13.3) and t(11;16)(q23;p13.3)]. This contig also encompasses the polycystic kidney disease 1 (PKD1), the familial Mediterranean fever gene (MEFV) and the syntenic breakpoint between mouse chromosomes 16 and 17. The average overlap between clones in the contig is about 25%. Our earlier sample sequencing (SASE) of this region had revealed that it is gene rich and G+C rich (>50% G+C), with the gene density approaching one gene/10 kb in some stretches. These observations are consistent with the cytogenetic designation of 16p13.3 as a G+C rich "T" band (Dutrillaux, 1973; Holmquist, 1992). Our strategy for sequencing involved nebulization to randomly break DNA, size selection of 3 kb fragments, double adapter cloning into bluescript KS+ plasmid, and sequencing of both ends to 6X random sequencing coverage. Sequencing reactions were predominately Big dye terminators (ABI). Assembly of sequence contigs was assisted by the inherent relationship of the end sequences being approximately 3 kb apart. Closure and finishing was achieved by a combination of primer walking, longer reads, and alternate chemistry reactions. Sequence analysis and annotation is semi-automated with use of the SCAN program (developed by Ricke). We have achieved 100% closure of all 58 clones which we have attempted to sequence from this region. Three gaps remain but clones have now been found which span these. One of these "gaps" in the cosmid contig map is in the same region of a breakpoint cluster in the CREB binding protein gene, which occurs in leukemias. This region was stably maintained in BACs however. The maximum G+C content found in a finished clone is 57%. Alu content has also been high, with up to 30 Alu's in a finished cosmid. Supported by the US DOE, OBER under contract W-7405-ENG-36.

3. Sequencing Human Chromosome 14 and the Mouse Major Histocompatibility Locus: A Progress Report

Lee Rowen, Anup Madan, Shizhen Qin, Lee Hood, and the Multimegabase Sequencing Group
Department of Molecular Biotechnology, University of Washington, Seattle, Washington
leerowen@u.washington.edu

To date, we have sequenced over 1.1 megabases of the mouse major histocompatibility locus and over 600 kb of chromosome 14. Our target region on chromosome 14 is 14q24.3-ter.

Based on a preliminary analysis of the mouse MHC sequences, and a comparison with the human sequence counterpart, we have drawn the following conclusions:

1) Evolutionarily conserved genes are interspersed with genes with no identifiable homologues in other species, suggesting that genes with both specialized and generalized (housekeeping) functions co-exist in the MHC.
2) The MHC class III region is the most gene-dense. In the human sequence, 17% of 263 kb contains the coding region of 20 genes (average of 1 gene per 13.2 kb). The average intergenic distance is 2.7 kb.
3) Expansion of gene family membership has occurred through the duplication of long repeats.
4) Gene content and order in human and mouse MHC is similar, although variation in the extent of gene duplication occurs both within and between species.

5) Conserved blocks between human and mouse correspond to the most gene-dense regions in each specie.
6) Isochore boundaries, based on GC content and genome-wide interspersed repeats, can be identified in the class II-III regions in both species.

4. Physical Mapping and Sequencing of Human Chromosome 16p12.1-11.2

Hyung Lyun Kang, Yicheng Cao, So Hee Dho[1], Diana Bocskai, Mei Wang, Xuequn Xu, Jun-Ryul Huh[1], Byeong-Jae Lee[1], Francis Kalush[2], Judith G. Tesmer[3], Eunpyo Moon[4], Norman A. Doggett[3], Mark D. Adams[2], Melvin I. Simon, and Ung-Jin Kim
Division of Biology, Caltech, Pasadena, CA 91125
[1] Institute of Molecular Biology and Genetics, Seoul National University, Seoul, Korea; [2] The Institute for Genomic Research, Rockville, Maryland; [3] Los Alamos National Laboratory, Los Alamos, New Mexico; and [4] Ajou University, Suwon, Korea
simonm@cco.caltech.edu

The first goal of the Human Genome Project is to determine the nucleotide sequences of the entire human genome. We have been mapping and sequencing the 6 Mbp region near the 16pCEN on the short arm of human chromosome 16 (16p12.1-11.2) jointly with The Institute for Genomic Research (TIGR) and Los Alamos National Laboratory (LANL). As shown by the complete sequences from the BACs derived from this region, the target region has many small and large peri-centromeric repeats. It has been theorized that due to these repeats, many of which consist of large numbers of short tandem repeats, near-centromeric regions are difficult to clone and map. In fact, most genomic libraries tend to have fewer clones covering the centromeric and telomeric regions. Our target region is relatively sparsely covered by STS markers. In fact, most genomic libraries tend to have fewer clones covering the centromeric and telomeric regions. Our target region is relatively sparsely covered by STS markers.

To provide large, contiguous stretches of BACs from the target region for high throughput shotgun sequencing at TIGR and JGI, Caltech has been developing BAC contigs using the STS and other ordered markers obtained from the YAC-STS map that was previously constructed by LANL. The 12X coverage human BAC libraries constructed at Caltech (A, B, and C) were screened by the combination of the STS-PCR screening on pooled libraries and the hybridization-based screening using probes that include cDNA inserts, BAC end clones, genomic DNA fragments and BAC inserts. Initially, a total of 46 STSs were screened against the libraries. More recently, Caltech has constructed a 7X coverage library D from approved human DNA samples, which has been screened by hybridization using the probes derived from STS-PCR products, BAC clone inserts (for BAC-to-BAC hybridization), and gel-purified YAC DNA (YAC-to-BAC hybridization). Thus far over 1,000 putative BACs from the target region have been identified. The clones are being built into overlapping contigs based on the analyses that include STS contents, BAC-to-BAC hybridization data, insert size, restriction fingerprint analysis, BAC end sequencing and BAC end sequence match with completely sequenced BACs, and FISH mapping on some selected BACs. Over 30 BACs from this region corresponding to approximately 4 Mbp in length have been sequenced at TIGR. To close the remaining gaps, we are currently designing new STS markers and OVERGO probes based on the BAC end sequence data along with the Alu-PCR products from the YAC clones covering the gaps. We also plan on screening new 4X coverage EcoRI BAC library. For the description, protocols, and data related to our projects, please visit our WEB site http://www.tree.caltech.edu.

5. Human Telomere Mapping and Sequencing

Han-Chang Chi[1], Deborah L. Grady[1], Harold C. Riethman[2], and Robert K. Moyzis[1]
[1]Department of Biological Chemistry, College of Medicine, University of California at Irvine, Irvine, CA 92697 and [2]The Wistar Institute, Philadelphia, PA 19104
hcchi@uci.edu

The Human Genome Project has undergone a dramatic shift this year to the goal of obtaining a 'framework' sequence of human DNA in just a few years. 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. A significant fraction (20%) of human DNA contains a high percentage of repetitive sequences, is unstable in most cloning vectors, and exhibits extensive polymorphisms both between individuals and populations. Producing quality maps and sequence in such regions, which faithfully represent human genomic DNA, will be a continuing challenge. One such region is represented by human telomeres. Following the discovery and cloning of the human telomere repeat $(TTAGGG)_n$ by our laboratory ten years ago, numerous investigations have implicated this sequence or genes near telomeres as likely targets for alterations during cellular aging and cancer progression.

Nearly all human telomeres have now been cloned as yeast artificial chromosomes by functional complementation. During the last year, our laboratory finished the 0.23Mb 7q telomere sequence (GenBank accession AF027390), the first RARE (RecA-Assisted Restriction Endonulease) cleavage confirmed telomere region to be sequenced directly up to the terminal $(TTAGGG)_n$ repeat. Nine overlapping cosmids and two PCR products obtained from the 7q telomere YAC clone HTY146 (yRM2000) were sequenced using a Sample Sequencing (SASE)-parallel primer walking strategy. In total, 18% of this telomeric sequence required extensive PCR and non-standard sequencing methods to finish. Confirmation of the sequence against human genomic DNA was conducted by PCR-sequencing, using primer sets picked every 20kb. The submitted sequence is a faithful representation of human DNA, containing less than one error in 10,000 bases. Computer and experimental analysis uncovered numerous open reading frames, expressed sequence tags (ESTs), and potential exons dispersed along the entire 226 kb region, as well as 6 single nucleotide polymorphisms (SNPs), 19 variable number of tandem repeats (VNTRs) and 20 microsatellite repeats. The first and second exons for the human vasoactive intestinal peptide receptor 2 (VIPR2) gene were localized approximately 191 kb internal to the $(TTAGGG)_n$ terminal repeat. This neuropeptide system is involved in a diverse set of physiological functions including smooth muscle relaxation, electrolyte secretion, and vasodilation. Primer pairs picked to amplify the regions of 7q containing VNTRs uncovered extensive polymorphisms in the limited numbers of individuals examined to date. We are nearing completion of mapping and sequencing two additional telomeres, 9q and 11q, chosen because these regions contain a limited amount of subtelomeric repeats. In addition, SASE analysis is being initiated on 14 additional telomeres that have been confirmed by RARE cleavage (1q, 2p, 2q, 6q, 7p, 8p, 8q, 12q, 13q, 14q, 17p, 18p, 18q, and 21q) in order to prioritize our next targets for finished genomic sequencing.

6. A Comparison of Sequence Gap Closure Strategies

Glenda G. Quan, Karolyn Burkhart-Shultz, Timothy Andriese, Andre Arellano, Long Do, Arthur Kobayashi, Brent Kronmiller, Madison Macht, Matt Nolan, David Ow, Hoan Phan, Melissa Ramirez, Warren Regala, Christina Sanders, Stephanie Stilwagen, Astrid Terry, Nelson Velasco, Vijay Viswanathan, Anthony V. Carrano, and Jane E. Lamerdin
Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550
quan1@llnl.gov

The goal of finish sequencing is to obtain high-quality, contiguous sequence of cosmid and BAC clone inserts. A major component of finish sequencing is gap closure. In order for the sequence to be contiguous, gaps in the initial sequence data,

obtained from random shotgun sequencing, must be closed. At the Joint Genome Center at Lawrence Livermore National Laboratory, we employ three main strategies for sequence gap closure: transposon "bombing", shatter library production, and custom primer walking. We currently use an in vitro transposon insertion strategy involving the random insertion of a yeast transposable element into a gap-spanning, circular plasmid. Using primers designed off both ends of the transposable element, new sequence can be obtained directing away from the insertion point. Transposon "bombing" allows us to identify new sequencing start points within the gap itself, and gives us the advantage of sequencing with two primers. In the shatter method, a double-stranded, linear fragment containing the gap sequence (e.g. a PCR product or restriction fragment) is sonicated into fragments of 300-500 bp in length. These short fragments are then sub-cloned into an M13 phage vector and sequenced using conventional ET-forward primers. These shatter libraries are particularly well-suited to regions of significant secondary structure which are recalcitrant to conventional sequencing chemistries, where the smaller inserts may contain only a portion of the hairpin in the original gap-spanning clone. Additionally, the data generated by these clones are very high in quality and can be assembled as 'mini' shotgun projects in those instances of very difficult assembly problems, such as long tandem repeats. Our third strategy utilizes the automated primer picking program in the sequence editor Consed for primer walking on existing clones that span a gap. The main advantage of primer walking is that it allows closure of small gaps with a minimum number of sequencing reads. We have used various combinations of these three strategies to increase our output of finished sequence by over 500% in the last fiscal year. Analyses are underway to evaluate the efficiency and cost of these three strategies in order to better tailor automated finishing protocols needed to achieve the ambitious sequencing ramps required to complete the JGI's portion of the human genome.

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

7. The SaF Finishing Tools

Matt P. Nolan, Jane E. Lamerdin, Glenda G. Quan, and Anthony V. Carrano
Joint Genome Institute, Lawrence Livermore National Laboratory, Livermore, California
nolan1@llnl.gov

Our modified shotgun sequencing effort has three phases. In the random phase we sequence a fixed number of plates resulting in 80%-95% of the cosmid bases meeting our quality-based, double-stranded, finish criteria (QbDsFc). During pre-finishing we resequence clones attempting in one round of forwards and reverses to meet the QbDsFc for 95% of the bases and close most gaps. During directed closure we close any remaining gaps and complete double-stranding. To reduce finishing costs and speed time to completion for our cosmid and BAC clone projects we created software to automate selection of finishing reads. We describe our SaF (Swedish and Finnish) software tools developed to 1) facilitate the specification of clones for resequencing and to 2) quantify the state of project contigs with respect to our QbDsFc.

We describe improvements to the SaF tools that helped us meet our ten-fold increase in sequence produced in the past year. In our production sequencing we use the SaF tools to fully automate clone selection in the pre-finishing phase and we require finishers to address each region identified during directed closure.

For a project assemblage our SaF tools identify bases not meeting the QbDsFc, then conglomerate these problem bases into problem regions using parameterized filtering and clustering algorithms.

They produce reports listing each problem region and a contig summary.

In prefinishing we are attempting to identify candidate clones for the creation of shatter libraries. Some simple improvements to our algorithm have helped target potential false joins resulting in fewer contigs coming out of the prefinishing stage. We are targeting more reverses at internal problem areas with higher error rate. Also, with a greater emphasis on sequencing BAC clones, we are hoping to more strongly target regions of adjacent ALUs as they are often the cause of gaps and false joins. Additionally, for the BACs we are trying to incorporate restriction enzyme map data to verify sections of properly aligned sequence order for the purposes of orienting contigs and identifying potential false joins.

New SaF tool features increase their usability in the directed closure phase. We have incorporated a feedback loop which identifies resequenced clones so that they don't get ordered redundantly so that we may use the automated clone selection in multiple passes and so we know when certain strategies have been played out. We describe our attempts to more tightly the SaF tools with consed. A greater emphasis is being placed in increasing the cost effectiveness of clone selection. For instance, we identify short clones so that we do not suggest sequencing their opposite ends.

Work performed under the auspices of the US DOE by Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48

8. Process Description of a 5 Mb / Year Finish Sequencing Operation Using 100% Plasmid Double End Sequencing

David C. Bruce, Leslie A. Chasteen, Donna L. Robinson, Myrona D. Jones, Jennifer Bryant, Nancy C. Brown, Beverly Parson-Quintana, Darrell O. Ricke, Mark O. Mundt, P. Scott White, Norman Doggett, and Larry L. Deaven
Human Genome Center, Los Alamos National Laboratory, Mail Stop M888, Los Alamos, NM 87545
bruce@telomere.lanl.gov

Beginning in Oct. 1997, the Center for Human Genome Studies (CHGS) at Los Alamos National Laboratory (LANL), as part of the Joint Genome Institute (JGI, http://jgi.doe.gov/), committed to first year finished sequencing of 2.84 MB of human sequence. A steady state finish rate of 5 MB / year, by the end of the first year is projected. An exhaustive description of the JGI Sequence quality standards and sequencing targets is available at http://jgi.doe.gov/Docs/JGI_Seq_Quality.html. Prior to Oct 1997, the CHGS had submitted 224 KB of finish sequence to GenBank using only Applied Biosystems software, minimal sample tracking, and little automation. As of Oct 4, 1998, the CHGS had finished 3.2 MB total and 2.8 MB unique sequence and reached a finish sequence output of 0.6 MB / month using phred/phrap/consed finishing tools, select automation, sample tracking system in conjunction with a fully redesigned process. The sequencing strategy is 6X plasmid end sequencing using dye terminator chemistry in production, quality gap closure using alternate chemistry / gel conditions in pre-finish and final gap closure using a combination of primer walking, transposon bombing and small insert libraries in finish. Implementation of this aggressive ramp will be presented including; sequencing strategy design, process analysis, personnel reorganization, automation, informatics, and quality / cost control with emphasis on the production phase of sequencing.

9. Automation of Finishing at JGI-LLNL

Stephanie Stilwagen, Matt Nolan, Andre Arellano, Karolyn Burkhart-Schultz, Long Do, Arthur Kobyashi, Brent Kronmiller, Madison Macht, David Ow, Hoan Phan, Glenda Quan, Melissa Ramirez, Warren Regala, Christina Sanders, Astrid Terry, Stephan Trong, Nelson Velasco, Vijay Viswanathan, Anthony Carrano, and Jane Lamerdin
Joint Genome Institute, Lawrence Livermore National Laboratory, 7000 East Ave, L-452, Livermore CA 94550
stilwagen1@llnl.gov

Lawrence Livermore National Laboratory has generated 10.5 Mb of highly accurate, finished genomic sequence of selected large insert clones (e.g.

cosmid, BAC, P1) from chromosome 19 of which 8.6 Mb has been completed within the last fiscal year. We were able to achieve this eight-fold increase with the introduction of a suite of finishing tools and web-based computer interfaces which are directly linked to automated robotic workstations. The LLNL sequencing strategy utilizes a 'shotgun' approach to generate our initial sequence data. The next stage of the process is pre-finishing which involves the selection of clones for re-sequencing for initial gap closure and ambiguity resolution. We have automated pre-finishing by utilizing the LLNL developed software tool, Swedish to select clones for re-sequencing with either the dye primer or dye terminator chemistry. After one round of pre-finishing, a project moves to the finishing phase and is assigned to a finisher.

A finisher makes use of multiple software tools in an iterative manner to obtain contiguous sequence that meets our standards for double-strand coverage and sequence quality. Finishing involves closing the remaining gaps, resolving ambiguities, and validating the assembly. Automation of the finishing process makes use of Consed, Swedish, Finnish, web interfaces, and robotic workstations to increase efficiency and throughput. While these tools have had a significant impact on our productivity, additional tools and automation are still necessary to decrease the amount of human intervention required for finishing to meet the challenge of completing the Human Genome by 2003.

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

10. Sequence Validation and Quality Assessment at the Joint Genome Institute

M. Bussod, N. Doggett, J. Fawcett, D. Ricke, K. Watson, O. Tatum, P.S. White, and M. Mundt
Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545
mira@telomere.lanl.gov

The Joint Genome Institute (JGI) is committed to producing high quality finished sequence data with fewer that 1 error in 10,000 bases. To ensure that we meet these strict criteria the JGI prescribes to a quality control process which requires that greater than 95% of all finished bases have Phrap scores greater than 40 and at least 95% of all bases are covered in reads from both strands (or 2 chemistries). In addition to these quality control criteria, the JGI has implemented post-sequencing Validation and Quality Assessment processes, which occur in 2 phases within the Joint Genome Institute. Sequence Validation occurs at each sequencing site prior to the submission of a sequence and involves comparing the final assembled sequence to 3 independent high-resolution restriction fingerprints. This pre-submission Sequence Validation process ensures that the finished sequence has been assembled correctly. The Quality Assessment process is a post-submission assessment of the sequence produced by the JGI. LANL has the responsibility for performing this Quality Assessment process for all of the sequence produced by the Joint Genome Institute. During the summer of 1998 our group successfully completed one round of sequence quality assessment sponsored by the NIH of 600 kb of finished sequence from 3 NIH centers, and we have recently begun a second round of this assessment involving greater than 1.2 Mb of finished sequence from three centers including the Sanger Institute.

Our strategy for the Quality Assessment process is to identify the poorest quality regions within each finished clone and target these for verification. Software tools are being developed to evaluate the quality of clone sequencing projects based on Phred

and Phrap scores. In addition, the techniques used and software modules written can be applied to the task of choosing optimal targets for resequencing. Base calling and structural assembly errors can be identified by using PCR, for example, and sequencing if necessary. Determination of sequence error probability is based on the Phrap values of the consensus bases where each base is given a P-value, the probability of the base being incorrect, depending on its quality. If the data is given in the form of a histogram, the calculation of the probability values for each clone project is dependent on the proportion of bases within each quality range. We used this technique to find good candidates for our JGI validation effort without requiring the full set of quality values. However, if the Phrap value of every base is available, a more accurate prediction of error rate is possible. In this case, sliding windows of consecutive bases can also be evaluated to detect regions with higher error rates and design targets for resequencing. In either case, correction factors can also be applied to the error calculations to account for the supposed conservative nature of the Phrap scoring system. The approaches described above are among those being compiled into a set of Java tools whose uses extend beyond just validation. Finishing requirements often mirror the needs of a quality assessment project. Right now, we use a similar version of a Java filter around a Primer3-based program to select oligonucleotide sequences for both finishing primer walks and validation PCR primers. In the NIH QA exercise, our success rate for getting PCR products was about 85%, even though we targeted more difficult regions to sequence. We recently received over 130,000 BAC end sequences from two centers to evaluate the DOE-funded BAC end sequencing effort. Studying these should be a new, exciting challenge with great potential benefit to the sequencing community. Supported by USDOE under contract W-7405-ENG-36.

11. LANL Finishing Team Accomplishments in FY98

J. Buckingham, L. Goodwin, C. Munk, L. Saunders, S. Thompson, S. Ueng, D. Ricke, and M. Mundt
Joint Genome Institute, Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545
buck@telomere.lanl.gov

In response to the Joint Genome Institute's goal of finishing 20 MB of high quality sequence, the Finishing Team was formed at Los Alamos National Laboratory's Center for Human Genome Studies. The task for this diverse group of biologists, computer scientists and mathematicians was to design an efficient process to quickly close clone projects and bring the sequence quality up to a high standard. The tools to do this job were largely untested and disorganized, and many new protocols and strategies had to be formulated to address problems, even as the "conditions of contest" changed over the past year. Timely feedback to reduce unnecessary work was also an important factor to our success.

Initially, LANL's sequencing capacity was directed at the double-ended plasmid SASE approach, using TAQ and later TAQ-fs. Two major improvements were switches to BigDye terminator reactions for production and ET dye primer chemistry on ABI 373's for finishing. The boost in quality from these two process changes was quite evident using both our own base caller and Phred. Phred, Phrap, and Consed had not previously been used at LANL, so several technicians went through UNIX training to become specialists at interacting with these programs. Auxiliary Java programs were also designed to analyze Phrap assembly structure and to suggest finishing reactions consisting of dye primer redos and primer walks. A paper trail system was converted to an automatic submission system for our oligo synthesizers. Following the lead of the production crew, we now use halfTERM (Genpak,Ltd.) with BigDye for our primer walks. In addition, we are adding DMSO to the formulation to improve the reactions. We have also used shatter libraries successfully and transposons not so successfully to close final difficult gaps that exist due to, for example, high GC content. Part of our automation

schemes included the use of Hydras and multichannel pipettors for setting up finishing reactions.

We are now streamlining our approach to efficiently address the issues involved with "draft" sequencing. We have defined a prefinishing step based on our "Strand Gap" report that will also help evaluate cost functions to feed back to our production team to determine level of shotgun required. Research plans include trying halfTERM with dye primer reactions and working with the Mermade oligo synthesizer that should be delivered in the next few months. We are also investigating the potential benefits of programming robots to select templates for reaction set ups and weighing these against potential disadvantages such as reduced quality. We will present relevant statistics to demonstrate the quality of our finishing reactions and their utility in alignments to our final consensus. This work contributed to the completion of 2.8Mb of sequence in FY98. Supported by US DOE under contract W-7405-ENG-36.

12. JGI-LANL Sequencing Cost Reduction and Quality Improvement: R&D Results

Owatha L. "Tootie" Tatum and P. Scott White
Los Alamos National Laboratory, Los Alamos, New Mexico
tootie@telomere.lanl.gov

With recent dramatic increases in JGI's sequencing effort, the need to improve efficiency and reduce costs while maintaining high quality standards is of utmost importance. To this end, JGI and other large-scale genome sequencing facilities have recognized that an active research and development team is vital to their success. Aspects of LANL sequencing R&D goals include improvements in sequencing reactions - in the form of modifications of existing systems and investigation into and development of new sequencing technologies and automation systems. As part of the sequencing effort for the JGI, LANL has placed sequencing research and development as an important priority.

In efforts to reduce costs, several modifications of existing chemistries have been examined, resulting in striking reductions in cost with actual improvements in read length and sequence quality. The protocols resulting from these R&D efforts have been implemented in the LANL production sequencing and finishing efforts with great success. Sequence obtained from difficult templates (i.e. BAC DNA) has been improved dramatically as a result of chemistry R&D as well. While improvements in chemistry have had the most immediate impact on cost, LANL has also focused on quality control and automation issues to further streamline the sequencing process. Commercially available automation equipment has been implemented into the production process line with a considerable saving of technician hands-on time. In addition to time/cost savings, high throughput automated systems have also been implemented to improve quality control early in the sequencing process. All aspects of sequencing R&D conducted by Los Alamos to date will contribute to the work at Production Sequencing Facility and may be of interest to other large-scale sequencing facilities as well.

13. Concatenation cDNA Sequencing and Analysis of 500 Human Brain cDNA Clones

Wei Yu, John Bouck, James H. Gorrell, Donna M. Muzny, and **Richard A. Gibbs**
Human Genome Sequencing Center, Department of Molecular and Human Genetics Baylor College of Medicine, Houston, Texas 77030
agibbs@bcm.tmc.edu

Using a shotgun based strategy entitled Concatenation cDNA Sequencing (CCS), we have completed sequencing of 503 random selected cDNA clones with a total length of 807 kb from *Homo*

sapiens brain cDNA library (1NIB). All sequence data have been annotated and submitted to GenBank. The statistics from completed projects have shown that CCS is as efficient as sequencing of single large DNA fragment, and the reads/kb range from 13-21 with an average of 16.8 and the number of primers/kb ranges from 0.62-1.8 with an average of 1.02. Computer analysis was performed to search for the similarity against the public database. Of the 471 clone sequences used for DNA similarity searches, 255 (54%) were not matched to any sequences in the non-redundant database. The remaining 216 were matched to previously defined sequences or known genes from human to other organisms. Of the 471 clone sequences, 230 clones (48.9%) possess putative complete and incomplete open reading frames with a minimal length of 100 amino acids. When all 471 cDNA sequences were compared to the protein sequences in the database, 255 were not assigned definitely to any known protein. For the remaining 216 clones, 145 displayed similarities to previously deposited protein sequences, providing a consistent search result between nucleic and amino acid data from each clone. There were 71 clones that failed to reveal any protein match despite their corresponding DNA similarity matches with database entries. To determine the amount of unique information that our cDNA clone sequences were adding to the database, we examined the distribution of 243 clones which have been incorporated into the unigene database maintained by the NCBI. When the 243 cDNA sequences were compared to the representative sequences from the unigene database, we found 10 cDNA sequences contained weak matches to representative clone, but were not included in unigene clusters. Of the 233 clusters that were matched, nearly all of them contained multiple sequences in each cluster. But when the same 233 clone sequences were used to compare to mRNA/gene sequences in each cluster, 143 (61%) clusters contained only one single mRNA/gene sequence, which is our cDNA sequences. The majority of the cDNA clones were found in small clusters with only a few other mRNA or EST.

14. Cosmid Finishing and Full Insert cDNA Sequencing Using Differential Extension with Nucleotide Subsets (DENS)

D. Zevin-Sonkin[1,2], H. Hovhanissyan[1], A. Ghochikyan[1], L. Lvovsky[1], A. Liberzon[1], M.C. Raja[1,3], E. Ben-Asher[2], G. Glusman[2], D. Lancet[2], and **L.E. Ulanovsky**[1,3]
[1]Dept. of Structural Biology and [2]Dept. of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, ISRAEL; [3]CMB, Argonne National Laboratory, Argonne, IL 60439-4833
levy@anl.gov

Differential Extension with Nucleotide Subsets (DENS) 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 presynthesized 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.

We use DENS for cosmid finishing and have tested it for full insert cDNA sequencing. The templates for cosmid finishing by DENS (7-8 overlapping fragments, ~ 5 kb each) were PCR amplified from the cosmid. The PCR products were made single-stranded using Lambda Exonuclease (Exo-PCR). If one of the two primers in the PCR is phosphorylated, the Exo digestion leaves the opposite strand single-stranded. The 8-mer primers for DENS sequencing were selected using our dedicated software. The DENS approach resulted in approximately a three-fold reduction in cost and time of finishing compared to the strategy used before: additional shotguns combined with custom synthesized primer walking on the whole cosmid and/or PCR fragments.

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. In a pilot experiment we used DENS and "Exo-PCR" 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 one per 10,000. The plotted distribution showed that base-calls with PHRED values less than 20 occurred only 1% of the time.

Sequencing Technologies and Resources

15. Structural Analysis of the T7 DNA Replication System and Further Development of its Use in DNA Sequencing and Amplification

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

DNA polymerases play an essential role in current methods of DNA sequencing, which require the efficient synthesis of DNA using the four natural nucleotides as well as analogs such as fluorescently-labeled nucleotides and chain-terminating dideoxynucleotides. We have been characterizing the structure and function of DNA polymerases in order to modify those properties that are important for DNA sequencing. Our work has focused on the DNA polymerases of the Pol I family, that includes T7 DNA polymerase and Taq DNA polymerase. We, in collaboration with Sylvie Doublié and Thomas Ellenberger, recently determined the 2.2Å crystal structure of T7 DNA polymerase locked in a replicating complex with a dideoxy-terminated primer-template, an incoming dNTP, and the processivity factor thioredoxin[1]. We are using this structure to design and characterize mutations in the active site of T7 and Taq DNA polymerases that have altered specificity for analogs with modifications in the sugar moiety (e.g. dideoxynucleotides, ribonucleotides and 3' fluoro derivatives) and bases containing bulky fluorescent substituents.

One property that distinguishes T7 DNA polymerase from the thermophilic DNA polymerases used for DNA sequencing and amplification is its high processivity. This is achieved by the binding of its processivity factor, *E. coli* thioredoxin, to a unique 74 residue domain that acts as a flexible tether to keep the polymerase bound to DNA. While this domain is unique to T7 DNA polymerase, it is modular in that it can be transferred to other homologous polymerases by gene fusion to generate hybrid enzymes that have dramatically increased processivity[2]. The crystal structure of the T7 DNA polymerase complex suggests that thioredoxin is acting to stabilize the region it binds to, allowing a number of basic residues to interact electrostatically with the DNA backbone to prevent dissociation. We are characterizing mutations in this region in order to further define the critical structural features and to engineer new DNA polymerases that have increased processivity.

The complex of T7 DNA polymerase and T7 helicase/primase synthesize DNA with high efficiency. We have been optimizing reactions carried out by these enzymes in combination with other T7 replication proteins. Conditions have been developed in which DNA synthesis is exponential. Using one pg

of plasmid DNA as template, a 15 min reaction can produce 10 μg of product DNA, corresponding to a 10 million-fold amplification. DNA synthesis is nonspecific; the entire plasmid is replicated. We are exploring the use of this amplification reaction to produce BAC and plasmid DNA for use in DNA sequencing reactions. This in vitro synthesis of DNA may be an attractive alternative to the current methods that rely on in vivo production in bacterial cells for the automated preparation and purification of DNA templates.

This work is funded in part by DOE grant DE-FG02-96ER62251 (Stanley Tabor, P. I.)

[1] Crystal Structure of Bacteriophage T7 DNA Polymerase Complexed to a Primer-Template, a Nucleoside Triphosphate, and its Processivity Factor Thioredoxin. Sylvie Doublié, Stanley Tabor, Alexander Long, Charles C. Richardson and Tom Ellenberger, Nature, 391, 251-258 (1998).
[2] The Thioredoxin Binding Domain of Bacteriophage T7 DNA Polymerase Confers Processivity on *Escherichia coli* DNA Polymerase I. Ella Bedford, Stanley Tabor and Charles C. Richardson, Proc. Natl. Acad. Sci. USA 94, 479-484 (1997).

16. Mutagenesis and Reaction Condition Studies of T7 RNA Polymerase Variants to Incorporate Deoxynucleotides

Mark Knuth, Scott Lesley, Heath Klock, Michelle Mandrekar, Ryan Olson, James Schaefer, and Kris Zimmerman
Promega Corporation 2800 Woods Hollow Rd. Madison, WI 53711
mknuth@promega.com

Our aim is to alter substrate specificity in T7 RNA polymerase for efficient incorporation of dNTPs and other nucleotide analogs. Such a promoter-directed polymerase could be used for DNA sequencing and creating hybridization probes without the need for initiating primer. Previously described mutants incorporate dNTPs, but not sufficiently for practical applications. We are undertaking a combined approach of site-directed mutagenesis and evaluating reaction conditions to create an efficient polymerase with altered nucleotide specificity. Results are shown for both efforts.

For mutagenesis, we superimpose all possible substitutions of individual target sites upon the previously described mutations. Approximately 200 sites, covering a large portion of the active site, were chosen for saturation mutagenesis. Evaluation is underway, and desirable substitutions will be combined and shuffled. Our previous results indicate that dNTP incorporation is inhibited by inefficient transition from the initiation to the elongation phase. In order to screen for improvement of this property, polymerase is purified from each mutant and a fluorescent assay used to determine its ability to incorporate mixtures of r/dNTPs.

Solution conditions, such as addition of organic solvents, have been reported to enhance dNTP incorporation but result in substantial reduction of activity. We have confirmed and extended these results and note that these agents lower the apparent denaturation temperature of T7 RNAP. Mutations which increase thermostability under these conditions might offset the activity decrease, and a screen for these is being incorporated in our mutagenesis approach. We also evaluated other agents and found several which also enhanced dNTP incorporation but with a lesser thermodestabilization. Our results suggest a conformational change in the protein rather than the DNA template may be responsible for this enhancement. Using our optimal conditions, incorporation of a mixture of 3 dNTPs / 1 rATP is 300% of the 4 rNTP value. Although transcript products are somewhat shorter (most <= 200 bp) than for 4 rNTPs, we are examining their performance in dye-terminator DNA sequencing.

17. Megabase and Gigabase Templates: Direct Automated Sequencing off Microbial and Eukaryotic Chromosomal DNA

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

Combination of the robust dye terminator and ThermoFidelase chemistries has provided solution for the automated sequencing directly from microbial Megabase-long templates. Novel approach streamlines gap closure in the large-scale projects and does not rely on extra subcloning or combinatorial PCR. Redesign of genome sequencing and gene hunting strategies promises to substantially reduce the volume of and eventually completely eliminate shotgun steps in microbial genome projects.

Our work indicates that Megabase sequencing chemistry provides results with high quality, sensitivity, reproducibility and speed. We will present data on the direct detection of single nucleotide polymorphisms (SNPs) and gross sequence variations between genomic regions from a number of closely and distantly related microorganisms. We will discuss technical and economic feasibility of the new strategy in large-scale projects on comparative genomics.

The development of sequencing chemistry for Gigabase templates such as human and other complex eukaryotic genomes is one of the most challenging tasks in technology development. Our initial work is focused on the specifics of preparation and handling of Gigabase templates, target selection, achievement of sufficient fluorescent signal strength and quality. We will review basic techniques used in the development of novel chemistry and present our first successful results on the automated sequencing directly from fish and human chromosomal DNA.

This work is supported in part by DOE grant DE-FG02-98ER82557 and NIH grant 2R44GM55485-02.

18. PCR Using Branched Modular Primers

Maura M. Devine, Mugasimangalam C. Raja, and **Levy E. Ulanovsky**
CMB, Argonne National Laboratory, Argonne, IL 60439-4833
levy@anl.gov

Here we present a novel PCR technique termed "branched primer PCR" which eliminates the need for custom primer synthesis by combining oligonucleotide modules selected from a pre-synthesized library. A branched primer involves two oligonucleotide modules that are physically linked by annealing to each other as well as to the target, forming a three-way junction. Branched primers were developed (initially for DNA sequencing rather than for PCR) as a type of modular primer whose modules anneal cooperatively to the template. This cooperativity is provided by mutually complementary segments in the two modules that bind to each other forming what is termed a "stem" region. Before actual PCR can take place, a branched primer is extended along the template. This extension strand is then used as the template for a reverse branched primer extension. The reverse extension product is then amplified using PCR primers homologous to the stems of each branched primer. These PCR primers are universal in that the stem sequence is the same in different branched primers. In contrast, the sequences of the stretches which are complementary to the template are variable throughout the presynthesized library of the oligonucleotide modules (each in a separate tube). Additional sequence-specificity of PCR is provided by nesting. Branched primer PCR is expected to be useful for applications such as resequencing closely related genomes (e.g. rodents and primates) which require a huge number of custom PCR primers. The

latter would then be conveniently replaced with a much smaller library of presynthesized oligonucleotide modules for branched primers (2,000 to 4,000 modules instead of millions of custom PCR primers).

19. Synthesis, Characterization, and Potential Applications of Biotinylated Energy Transfer Oligonucleotides

Jin Xie[1], Richard A. Mathies[2], and Alexander N. Glazer[1]
Departments of [1]Molecular and Cell Biology and [2]Chemistry, University of California, Berkeley, CA 94720
Glazer@uclink4.berkeley.edu

Energy transfer (ET) fluorescent dye-labeled primers have provided a decadic improvement in the performance of DNA sequencers for high-throughput sequencing[1,2]. The acceptor emissions of high spectral purity also make ET primers ideal for diagnostic applications, such as forensic identification and genetic typing of short tandem repeats[3]. Biotin has an extraordinarily high affinity for streptavidin with a reported dissociation constant of ~10-15. This very strong binding affinity has made the biotin-streptavidin system very attractive for a multitude of in vitro labeling applications. We describe here the synthesis and characterization of biotinylated fluorescent ET reagents. Hung et al. have shown that CYA-ROX primers with a donor-acceptor spacing of 8-10 nucleotides offer excellent acceptor emission intensities coupled with negligible donor emissions[4-7]. We have synthesized oligonucleotides with the sequence 5'-CYA-NNNNNNNNNTROXNNTBNNNNNNN-3' with donor-acceptor fluorophore pairs separated by 10 intervening nucleobases, but varying in the location of TB, in this example introduced two bases 3' to the base carrying the acceptor ROX. Biotin-labeled T (TB) was introduced by the use of biotin-dT phosphoramidite at different locations in the oligonucleotides. CYA, 3-(e-carboxypentyl)-3'-ethyl-5,5'-dimethyloxacarbo-cyanine, a dye with a high absorption cross-section but a low fluorescence quantum yield, was chosen as an energy donor at the 5'-end of the oligonucleotides, and ROX as an acceptor was attached to a modified thymidine (TROX). We have compared the quantitative spectroscopic properties of four biotinylated ET reagents differing in the spacing between donor-biotin pairs and acceptor-biotin pairs. CYA10ROX-2-Biotin (where 2 is the number of nucleotides between the acceptor and biotin) reagent offers the best combination of acceptor fluorescence emission intensity and spectral purity. With 488-nm excitation, the fluorescence emission intensity of C10R-2-Biotin is 16-fold stronger than that of the corresponding oligonucleotide labeled with the acceptor ROX as the only dye. These biotinylated ET reagents have a broad range of potential applications, e.g., affinity purification and detection in DNA mapping applications on chips, and in cell sorting[8]. For such purposes, we have prepared and characterized ET-oligonucleotide-streptavidin conjugates for use in multiplexed assay systems.

[1]J. Ju, A.N. Glazer, and R.A. Mathies Nature Medicine 2, 246-249 (1996).
[2]A.N. Glazer and R.A. Mathies Curr. Opinion Biotechnol. 8, 94-102 (1997).
[3]Y. Wang, S-C. Hung, J.F. Linn, G. Steiner, A.N. Glazer, D. Sidransky, and R.A. Mathies Electrophoresis 18, 1742-1749 (1997).
[4]S-C. Hung, J. Ju, R.A. Mathies, and A.N. Glazer Anal. Biochem. 243, 15-27 (1997).
[5]S-C. Hung, J. Ju, R.A. Mathies, and A.N. Glazer Anal. Biochem. 238, 165-170 (1996)
[6]S-C. Hung, R.A. Mathies, and A.N. Glazer Anal. Biochem. 252, 78-88 (1997).
[7]S-C. Hung, R.A. Mathies, and A.N. Glazer Anal. Biochem. 255, 32-38 (1998).
[8]See abstract "Integrated Sequencing Sample Preparation on CE Microplate" by Y. Shi, I. Kheterpal, J. Xie, A.N. Glazer and R.A. Mathies.

20. Development of a Multilabel DNA Mapping Technique Using SERS Gene Probes

Tuan Vo-Dinh[1], David L. Stokes[1], Guy D. Griffin[1], Jean-Pierre Alarie[1], Edward J. Michaud[1], Terry Bunde[1], Ung-Jin Kim[2], Melvin I. Simon[2]
[1]Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6101, USA

[2]Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA
tvo@ornl.gov

We report the development of a novel approach for use in DNA mapping and bacterial artificial chromosomes (BAC) colony hybridization using a unique type of DNA gene probe based on surface-enhanced Raman scattering (SERS) labels. An important step toward sequencing the human genome involves assembling ordered, overlapping sets (contigs) of clones that have been mapped and well characterized. A unique approach that would greatly facilitate large-scale genomic sequencing involves building genome-wide BAC-based contig maps.

In this work, we have developed various types of SERS-active substrates that can be used to provide this "label-multiplex" capability, thereby reducing the time for genome characterization. We have developed various schemes for binding SERS labels onto DNA targets for use in BAC clone mapping. We have demonstrated the feasibility of a multi-label SERS detection scheme, whereby multiple labels can be detected simultaneously in each multiplex probing cycle. Raman spectroscopy is an important analytical tool due to its excellent specificity for chemical group identification. With the use of the SERS effect, Raman scattering efficiency can be enhanced by factors of up to 108 when a compound is adsorbed on or near special metal substrates[1]. The surface-enhanced Raman gene (SERGen) probes do not require the use of radioactive labels and have a great potential to provide both sensitivity and selectivity for DNA mapping and sequencing. The method is aimed at simultaneous detection of multiple probes for DNA mapping/sequencing using BAC clone applications. The technology is designed to be versatile and broad-based, and to allow a rapid and shortcut approach to significantly improve the speed of BAC colony hybridization. cDNAs are an excellent resource to rapidly build genome wide BAC contigs. They represent inexpensive, efficient probes to screen BAC libraries by colony hybridization. However, the conventional approach relying on 32P-labeled probes is laborious and time consuming. Multiplexing probes with non-radiative chemicals that can be efficiently distinguished after hybridization will greatly reduce the time and effort for establishing the large-scale BAC library that is required for characterizing the entire genome of human, mouse and other organisms.

ACKNOWLEDGMENTS
This research is sponsored by the Office of Biological and Environmental Research, U.S. Department of Energy under contract DE AC05-960R22464 with Lockheed Martin Energy Research Corporation.

REFERENCES
[1]T. Vo-Dinh, "Surface-enhanced Raman spectroscopy using metallic nanostructures" in Trends in Analytical Chemistry, 17, 557 (1998)

21. Vectors for Using Nested Deletions to Sequence Either Strand of Cloned DNA

John J. Dunn, Laura Praissman, Laura-Li Butler-Loffredo, John J. McNulty, and F. William Studier
Biology Department, Brookhaven National Laboratory, Upton, NY 11973-5000
jdunn@bnl.gov

Regions of highly repeated DNA are encountered frequently in human DNA and are likely to be particularly troublesome near centromeres and telomeres. Highly repeated regions are difficult to assemble correctly by shotgun sequencing, but cloned fragments at least 10 kilobase pairs long can be sequenced and assembled easily by generating an ordered set of nested deletions whose ends are separated by less than the length of sequence read from a single priming site within the adjacent vector. Assembly of the overlapping sequences is guided by knowledge of the relative length of the portion of the fragment remaining in the clone, as determined by gel electrophoresis.

We have made a series of plasmid vectors, the pZIP series, which allow rapid generation of an ordered set of nested deletions from either strand of a cloned DNA fragment. The vectors are based on the low-copy F replicon. The size of the vector DNA has been reduced to the 4.5-kbp range by removing the 2.5-kbp sop (stability of plasmid genes) region. The resulting plasmids have the low copy number typical of F plasmids and remain stable enough to be easily maintained by growth in the presence of kanamycin, the selective antibiotic. DNA in amounts convenient for sequencing is readily obtained by amplification from an IPTG-inducible P1 lytic replicon.

Nested deletions are generated by cleavage near one end of the cloned fragment, using commercially available site-specific endonucleases (PI-PspI, I-CeuI or I-SceI) whose recognition sites span 18-30 bp and are therefore unlikely to occur in cloned DNA fragments. Cleavage by these nucleases generates four-base 3' overhangs that are resistant to digestion by *E. coli* exonuclease III. A second cleavage by one of several nucleases with 8-base recognition sites leaves the end adjacent to the cloned fragment susceptible to ExoIII digestion, permitting unidirectional 3' to 5' digestion across the cloned fragment. The resulting single-strand tails are digested with S1 nuclease, the ends are repaired and ligated with T4 DNA polymerase and ligase, and clones are obtained by electroporation. An range of digestion times with ExoIII can easily produce a distribution of deletion lengths extending across the entire cloned fragment. Cleavage sites for the site-specific endonucleases are positioned in the vectors so that nested deletions can be generated from either end of an individual cloned fragment.

Conditions for routinely generating ordered sets of nested deletions and using them to sequence both strands of cloned fragments in the 5-kbp to 15-kbp size range are being developed by sequencing fragments of human DNA from BACs.

22. Direct Conversion of PCR Products into Bidirectional Sequencing Fragments

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

The search for more efficient and direct approaches to genomic sequencing continues to gain attention, particularly in gap-filling, finishing, and diagnostic applications. We have 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 described[1]) or chemically. Both approaches take advantage of differences in reactivity of the normal and modified 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 boranophosphate nucleoside. However, normal and boranophosphate 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. The 5-ethyl-alpha-borano-dCTP analog was found to exhibit an increased resistance to exonuclease III compared to the alpha-borano-dCTP used previously

in our method, without affecting incorporation, and resulted in more even banding patterns. 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.

We are also developing a chemical method for generating sequencing fragments, as an alternative to exonuclease chew-back. In the chemical approach, we have identified reagents that selectively cleave the backbone of the PCR product at boranophosphate linkages, while leaving the normal phosphodiester linkages intact. We anticipate that chemical cleavage following incorporation of fluorescently labeled borano-dNTPs may result in a more efficient method of sequencing. Also under investigation are agents that can result in colorimetric detection of boranophosphate.

Direct sequencing of PCR products simplifies mono- and bidirectional sequencing and provides 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).

23. Analysis of Gradients of Polymer Concentration or Ionic Strength

Mark A. Quesada, David J. Fisk, and F. William Studier
Biology Department, Brookhaven National Laboratory, Upton, NY 11973
quesada@bnl.gov

We are investigating whether gradients of polymer concentration or ionic strength can extend read lengths into the 1000-2000 base range when analyzing DNA sequencing reactions by capillary electrophoresis. Longer read lengths would increase sequencing efficiency and reduce the effort needed in the assembly and finishing stages of genome sequencing.

Gradients of polymer concentration or ionic strength along the length of the capillary are generated by merging two solutions in the capillary, using programmable syringe pumps. Parameters important for obtaining reproducible gradients were identified and controlled with the aid of fluorescent dye to analyze the distribution of one of the solutions along the length of the capillary. Because entangled polymer solutions have high viscosity, balancing hydrostatic conductance at the junction between the merging solutions is critical for producing well defined, reproducible gradients. A smooth gradient with uniform composition in the radial direction at each capillary cross section is established by radial diffusion, primarily of water and other low molecular weight components within the capillary (polymer swelling). Production of reproducible gradients requires merging the solutions in a controlled way at a sufficiently low flow rate, and allowing sufficient time for diffusion to create a smooth and uniform gradient before the capillary is used for analysis.

Once conditions were established for preparing reproducible and useful gradients of polymer concentration in capillaries, the effects of different gradient configurations on read length were examined. We expected that a gradient of increasing polymer concentration could counter the peak-broadening effects that are responsible for loss of resolution at long read lengths, and this appears to be the case. Certain gradient configurations yield single-base resolution near 800 bases in standard, room-temperature analyses with separations continuing well beyond 1000 bases. Gradients of increasing ionic strength (salt concentration) might be expected to have a similar band sharpening effect,

and we are beginning to explore resolution in different combinations of polymer and salt gradients.

If the band sharpening effects of polymer or salt gradients can increase resolution and extend DNA read lengths in capillary electrophoresis, the same concepts may also find application for improving the performance of microchannel systems and disposable chips.

24. Design and Assembly of a Turnkey, High Throughput Oligonucleotide Synthesis Facility for Use on the Human Genome Project

J. Shawn Roach and Harold R. Garner
Center for Biomedical Invention, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd. NB11.102B, Dallas, TX 75235-8573
roach@ryburn.swmed.edu

The objective of this project was to design and assemble a highly automated, high throughput oligonucleotide synthesis facility that requires as little operator intervention as is practical. Each of the pre-processing and post-processing steps required for high throughput oligosynthesis were examined for opportunities to streamline and automate if practical. These steps include 1) loading the solid supports into the filter plates prior to synthesis, 2) cleaving the oligos from the solid supports post synthesis, 3) rapid evaporation of the residual solutions after chemical deprotection, 4) optical density evaluation of the oligo concentration after resuspension, 5) gel electrophoresis evaluation of the oligo quality, and 6) automated dilution of the oligos to a normalized concentration after resuspension.

The heart of the system is two MerMade high throughput oligonucleotide synthesizers that are each capable of synthesizing up to 192 oligos a day. The oligos are synthesized on solid supports loaded into two 96-well filter plates. Standard phosphoramidite chemistry is used to perform the synthesis. The MerMade was designed by Dr. Harold R. Garner and Dr. Simon Rayner of the University of Texas Southwestern Medical Center. The MerMades used on this project were built by Avantech Automation Corporation (New Braunfels, TX) according to the Garner-Rayner design with minor modifications. Additional devices in the facility include a Biomek 2000 Workstation (Beckman Instruments, Fullerton, CA) for performing several of the liquid transfer and filtration steps necessary, a Jetstream evaporator and Scirocco gas heater system (Helix Scientific, Warminster, PA) for rapid sample concentration, and a Spectramax 190 Plus UV-Vis spectrophotometer and 96-well plate reader (Molecular Devices, Sunnyvale, CA) for optical density analysis.

Data from oligonucleotides synthesized by the facility will be presented along with background information on the turnkey synthesis facility.

25. Prep Track I - A Dynamic Approach to Liquid Handing Robotics

D. Humphries, M. Pollard, J. Bercovitz, C. Reiter, and B. Gray
Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California
DEHumphries@lbl.gov

Prep Track I is a modular liquid handling robot that was developed at LBNL for the Human Genome Project. This machine has been in productive operation for more than a year. The basic function of this machine is to process solutions in micro-titer plates that are removed from input cassettes and fed to two conveyor belts where various liquid processes are performed by 96 syringe multi-dispensers or hydra heads. After processing, the plates are automatically loaded into output cassettes. The hydra heads of each module have access to automated wash and rinse baths and, for some modules, a cold plate and rinse bath system. Production activity on Prep Track I has indicated a need for increased capability and flexibility in the physical infrastructure of the device. Towards this end, innovations from the newer Prep Track II are being applied to Prep Track I to allow a flexible three bath system with an interchangeable cold plate arrangement. A programmable valve manifold and multiple reservoirs will allow rapid reconfiguration of the bath/supply system by simply changing protocols and in some cases switching the physical order of baths and cold

plates. The detail design process for Prep Track I is in turn feeding forward into Prep Track II to increase its capability and flexibility.

26. PrepTrack II Design: Lessons Learned from PrepTrack I

John Bercovitz, Martin Pollard, David Humphries, Mario Cepeda, Charlie Reiter, and Bruce Gray
Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California
JHBercovitz@lbl.gov

The PrepTrack is an automated high-throughput microtiter plate liquid handling machine. It is a linear modular system in which microtiter plates are transported from module to module by means of a conveyor belt. The instrument is designed to make automated efficient use of the Robbins Scientific pipetting head (96 or 384 syringes). All of the modules can operate simultaneously to perform liquid handling operations on individual microtiter plates or liquid transfers between microtiter plates. Each pipetting module is equipped with automated self-cleaning water and bleach baths. Efficiencies are realized by the parallelism of the Robbins pipetting head, the parallelism of multiple modules operating simultaneously, and the ability to continue pipetting operations on available modules while other modules are occupied with self-cleaning procedures.

The first PrepTrack machine has been in production use for one year. We have since completed a second PrepTrack machine. There were several design changes based on lessons learned from use of PrepTrack I. The design changes and reasoning behind the new designs will be discussed.

27. Adapting the Tecan Genesis 2 Meter Workstation for High Density Agarose Gel Loading

Linda Sindelar, John Bercovitz, Mario Cepeda, and David Humphries
Joint Genome Institute, Lawrence Berkeley Laboratory, Berkeley, California
LESindelar@lbl.gov

We have modified the Tecan Genesis Workstation to load high density agarose gels that are used for transposon mapping. Each Gel contains 210 wells that are 1mm wide and have a center to center spacing of 2.25mm. To achieve the positional accuracy required for this application we have designed modified pipetting tips and elevated work decks that hold four gels, source microtiter plates, markers, and a custom tip calibration fixture. We have also designed and fabricated casting trays and combs which precisely locate the gels on the work deck. Four gels are loaded with buffer and sample in 32 minutes. The application was written in the Tecan Logic Software.

28. Technology Development for the Human Genome Project

Chris Robinson[1], Todd Brooks[1], Travis Crane[1], Chris Elkin[2] and **Trevor L. Hawkins**[1,2]
[1]College of Medicine, University of Florida, Gainesville, Florida and [2]CuraGen Corporation, Gainesville, Florida
thawkins@fl.curagen.com

We are continuing our work on integrated robotic systems and technology development to aid the genome program. The integrated robotic approaches follow on from our development of the Sequatron robotic systems some years ago. Now, we are focusing on reducing volumes of the amplification and sequencing reactions as well as using higher density plates to perform these reactions. This all

requires the development of new or modified hardware, such as thermal cyclers, which once developed will be modules for a new fully integrated system.

We are also working on improvements to the existing bottlenecks in high throughput DNA sequencing. One has been the development of a very low cost adaptation to the ABI 377 system that allows 96 lane gels to be run with the same results as found with the commercially available upgrade. Another, is the automated loading and pre running of ABI 377 gels for use in a high throughput facility.

Lastly, we are exploring the use of MALDI Mass spectrometry as a tool for the analysis of DNA extension products, specifically the resolution of compressions and error detection in genomic sequencing projects.

29. Automation for High Throughput Genomic DNA Sequencing

Ronald W. Davis
Biochemistry Department, Stanford University, Stanford, California
princed@leland.stanford.edu

Stanford has developed several devices that can be used as elements in a high throughput sequencing environment. Among the instruments in development are thermal cycler and plasmid purification devices. Stanford is working with the Joint Genome Institute on the incorporation of selected instruments into their production environment.

30. Co-Development of High Throughput Sequencing Systems with the Joint Genome Institute

Eric Lander
Whitehead Institute, Cambridge, Massachusetts
lander@wi.mit.edu

The Joint Genome Institute (JGI) and Whitehead Institute will establish a Co-Development Program to produce an automated DNA sequencing production line with a capacity of 200 Mb per year. The production line will consist of:

- Automated devices for sampling processing. The devices are based on existing systems used at Whitehead, but are comprehensively re-designed and re-engineered for the requirements of a factory production line. The design involves a universal "base system" that is customized for four specific applications. The devices will be constructed by Intelligent Automation Systems (IAS), with which Whitehead worked successfully on the construction of its Genomatron system.

- Informatics system. The accompanying informatics system will consist of a comprehensive database, workflow pipeline, and analytical software.

Production lines will be installed at both JGI and Whitehead. The Co-Development Program will implement, evaluate, and modify the production line. The evaluation will include using the system to sequence a total of 20 Mb of genomic sequence from human chromosome 19, consisting of 10 Mb at Whitehead and 10 Mb at JGI.

31. Laboratory Automation for Finish Sequencing at LLNL

Stephan Trong, Arthur Kobayashi, David J. Ow, Matt P. Nolan, Tom Slezak, Stephanie A. Stilwagen, Glenda G. Quan, and Jane Lamerdin
Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California
trong1@llnl.gov

The Human Genome Center at LLNL is performing high throughput DNA sequencing of the human genome. In the past year, we have contributed over 8.6 MB of high-quality finished sequencing to the Joint Genome Institute's total of 20.9 MB. This ramp represents an increase of more than 500% over the amount finished by LLNL last year (1.5 MB). One of the contributing factors in achieving this goal was the

automation of sample processing through the use of robotic workstations.

In the finishing phase, we are currently processing an average of 9,000 samples per month with expectations of a 25-50% increase in the coming year. To meet this high volume of sample processing, we have employed the use of Tecan Genesis 150 liquid handling robots to rearray DNA templates into 96-well plates and for setting up sequencing reactions for the rearrayed clones/templates. To integrate this process with our sample-tracking system, we have developed a web-based system to perform the following functions:

- Finish/Pre-finish clone request and batching for rearray.
- Oligonucleotide ordering and DNA template request and batching for rearray.
- Clone resubmission and rearraying from on-line pcr gel images.
- Shatter library request and batching.
- Transposon bombing request and batching.

To meet our aggressive ramp over the next few years, we will continue to expand our automation effort and add new functions to our system including automating the rearray of custom oligos for pcr and end-sequencing reactions, processing samples in 384-well plates and automating plate handling using robotics.

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

32. Sheath-Flow Capillary Array DNA Sequencer Development at JGI/LBNL

Jian Jin, William F. Kolbe, Yunian Lou, Earl W. Cornell, Alex Cheung, and Joseph M. Jaklevic
Ernest Orlando Lawrence Berkeley National Laboratory, University of California, Engineering Science Department, 1 Cyclotron Road, Berkeley, CA 94720
Jian_Jin@lbl.gov

We have developed a 96-channel capillary electrophoresis system capable of production-level sequencing at increased rates and, more importantly, with improved automation. The system is based on an adaptation of the best available technology developed by several laboratories. In particular, we employ the sheath-flow excitation/detection geometry and a DNA sequencing protocol using linear polyacrylamide as sieving media. In addition, we have developed an effective base-calling software platform using a combination of algorithms. The sequencing system is fully integrated and includes a fixture for off-line capillary-array coating, gel-replacement and sample loading. The four-color sequencing instrument employs a cooled-CCD camera for data acquisition. Our custom base-calling software has been fully integrated with existing assembly algorithms including calibrated Phred scores and Phrap-based assembly. Currently we achieve a total run time of less than two hours, separating 750 bases per channel, with an average read-length (defined as having a Phred score > 20) of 350-400 bases/trace. The turn-around time between runs is less than 5 minutes. Over the past 9 months we have conducted a full evaluation of the system using sequencing templates taken directly from JGI production runs. During this time we generated more than 6 Mb of raw sequence data for system performance evaluation and protocol development. Those results have demonstrated that our system produces sequencing data with a quality comparable to commercial slab-gel systems currently employed in production sequencing. Detailed comparison data will be presented and future plans and discussed.

This work was supported by the Director, Office of Energy Research, Human Genome Program, of the U.S. Department of Energy under Contract N0.DE-AC03-76SF00098

33. Fully Automated DNA Sequencing with a Commercial 96-Capillary Array Instrument

Qingbo Li, Thomas E. Kane, Changsheng Liu, Harry Zhao, Robert Fields, and John Kernan
SpectruMedix Corp., 2124 Old Gatesburg Rd., State College, PA 16803
qbli@spectrumedix.com

A commercial high-throughput DNA sequencer has been developed based on a robust multiplexed 96-capillary electrophoresis system and a high-performance replaceable gel matrix. The instrument is fully automated with all the operation steps carried out and controlled by the instrument computer. The detector system employs on-column laser-induced fluorescence detection with an air-cooled argon ion laser as the excitation source. The instrument allows automatic processing of many 96-well sample trays without human intervention.

Including the time for sample introduction, separation, and capillary reconditioning, the instrument is capable of one complete run within two hours. The sequencing throughput of half million bases per day is readily achievable using this 96-capillary instrument. The instrument is also compatible with other DNA fragment analysis. With the experience of successfully building the 96-capillary instrument, the team is taking a further step toward developing a 384-capillary instrument.

Testing results will be presented for a fully functional 96-capillary instrument. Performance data of the instrument will be discussed.

34. Automation and Integration of Multiplexed On-Line Sample Preparation with Capillary Electrophoresis for High-Throughput DNA Sequencing

Edward S. Yeung, Hongdong Tan, and Nanyan Zhang
Ames Laboratory, Ames, Iowa
yeung@ameslab.gov

An integrated and multiplexed on-line instrument starting from DNA templates to their primary sequences has been demonstrated based on multiplexed microfluidics and capillary array electrophoresis. The instrument automatically processes 8 templates through reaction, purification, denaturation, preconcentration, injection, separation and detection in a parallel fashion. A multiplexed freeze/thaw switching principle and a distribution network were utilized to manage flow and sample transportation. Dye-labeled terminator cycle-sequencing reactions are performed in an 8-capillary array in a hot-air thermal cycler. Subsequently, the sequencing ladders are directly loaded into separate size exclusion chromatographic columns operated at ~60 °C for purification. On-line denaturation and stacking injection for capillary electrophoresis is simultaneously accomplished at a cross assembly set at ~70 °C. Not only the separation capillary array but also the reaction capillary array and purification columns can be regenerated after every run. The raw data allow base calling up to 460 bp with an accuracy of 98%. The system is scalable to a 96-capillary array and will benefit not only high-speed, high-throughput DNA sequencing but also genetic typing.

An automated and integrated system for DNA typing directly from blood samples has been developed. The multiplexed eight-array system is based on capillary microfluidics and capillary array electrophoresis. Three short-tandem-repeat loci, vWA, THO1 and TPOX, are co-amplified simultaneously in a fused-silica capillary by a hot-air thermocycler. Blood is directly used as the template for polymerase chain reaction. Modifications of standard protocols are necessary for direct PCR from blood. A programmable syringe pump plus a set of

multiplexed liquid nitrogen freeze/thaw switching valves are employed for liquid handling in the fluid distribution network. The system fully integrates sample loading, PCR, addition of an absolute standard, on-line injection of sample and standards, separation and detection. The genotypes from blood samples can be clearly identified in eight parallel channels when the electropherograms are compared with that of the standard allelic ladder by itself. Regeneration and cleaning of the entire system prior to subsequent runs are also integrated into the instrument. The system can be expanded to hundreds of capillaries to achieve even higher throughput.

35. Long-Read DNA Sequencing by Capillary Array Electrophoresis

Oscar Salas-Solano, Lev Kotler, Zoran Sosic, Arthur W. Miller, Yongwu Yang, Haihong Zhou, and **Barry L. Karger**
Barnett Institute and Department of Chemistry, Northeastern University, 360 Huntington Avenue, Boston, MA 02115
bakarger@lynx.neu.edu

The increasing prominence of capillary array electrophoresis for DNA sequencing raises the importance of being able to obtain long reads on a regular basis with such instruments. We have recently reported the use of capillary electrophoresis (CE) for routine DNA sequencing of 1000 bases in less than one hour using replaceable linear polyacrylamide solutions (Salas-Solano et al., Anal. Chem. 1998, 70, 3996-4003). These results have now been extended to a multiple-capillary array. Cycle-sequencing reactions, and most steps of subsequent sample purification, are performed in 96-well microtiter plates with a system built around a Biomek 2000 robot. Before each run, the array of polyvinyl alcohol-coated capillaries is refilled with linear polyacrylamide solution. The 488 nm output of an argon laser is directed into an optic fiber, and then into a line generator, which focuses the beam into a 35-micrometer wide line extending across the entire bank of capillaries. The emitted dye fluorescence passes through notch filters, and through a transmission grating for spectral dispersion, and is imaged onto a CCD that is read at 3 Hz. This design has no moving parts and is easy to align. Base-calling is done by an expert system (see separate abstract of A. W. Miller and B. L. Karger), and read lengths of 1000 bases and above at 98-99% accuracy are routinely obtained. We will also report on our latest results for achieving long read length sequencing using capillary electrophoresis with replaceable polymer solutions.

This work is being supported by DOE grant DE-FG02-98ER 69895.

36. DNA Sequencing Using Capillary Array Electrophoresis

Indu Kheterpal[1], Gary T. Wedemeyer[1], Yuping Cai[2], Alexander N. Glazer[2], and Richard A. Mathies[1]
[1]Departments of Chemistry and [2]Molecular and Cellular Biology, University of California, Berkeley, CA 94720
indu@zinc.cchem.berkeley.edu

Capillary array electrophoresis has emerged as a valuable tool for DNA analysis. We are developing methods for obtaining high quality sequencing separations using confocal fluorescence CAE instruments[1,2] and energy transfer (ET) primers[3]. In practice replaceable separation matrices and base calling programs have been evaluated and optimized for high throughput sequencing separations of genomic DNA fragments from the cyanobacterium *Anabaena*.

We have evaluated the available replaceable gels using the same sequencing samples, temperature, detection system, injection and separation conditions. These gels can be pumped into the capillaries allowing the use of capillaries for potentially 100 runs. The three gels evaluated were linear polyacrylamide (LPA), hydroxyethylcellulose (HEC)

and a mixture of HEC and polyethylene oxide (PEO). We have found LPA to provide the best sequencing separations with the longest read lengths of 1000 bases in the least amount of time. We have also evaluated several base calling packages for their ease-of-use, ability to batch process and base-calling performance. BaseFinder[4] has emerged as the leading program for our data and is being used for all of the sequencing data analysis.

We are currently validating methods by incorporating them into our *Anabaena* sequencing project performed by undergraduates. Several 6-15 kbp libraries of *Anabaena* genome potentially involved in the biosynthesis and control of phycobiliproteins have been constructed. The templates for sequencing are prepared by partial digestion of the genomic DNA utilizing restriction enzymes cocktails. Fragments in ~0.5 kb range are cloned into pUC 19 for bidirectional sequencing. The sequencing samples are generated using the Sanger dideoxy method and cycle sequencing. The separations are performed using our planar CAE instruments[2] and replaceable gels. The data are analyzed using BaseFinder and assembled using Phred, Phrap and Consed. Two libraries p69 and p74 have now been completely assembled and sequencing on five other libraries is near completion. These libraries total over 70,000 bases and the fragments are being sequenced with ~5-fold redundancy to ensure complete and accurate assembly.

[1] Huang, X. C.; Mathies, R. A. (1992) Nature (London), 359, 167-169.
[2] Kheterpal, I.; Scherer, J. R.; Clark, S. M.; Radhakrishnan, A; Ju, J.; Ginther, C. L.; Sensabaugh, G. F. and Mathies, R. A. (1996) Electrophoresis, 17, 1852-1859.
[3] Ju, J.; Glazer, A. N.; Mathies, R. A. (1996) Nature Medicine, 2, 246-249.
[4] Giddings, M. C.; Severin, J; Westphall, M; Wu, J. Z; and Smith, L. M. (1998) Genome Research 8, 644-665.

37. Focused Single Molecule DNA Detection in Microfabricated Capillary Electrophoresis Chips

Brian B. Haab and **Richard A. Mathies**
Department of Chemistry, University of California, Berkeley, CA 94720
rich@zinc.cchem.berkeley.edu

Single-molecule fluorescence burst counting is a highly sensitive method for detecting electrophoretic separations of ds-DNA fragments[1] with applications in environmental monitoring and health care diagnostics. We previously presented methods for optimizing dye labeling, laser power and data analysis, and conventional CE separations of ds DNA fragments in the 100-1000 bp range were detectable when only 50-100 molecules passed through the probe volume.[2] We have now performed single DNA molecule detection in glass capillary electrophoresis (CE) chips which offer improved optics, faster separations, and increased molecular detection efficiency compared to conventional capillaries.[3] Chips were fabricated with a 145 mm thick top plate that was matched to the design specifications of the 100X, 1.3 NA objective, yielding a two-fold increase in light collection efficiency. The channels were designed to focus a greater number of molecules through the laser beam to achieve enhanced detection sensitivity. The sample was constricted in the region of the 1 mm diameter focused laser beam by physical narrowing of the separation channel and by electrokinetic focusing caused by additional side channels in the detection region. The sample stream width decreased and the single molecule count rate increased linearly with the focusing current density. A four-fold improvement in molecular detection efficiency was achieved while maintaining single molecule sensitivity. The CE separation of a 500 bp PCR product was then detected using molecular focusing, which showed a two-fold increase in signal compared with conventional detection. A 300 fM sample was easily detectable with a signal-to-noise ratio of eight. These developments will enhance our ability to use CE separations to detect trace pathogen contamination or DNA mutation.

[1] B. B. Haab and R. A. Mathies, Anal. Chem. 34, 3253-3260 (1995)

[2]B. B. Haab and R. A. Mathies, Appl. Spec. 51, 1579-1584 (1997)
[3]B. B. Haab and R. A. Mathies, Proc. SPIE 3259, 104-112 (1998)

38. Ultra-High Throughput DNA Genotyping and Sequencing on Radial Capillary Array Electrophoresis Microplates

Peter C. Simpson, James R. Scherer, Yining Shi, and Richard A. Mathies
University of California, Berkeley, CA 94720
peter@zinc.cchem.berkeley.edu

The microfabrication of DNA sample preparation, electrophoretic separation and detection devices is making possible a new generation of high-speed, high-throughput DNA analysis systems. Our research is focused on the ultra-high throughput analysis of PCR products for genotyping applications as well as DNA sequencing on microfabricated capillary array electrophoresis (CAE) microplates. These CAE microplates perform high speed analysis of multiple samples in parallel increasing the throughput by several orders of magnitude over conventional slab or capillary array systems[1]. Several generations of CAE microplates have been developed to optimize layout and performance. Our current design uses a circular scanning confocal fluorescence detection system together with radially symmetric channel layouts. The design consists of a common anode reservoir in the center of a circular 4" or 6" diameter wafer and an array of 96 channels extending radially outward towards injector units at the perimeter of the wafer. This radial design gives quality high speed separations by eliminating resolution reducing turns and allows the analysis of 96 samples in parallel on a single microplate. The confocal rotary scanner can measure fluorescence from DNA fragments in all channels at a rate of up to 15 samples/sec. The major advantage of a rotary scanner over linear scanners is that the motion of the scanner is continuous, making it easier to control the velocity at high sample rates. High sample rates are necessary to ensure good resolution of electrophoretic bands. The scanner is capable of collecting 2880 data points/revolution at 23.15 ms intervals. The scanner utilizes four independent ADCs to simultaneously acquire data from four color electrophoresis runs.

The operation and capabilities of the radial CAE microplates with the rotary scanning system were first demonstrated by performing high speed electrophoretic separations of 96 pBR322 MspI DNA samples in 40 seconds. Genotyping of methylenetetrahydrofolate reductase (MTHFR), a candidate gene for vascular disease and neutral tube defects, was also performed on 4" diameter radial CAE microplates to demonstrate the rapid analysis of biologically relevant samples (in collaboration with Prof. M. Smith and C. Skibola in the School of Public Health, UCB). Two-color multiplexed fluorescence detection of the MTHFR genotypes was accomplished by prelabeling standard pBR322 MspI DNA ladder with a red emitting bisintercalation dye (butyl TOTIN) and prelabeling of the MTHFR DNA with a green emitting bisintercalation dye (TOTO)[2]. Using this two-color multiplexing method, 96 MTHFR DNA samples were genotyped in less than 2 minutes with 4 bp resolution. Radial CAE microplates fabricated on 6" wafers are currently being developed for ultra-high throughput DNA sequencing applications.

[1]P.C. Simpson, D. Roach, A.T. Woolley, T. Thorsen, R. Johnston, G. F. Sensabaugh, and R.A. Mathies, Proc. Nat. Acad. Sci., USA, 95, 2256-2261 (1998)
[2]S.M. Clark, and R.A. Mathies, Anal. Chem., 69, 13354-1363 (1997)

39. Integrated Sequencing Sample Preparation on CE Microplates

Yining Shi[1], Indu Kheterpal[1], Jin Xie[2], Alexander N. Glazer[2], and Richard A. Mathies[1]
[1]Departments of Chemistry and [2]Molecular and Cellular Biology, University of California, Berkeley, CA 94720
rich@zinc.cchem.berkeley.edu

Microfabricated devices are revolutionizing the field of DNA electrophoresis because DNA fragments can now be separated in less than 1 minute[1,2] and sequencing separations are achieved in ~10 minutes[3]. Furthermore, Microfabricated devices allow the integration of sample preparation, clean-up, separation and detection. To achieve this goal we have performed high quality sequencing separations on microchannels and are developing solid-phase microfluidic methods to concentrate and clean up DNA samples for efficient injection into the separation columns.

The quality of separation of DNA fragments is highly dependent on the injection and separation conditions. We have optimized four-color sequencing on microfabricated capillary electrophoretic devices for separation matrix, temperature, channel dimensions, injector size and injection parameters. Linear polyacrylamide (LPA; 4%) matrices were used to achieve sequencing separations of 600 bases on 7 cm long channels in ~20 minutes. The sequence data were analyzed and base-called using BaseFinder[4] and an accuracy rate of 99.4% was obtained to 500 bases[3].

We are now developing methods to integrate these excellent separations with sample preparation methods on a single device. We have synthesized biotinylated energy transfer primers for fragment amplification and sequencing[5]. The presence of the biotin allows us to utilize solid-phase surface chemistry to purify and concentrate DNA samples before introducing them into the separation columns. We have successfully constructed sandwich structures of biotin-streptavidin-biotin on the glass surface. The biotinylated PCR products are pumped through a capture chamber and concentrated onto the surface containing biotin-streptavidin. The products captured in the reaction chamber are cleaned, denatured with formamide at 90°C and injected directly into the separation columns. These sample clean-up methods are relevant to developing fully integrated microdevices.

[1] Woolley, A. T. and Mathies, R. A. (1994) Proc. Natl. Acad. Sci. U.S.A. 91, 11348-11352.
[2] Simpson, P. C.; Roach, D.; Woolley, A. T.; Thorsen, T.; Johnston, R.; Sensabaugh, G. F. and Mathies, R. A. (1998) Proc. Natl. Acad. Sci. U.S.A. 95, 2256-2261.
[3] Liu, S.; Shi, Y.; Ja, W.; Mathies, R. A. (1998) Anal. Chem. in press.
[4] Giddings, M. C.; Severin, J; Westphall, M; Wu, J. Z; and Smith, L. M. (1998) Genome Research 8, 644-665.
[5] See abstract "Synthesis, Characterization, and Potential Applications of Biotinylated Energy Transfer Oligonucleotides" by J. Xie, R. A. Mathies, A. N. Glazer.

40. Integrated Electrochemical Detection with Microfabricated Capillary Electrophoresis Chips

Pankaj Singhal[1], Jin Xie[2], Alexander N. Glazer[1], and Richard A. Mathies[2]
Departments of [1]Chemistry and [2] Molecular and Cellular Biology, University of California, Berkeley, CA 94720
pankaj@zinc.cchem.berkeley.edu

Microfabrication technology has enabled the development of miniaturized capillary electrophoresis (CE) chips or microdevices that can perform preparation, amplification and electrophoretic separation of a wide variety of analytes on a very short time scale[1-3]. However, nearly all microchip analyses to date have utilized laser excited fluorescence detection. While fluorescence detection is very effective, it is not easy to integrate the laser and optical system into the microfabricated chip to make a completely miniaturized analysis system. We have therefore been exploring the development of microfabricated electrochemical detection systems for microchip CE analyses because of the high sensitivity of this method and the ease of integration.

In our first studies, platinum electrodes were microfabricated on glass CE-chips to demonstrate the feasibility of integrated electrochemical detection[4]. Although, redox-active neurotransmitters were detected directly with high sensitivity, non-electroactive DNA could only be detected using indirect detection. In order to make electrochemical detection more universal for chip-based analyses, redox-active labels can be attached to inherently non-electroactive compounds. Specifically, we have synthesized an M-13 primer with hydroquinone and ferrocene labels to demonstrate the feasibility of attaching labels to DNA. We have worked out the synthetic routes to prepare active N-hydroxysuccinimide esters of these labels. Activated esters of 1, 4-dihydroxy-2-naphthoic acid or a ferrocene were coupled with a 5'-aminohexyl terminated M-13(-40) universal primer sequence to make two different electroactive DNA primers. We have been able to detect these labeled DNA primers down to zeptomole levels using our micro-fabricated CE-chips with integrated electrochemical detection.

As a number of different labels are available for attachment to various analytes, simultaneous detection of multiple samples is conceivable with very high selectivity. To demonstrate this concept, we selected various labels which exhibit different redox-properties and are therefore readily distinguishable. These labels were detected with high selectivity using CE-chips with integrated electrochemical detection. A matrix-coding method was developed to collect the electrochemical signals from each label. This method also uniquely addresses each signal, so that the labels were detected without any overlap from each other. CE-chip designs using this approach for multiplex analyses in a single separation will also be presented. To further highlight the potential of integrated electrochemical detection, we will present a fully portable version of our microchip based system. This instrument validates that integrated electrochemical detection allows CE-chip based analyses to be miniaturized and portable.

[1]Woolley, A. T. and Mathies, R. A.; (1994) Proc. Natl. Acad. Sci. U.S.A. 91 11348-11352.
[2]Woolley, A. T., Sensabaugh, G. F. and Mathies, R. A.; (1997) Anal. Chem. 69 2256-2261.
[3]Simpson, P. C., Roach, D., Woolley, A. T., Thorsen, T., Johnston, R., Sensabaugh, G. F. and Mathies, R. A.; (1998) Proc. Natl. Acad. Sci. U.S.A. 95 2256-2261.
[4]Woolley, A. T., Lao, K., Glazer, A. N. and Mathies, R. A.; (1998) Anal. Chem. 70 684-688.

41. Integrated Microchip Devices for DNA Analysis

R. S. Foote, W. C. Dunn, J. Khandurina, N. Kroutchinina, T. McKnight, L. C. Waters, S. C. Jacobson, and J. M. Ramsey
Chemical & Analytical Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37830-6142
footers@ornl.gov

Microfabricated microfluidic devices are being developed for integrated processing and analysis of DNA samples. The steps of DNA extraction, amplification, preconcentration and electrophoretic analysis can be carried out on monolithic devices. We have previously demonstrated integrated cell lysis, multiplex PCR and capillary electrophoretic (CE) size analysis on microchips using bacterial samples. Integrated PCR-CE microchips are now being used for the analysis of simple sequence repeat (SSR) loci in mammalian genomes. To demonstrate the potential utility of this technology for rapid PCR-based gene mapping, SSR polymorphisms (SSRPs) at two mouse genome loci, D4Mit141 and D8Mit9, were identified and compared for *Mus musculus* (C57/Bl or C3H), *Mus spretus*, a C57/Bl x spretus hybrid, and progeny from an interspecies (C3H x spretus) backcross used to create genetic maps. For both loci, microchip electrophoresis patterns of SSR fragments generated from animals heterozygous for musculus and spretus alleles were clearly distinguishable from those of homozygotes and PCR product sizes were

determined for respective SSRs by co-electrophoresis with marker DNA. Integrated microchip analysis of human forensic samples has also been demonstrated by DNA fingerprinting at the CSF1PO, THO1, TPOX and vWA loci.

The microchip CE analysis time for PCR products is typically less than 5 minutes, so that the throughput for integrated PCR-CE microdevices is primarily determined by the PCR thermal cycling time. Approaches to increasing the throughput of these devices include the use of multiple reaction chambers for parallel PCR, fast thermal cycling using thermoelectric heating and cooling, and on-chip concentration of products from low cycle number PCRs. The last approach is being explored by incorporating a DNA concentration region into the microchip architecture between the reaction chamber and the separation channel. A porous membrane between two parallel channels is incorporated into the channel manifold using a silicate adhesive to bond the cover plate to the substrate. The thin silicate layer serves as a semi-permeable membrane allowing ionic current to pass between the separated channels but retaining large DNA molecules. Preconcentrated sample is then injected into the separation channel and electrophoretically analyzed. DNA fragments were concentrated on-chip from PCR amplifications by up to 2 orders of magnitude by this method, allowing product analysis at a reduced number of thermal cycles.

42. Single Nucleotide Polymorphism Detection and Identification Directly from Human Genomic DNA by Invasive Cleavage of Oligonucleotide Probes

Victor Lyamichev, Andrea L. Mast, Jeff G. Hall, James Prudent, Tamara Sander, Monika de Arruda, David Arco, Bruce P. Neri, and **Mary Ann D. Brow**
Third Wave Technologies, Inc., 502 S. Rosa Rd.
Madison, WI 53719
madbrow@twt.com

Detection of DNA by invasive cleavage arises from the coordinated action of a pair of overlapping synthetic oligonucleotides hybridized to adjacent regions on a DNA target. At the point of overlap, i.e. where one or more nucleotides from each oligomer compete for same complementary site on the target, a unique secondary structure forms when the 3' end of one oligomer (invasive probe) displaces the 5' end of the other (signal probe). This displaced "flap" is in turn recognized by a structure-specific endonuclease and cleaved to release a fragment.

The specific cleavage of a downstream flap has been employed as an extremely sensitive, quantitative, and highly specific assay for the detection of target DNA both alone and in mixture of extraneous DNA. Because cleavage depends on the correct alignment of the oligonucleotides, the cleavage is sufficiently specific to enable discrimination of single nucleotide polymorphisms (SNPs) and can readily differentiate homo- from heterozygotes in single-copy genes in human genomic DNA. Moreover, we have defined reaction conditions that allow multiple copies of the downstream oligonucleotide probe to be cleaved for each target sequence without temperature cycling, thereby amplifying the cleavage signal and allowing quantitative detection of target DNA at sub-attomole amounts.

The analysis of nucleic acids in this fashion has several advantages over existing methods of oligonucleotide-based detection. First, by requiring two oligonucleotides, the reaction is highly specific for the intended target sequence. Second, the specificity of the enzyme requires precise alignment of the probes for cleavage to occur, providing a much higher level of specificity than can be achieved by hybridization alone, and allowing single-base discrimination of multiple alleles present in a mixed sample. Third, the products of this cleavage reaction can be analyzed via indirect readouts that utilize the nucleotide sequence of products, such as by capture on solid supports, thus simplifying the equipment needed to perform the procedure and adding yet another level of discrimination for the desired cleavage products. Fourth, amplification of a target-dependent signal, rather than the target itself, means that traces of product "carried over" from a completed detection reaction cannot themselves be amplified to lead to false positive results. Finally, detection of specific sequences directly from genomic DNA without intervening DNA amplification avoids false negative or false positive SNP detection that

may arise due to low fidelity replication during an amplification step.

43. High Throughput SNP Discovery and Scoring Using Bead-Based Flow Cytometry

P. Scott White, Hong Cai, and John P. Nolan
Los Alamos National Laboratory, Los Alamos, New Mexico
swhite@telomere.lanl.gov

There is a pressing need for SNP discovery and analysis capabilities that are rapid and robust. We are developing approaches using microsphere-based flow cytometry to address these needs.

For SNP discovery, we have developed a system that uses immobilized mismatch-binding proteins (IMBP) to detect SNPs in heteroduplexes. IMBP-coated microspheres are added to fluorescently labeled PCR amplicons, and analyzed by flow cytometry. The detection of fluorescence associated with the beads indicates the amplified region contains one or more SNPs, which are then sequenced.

Bead-based minisequencing or oligo ligation using flow cytometry is used to score SNPs. A novel system for multiplexed analysis enables simultaneous scoring of 64 or many more different SNPs/sample. Furthermore, because of the quantitative nature of flow cytometry, pooling amplicons from large numbers of individuals will allow for the determination of the frequencies of each SNP in populations.

These microsphere-based flow cytometric analyses have the following general advantages: 1) Intrinsic resolution between free and microsphere-bound probe, allowing homogeneous assays with no wash steps; 2) Quantitative, multicolor fluorescence detection with sensitivity surpassing microplate or microscope-based detection; 3) Soluble solid phase that can be prepared, pipetted, and handled by conventional fluidics systems; and 4) An instrument that is already available in core facilities at the vast majority of research universities, medical schools, pharmaceutical companies, and clinical diagnostic laboratories. Furthermore, the potential for multiplexing these assays will greatly enhance throughput and allow for the scanning of over one megabase/day for new SNPs, or for scoring thousands of individuals for hundreds to thousands of known SNPs/day.

44. DNA Characterization by Electrospray Ionization-Fourier Transform Ion Cyclotron Resonance Mass Spectrometry

David S. Wunschel, Ljiljana Pasa Tolic, Bingbing Feng, James E. Bruce, Harold R. Udseth, and **Richard D. Smith**
Environmental Molecular Sciences Laboratory, Mail Stop: K8-98, Pacific Northwest National Laboratory, Richland, WA 99352
dick.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) products [1], 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 the 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 methylation[2]. 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[3]. 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 PCR. This would allow for direct selection and analysis of individual components from within mixtures that may share a high degree of similarity without cloning. Alternatively, DNA species that cannot be identified through traditional sequencing methodologies, those containing base modifications, can be isolated with the nature and position of the modification identified. Most importantly this potentially allow identification of low abundance products containing modifications where few if any alternatives for their detection exist. These and related recent advances will be described.

[1] "Characterization of PCR products from bacilli using electrospray ionization FTICR mass spectrometry", D. C. Muddiman, D. S. Wunschel, C. L. Liu, L. Pasa Tolic, K. F. Fox, A. Fox, G. A. Anderson and R. D. Smith, **Anal. Chem.** 68, 3705-3712 (1996)
[2] "Mass measurement of a PCR product from the Lambda bacteria phage at the 223 base pair level by ESI-FTICR", D. S. Wunschel, B. Feng, L. Pasa Tolic, and R. D. Smith.
[3] "Heterogeneity in *Bacillus cereus* PCR Products Detected by ESI-FTICR Mass Spectrometry", D. S. Wunschel, D. C. Muddiman, K. F. Fox, A. Fox, and R. D. Smith, **Anal. Chem.** 70, 1203-1207 (1998)

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

45. Laser Desorption Mass Spectrometry for DNA Sequencing and Analysis

C. H. Winston Chen, N. R. Isola, N. I. Taranenko, V. V. Golovlev, and S. L. Allman
Life Science Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831
chenc@ornl.gov

During the past few years, rapid progress has been achieved for both slab gel electrophoresis and capillary gel electrophoresis. Many experts in the field expect that most parts of human genome can be sequenced within next 3 to 7 years. However, some portions of DNA in human genome which has long repeats and with secondary hairpin structures are still very difficult to be sequenced by conventional gel electrophoresis with Sanger's enzymatic method to produce DNA ladders. The band compression often occur for DNA segments with high GC ratio and/or with secondary structures. PCR process may not faithfully replicate the DNAs which have hair pin structures. Thus, a new and reliable approach to sequence these "difficult" templates is critical for completing the sequencing of the entire human genome. Recently, we tried to couple laser desorption mass spectrometry with Maxam Gilbert chemical degradation method to produce DNA ladders to achieve sequencing of DNA templates with high GC component. DNA templates were first bound with biotin so that DNA ladders produced by the chemical degradation method can be isolated from the solution by magnetic bead streptavidin separation. Then these isolated DNA segments are released from streptavidin and subsequently analyzed by laser desorption mass spectrometry. Since the sequencing by laser desorption mass spectrometry is based on the measurement of molecular weights, band compression is no longer a problem. Since no PCR is required, non-faithful replication by PCR due to the secondary structures or a large number of repeat can be eliminated.

In addition to the sequencing by Maxam Gilbert's approach, we also used laser desorption mass spectrometry for DNA sequencing for DNA ladders produced by Sanger's method. ss-DNA templates larger than 100 nt were successfully sequenced. ds-DNAs larger than 200 bp were also sequenced.

However, mass resolution and detection sensitivity still need more improvement for sequencing longer DNAs. In addition to sequencing DNA with ladders produced by chemical methods, we also developed a technology to produce DNA ladders during the laser desorption process. By controlling the pH value and selecting the right matrices, direct sequencing of short DNA was obtained without the need of the preparation of DNA ladders. Since the sequencing process with this approach is very fast, it can be used to sequence a large number of probes which are often used for diagnosis by hybridization.

* Research has been supported by DOE/OBER

46. PCR Product Size Measurement using MALDI Mass Spectrometry

G.B. Hurst, Y. Kim, K. Weaver, and M.V. Buchanan
Organic and Biological Mass Spectrometry, Oak Ridge National Laboratory, Oak Ridge, Tennessee
hurstgb@ornl.gov

Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) has considerable potential as a technique for rapid and accurate analysis of PCR products. We are pursuing two specific end-uses of this technique: mapping mutant phenotypes to chromosome regions and determining the extent of chromosomal rearrangements that are targets for mutagenesis in collaboration with ORNL's Laboratory for Comparative and Functional Genomics and screening of endogenous bacterial to assess genetic potential for bioremediation, in collaboration with Mary Lidstrom at the University of Washington. Current capabilities include virtually routine analysis with near single-base resolution up to 100 bases, and the less routine ability to measure 200-mers or larger products. We are working to further extend this technique to larger PCR products (and similarly-sized DNA derived from other sources), as well as to develop schemes for scaling up the applicability of MALDI to larger numbers of samples.

The salts and buffers necessary as reagents for the PCR act as interferences for the MALDI process, and therefore must be removed prior to MALDI-MS analysis. To overcome this problem, we have developed a rapid reverse-phase method for purifying PCR products [1], and have demonstrated the parallel implementation of this procedure in a 96-well microtiter-format using a filter plate loaded with the appropriate reverse-phase resin. Manual implementation of the 96-well purification method, using an 8-channel pipettor and a vacuum manifold, requires approximately 20-30 minutes for 96 samples.

For MALDI-MS analysis, PCR products must be combined with a matrix material and allowed to dry on a sample plate. The resulting inhomogeneous spot is sparsely dotted with regions that yield useful spectra when interrogated with the desorption laser ("sweet spots"), and therefore requires either human expertise for laboriously choosing promising locations across the sample, or inefficient automated positioning of the laser at numerous positions in hopes of locating a sweet spot. For this reason, we are developing methods for preparing more homogeneous spots containing the mixture of PCR product and matrix, using polymeric substrates and/or additives. Fluorescently-labeled primers or PCR products may allow us to correlate MALDI-MS results with fluorescence microscopy imaging of the DNA distribution in these samples.

References

[1] "MALDI-TOF Analysis of Polymerase Chain Reaction Products from Methanotrophic Bacteria," G.B. Hurst, K. Weaver, M.J. Doktycz, M.V. Buchanan, A.M. Costello, and M.E. Lidstrom, *Anal. Chem.* **70**, 2693-2698 (1998).

Research supported by the Environmental Management Science Program, Office of Biological and Environmental Research, U.S. Department of

Energy, and the Oak Ridge National Laboratory Director's Research and Development Funds. Oak Ridge National Laboratory is managed for the United States Department of Energy by Lockheed Martin Energy Research Corp. under contract DE-AC05-96OR22464.

47. Analyzing Genetic Variations by Mass Spectrometry

Lloyd M. Smith, Travis Berggren, Tim Griffin, Zhengdong Fei, and Mark Scalf
Department of Chemistry, University of Wisconsin, Madison, WI 53706-1396
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.

48. DNA Sequencing by Single Molecule Detection

James H. Jett, Peter M. Goodwin, James H. Werner, Hong Cai, and Richard A. Keller
Chemical Sciences and Technology Division and Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
jett@lanl.gov

We are developing a DNA sequencing technique that is based upon single fluorescent molecule detection. The goal is to sequence long strands of DNA approaching 40 kb in length at rates of 50 bases per second. The approach being pursued is to: 1) fluorescently label a strand of DNA by enzymatic incorporation of prelabeled nucleotides; 2) attach a single labeled strand of DNA to a microsphere; 3) suspend the microsphere in the flow stream of a flow cytometer capable of single molecule detection and identification; 4) add an exonuclease with activating cofactors that will cleave sequentially the labeled nucleotides; and 5) identification of the cleaved nucleotides by analysis of laser-induced fluorescence from the label attached to the nucleotides. We have made considerable progress towards demonstrating this approach to DNA sequencing. Briefly, the current status of each of the steps above is as follows. Up to three base types in strands of DNA 2-7 kbp long have been labeled base identifying fluorophors. Multiple strands of labeled DNA have been attached to microspheres and individual microspheres suspended by a laser optical trap in the flow stream of the detection apparatus. Enzymatic activity of several exonucleases on DNA attached to microspheres under flowing conditions has been observed. Fluorescent molecule identification at the single molecule level has been demonstrated by correlated measurements of fluorescent burst intensity and fluorescence lifetime with a single excitation wavelength and a single detection channel. Details of the progress made in each of the steps will be discussed.

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

49. Manipulation of Single DNA Molecules by Induced-Dipole Forces in Micro-Fabricated Structures

Chip Asbury, Paolo Prati, and Ger van den Engh
Department of Molecular Biotechnology, University of Washington, WA 98195
asbury@biotech.washington.edu

We are exploring the use of induced-dipole forces in oscillating, divergent electric fields, for trapping, moving and stretching DNA molecules. These forces,

which are distinct from electrophoretic forces, can be generated by means of microscopic metal patterns on quartz substrates. Micro-fabrication techniques can be employed to generate large numbers of traps on a single wafer. This technology lends itself well for massive automation and parallelization of DNA sample preparation.

Unlike electrodes for electrophoresis, DNA trapping can be achieved with floating electrodes. Voltage applied by external wires to the fluid is passively distributed among hundreds of microelectrodes. The electric field lines concentrate on the electrode edges exerting strong attractive forces on DNA molecules in solution. If the electrode gaps are small, significant trapping forces can be obtained without inducing electrolysis. The magnitude of these forces appears to vary with DNA molecular weight.

We have built several devices for manipulating small quantities of DNA, such as shift registers, DNA concentrators, etc.. We are now developing more complex structures that combine DNA dipole traps with more elaborate manipulations. We have made electrophoresis capillaries with concentrating traps at the entrance and exit. We have constructed capillaries with a series of traps along their length for size dependent DNA separation. We are working on a device that guides DNA molecules along a precise trajectory past a fluorescence detector. We are also exploring the use of traveling waves to concentrate DNA from a large area.

We will describe these and other structures and will present the conditions under which these devices are most efficiently used.

50. A Quantitative Analytical Tool for Improving DNA-Based Diagnostic Arrays

Tom J. Whitaker
Atom Sciences, Inc., 114 Ridgeway Center, Oak Ridge, TN 37830
whitaker@atom-sci.com

Sequence analysis using hybridization on ODN arrays is particularly well suited for genetic diagnostics, sequencing cDNAs, and partial sequencing of clones to allow mapping. In spite of this, quality control issues have hampered the full acceptance of these arrays, which are often called "gene chips". A high dynamic range, quantitative measurement method is needed to study parameters that increase efficiency and new methods of array production. With such a tool, systematic studies of hybridization strategies could be undertaken which would almost certainly lead to improved efficiencies and lower array manufacturing costs. Although fluorescence detection has been adequate for analysis of the hybridized chips, it does not have the spatial resolution or dynamic range to image the surface density of small (e.g. 20μm spot size) bound probe ODNs or to perform hybridization kinetics studies.

We have just begun a new project to develop and utilize a high-resolution, quantitative method to analyze ODNs on an array. The technique involves detection of tin-labeled ODNs by sputtering them from the surface with an energetic ion beam, selectively ionizing the resulting tin atoms with wavelength-tunable lasers, and analyzing the ions with time-of-flight mass spectrometry. We have previously shown that this sputter-initiated resonance ionization microprobe (SIRIMP) technique can have sub-mm resolution and is highly quantitative in measuring a wide range of concentrations of elements in semiconductors. We have also shown that SIRIMP can detect DNA fragments labeled with stable isotopes [1,2]. We now plan to apply SIRIMP to measurements of tin-labeled ODNs immobilized on a surface and to tin-labeled ODNs synthesized in situ

on the surface. The initial phases will be used to develop and demonstrate the technique and calibration procedures. In later stages, we will work closely with Affymetrix to analyze and image in situ arrays with very small (20μm) features to determine the homogeneity of binding. Additionally, hybridization experiments will be performed with two different stable isotopes of tin labeling the probe and target ODNs to determine the correlation between the surface density of the immobilized probes and hybridized targets.

The research reported here was funded, in whole or in part, by DOE grant #DE-FG02-98ER82536. Such support does not constitute an endorsement by DOE of the views expressed in this abstract.

[1]H.F. Arlinghaus, M.N. Kwoka, X.-Q. Guo and K.B. Jacobson, Analytical Chemistry 69, 1510 (1997).
[2]H.F. Arlinghaus, M.N. Kwoka, and K.B. Jacobson, Analytical Chemistry 69, 3747 (1997).

51. A Light-Directed DNA/RNA-Microarray Synthesizer

Xiaochuan Zhou[1], Robert Setterquist[1], Xiaolian Gao[2], Peilin Yu[2], Eric LeProust[2], Laëtitia Sonigo[2], Jean Philippe Pellois[2], Hua Zhang[2], Erdogan Gulari[3], and Ning Gulari[3]
[1] Xeotron Corporation, Houston, TX 77030
[2] University of Houston, Department of Chemistry, Houston, TX 77204-5641
[3] University of Michigan, Department of Chemical Engineering and Center for Display Technology and Manufacturing, Ann Arbor, MI 48109
xczhou@email.msn.com

Practical advancement in biochip technologies for routine use in drug discovery and genomic applications will require a flexible and affordable chip-fabrication technology. To address this need, a multidiscipline project has been initiated. A programmable, light-directed DNA/RNA array synthesizer that uses solution-based photochemical synthesis is being developed for efficient production of high-density DNA/RNA chips. This presentation reports on the latest results of instrument development and photochemistry effort.

The project consists of three main integrated tasks: (1) design and construction of a programmable photolithographic system, (2) development of a novel solution photochemistry for nucleic acid synthesis, and (3) design and fabrication of synthesis microreactors. At the heart of the photolithographic system is a commercially available digital spatial optical modulator, which accurately produces light patterns that are used for initiating parallel high-density nucleic acid synthesis. The digital spatial optical modulator effectively replaces the necessity for using photomasks technologies. The solution photochemistry under investigation is a modification of well-established conventional synthesis protocols. Microreactors are being developed using standard microfabrication processes in order to implement the DNA/RNA synthesis photochemistry. Each reactor contains an array of microfabricated reaction wells (synthesis sites). Each microwell serves to individually isolate each reaction during the light-directed parallel sequence syntheses.

The outcome of the undertaken project will lead to a prototype DNA/RNA array synthesizer. The prototype instrument will be further developed into a commercial model. The envisioned instrument will allow researchers to make high-density and high fidelity DNA/RNA-chips of their own designs at an affordable cost. In addition, there is obvious potential to expand the light-directed chemical approach in this project for synthesis of other combinatorial arrays (peptide, carbohydrate, and small molecule).

52. Development of Flowthrough Genosensor Chips

Mitchel J. Doktycz and Kenneth L. Beattie
Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6123
okz@ornl.gov

A flowthrough genosensor chip is under development at ORNL. The core of this technology is a microchannel hybridization array, containing numerous specific DNA sequences, immobilized within individual cells of densely packed straight, smooth channels traversing a thin silicon or glass substrate. When a nucleic acid sample is labeled and

passed through the microchannel genosensor chip, hybridization occurs at porous cells bearing immobilized DNA probes complementary to the target sequence. The quantitative binding pattern reflects the relative abundance of specific target sequences within the nucleic acid analyte. The flowthrough chip configuration has several important advantages over flat surface DNA chips being developed elsewhere: faster hybridization kinetics, superior binding capacity, improved ability to analyze dilute solutions of nucleic acids, including both strands of a heat-denatured PCR fragment.

Related technology for taking advantage of the benefits of the flowthrough genosensor include the development of micromachining techniques for the construction of flowthrough silicon chips to complement those constructed using channel glass. A customized robotic spotting system has been developed that includes a high resolution positioning system, sapphire dispensing tips for touch-off dispensing, and, more recently, solenoid-controlled ink jets for remote droplet delivery. A prototype fluidics system has been developed that involves syringe pump-driven fluid flow, a custom chip holder attached to the stage of a Zeiss Axiovert fluorescence microscope and a CCD camera for real-time quantitative detection of hybridized fluorescent-labeled strands. A software package for intelligent selection of oligonucleotide probes for a given chip application has been developed.

The flowthrough genosensor system is now being used to develop applications in the areas of genotyping and mRNA profiling, in collaboration with various laboratories. Gene expression profiling in mammalian systems, including mouse and sheep, is being pursued as well as bacterial systems for evaluating expression patterns in soil microorganisms as an indicator of genotoxic response in the environment. Another application being developed is high throughput genotyping. In this work miniature flowthrough genosensors are used to simultaneously analyze numerous single nucleotide and short insertion-deletion polymorphisms. In another application area, the ultrahigh surface area of channel glass is being exploited to create arrays of "microreactor cells" containing immobilized BAC DNAs, for use in repetitive reactions needed for genome mapping and sequencing, including cycle sequencing reactions, PCR, and hybridization mapping of expressed sequences to their genomic clones.

53. Sequence Analysis and Thermodynamic Studies of Short DNA Duplexes on Oligonucleotide Generic Microchip

A. Fotin, D. Proudnikov, **E. Timofeev**, G. Yershov, Eu. Kirillov, A. Drobyshev, E. Khomyakova, A. Zasedatelev, and A. Mirzabekov
Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, 117984 Moscow, Russia; Argonne National Laboratory, Argonne, IL 60439, USA; and Moscow Institute of Physics and Technology, 141700 Dolgoprudny, Russia
timofeye@everest.bim.anl.gov

Generic microchip — an oligonucleotide microchip containing a complete set of hexanucleotide probes — has been manufactured and applied for sequence analysis and thermodynamic studies. Isothermal hybridizations as well as thermal denaturation experiments allowed performing sequencing of model synthetic oligomers up to 70 bases long. Melting experiments on the chip using fluorescent microscope have demonstrated reliable discrimination between perfect and mismatched duplexes. This technique was succesively applied for identification of mutations. Studies of thermodynamics of DNA duplex and triplex formation and the effect of modified bases and minor groove binding ligands on duplex stability have been carried out on the generic microchip.

Mapping

54. Third-Strand Binding Probes for Duplex DNA in Particles of Varying Size

Marion D. Johnson III and Jacques R. Fresco
Princeton University, Department of Molecular Biology, Princeton, New Jersey
mjohnson@molbio.princeton.edu,
jrfresco@princeton.edu

Third-strand oligonucleotide binding to native double-stranded DNA in some free form or in chromatin via triple helix formation provides a sequence specific way to attach ligands, including cytochemically detectable ones, to DNA-containing structures. Last year we described the development of an in situ methodology for identifying metaphase chromosomes. We particularly demonstrated binding of a highly specific fluorescent third-strand probe directed to a 16 base pair DNA α-satellite target of human chromosome 17. We continue to exploit this type of interaction for various purposes of relevance to the Human Genome Project.

Some of our recent efforts have been directed at:

1. Developing probes specific for the a-satellite regions of human chromosomes X and 16. Such probes expand our capacity to cytogentically identify and isolate individual human chromosomes.

2. Exploiting such third strand probes to greatly facilitate isolation by flow sorting of individual human chromosomes and possibly chromosomes of other species.

3. Developing fluorescent-labeled probes to identify accessible multicopy sequences within the *Drosophila* histone gene cluster. These probes are being employed to locate such target sequences in *Drosophila* ovarioles.

To assure these ends, we have also investigated the consequences of DNA structure and size on the mechanism and equilibria of probe binding.

55. Optical Mapping: A Complete System For Whole Genome Shotgun Mapping

D.C. Schwartz[1,2], T. Anantharaman[2], J. Apodaca[1], C. Aston[1], V. Clarke[1], D. Gebauer[1], S. Delobette[1], E. Dimalanta[1], J. Edington[1], A. Evenzehav[1], J. Giacalone[1], V. Gibaja[1], C. Hiort[1], E. Huff[1], J. Jing[1], Z. Lai[1], D. Lazaro[1], E. Lee[1], J. Lin[1], K. Lin[1], B. Mishra[2], L. Ni[1], S. Paxia[2], B. Porter[1], R. Qi[1], A. Ramanathan[1], Y. Skiadis[1], J. Vafai[1], W. Wang[1], H. Zhao[1]
[1]W. M. Keck Laboratory for Biomolecular Imaging, Department of Chemistry, and [2]Courant Institute of Mathematical Sciences, Department of Computer Science, New York University, NY 10003
schwad01@mcrcr6.med.nyu.edu

Optical Mapping is a single molecule approach for the rapid production of ordered restriction maps from individual DNA molecules. Fluorescence microscopy is used to directly image individual DNA molecules bound to derivatized glass surfaces, and cleaved by restriction enzymes. Fragments retain their original order, and cut sites are flagged by small, visible gaps. The Optical Mapping system has advanced in several critical areas to emerge as a means for the detailed mapping of both clones and entire genomes (*Deinococcus radiodurans* and *Plasmodium*

falciparum). We mapped these entire microbial genomes using megabased-sized genomic DNA molecules (600-10,000 kb). Because large fragments of randomly sheared DNA are mapped with high cutting efficiency, many overlapping restriction site landmarks allow contigs to be assembled and a shotgun mapping strategy can be employed. High resolution whole genome maps can therefore be assembled without library construction and associated cloning artifacts. Because ensembles of single molecules are analyzed, small amounts of starting material are required enabling mapping of microorganisms, which are problematic to culture. Whole genome maps enable the size of the genome to be accurately determined, an important prelude to any sequencing endeavor. Most importantly, whole genome maps from genomic DNA provide an in situ picture of the architecture of the entire genome, revealing the number of chromosomes, existence of extrachromosomal elements etc. Populations can be potentially be characterized by comparing maps from different strains. Recent efforts have been to create high resolution maps of *E. coli* O157 strain (5.4 mgb) as a scaffold for facilitated sequence assembly and verification (Collaborator: F. Blattner, U. Wisconsin). We will compare maps generated "in silico" from the sequence of *E. coli* K12 (4.6 mgb) to identify regions that are unique to O157 and could be targeted for sequencing. Given the success we enjoyed in the restriction mapping of whole microbial genomes, and the proven reliability of the contig assembly algorithms developed for these efforts, we decided to construct a reference restriction map of the entire human genome. In four weeks our laboratory mapped 0.6 human genome equivalents at 40 kb resolution, using genomic fragments with average size of 2.1 mb. Our analysis of the contigs formed showed good correspondence with suitably modified Lander-Waterman physical mapping criteria in terms of the number and depth of overlapped genomic fragments. Goals are to simultaneously complete the human reference map to include 10-15x coverage and to link with other physical maps by the alignment of restriction mapped BAC contigs. The utility of this map will be to facilitate large scale sequencing projects and to provide a novel resource for the analysis of large populations.

56. Verifying Sequence By Atomic Force Microscopy

David P. Allison and Peter R. Hoyt
Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6123
allisondp@ornl.gov

Atomic force microscopy (AFM) technology can be developed, as a sequencing alternative, to verify homologies between DNA species, accurately, inexpensively, and with high-throughput.
By forming heteroduplexes between sequenced and test molecules, deletions, substitutions, and perhaps even point mutations, can be imaged and precisely located by AFM.

Using AFM imaging we have identified deletions of 22 to 450 bp in heteroduplexes of linearized mutant and wildtype pSV-ß-Galactosidase plasmid (6821 bp). Additionally, utilizing an AFM technique we developed for mapping cosmids, which employs imaging a cleavage deficient mutant E*co*RI endonuclease site-specifically bound to active sites, we have simultaneously located deletions relative to the E*co*RI sites on the pSV-ß-Galactosidase heteroduplexes.

We have imaged the specific binding of mismatch repair enzymes to heteroduplexes between wildtype and plasmids with point mutations. Conditions to maximize binding efficiency and define specificity for all combinations of mismatches are being evaluated on plasmid constructs.

When developed this technology could provide high-throughput sequence verification of heteroduplexes generated from long range PCR or clone libraries. Furthermore these procedures can be accomplished by technicians, use readily available relatively inexpensive instrumentation, and should be fully transferable to most laboratories.

57. Molecular Cytogenetics Comes of Age: A Resource that Extends From "T" to Shining "T"

J. R. Korenberg[1], X. N. Chen[1], D. Noya[1], X.Wu[2], B. Birren[2], T. Hudson[2,3]
[1] Medical Genetics Birth Defect Center, Cedars-Sinai Medical Center, University of California at Los Angeles, Los Angeles, California; [2]Whitehead Institute, Massachusetts Institute of Technology, Massachusetts; and [3]McGill University, Montreal, Canada
jkorenberg@xchg.peds.csmc.edu

With the maturation of the DNA sequence of the human genome, it becomes necessary to link this sequence to the language of clinical medicine. Therefore, to provide this bridge and at the same time, to anchor the genetic map to the chromosomal map, a genome-wide resource of bacterial artificial chromosomes (BACs) carrying defined Genethon polymorphic markers has been defined.

Using PCR, 17,200 BAC DNAs were screened using a five-dimensional pooling scheme. Positives were confirmed by PCR, linked to a mapped BAC array and/or streaked and re-confirmed by FISH (fluorescence in situ hybridization) using fluorescence reverse banding at the 500-700 band stage. All data were recorded in a Fourth Dimension relational database and images archived on a gigabyte optical disc system.

The resource is now represented by a BAC/STS map representing all human chromosomes, and includes 860 STS/BAC combinations, representing 882 total markers, each BAC mapping to a single chromosome sub-band. Of these 648 carry a Genethon markers, 84 carry ESTs, 42 carry known genes, 9 carry markers that were unmappable by any previous technique, and 108 carry random markers. A further 163 STS/BAC pairs mapped to one of multiple sites defined by FISH. Of the total 1122 marker/BAC pairs tested, 1023 of the chromosome assignments were in agreement; 99 were not.

This framework resource uniquely provides integration with all existing maps. It anchors early sequencing, provides rapid access to cancer and prenatal breakpoints and their candidate genes, can map new genes to single bands without FISH when integrated with RH data, and can be used as targets for CGH (comparative genome hybridization) on slides, chips or filters. The resource integrates genome with genetics and medicine. It should speed solutions for diagnosis, prognosis and ultimately treatment.

It may be viewed on http://www.csmc.edu/genetics/korenberg/korenberg.html and is available.

58. Automated Purification of Blood, or Bacterial Genomic DNA

Dan P. Langhoff, Tuyen Nguyen, and **William P. MacConnell**
MacConnell Research Corporation, San Diego, California
macres@macconnell.com

In Phase I of the SBIR research we have developed a prototype of a fully automatic high-throughput blood or bacteria genomic DNA isolation instrument. Unlike any other process currently used for genomic DNA isolation, this instrument uses a derivative of electrophoretic separation technology that was developed by our company for automated purification of plasmid DNA. The separation technique is novel and powerful in that it requires no moving parts and can be performed with the combination of a simple disposable sample cassette and an inexpensive processing instrument. The process purifies high molecular weight genomic DNA direclty from a cell lysate through the use of electrophoretic movement of the DNA which is placed in between barriers of agarose medium that are contained in a disposable cassette device. The purification method results in highly pure genomic DNA over a wide range of input sample quantities, including as low as 1000 starting cells, and it gives high DNA yields while not relying on chromatography adsorption or solvent

precipitation at any step. The purified DNA is suitable for use in PCR sequencing and in RFLP analysis. The disposable sample cassette is designed to process twelve or more samples in parallel and the processing instrument holds up to several of these cassettes. The instrument will inexpensive to manufacture and it will occupy less then one square foot of laboratory bench space.

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, as well as in research applications. The ability to rapidly and reproducibly isolate DNA from blood and other bodily samples is required to identify, characterize and treat factors involved in human disease and disorders.

DOE SBIR PHASE I GRANT
#DE-FG-03-98ER82612

59. New Host Strains for Stabilization and Modification of YAC Clones

Natalay Kouprina, Maxim Koriabine, and **Vladimir Larionov**
Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709
larionov@niehs.nih.gov

The recent development of a new approach (TAR cloning) for the selective isolation of specific regions and genes from complex genomes as large linear or circular YACs greatly advanced YAC cloning technology[1,2]. While TAR cloning provides many opportunities for studying mammalian genomes, some YAC isolates containing multiple repeats may be mitotically unstable. Recently we systematically studied the contribution of several *RAD* genes to the stability of human YAC clones in yeast. Using a variety of linear and circular internally marked YACs, we demonstrated that *rad52* substantially stabilizes human DNA inserts, decreasing YAC instability 25- to 400-fold compared to a recombination-proficient host strain. In contrast, other *rad* mutant strains analyzed *(rad1, rad50, rad51, rad54* and *rad55*) had a minor affect (2- to 5-fold reduction) on YAC instability. Thus, if YAC stabilization is desired, propagation in a *rad52*-deficient strain is strongly advisable. However, there is no opportunity to manipulate YACs by recombination in *rad52* strains. Moreover, *rad52*-deficient strains cannot be used for specific gene isolation by TAR cloning. Therefore, we chose to develop a *rad52*-based system that could be used for stable maintenance of any YAC, while providing the opportunity for recombinational manipulation. We constructed a set of *kar1* strains that have a conditional *RAD52* gene under the control of the galactose-inducible *GAL1/GAL10* promoter. These strains are *rad52*-deficient on glucose-containing medium and recombination-proficient on medium containing galactose. A YAC from any genetic background can be efficiently and accurately transferred into new hosts during mating with karyogamy-deficient *kar1* strains[3]. To expand more the utility of a new YAC transfer system, the RNA telomerase gene *TLC1* in *RAD52*-conditional strains was modified to produce $(TTAGGG)_n$ repeats specific to human telomere sequences[4]. The *kar1*-induced transfer of a YAC into the strains with the modified *TLC1* gene resulted to replacement of yeast-specific telomeric repeats by human-specific repeats in the YAC.

[1]Larionov *et al.* (1997) *Proc. Natl. Acad. Sci. USA* **94**: 7384-7387.
[2]Kouprina *et al.* (1998) *Proc. Natl. Acad. Sci. USA* **95**: 4469-4474.
[3]Spencer *et al.* (1994) *Genomics* **22**: 118-126.
[4]Henning *et al.* (1998) *Proc. Natl. Acad. Sci. USA* **95**: 5667-5671.

60. Direct Isolation of a Centromeric Region from a Human Mini-Chromosome by in Vivo Recombination in Yeast

Natalay Kouprina, Motonobu Katoh, Mitsuo Oshimura, and Vladimir Larionov
Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, NC 27709
kouprina@niehs.nih.gov

Isolation of specific chromosomal regions and entire genes has typically involved cloning of random fragments as BACs or YACs followed by a long and laborious process to identify the region of interest. Using the recently developed TAR cloning technique in yeast[1], it has been possible to directly isolate specific chromosomal regions and genes from complex genomes as large linear or circular YACs. In this study we applied a modified version of this technique[2] for isolation of a centromeric region of the human mini-chromosome D1 containing 5 Mb of the human chromosome Y[3]. This mini-chromosome was generated by two rounds of telomere-directed chromosome breakage leading to a loss of sequences from both arms of the chromosome. Despite the small size and loss of a significant part of centromeric repeats (there is only 140 kb of alphoid DNA left), the D1 mini-chromosome segregates accurately in mitosis, suggesting that a 140 kb block of alphoid DNA alone or along with the short arm flanking sequences is sufficient for a centromere function. Taken in advantage that the first round of chromosome Y breakage resulted in truncation of the chromosome within a block of alphoid DNA (i.e. a new telomere and a block of alphoid DNA became physically linked), we developed a scheme to isolate a centromeric region from the mini-chromosome D1. Direct transformation of genomic DNA isolated from hybrid cells carrying the mini-chromosome into yeast spheroplasts resulted in a rescue a centromeric region as a set of linear YACs with sizes from 50 kb to 300 kb. To prevent YAC rearrangements due to the presence of multiple repeats, the isolates were maintained in the host strain with the conditional RAD52. Each YAC isolate containing an entire block of alphoid DNA (i.e. YACs bigger than 140 kb) was circularized, retrofitted into BAC with the NeoR mammalian selectable marker and accurately transferred into the *E. coli* cells. Since no detectable changes in YACs were observed after retrofitting to BACs, the BAC DNAs were transferred into human cells for further functional analysis.

[1]Kouprina et al. (1998) Proc. Natl. Acad. Sci. USA 95: 4469-4474.
[2]Kouprina et al. (1998) Genome Research 8: 666-672.
[3]Brown et al. (1996) Proc. Natl. Acad. Sci. USA 93: 7125-7130.

61. Insert Clone Selection by Sorting GFP-Expressing *E. coli*

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

At a previous DOE contractors meeting (Santa Fe, 1995), we proposed a method for insert clone selection by sorting Green Fluorescent Protein-containing *E. coli* into individual culture wells. Direct selection of insert-containing bacteria with a cell sorter abolishes the need for clone picking. We are now using this approach in an automated, integrated process for large-fragment DNA subcloning.

The process utilizes a vector in which insertion of DNA at a cloning site causes GFP expression. Insert-containing bacteria are selected in a cell sorter and deposited into 10 microliter wells. The vector may be amplified either by culturing the bacteria or by PCR. The amplified product may be used as a template for fluorescent dye-based sequencing reactions. Amplification by PCR avoids the need for a DNA extraction step.

We have now demonstrated proof-of-principle for several steps of the process. We have created a vector with a GFP gene downstream of a strongly regulated promoter. The plasmid contains a Lac repressor gene. Insertion of a DNA fragment inside the repressor gene disrupts its function, resulting in GFP expression. We currently have vector strains with Blue as well as Green Fluorescent Proteins. The green protein offers the best signal to noise ratio of the two. Individual bacterial clones containing inserts can easily be detected and sorted. Results of insert amplification by PCR from single-sorted bacteria will be presented.

62. A Resource of Mapped BAC Clones for Identifying Cancer Chromosome Aberrations

Norma J. Nowak[1], Jeffrey Conroy[1], Greg P. Caldwell[1], Joseph Catanese[1], Barbara Trask[2], John D. McPherson[3], David R. Bentley[4], Grace Shen[5], and Pieter J. de Jong[1]
[1]Department of Cancer Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263; [2]Department of Molecular Biotechnology, University of Washington School of Medicine, Seattle, WA 98195; [3]Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108; [4]Sanger Centre, Hinxton, Cambridge, CB10-1RQ, UK and 5CCAP Program; and [5]National Cancer Institute, Bethesda, MD
nowak@dejong.med.buffalo.edu

We are generating a resource of mapped BAC clones from the arrayed human BAC library (RPCI-11) for use in fluorescent in situ hybridization (FISH) analysis of chromosomal rearrangements in human tumors. This work is performed under the auspices of the NCI Cancer Chromosome Aberrations Project (CCAP). Our goal is to establish mapped BAC clones by screening 6-fold redundancy of the BAC library with 4,000 markers mapped at high resolution through radiation hybrid (RH) panels. Markers are judiciously selected spaced less than 1 Mb with preference for markers mapped to both high and low-resolution RH panels. To increase the success rate of the marker to BAC correlation, we are employing overlapping oligonucleotide probes ("overgo") based on the EST and STS sequences for the markers. The overgos are designed with the same average melting characteristics and are labeled by replicating the 5'overhangs using P-32 nucleotide triphosphates. To increase the throughput of the screening process to high density BAC colony membranes, the probes are pooled in mixtures of 36 probes each. Informative probe mixtures are prepared through the use of a 3-dimensional (6x6x6) probe pooling strategy consisting of 216 distinct probes. All predicted probe-BAC pairs are being confirmed by PCR and the expected 4-6 overlapping BACs for each marker are subsequently validated by restriction digest fingerprinting. In our initial mapping effort, we recovered BAC clones for 586 genetic markers on chromosomes 1, 5, 6, 18, 19, 21, 22 and Xp. After completion of three rounds of screening (736 markers), we have attained an average 2.5 Mb level of resolution and 4.2 BACs per marker. Overgos for the remaining Sanger framework markers are being designed and all mapped clones along with corresponding mapping information will be deposited in the public domain. Up to one BAC clone per marker will be characterized by in situ hybridization experiments to establish its usefulness as a FISH probe and to provide additional independent confirmation of the map location.
Work supported by NCI, DOE and the Wellcome Trust.

63. Preparation of New BAC Vectors for BAC Cloning and Transformation-Associated Recombination ("TAR") Cloning

Changjiang Zeng[1], Yu Wang[1], Kazutoyo Osoegawa[1], Natasha Kouprina[2], Vladimir Larionov[2], and Pieter J. de Jong[1]
[1]Department of Cancer Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263 and [2]Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, NC 27709
zeng@dejong.med.buffalo.edu

Recently, efficient procedures have been reported for the re-cloning of large genomic DNA fragments in yeast as circular YACs. This approach of

"Transformation Associated Recombination" cloning utilizes homologous recombination between a linear YAC vector and homologous sequences in complex genomic DNA to generate circular yeast artificial chromosomes. For this purpose, a specific YAC vector equipped with short (unique) genomic sequences from the targeted region needs to be constructed. The sequences are positioned at the ends of the linear YAC fragment used for transformation into yeast spheroplasts. To make the process of TAR rescue of genomic DNA more universally applicable, we constructed several hybrid BAC/YAC vectors designated as pTARBAC-1, -2 and -4. These vectors differ from our earlier BAC vector (pBACe3.6) by the presence of a yeast centromere (CEN3) and a yeast-selectable marker (his3). The TARBAC vectors have been used to prepare BAC libraries for several species, including *Trypanosoma brucei*, Giardia and Cat. Clones lacking inserts do not generate viable yeast colonies after transformation of yeast spheroplasts and selection for his-function. However, most (25out of 30) of the Trypanosome BACs with 130 kb average inserts, transform yeast at high efficiency, indicating the presence of ARS elements in most of the genomic insert fragments. Most of the insert sequences in the Trypanosome BACs can be deleted by treating the BACs with a restriction enzyme (e.g. EcoRI) which lacks corresponding sites in the TARBAC vector. Such EcoRI-deleted BAC clones are functionally similar to the previous TAR-rescue vectors because they have (unique) genomic sequences at the ends of a hybrid BAC/YAC vector. Hence, most of the BAC clones in a TARBAC library can be used to generate TAR-rescue vectors to (re-)clone genomic segments from different haplotypes or from related species. We have confirmed that most of the deleted TARBACs have also lost the ARS elements as indicated by the loss of their capability to successfully transform yeast spheroplasts. We are currently exploring the re-isolation of the deleted sequences by co-transformation of deleted BAC DNA with genomic Trypanome DNA. The less-complex genomes of unicellular eukaryotes are used to model future work with mammalian TARBAC libraries.

Information on our current libraries can be obtained from our Web page: http://bacpac.med.buffalo.edu.
* Supported in part by grants from the U.S. DOE , NHGRI and the German Forschungs Gemeinschaft (DFG).

64. "RPCI" Human and Mouse Bacterial Artificial Chromosome Libraries: Construction and Characterization

Kazutoyo Osoegawa[1], Baohui Zhao[1], Minako Tateno[2], Eirik Frengen[1], Joseph J. Catanese[1], Yoshihide Hayashizaki[2], and Pieter J. de Jong[1]
[1]Department of Cancer Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263 and [2]Genome Science Laboratory, Riken Tsukuba Life Science Center, Japan
kazutoyo@dejong.med.buffalo.edu

Human male and female bacterial artificial chromosome (BAC) libraries have been constructed in the pBACe3.6 vector in compliance with the new NIH-DOE guidelines on anonymous donor selection and informed consent. A 25-fold genome equivalent male BAC library (RPCI-11) was constructed by partial digestion with a combination of EcoRI and EcoRI methylase. The library has been distributed to 30 genome centers as a major resource for the human genome sequencing effort. The average insert size was estimated to be 173 kb. The average redundancy of the library was determined at 23.9 positives per marker by hybridization of 45 single locus probes. All 1,076 marker-positive BAC clones were confirmed to be part of 45 single-marker contigs by restriction-fingerprinting. An additional 123 BAC clones were identified using probes derived from a minimal overlapping set of PAC clones on 14q24.3. The resulting 1.5-Mb BAC contig has been assembled by hybridization using BAC-end probes and PCR with STS markers, thus allowing the mapping of 264 markers within the contig. The genomic sequence for the contig region generated at Washington University, facilitates the analysis of the contig-integrity and allows the determination of the

BAC clone fidelity. A total of 87 clones were confirmed to be non-chimeric because both insert-ends map back to the contig. No rearranged clones have been observed within the 5.5-kb STS-resolution contig map. More recently, a second BAC library (RPCI-13) has been constructed from an anonymous female donor. This library has been generated from genomic DNA partially digested with EcoRI (segment 1 & 2, 10x redundant) and partially digested with DpnII (segment 3 & 4, 10x redundant). In addition to the human BAC libraries, two approximately 10-fold redundant murine BAC libraries have been prepared and extensively characterized. The source DNA for these libraries was obtained from female mice from two inbred strains: 129SvEvTAC and C57B6 for the RPCI-21 & 23 libraries, respectively. The information on the current libraries can be obtained on our Web page at: http://bacpac.med.buffalo.edu.

* Supported by grants from the U.S. DOE (#DE-FGO3-94ER61883), NIH (#1RO1RGOl 165).

65. Characterization of a BAC Clone Resource for Human Genomic Sequencing: Analysis of 150 Mb of Human STCs and Implications for Human Genomic Sequencing

G. G. Mahairas, J. C. Wallace, J. Furlong, K Smith, S. Swartzell, A. Keller, HTSC Staff and L. Hood
High Throughput Sequencing Center, University of Washington, Seattle, WA 98109
gmahaira@u.washington.edu

Together with The Institute for Genomic Research (TIGR), we have sequenced the BAC ends or sequence tagged connectors (STCs) from 160,000 BAC clones. We have also generated a HinDIII restriction digest for each BAC whose end sequences have been determined at the University of Washington and developed strategies and tools for using this resource in support of large-scale genomic sequencing. We have demonstrated proof of concept for its use. Together with TIGR, we propose to complete the characterization of an STC clone resource from two IRB-approved human BAC libraries to 22.5-fold clone (BAC) coverage (e.g. 450,000 BAC clones assuming an average insert size of 150 kb). These data are available on the world wide web through dbGSS and our web sites (orcas.htsc.washington.edu and www.tigr.org) and the clones are available for distribution to the scientific community through Research Genetics. Nine hundred thousand STC sequences will provide a sequence marker of 300 to 500 base pairs (bp) on average every 3,100 bp across the genome. The BAC libraries and the data pertaining to them will enable the facile selection of minimum tiling paths of BAC clones across each of the human chromosomes for large-scale sequence analysis. Here we present data to support the STC approach for sequencing of the human genome and other moderate to large genomes. The STC approach eliminates the need for up front physical mapping and uses BAC clones as the basic sequencing reagent. The major advantages of the STC approach are: (i) reduced cost and effort to obtain complete low and high resolution maps and front end automation is greatly simplified. (ii) The BAC clones are readily available through Research Genetics. (iii) As improved techniques for generating BACs or other yet to be developed libraries appear, reasonable numbers of these new clones could easily be added to the database and clone collection. (iv) This approach will obviate the significant problem of closure for high resolution physical mapping. (v) The existing chromosomal landmarks, STS, PCR-specific sites, EST, or partial cDNA sequence, can be easily placed on the BAC clones, adding additional markers for BAC clones and taking significantly advantage of any associated biological information. (vi) The 10% of the genome obtained in the STCs can be searched against the sequence data base to identify many interesting landmarks (e.g. genes, STSs, EST, etc.) that could locate the BAC clone on the preexisting chromosomal maps. (vii) Chromosomal regions of key biological interest can be identified and sequenced first. (viii) The human genome can be sequenced earlier and for less cost. (ix) The STC approach will provide useful clones for biological studies even at the very early STC sequencing stages when only 3- to 4-fold coverage is achieved. The STC approach streamlines the task of clone selection by doing much of the work up front and by using sequence alignment and computers as the primary tools to identify sequencing targets. Additional major advantages of the STC strategy are that it is rapid in that clone selection is automated, STC data directly

correlates with a clone which can be used for shotgun sequencing without further evaluation, surveys the entire genome and is more dense allowing greater versatility in the use of the data including genotyping analysis. Perhaps the greatest advantage of the STC resource is that it can be used by any investigator for clone or sequencing target selection via the World Wide Web. The STC clone library also serves as a large scale genomic survey tool and provides access to many characterized clones in any part of the genome. The implications of this type of resource transcend simple genomic sequencing. Additionally, we will describe the University of Washington High Throughput Sequencing Facility capable of producing 2 million BAC end sequences per year.

66. Human BAC End Sequencing

Shaying Zhao, Mark Adams, Bill Nierman, and Joel Malek
TIGR, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville MD 20850
szhao@tigr.org

BAC end sequences (BESs) provide highly specific markers. In genomic sequencing, the clones to be sequenced next can be selected by searching the completed sequence against a BES database. The average insert size of BAC clones is about 150 kb and therefore BESs are useful in chromosomal walking and assembly. End sequences from 300,000 clones (15x clone coverage) will be generated by TIGR and UofWashington. At TIGR, we have sequenced BESs from both CalTech and Pieter de Jong libraries with a successful rate >80% and an average read length of 450. The pair percentage is >65% and the average phred score is 28. We also resequence both ends of one-end-failed clones for higher pair % and quality control. For those clones we resequenced so far, the redo sequences always match the original ones. The average cost is about $4.50 per BES and $0.10 per base. We continue improving our protocol to decrease the cost and increase the successful rate and read length. Up to date, we have submitted more than 130,000 BESs to GenBank. We have collected more than 300,000 BESs from TIGR, U of Washington and CalTech for our search database at http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_search.html and ftp site (ftp://ftp.tigr.org/pub/data/h_sapiens/bac_end_sequences/).

The finished 600,000 BESs will cover about 10% human genome and provide a sequence marker every 5 kb across the genome. BESs can be used to survey the whole genome. We searched BESs against existing databases of repeats, STSs and ESTs and the results are presented at our web site (http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_anno.html). On average 50% of BESs contain known repeats and the length ranges from 21 to 806 with an average of 185 bases. And 30% bases are repeats masked. With identity >=95%, 3 % BESs match ESTs while 0.2% match STSs which are used to locate some of the BACs on the preexisting chromosomal maps. BESs are also used to assess the representative of BAC libraries and tie up the existing contigs. We are collaborating with other institutes to map some of the BESs and the results will be presented on web.

67. Construction of a Genome-Wide Human BAC-Unigene Resource

Bum-chan Park[1], Robert Xuequn Xu, Chang-Su Lim, Mei Wang, Aaron Rosin, Steve Mitchell, Hee Moon Park[1], Eunpyo Moon[2], Ung-Jin Kim, and Melvin I. Simon
Division of Biology, Caltech, Pasadena, CA 91125
[1]Chungnam University, Taejon, Korea and [2]Ajou University, Suwon, Korea
simonm@cco.caltech.edu

With the availability of high quality BAC libraries with stable, large inserts, it is now feasible to rapidly develop genome-wide physical BAC contig resources to cover the large mammalian genomes. For this purpose, we have tried to screen human BAC

libraries using mapped Unigene cDNA clones as probes. Currently, over 52,000 mapped Unigenes (non-redundant, unigene sets of cDNA representing EST clusters) are available for human alone. A total of 44,000 Unigene cDNA clones have been supplied to us by Research Genetics. We have currently deconvoluted over 10,000 Unigene probes against a 4X coverage human BAC library D using high density colony hybridization filters. 10,000 batches of Unigenes are arrayed in a logical array of 100 X 100 matrix from which 100 row pools and 100 column pools are derived. Library filters are hybridized with pooled probes, thus reducing the number of hybridization required for addressing the positives for each Unigene from 10,000 to 200. Details on the experimental scheme as well as daily progress report is posted on our WEB site (http://www.tree.caltech.edu). Initial assessment of the deconvolution data indicates that over 95% of the Unigenes have been deconvoluted so that we could have made a BAC-Unigene resource for them. 800 additional Unigene probes and 1,200 Unigene probes which were already deconvoluted by 100x100 have been re-screened by 20x20 to determine the accuracy and to estimate the rate of false positive hits as a function of probe complexity and improve the accuracy. To circumvent the cross-hybridization problems inherent to some Unigene probes, we are also designing OVERGOes from sequences derived from mapped, well annotated genes. Human BAC-Unigene resources generated in this effort will contribute toward the realization of the "whole genome" approaches for human and other model organisms.

68. A New Bacterial Artificial Chromosome (BAC) Vector, a Large-Insert (Average of Over 200 kb) BAC Library of the Human, and an Improved Method of Construction of BAC Libraries

Sangdun Choi, Yu-Jiun Chen, Mel Simon, and Hiroaki Shizuya
Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA
schoi@cco.caltech.edu

BAC (bacterial artificial chromosome) cloning has served an important role in human and mammalian genomics since its introduction in 1992 by our laboratory. BAC libraries are currently in use or under development for virtually every important genome. The primary reasons why BACs are so useful is that they can stably maintain large DNA inserts (up to 350 kb) in *E. coli*, and are amenable to virtually all of the sophisticated molecular biology techniques developed for *E. coli*. We have been constructing BAC libraries of human and mouse in the last several years. Total number of human BAC clones generated is now close to one million. Recently we developed a new BAC vector and an improved method of construction of BAC libraries, and began constructing a series of BAC libraries with much larger insert size (182 - 202 kb) from human and a variety of organisms including *Arabidopsis*, maize, and rice. The larger insert genomic BAC libraries will provide significant improvement to applications in physical mapping, positional cloning, and DNA sequencing.

69. One Tier Pooling of a Total Genomic BAC Library

D.C. Torney, J.L. Longmire, D.C. Bruce, J. Fawcett, M. Campbell, J. Tesmer, M. Maltbie, B. Taggett, T. Tatum, P. Jewett, J. Meyne, N. Lenhert, Y. Valdez, S. Bailey, A. Schliep[1], L.L. Deaven, and **N.A. Doggett**
Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545 and [1]University of Cologne, Cologne, Germany
doggett@gnome.lanl.gov

We have developed a single-tier pooling approach which enables the screening of a 12X diploid human genomic BAC library of 221,184 clones, 165 kb average insert size (one-half of the total 24X RPCI-11 library, http://bacpac.med.buffalo.edu) in a single screen of 376 PCR reactions, followed by confirmatory reactions of the predicted positive BAC clones. Prior to pooling of the library, the 384 well library stock plates were translated to 96 well plates. Pooling of the library was performed in increments of a quarter at a time: each a 3x coverage containing

55,296 clones. Ninety-four pools were made from each quarter of the library including two sets of plate pools (11 pools each), two sets of row pools (12 pools each), two sets of column pools (12 pools each), and one set of left and right diagonal pools (12 pools each). Each of these eight sets of pools was made from a plate-rearranged 24x24 array of 96-well microtitre dishes. Thus, most pools contain 4,608 clones. The plate rearrangements were selected to minimize the numbers of co-incidences of pairs of clones in pools. For the row, column, and diagonal pools, the "lines" of clones from the array are combined to make the final 12 pools, with no pool containing more than one line incident on any plate. Pool construction was accomplished, straightforwardly, with manifolds which have been designed to work on the Robbins Hydra (manuscript in preparation). Prior to pooling, the performance of the single-tier design (in the presence of experimental errors) was simulated. As with real data, the Markov chain Monte Carlo procedure was used to rank the candidate positives. Results compared favorably with a random 8-sets pooling design (Knill et al., J. Comp. Biol., 3, 395-406 (1996), and Bruno et al., Genomics, 26, 21-30 (1995)). We will present results from screening the pools with STS primer pairs that demonstrate that these single-tier pools will serve as valuable resources for rapidly isolating BAC clones for mapping and sequencing. Supported by the U.S. D.O.E. Office of Biological and Environmental Research under contract W-7405-ENG-36.

70. High Density Colony Filter Production and Automated Data Analysis for Efficient Hybridization Screening of BAC Libraries

Anca Georgescu, Laura Kegelmeyer, Bernadette Lato, Hummy Badri, Matthew Groza, and Anne Olsen
Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551
olsen2@llnl.gov

Bacterial Artificial Chromosomes (BACs) have proven to be excellent reagents for construction of sequence-ready maps. As the demand for mapped BACs continues to increase, efficient methods for screening these libraries are needed to identify clones at the required throughput. We describe a format for high-density colony filter production in conjunction with a program for automated analysis of hybridization results, which has greatly improved the accuracy of identifying positive signals.

Colony filters are plated at a density of 6 x 6 x 384, or 13,824 colonies per 8 x 12 cm filter. All colonies are plated in duplicate in a unique offset pattern optimized to prevent ambiguity in identification of positive offsets. A positive control is plated in the first and last offsets of each subgrid to enable automated drawing of major grid lines by the analysis program. Hybridization probes consist of Alu-PCR products of cosmid or BAC clones, or overgos (J. McPherson, Washington Univ.) designed from cosmid or BAC end sequences. Multiple probes are pooled in a single hybridization, and positive colonies are re-arrayed for hybridization with individual probes. Hybridization signals are analyzed by the "blot-score" program developed at LLNL. The program currently operates in semi-automated manner, with the potential for full automation in the near future. In the current mode, the user loads a phosphorimager file of hybridized filters and selects the filters to be analyzed. Using the positive control offset signals, the program dynamically draws gridlines to indicate the 384 major subdivisions on the filter. When a subgrid containing a positive is selected, the program displays an enlarged view of the positive subgrid with the 6 x 6 minor gridlines drawn. True positives appear in unique duplicate patterns. The program highlights the position of the expected duplicate signal when the user selects a positive, thus providing immediate feedback on the validity of a given pair of observed signals. The program is linked to the LLNL mapping database to facilitate entry of hybridization results into the database and retrieval from the database of map information relevant to positive clones identified.

Work performed under the auspices of the US DOE by Lawrence Livermore National Laboratory under contract W-7405-ENG-48.

71. Systematic Conversion of a YAC/STS Map into a Sequence Ready BAC Map

C. Han and N.A. Doggett
Joint Genome Institute, Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545
chan@telomere.lanl.gov

We are starting with a previously constructed integrated physical map of human chromosome 16 (Doggett et al., Nature 377:Suppl:335-365, 1995) and converting this to a new sequence-ready BAC of this chromosome. The YAC/STS component of the integrated map consists of 900 CEPH megaYACs, and 300 flow-sorted 16-specific miniYACs that are localized to and ordered within somatic cell hybrid breakpoint intervals with 1150 STSs. This YAC/STS map provides nearly complete coverage of the euchromatic arms of the chromosome and provides STS markers on average every 78 kb. The integrated map also includes 470 genes/ESTs/exons, 400 genetic markers, and 530 cosmid contigs (110 kb average size, and covering 60% of the chromosome). To create large sequenceable targets of this chromosome we are using a systematic approach to screen high density BAC filters with evenly spaced probes. Probes are either pooled overlapping oligonucleotides (overgos, method developed by John McPherson, Wash U.). In order to select evenly spaced probes we first identified all available sequences in the integrated map. These include sequences from genes, ESTs, STSs, and cosmid end sequences. Since the integrated map was constructed on a physical scale we are able to select for sequences at a spacing of 50 kb - 100 kb when these were available. We then used BLAST to identify 36 bp unique fragments of DNA for overgo probes. Up to 236 overgos have been pooled in a single hybridization against a 12X coverage human BAC library (RPCI-11). Positive BACs that are identified from the pooled overgos are rearrayed on membranes and hybridized with either two-dimensional subpools of overgos to determine which BAC clones are positive for individual overgos. Probe-content BAC contigs are constructed in this manner. BAC contigs are then restriction mapped to select the optimal tiling sets for sequencing. Thus far we have identified over 6000 BACs from the chromosome 16 long arm and from an 11 Mb region of the short arm by the hybridization of 1,00 3 overgos. 35 Mb of BAC probe-content maps (by completion of the 2-dimensional hybridizations) and 10 Mb of sequence ready restriction maps have been constructed from these targets. Supported by the US DOE, OBER under contract W-7405-ENG-36.

72. An Arrayed BAC Resource for the High Resolution Mapping of Cancer-Related Chromosome Aberrations

Eunpyo Moon[1], **Jonghyeob Lee**[1], Mei Wang, Bum-Chan Park, Ken Myambo[2], Colin Collins[2], Melvin Simon, Ung-Jin Kim
[1]Ajou University, Suwon, Korea
[2]Lawrence Berkeley National Laboratory, Berkeley, California
simonm@cco.caltech.edu

Numerous human chromosomal aberrations known to be related to cancer phenotypes have been catalogued to date. 27,000 such aberrations have been documented and mapped to chromosomal subregions at the resolution of cytobanding technologies. Mapping and characterizing these aberrations at high resolution will provide clues to the underlying molecular nature of many of these cancers.

We are currently establishing a BAC resource that will cover the known cancer-related regions. The resource will also include 2-3,000 BACs spread over the genome. These BACs are being identified by cDNA inserts or OVERGO probes designed from published ESTs that have been well annotated and mapped to genomic locations. To date, over 200 cDNA probes have been selected from I.M.A.G.E. cDNA library and used for screening against the approved Caltech Human BAC library D. A total of 800 BACs have been selected by these probes and have been deconvoluted against the individual probes. Numerous OVERGOes are being designed and are

used for library screening. The resulting BAC resource will provide a roughly 1 Mb resolution BAC array and will serve as a framework for high resolution FISH mapping of the chromosome aberrations. The BACs and BAC contigs from the array will also be used for the identification of the culprit genes, the molecular basis of the cancers incurred by the aberrations, and for the development of the "oncochip" to be used for efficient diagnosis of chromosome aberrations by the use of Comparative Genome Hybridization (CGH) technique.

73. A 12 Mbp Completely Contiguous Sequence-Ready BAC Contig in Human Chromosome 16p13.1-11.2

Yicheng Cao, Hyung Lyun Kang, So Hee Dho[1], Diana Bocskai, Mei Wang, Xuequn Xu, Jun-Ryul Huh[1], Byeong-Jae Lee[1], Francis Kalush[2], Judith G. Tesmer[3], Eunpyo Moon[4], Norman A. Doggett[3], Mark D. Adams[2], Melvin Simon, and Ung-Jin Kim
[1] Institute of Molecular Biology and Genetics, Seoul National University, Seoul, Korea; [2] The Institute for Genomic Research, Rockville, Maryland; [3] Los Alamos National Laboratory, Los Alamos, New Mexico; and [4] Ajou University, Suwon, Korea
simonm@cco.caltech.edu

Here we present a 12 Mbp of BAC contiguity in the centromeric half of the chromosome 16p arm. The work initially involved extensive screening of deep human BAC libraries developed at Caltech using the STS markers that have been mapped to the target regions by Los Alamos National Laboratory. The positive BACs were characterized by sizing, FISH mapping, BAC end sequencing and sequence matches, and restriction fingerprint analysis. For the clones submitted for complete sequencing, genomic Southern blot analysis was performed to confirm the colinearity of the clones with the genomic DNA. 51 BAC clones in this region have been completely sequenced by TIGR. We post a comprehensive summary of the screening and characterization data for this and all other projects through our WEB site http://www.tree.caltech.edu.

We use the AceDraw program developed by the CS140 team at Caltech for the construction and updating of the contig map. This tool allows freehand, real scale map drawing using various data and human judgement, and is capable of communicating with databases including ACeDB. We extensively utilize STS contents, restriction fingerprint data, and BAC end sequence matches for the establishment of clone-to-clone contiguity. For the gap closure, we also utilized BAC end probes, OVERGOes and additional BAC libraries such as RPCI.

74. Completing the Sequence-Ready Map of Chromosome 19

Laurie Gordon, Anca Georgescu, Mari Christensen, Sha Hammond, Hummy Badri, Bernadette Lato, Matthew Groza, Linda Ashworth, Mark Wagner, and Anne Olsen
Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551
olsen2@llnl.gov

Chromosome 19 is the most GC-rich chromosome, suggesting an especially high gene density. This prediction is supported by transcript mapping results (Deloukas et al., Science 282, 744-746, 1998), that indicate chromosome 19 has the highest number of gene-based STSs relative to size of all the chromosomes. Thus this chromosome should be an extremely rewarding sequencing target in terms of gene discovery and elucidation of gene structure and organization.

We are nearing completion of a sequence-ready map of chromosome 19. The current map consists of 72 BAC/cosmid contigs with an average size of 710 kb. The contigs have been ordered along the chromosome by high resolution FISH, so their location is well

defined relative to the cytogenetic map. The ordered contigs span a total of 51 Mb, or 93% of the non-centromeric portion of the chromosome. The average size of remaining gaps is an estimated 80 kb. For gap closure, probes developed from the ends of contigs are hybridized to high-density BAC colony filters, and positive BACs are incorporated into the existing map by analysis of restriction digests.

All contigs have been restriction mapped with EcoRI, resulting in a high-resolution restriction map of almost an entire chromosome. The distribution of EcoRI sites varies along the chromosome, with relatively larger fragments more common in light band regions. Several EcoRI polymorphisms between clones from different sources have been detected in the process of assembling restriction maps. The average depth of coverage of restriction mapped contigs is 8.5-fold. The high depth of coverage and mix of cosmid and BAC clones generally enables selection of an optimum set of spanning clones with minimum overlap for sequencing. About 30 Mb of chromosome 19 have been sequenced or are currently in the sequencing queue. The average size of contigs being sequenced is 830 kb. An average overlap between sequence tiling path clones of 10% is estimated from the map. All sequence tiling path clones are digested with three additional restriction enzymes to provide data for confirmation of final sequence assembly. Updated chromosome 19 data, including all restriction maps, are available on the LLNL Genome Center web site at http://www-bio.llnl.gov/bbrp/genome/html/chrom_map.html.

Work performed under the auspices of the US DOE by Lawrence Livermore National Laboratory under contract W-7405-ENG-48.

75. High-Throughput Multiplexed Fluorescent-Labeled Fingerprinting of BAC Clones

Yan Ding[1], Martin D. Johnson[2], Wang Q. Chen[3], Gigi E Park[1], Yujin Chen[1], and Hiroaki Shizuya[1]
[1]Beckman Institute, Division of Biology, California Institute of Technology, Pasadena, CA 91125, U. S. A; [2]PE Biosystems, 850 Lincoln Center Drive, Foster City, CA 94404, U. S. A.; and [3]Paracel Inc., 80 S. Lake Ave #650, Pasadena, CA 91101-2616, U. S. A.
yding@cco.caltech.edu

Human Genome Project has entered in a large scale sequencing stage. Currently, numerous maps, for example, STS or EST content maps, and YAC contig maps, have been constructed across all human chromosomes. However, building of physical maps that involve large contigs of sequencing units such as BACs still falls far behind. In order to fill this gap in a timely fashion, we have been worked on high-throughput contig assembly of BAC clones through multiplexed Fluorescent-labeled fingerprinting. Projects that rely on restriction fragment size lists to establish relationships between clones require extensive overlaps due to the limited resolution inherent in these strategies. We are currently developing a fingerprinting method using certain class IIS restriction enzymes, which cut DNA a few basepair away from their recognition sites and generate 5' overhangs consisting of 1 to 5 unknown bases. With the recessed strand serving as primer, these overhangs can be sequenced using modified fluorescent dideoxy terminator sequencing reagents[1]. When a fifth dye is used for an internal lane size standard, each fragment can be characterized by both size and end sequence of it's terminal 1 to 5 bases. This enhanced detail greatly increases the power to detect minimum overlap. Using this method, it is theoretically possible to identify overlaps of 15% for a project with 10,000 clones. The increased information content of each fragment also assists assembling accurate overlaps and establishing minimal tiling path with fewer clones. We will report our latest effort on optimizing the fingerprinting technique, software development to interpret the data, and a test on 500 to 1000 BAC clones, which have been identified with markers located on a 20 Mb region from 16p13.1 to 16p11.2, to assemble contigs.

[1] Brenner, S. and Livak, K. J. 1989, Proc. Natl. Acad. Sci. USA, 86:8902-8906.

76. Progress Towards a High Resolution Sequence-Ready Map of Human Chromosome 5

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

The high resolution map of chromosome 5 at JGI/LBNL began at the distal portion of the long arm. The region was chosen because it contains a cluster of cytokine growth factors (IL3, IL4, IL5, IL9, IL12, IL13, GM-CSF, FGFA, M-CSF) and receptor genes (GRL, ADRB2, M-CSFR, PDGFR) and was thought likely to yield related genes through full sequence analysis. The expanded region also contains a number of disease genes. These include genes associated with susceptibility for asthma, schizophrenia, corneal dystrophies, low-frequency hearing loss, Treacher-Collins syndrome, various types of myeloid disorders including acute myeloid leukemia, Cockayne syndrome, spinal muscular atrophy, split hand/split foot (DSS1), polyposis coli. The putative colorectal cancer tumor suppressor MCC, Zinc-finger Protein 131 associated with lymphadenogenesis, and the Leukemia Inhibitory Factor Receptor (LIFR) are other disease associated genes in the region.

The isolation of BACs is based on a combination of colony hybridization and PCR approaches using STSs obtained mostly from public databases. Contigs are expanded by end-sequence STS walking. Contigs are oriented using STSs developed from known genes and ordered genetic and RH markers. All clones are sized by pulsed-field gel electrophoresis, and their map locations are confirmed by fluorescent in situ hybridization. The size of overlaps between BACs is determined by comparison of restriction fragments from a single endonuclease digest.

We have so far mapped 2463 clones to 5q. Ninety-four percent of these are BACs. A total of 2341 STSs have been employed in the contig forming process. Over 50% of the STSs were derived from clone ends. We have in excess of 120 contigs from the distal 65 Mb of 5q ranging in size from 200 Kb to 4.2 Mb.

Clones with minimal overlap that form contigs as determined by the STS content and restriction maps are selected for sequencing. To date, 390 clones on the q arm of chromosome 5 have entered the sequencing pipeline, totaling approximately 45 Mb of unique target or 71% of the clone insert total.

Detailed information on STS and restriction maps can be found at our Web site: (http://www-hgc.lbl.gov/human-maps.html)

77. High Throughput Fingerprinting and Contig Assembly to Supply Sequence Ready Templates to the JGI-PSF

Linda Meincke, Robert Sutherland, Connie Campbell, Joe Fawcett, Phil Jewett, Lynn Clark, Cliff Han, Larry Deaven, and Norman Doggett
Joint Genome Institute, Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545
meincke@telomere.lanl.gov

LANL's mapping goal for FY99 is to produce 48 Mb of sequence ready maps (40% of the JGI production goal). Our primary target is the q arm of chromosome 16 for which approximately 5000 BAC clones have been identified. A probe-content map is first constructed with the use of overgo probes (see abstract by Cliff Han for details). Gaps in this map are closed by PCR-screening with BAC-end STSs of a single-tier pooled BAC library (see abstract by

David Torney for details). Clones are selected for fingerprinting based on map order and DNA is prepared in 96 well deepwell dishes. This DNA is restricted with EcoR1, also in a 96 well format, and is run on twelve fingerprinting gels. Gels are stained with ethidium bromide and images are captured on the Biorad Fluor-S MultiImager system and scored using the Bio Image Advanced Quantifier software. Fragment data is generated and organized into Excel spreadsheets and processed into a Sybase mapping database. Input files for contig assembly are generated from the Sybase database and passed to GRAM, a program that provides graphical representation and utilizes algorithms to assist in contig assembly. Minimal tiling sets of clones are then selected from the GRAM contigs and these are fingerprinted with two additional enzymes. Overlap among the tiling set is confirmed with the additional fragment data using GRAM. These confirmed clones are then released for sequencing.

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

Informatics

78. The Genome Annotation Collaboration: An Overview

Jay R. Snoddy, Morey Parang, Sergey Petrov, Richard Mural, Manesh Shah, Ying Xu, Sheryl Martin, Phil LoCascio, Kim Worley[1], Manfred Zorn[2], Sylvia Spengler[2], Donn Davy[2], Chris Overton[3], **Edward C. Uberbacher**, and the Genome Annotation Consortium
Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; [1]Baylor College of Medicine, Houston, Texas; [2]Lawrence Berkeley National Laboratory, Berkeley, California; and [3]University of Pennsylvania, Philadelphia, Pennsylvania
ube@ornl.gov
http://compbio.ornl.gov/gac

The Genome Annotation Consortium is organizing software and database development projects toward a common goal of providing as much value-added annotation as possible on a genome sequence framework. The consortium is applying computational analysis modules and information technologies to the output of genome sequencers. We have developed a prototype system and process that will be presented at the Oakland workshop. We are also interested in forging new collaborations to add value to the genome sequence and annotation framework. Desired collaborations should improve the analysis process or the underlying technologies that are required for this analysis. This basic annotation process includes the following steps:

1. Acquisition of genome sequence data and other data that can be readily attached to genome sequences;
2. Assembly of sequence data into a consensus genome sequence framework;
3. Genome-scale analysis of sequence and other data that predict genes, gene products, or other features and integration of existing experimental information onto that genome sequence framework; and
4. Large-scale analysis of genome-based catalogs of genes and proteins that add homologous, functional, phylogenetic, and other types of relationships among the genes and proteins.

The outputs of our desired process include:

1. An assembled genome sequence framework;
2. Genes and features attached to that framework;
3. Catalogs of genes and proteins encoded by genomes;
4. Links among genes, proteins, genome maps, homologous relationships, phylogenetic trees, and other relationships for computational and experimental genome data.

Our current prototype is being applied to the output of all the large-scale genome sequencing centers for human sequences. We are adding genome mouse and microbial sequences to our prototype (see abstract of Larimer et al. for microbial analysis). As part of the initial prototype, we have established a data-acquisition component that retrieves data from genome center web sites and GenBank. This acquired data, for example, includes clone-contig overlap that is not always in the GenBank/EMBL/DDBJ entry. We have established a sequence-assembly component that creates a consensus genome sequence framework by assembling the different clone sequences. In addition, we acquire other experimental observations that can be linked to that genome-sequence framework during annotation (e.g., ESTs, STSs, cDNAs).

We have developed a number of analysis modules, including GRAIL-EXP modules (see abstract of Xu et al.). We have integrated these analysis modules in a data-analysis process that creates a comprehensive genome-wide analysis (see abstract of Shah et al.). This comprehensive analysis process will be updated to ensure that new data can be added to the genome sequence framework. We have made progress in adding navigation and summary reports (see abstract of Snoddy et al.).

We also have made progress on the difficult issue of data storage and management that can organize this diverse experimental and computational data (see abstract by Petrov et al.). We have produced different catalogs of genes and proteins including (1) GenBank annotated genes, (2) Genscan-predicted genes, and (3) GRAIL-EXP-predicted genes (including a subset of genes that have some EST evidence for expression). We have produced a Java-based interface (the Genome Channel Browser v. 2.0) and an HTML-based data-access method. These interfaces, other planned interfaces, and other progress will be presented at the Oakland meeting.

The analysis modules used in the comprehensive genome-analysis processes also will be available as public servers (see abstract of LoCascio et al.). These servers would permit users to analyze their new data or subsets of public data. Some of these analysis modules also will be portable and could be applied at a number of sites beyond the consortium member sites, including genome centers. We expect that our data-analysis process and computational infrastructure will also foster other genome-based, large-scale computational biology, including prediction of protein structure and modeling of biological systems.

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

Morey Parang, Richard Mural, Manesh Shah, Doug Hyatt, Miriam Land, Jay Snoddy, Edward C. Uberbacher, and the Genome Annotation Consortium
Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee
ube@ornl.gov
http://compbio.ornl.gov/gac

We are developing and deploying a series of interface tools for visualizing and querying the reference human genome and other genomes assembled and annotated by the Genome Annotation Consortium (see related abstracts). The Genome Channel Browser is a Java 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 provides a diverse set of functional features in a DNA sequence including PolyA sites, CpG islands, repetitive DNA, simple repeats, STSs, GRAIL2 and Genscan exons, as well as GRAIL-EXP and Genscan gene models and their respective protein translations. The underlying information and evidence for genes and other features is presented in a variety of text windows, graphics windows, and summary reports.

The new version of the Genome Channel Browser (v2.0) offers such improved user interface and additional capabilities as the following:

- 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.
- Pattern searches against genomes that return genome location and context of related sequences (Visual Genome Search Server).

In addition to Java-based browsing, several HTML-based interfaces to Genome Channel data are being developed. Genome, chromosome, contig, and clone summary reports, gene and protein lists, homologies, and other features are available for browsing and querying. For example, the chromosome summary report includes the following for each chromosome: (1) the acquired genomic sequences (including clones from GenBank/NCBI/DDBJ and genome centers); (2) the assembled sequences (number of contigs, nucleotide length of assembled contigs, and estimated percent completion of contiguous sequence for each chromosome); and (3) genes (including number of GenBank human annotated genes, Genscan predicted genes, total GRAIL-EXP genes, and those that have at least partial EST evidence); and other features. Browsable and queryable lists of genes, their protein translations, and homologs are compiled with links to related information in multiple databases. Compilations of other features, including STS sequences that are attached to genome sequence contigs, link the genomic sequence to genetic and RH maps.

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 polymorphisms and mutations.

80. Genome Annotation Data Management and Data Administration: Developing Summary Results for User Navigation, Genome Research, Improved Data Processing, and Quality Metrics

Jay R. Snoddy, Miriam Land, Sheryl Martin, Morey Parang, Inna Volker, Denise Schmoyer, Manesh Shah, Sergey Petrov, Edward C. Uberbacher, and the Genome Annotation Consortium
Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory,
Oak Ridge, Tennessee
v8v@ornl.gov
http://compbio.ornl.gov/gac

Summary reports of the genome annotation data and the underlying data management required to generate them are being constructed. A goal is to create these reports from a robust, queryable, and scalable data management system (see abstract of Petrov et al.). Some of these summary reports will be available as online HTML documents. These summaries can help improve four primary areas.

- **User Navigation:** Some reports are being constructed that can allow users to understand the available annotation data of the Genome Annotation Consortium and to easily allow users to navigate among that data. In addition, query interfaces will be needed and constructed (also see abstract of Parang et al.).

- **Research:** Some reports and queries can allow GAC and external researchers to analyze the characteristics of the developing genome sequence framework and the annotation attached to that framework. This allows for a comprehensive analysis of the currently available genome annotation since the analysis is consistently applied across all the available genome data.

- **Data Processing:** Some reports can help ensure that data flows smoothly through the several automated and manual steps in the annotation process. These current steps include data processing through different analysis modules (see abstract of Shah et al.). There is also a need to support data acquisition, cleansing, and curation; while these steps should be as automated as possible, there will be some manual data administration. Reports will be needed to assist the GAC personnel in the acquisition of new data. Summary reports will be needed to assist the GAC collaborators in flagging potential inconsistencies in the underlying data for data cleansing and curation. Reports could be generated so GAC collaborators could obtain the sets of annotation results applicable for research or integrating into a subsequent analysis process.

- **Quality:** Some reports can provide metrics for the Genome Annotation Consortium that can assist in quality control, quality assurance, and quality improvement of data analysis, data administration, and processing. These reports can help us measure, monitor, and improve the annotation quality and integration of the overall system.

For each genome, chromosome, sequence contig, and clone, a set of summary reports is being created. Hyperlinks to underlying data and more details will be linked to these reports.

Several general observations can be made now from the current snapshot of the data, and details from a later snapshot will be presented at the Oakland workshop. There are 7 to 10 times more predicted gene models (both GRAIL-EXP and Genscan gene models) than gene models annotated in GenBank. The majority of the predicted GRAIL-EXP genes do have one or more ESTs that are used in the gene modeling. A third to half of the gene models that predict putative protein sequences have a reasonable BLAST hit to known proteins in Swiss-Prot (BLAST with an Evalue <=1.0e-4). By this BLAST hit criteria, about 3 times more predicted genes appear to have good homolog candidates than there are annotated genes in the GenBank archival record.

By the time of the Oakland workshop, we hope to display several online summary reports that can demonstrate the current state of genome annotation for genomes, chromosomes, contigs, and sequenced clones. This should provide users with the results of the different but integrated data-management and processing steps that we employ in genome annotation. We would be interested in suggestions for other reports or queries that others may find useful.

81. Data Management for Genome Analysis and Annotation: Engineering a Fundamental Infrastructure for Data that Supports Collaboration in Genome Annotation

Sergey Petrov, Jay R. Snoddy, Michael D. Galloway, Sheryl Martin, Miriam Land, Morey Parang, Tom Rowan, Denise D. Schmoyer, Manesh Shah, Inna E. Vokler, Edward C. Uberbacher, and the Genome Annotation Consortium
Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee
ptv@ornl.gov
http://compbio.ornl.gov/gac

The GenomeDataWarehouse, GDW, is a heterogeneous information system created to store and support data from multiple sources. GDW currently is being developed and filled with data. The purpose of GDW is to provide data management to highly diverse and distributed sets of data. One goal is to create and support a new form of on-line analytical processing (OLAP) that is suitable for the complex world of genome research data. Another goal is to provide the right kind of data management that can encourage several groups in the Genome Annotation Consortium to collaborate in adding experimental and computational data to a developing genome sequence framework. The data management will need to provide data to the different data-analysis modules that ORNL and other collaborators are creating and support the linkages among the underlying experimental data and computational data produced from those different modules. This data warehouse will assist in the production of several user interfaces that are described in other abstracts or

are under development; there will not be one monolithic interface to GDW.

Conceptually, GDW consists of three parts: archival data sources, the kernel, and data marts. Currently, GDW is based on two Sybase servers, an SRS server, and data files running on networked Sun workstations. One Sybase server is dedicated to a copy of the Genome Database and, the second to kernel databases and a developing Genome Channel database. The archival data sources include sets of data from community databases (e.g., GenBank, SwissProt, Prosite, and GDB). The GDW Kernel is a set of databases used to store identification data on biologically meaningful objects and cross-references regardless of object origin, structure, and representation. The data marts are precompiled data sets reflecting the logic of a particular interface.

For archival data sources, we occasionally must internalize and manage community data within ORNL computers for a variety of performance, update, and querying reasons. These archival data are attached to the evolving genome sequence framework. We are using an SRS server (a product of the EMBL/EBI) for archiving and maintaining much of these data (http://ash.lsd.ornl.gov/srs5/). Our implementation of the SRS server at ORNL provides access to 31 community flatfile databases. We are evaluating the use of SRS and other mechanisms for serving annotation data that we are creating.

The data warehouse kernel provides the underlying mechanism that manages data in this complex area where we cannot enforce transactional control and data integrity of some underlying archival data. Given the difficulties and constraints in technology and available resources, our warehouse doesn't require integration of all data from different sources and does not completely enforce global integrity; data are stored "as is". At the same time, a mechanism is needed to provide cross-references between information on same objects in different data marts and the relationships that originate from the data sources and our analyses. The Kernel consists of several databases storing IDs of objects found in archival data sources, their classifications, and relationships — including cross-references. The structure of the kernel databases itself doesn't depend on structure of objects and relationships found in original sources. All databases in the kernel have almost-identical logical structure; data was divided among several databases to improve kernel performance only. Each database represents relationships between objects and their classes in a meta-closed way; every class and relationship is represented as an object and, therefore, information expressible in the database can include relationships between classes and relationships, as well as classification of relationships. Flexibility of chosen data representation allows us to include new data sources on the fly and to represent new classes of objects and new relationships found in genomic data. This approach comes at the cost of performance, but these databases are not meant to be routinely accessed by users.

The data marts are the read-only data sources that users routinely access to analyze and navigate the data. Data marts are compiled under the control of the GDW kernel. Each data mart reflects the internal logic of user interfaces and software systems that are focused around one aspect of genome data. Our current Genome Channel data system is the first example of these data marts. The Genome Channel data mart organizes data around genome structures and features on the assembled genome sequence framework. In the future, we will be developing other data marts, including a gene and protein catalog. Although this may contain data similar to parts of Genome Channel, the data system and interface are to be organized around genes, proteins, and the relationships among genes and proteins (including what is known and what we can predict about homology, phylogeny trees, protein families, and function).

We are developing several interfaces that will allow users access to data marts and mechanisms to query and navigate among data marts and other data

sources. We are trying to give the user some flexibility in altering the view of the data. We are also exploring the application of the Internet standard, XML, as a method of expressing some annotation data; this should allow the user a lot of flexibility in altering data presentation at the client browser without going back to the data mart and altering the underlying content. We anticipate that a number of initial HTML prototypes will be available by the time of the Oakland meeting and hope to acquire more feedback on our efforts.

82. Genome Channel Analysis Engine: A System for Automated Analysis of Genome Channel Data

Manesh Shah, Morey Parang, Doug Hyatt, Michael Galloway, Richard Mural, Kim Worley[1], Edward C. Uberbacher, and the Genome Annotation Consortium
Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee and [1]Baylor College of Medicine, Houston, Texas
x9m@ornl.gov
http://compbio.ornl.gov/gac

The Genome Channel Annotation Toolkit currently incorporates several exon- and gene-prediction programs as well as other kinds of feature-recognition systems and database homology search systems. Most of these analysis systems have been developed by the Genome Annotation Consortium collaborators, while some systems have been obtained from other researchers who make their code available but are not currently consortium members. The exon- and gene-recognition systems include GRAIL, GRAIL-EXP, Genscan, and Genie. Feature-recognition systems include the GRAIL suite of tools: CpG island, PolyA sites, simple repeats and repetitive DNA elements. Database homology systems include NCBI BLAST and Beauty postprocessing.

The Genome Channel Analysis Engine is an automated system that facilitates the analysis of contig sequences contained in the Genome Channel repository. It schedules and distributes the various processing tasks on several networked computer systems in a concurrent, pipelined mode to best utilize the available computer resources and to achieve optimal throughput. The scheduling is organized in terms of analysis epochs. At the start of each cycle, a data-refresh procedure is executed to detect and compile a list of all new and updated contig sequences that the sequence data-retrieval engine has incorporated in the Genome Channel staging area since the previous cycle's data refresh. It also checks the database source ftp sites for updated versions of all databases required by various analysis tools and updates local copies as necessary.

A master process then starts up servers on the available machines, including a PVM process for GRAIL analysis modules. Using a combination of Perl scripts and C programs, the analysis engine automatically runs the analysis tools on new contigs. The master process distributes the tasks as required to servers running on other machines. Some tasks are performed in parallel, including the GRAIL analysis tools (PVM) and the GRAIL-EXP Blast search (MPI). In addition, once protein translations have been obtained for the predicted genes and exons in a contig, they are immediately piped to a Beauty postprocessing server for a detailed homology search. The ultimate goal of this scheduling and distribution scheme is to reduce the time required to process 100 Mb of data to under 24 hours. Using the currently available resources, the analysis engine can process 100 Mb in about 72 hours.

Analysis processing is currently being performed at ORNL and is also being deployed at Lawrence Berkeley National Laboratory (LBNL). The computational infrastructure at ORNL consists of a cluster of 15 DEC Alpha workstations, 2 Sun HPC 450 UltraSparc servers, and a 200 GB Network Appliance RAID disk storage unit. These resources are barely adequate for handling the current rate of growth of sequence data. We are pursuing several strategies to deal with the anticipated rate of sequence generation and the consequent growth in compute and storage requirements. Some of the most compute-intensive tasks are being ported to the Paragon supercomputer at ORNL and to the supercomputers at LBNL. We also plan to evaluate the High Performance Storage System (HPSS) at ORNL for data storage.

83. GRAIL-EXP: Multiple Gene Modeling Using Pattern Recognition and Homology

Ying Xu, Manesh Shah, Doug Hyatt, Richard Mural, **Edward C. Uberbacher**, and the Genome Annotation Consortium
Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee
ube@ornl.gov
http://compbio.ornl.gov/gac

GRAIL-EXP is a multiple gene-modeling system that combines information from the analysis of EST homology with pattern recognition to construct accurate gene models. We believe that these current improvements in GRAIL-EXP represent fundamental advances in gene-modeling accuracy and computational performance. Currently, the system is being used extensively by the Genome Annotation Consortium to provide comprehensive genome-wide annotation for genomic DNA sequence from human, mouse, and other model organisms as well as several microbial organisms. GRAIL-EXP is used in this context to analyze long stretches of human and mouse DNA sequences (contigs that span tens of thousands to more than a million bases) to correctly identify and characterize the large numbers of genes found in such sequences.

Computational methods for gene identification in human genomic sequences typically consist of two phases: coding-region recognition and gene modeling. Although several effective methods for coding-region recognition are available, parsing the recognized coding regions into appropriate gene structures remains a difficult problem. GRAIL-EXP addresses the problem of multiple gene identification, using a set of biological heuristics and information available from sequence homology with available EST and mRNA sequences.

GRAIL-EXP uses GRAIL for predicting exons in a sequence. GRAIL evaluates all possible exon candidates in a DNA sequence and groups the high-scoring candidates into overlapping clusters. Those containing repetitive DNA elements are filtered out based on BLAST alignments of the exon candidates with a repetitive DNA database.

In the next phase, the system uses BLAST to identify all EST and mRNA sequences (obtained from GenBank dbEST and TIGR's human transcript sequence database) that have a sufficiently high BLAST alignment score with the candidate exons. The system also extracts information useful for the subsequent gene-modeling phase from each matched entry in the database. This results in a set of alignments for each exon candidate.

In the gene-modeling phase, an optimal gene model is constructed from the predicted exon candidates and the alignment information using dynamic programming. A set of nodes, one for each exon candidate or aligned-EST-sequence pair, is created. Each node is assigned a score based on its GRAIL score and the BLAST score for that alignment. The best-scoring gene model ending at each node is calculated using a recursive algorithm. Each exon is examined in three possible roles (as being the initial, middle, or terminating exon of a gene model). The algorithm assigns penalties and rewards at each step based on reading-frame mismatch, existence of in-frame stop codon, and terminating exon not ending in a stop codon. A node that uses the same EST as the previous node is assigned a reward that significantly outweighs the penalties. This guarantees that an EST that matches multiple exon candidates will have overriding influence on the gene model.

If the optimal gene model incorporates a set of one or more matched ESTs, then the system determines if any regions of the ESTs were not covered by the gene model exons. In this case, the system tries to locate the missing EST fragments in the appropriate intervals of the genomic sequence. If located, that region is added to the gene model as an exon.

GRAIL-EXP, a complex system with several logical components and numerous subcomponents, has been designed and implemented as a modular system. This is convenient for distributing various analysis tasks on multiple computers to achieve higher throughput. The system currently runs on a cluster of 10 DEC Alpha workstations and is able to analyze around 1 Mb of genomic sequence in about 15 minutes. Work is under way to achieve significant speedup by porting the most computationally-intensive modules to the Paragon supercomputer in the Center of Computational Sciences at ORNL and to similar platforms at Lawrence Berkeley National Laboratory. A Java-based graphical user interface has been developed to provide an interactive environment for the analysis of user-supplied DNA sequences. The system will be made available to the genome community via the public GRAIL server at ORNL in early 1999.

84. High-Performance Computing Servers

Phil LoCascio, Doug Hyatt, Manesh Shah, Al Geist, Bill Shelton, Ray Flannery, Jay Snoddy, Edward Uberbacher, and the Genome Annotation Consortium
Life Sciences Division and Computer Sciences and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee
locasciop@ornl.gov
http://compbio.ornl.gov/gac

Advances and fundamental changes in experimental genomics brought about by an avalanche of genome sequences will pose challenges in the current methods used to computationally analyze biological data. We are constructing a computational infrastructure to meet these new demands for processing sequence and other biological data within the Genome Annotation Consortium project, for genome centers and for the biological community at large. To cope with this 20-fold data increase, we have been developing the necessary high-performance computing tools to address this scaling challenge. As part of the Department of Energy's Grand Challenge in computational genomics, we have developed a number of applications to form part of a high-performance toolkit for the analysis of sequence data.

The initial tools we wish to include in the toolkit are high-performance biological application servers that include BLAST codes (versions of BLASTN, BLASTP, and BLASTX), and codes for sequence assembly, gene modeling (e.g., GRAIL-EXP), multiple sequence alignment, protein classification, protein threading, and phylogeny reconstruction (for both gene trees and species trees).

The tools and servers will be 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 this high-performance toolkit 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 that the GAC and our users can ask are not greatly affected by hardware limitations.

The system'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 the administrator having policy control over agent behavior, and namespace profile.

At the atomic transaction level of detail, clients and servers behave as expected, with clients issuing requests and servers responding. A higher level of transaction detail permits a much more complex model of operation where clients can be operated from within servers, and servers can be directed to propagate replies. This nested transaction model is very powerful for developing decoupled calculation and query facilities. The complex interaction is completely transparent to the user because all transactions are controlled by a query agent.

The GRAIL-EXP gene-recognition application will be deployed as a server that uses this model. The application derives alignment services from BLAST servers elsewhere. Internally, the GRAIL-EXP server is composed of a number of independent components that interact as a nested set of transactions. The

ability to assign different resources to different components is an extremely important feature for maintaining a credible load-balancing scheme.

Due to the logical decoupling of the query infrastructure, we are able to produce a model with both excellent scaling abilities and fault-tolerant characteristics. In testing the ability to run multiple instances of GRAIL-EXP and 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.

The overall software engineering design has been constructed very carefully to provide a nonspecific use of distributed resources, through the neutral application programming interface (NAPI) layer. NAPI is used to encapsulate the functionality required for distributed operation while utilizing the currently available resources. The underlying infrastructure is subsumed using PVM (parallel virtual machine) for robust heterogeneous operation and MPI (message passing interface) for homogeneous application development, optional where available. Other infrastructure ports can be accommodated (e.g., JAVA RMI), but the focus is now on the design of the high-level component model functionality and semantics.

Located at Oak Ridge National Laboratory within both the Center for Computational Sciences and the Computational Biosciences section, the development testbed consists of three super computers (Intel Paragons), some SGI SMP machines, and a DEC Alpha Workstation cluster. We are rapidly approaching alpha-stage deployment testing; after testing performance and stability, we can deploy the framework to NERSC, other high-performance computing sites, and other collaborators.

85. DOE Joint Genome Institute Public WWW Site

Robert D. Sutherland and Linda Ashworth
Los Alamos National Laboratory, Los Alamos, New Mexico, and Lawrence Livermore National Laboratory, Livermore, California
rds@lanl.gov

The Joint Genome Institute (JGI) has rebuilt its public WWW site to improve presentation of sequence and mapping data, and make data access more efficient. This site includes: (1) combined access to data from LANL, LBNL, LLNL, and the new Production Sequencing Facility (PSF), (2) information about the JGI and its member institutions, (3) links to member institution's WWW pages and other relevant WWW sites, (4) physical maps for regions being sequenced by the JGI, (5) links to entries submitted to public databases, (6) links to the JGI FTP server, (7) finished sequence data with associated quality information, and (8) information and WWW-links to promote public education and understanding of Genetics and the Human Genome Program. We have also upgraded our extremely fluid internal site which supports data sharing and communication between the four sites. This work is funded by the United States Department of Energy. URL: http://jgi.doe.gov

86. JGI Informatics and the PSF Network

Tom Slezak, Mark Wagner, Lisa Corsetti, Sam Pitluck, Arthur Kobayashi, Mimi Yeh, Brian Yumae, and Peg Folta
Joint Genome Institute; Lawrence Livermore National Laboratory, Livermore, California and Lawrence Berkeley National Laboratory, Berkeley, California
slezak@llnl.gov

The component labs of the JGI sometimes appear to have little more than their funding source in common,

both in biology and informatics. Meeting the stiff FY98 production goals was more of a miracle of elasticity of existing systems and people than any sort of fusion of component parts or philosophies. It is only now that we are actually working together in the PSF that we are being truly compelled to develop a new common culture. In some senses, informatics is being used as both the carrot and stick with which to overcome many of the differences, gratuitous or not, that have evolved at the JGI member labs.

Consolidation in informatics can only occur if there is corresponding consolidation or similarity in the underlying biological methods. Systems developed for transposon-based sequencing are highly dissimilar from those developed for shotgun strategies. Similar differences occur in mapping systems, which in the JGI have ranged from Claris Draw to Sybase. It is not feasible to make dramatic changes to processes that are under extreme production pressures; we are challenged to provide gradual, seductive improvements that bring about unification without derailing production ramps.

The long-term vision for JGI informatics is that information should be available in a JGI-centric view throughout the entire process: from mapping (via several methods), to production sequencing (allowing for multiple methods), to annotation and submission (allowing for several styles of manual and semi-automated annotation), and eventually to a range of functional genomics and structural biology. We are in the very earliest days of implementing this dream, as we struggle with production ramp rates unmatched anywhere that demand full effort from our best people who would otherwise be building our new systems. We will discuss our early efforts and our aspirations for this important first year of true "Joint-ness".

The PSF network has been designed to accommodate growth and flexibility in light of uncertain future demands. It features a cable plan that allows any jack to be a network, digital voice, or analog fax line as needed. The network is segmented into a high-reliability subnet for DNA sequencer data acquisition and primary server/storage functions, and another subnet for office and lab computers. We acknowledge the excellent work and contributions of the LBNL Communications (Sig Rogers), LBLnet (Ted Sopher), and ESnet (Jim Leighton) staffs in making the PSF network happen on time and on budget.

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

87. Verification of Finished Sequence at JGI-LLNL

Karolyn J. Burkhart-Schultz, Amy M. Brower, Arthur Kobayashi, Matt Nolan, Melissa Ramirez, and Jane E. Lamerdin
Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550
burkhartschultz1@llnl.gov

The JGI-LLNL sequencing group submitted 8.6 Mb of genomic sequence to the NCBI database in the 1997-1998 fiscal year. This accomplishment represents an 8 fold increase over our submissions for the previous year. An integral part of our finishing process is verification. The verification process allows an independent assessment of the validity of the finished assembly and final consensus sequence of each large insert clone (i.e. cosmid or Bac) project. Verification involves: 1) re-checking the finisher's validation of the assembled clone/project; 2) independent re-assembly of all the reads in the project with all finisher edits removed and comparison of this "no-edits" consensus to that submitted by the finisher; and 3) analysis of the extent and quality of the overlap of the finished clone with adjacent clones in the sequencing tiling path.

A finished project is submitted for verification along with a validation report prepared by the finisher. The verifier uses this report, as well as Consed and LLNL-developed tools to identify regions where strict standards for quality and double stranding may not be met. The verifier "re-checks" any "problematic" or difficult regions encountered during the assembly process. In addition, the final consensus is "digested" and the fragment sizes compared to those obtained by restriction mapping data compiled for each cosmid or

Bac. At least three digests (e.g. BamH1, BglII, EcorI, EcoRI/BglII, or Xho1) are used in these comparisons and any significant deviations between the map and sequence data are flagged.

The purpose of the "no-edits" re-assembly of a project is to remove any possible biases introduced by the finisher in the process of obtaining contiguous sequence. Currently this assembly is performed using an earlier version of the Phrap assembly engine. If the product of the "no-edits" re-assembly is not contiguous, the reasons for any breaks are examined and explained. Similarly, base discrepancies between the "no-edits" and finished consensus sequences are examined to determine which contains the valid basecall. If the contig breaks or sequence discrepancies cannot be resolved, the verifier may request that further data be generated.

Completion of the verification process requires resolution of all issues discussed above and that the final assembly and consensus sequence are supported by the data. A verification report with explanations of various issues is added to the validation report in the project directory. Rigorous verification of finished sequence assures the integrity and the quality of the final submitted sequence in the public database.

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

88. Informatics for Production Sequencing at LLNL

Arthur Kobayashi, David J. Ow, Matt P. Nolan, Stephan Trong, Tory Bobo, Tom Slezak, Mark C. Wagner, T. Mimi Yeh, Lisa Corsetti, Jane Lamerdin, Paula McCready, Evan W. Skowronski, and Anthony V. Carrano
Lawrence Livermore National Laboratory, Livermore, California
kobayashi1@llnl.gov

This past year, LLNL contributed over 8.6 MB of high-quality finished sequence to the Joint Genome Institute total of 20.9 MB. This represents an increase of more than 500% over the amount finished by LLNL last year (1.5 MB). We have managed to support this increased throughput through incremental improvements to our existing informatics infrastructure.

Our current system features extensive sample tracking, quality-control checks and reporting on every sequence run on our ABI sequencers, automated prefinishing, and an integrated suite of robotic workstations which perform rearraying, sample preps, etc. Most of our interfaces have been converted to WWW-based forms, and all of our sample information is stored in a Sybase relational database. Data is transferred through an automated sample sorting system over a dedicated 100 MB/sec ethernet local area network segment.

Over the next few years, our throughput will continue to increase at an aggressive rate. We are also faced with the challenge of relocating our core sequencing facility to Walnut Creek while developing our finishing capabilities at LLNL. We are currently working on improving our existing system through increased automation, barcode labeling of samples, and a new database schema. Some of these projects are described in more detail in other posters.

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

89. A Workflow-Based LIMS for High-Throughput Sequencing, Genotyping, and Genetic Diagnostic Environments

Peter Cartwright
Cimarron Software, Salt Lake City, Utah
pc@cimsoft.com

The widespread application of new genetic and DNA-based technologies and techniques to important medical and biological problems has resulted in an explosion of data and information. Advanced information systems are needed to comprehensively collect, analyze, and manage these critically important data.

This project has developed, marketed, and supported a family of workflow-based software products that addresses the specific challenges and needs of managing genetic and DNA-based information for the gene and drug discovery processes. The Cimarron Workflow System will accelerate the creation, modification, automation, and re-use of laboratory data collection and analysis activities.

Cimarron has augmented its existing Activity Model and LIMS tool kit with a central, well-defined Workflow Modeling and Management capability. The resulting Workflow Model and augmented tool kit have been used successfully in customer systems within both academic and industry.

Cimarron has continued to augment the Cimarron Workflow Model and the LIMS tool kit with advanced routing and queue management capabilities, a graphical workflow modeling application by which a domain expert can design a lab system, and other advanced information system capabilities, and will continue to market this tool kit to customers within genome labs and production facilities. This project will result in a commercial tool for building workflow-based laboratory information management systems. A large market exists for such a tool within the domain of high-throughput facilities for sequencing, genotyping, and genetic diagnostics.

90. A Simulation Extension of a Workflow-Based LIMS

Peter Cartwright
Cimarron Software, Salt Lake City, Utah
pc@cimsoft.com

The Human Genome Project is a complex scientific enterprise of national significance whose success will be greatly accelerated by effective project management and planning tools. This SBIR project is delivering such a tool by integrating simulation (computer modeling) software with laboratory information management databases. Central to the HGP are high throughput molecular biology laboratories which critically depend on cost effective management of complex experimental and production workflows. This project is developing software to simulate laboratory workflows under real and what if scenarios. This software is unique in that it derives its workflow model and configuration parameters from the real laboratory workflow, as stored in its operational laboratory information management system.

Phase I of this project was devoted to feasibility evaluation, design finalization, and technology assessment for an integrated laboratory information management / simulation facility. The findings and results confirmed that the integration was feasible, powerful, and conceptually elegant. Customers of Cimarron Software have helped shape and evaluate a prototype system, and confirmed the marketability of a fully engineered product.

Phase II of this project is building a fully engineered product for specifying, launching, monitoring, saving and comparing simulation runs. The interactive system is being packaged as a Java applet, and hence web enabled. This simulation facility is a natural and commercially valuable extension of Cimarron's core technology. Essentially all of Cimarron's current customers have indicated they would purchase such an extension if it were well integrated into their systems. Longer term, it is expected that this simulation / database combination will have similar benefits to the broader workflow management market.

91. A Graphical Work-Flow Environment Seamlessly Integrating Database Querying and Data Analysis

Dong-Guk Shin[1], Lung-Yung Chu[1], Lei Liu[1], Nori Ravi[1], Joseph Leone[2], Rich Landers[2], and Wally Grajewski[2]
[1]Computer Science & Engineering, University of Connecticut, Storrs, CT 06269-3155 and
[2]CyberConnect EZ, LLC, Storrs, CT 06268
shin@engr.uconn.edu

In the past, we have been has been very successful in developing a graphical ad hoc query interface capable of accessing heterogeneous public genome databases. This project aimed at developing a suite of user-friendly software designed to aid computational biologists in accessing various independently managed genome databases. This software makes the SQL query syntax manageable for the novice user and makes unfamiliar complex genome database schemas quickly understandable for less experienced persons. Furthermore, this software aids users in quickly expressing semantically-correct ad hoc queries. The impact of wide distribution of this software is expected to be significant. Computational biologists who have been reluctant to use genome databases themselves would begin to query the databases themselves, thanks to the numerous user-friendly features built-in the easy-to-use graphical interfaces. Most distinctively, the computational biologists will be able to ask cross-database queries against multiple genome databases that are springing up within the genome community.

We are currently investigating ways of embedding this user-friendly database access tool into a graphical work-flow management environment. Although being able to query various genome databases easily and being able to make associations between remotely located data is essential, we consider that it is imperative to produce an integrative environment in which both database querying and data analysis activities can be carried out seamlessly in a cohesive manner. This requirement is critical because many biologically significant questions are centered around performing analysis programs. In the proposed scenario, a computational biologist should be able to store persistently results of a Blast search into a database and subsequently should be able to query and cross-link filtered Blast result with existing genome databases. Similarly, a computational biologist should be able to perform a database query and funnel the query results into a subsequent Blast or Fasta search, etc. Furthermore, the user should also be able to conveniently change analysis or query results into a certain data format to be input to tree alignment programs, like CLUSTALW, or tree building programs, like Phylip and Puzzle, for final stages of analysis visualization. The ultimate goal of this project is to produce an easy-to-use work-flow editing environment in which the user can easily specify data flow involving both database querying and data analysis. This project is being pursued in collaboration with JGI.

Acknowledgments
[1]The author's work was supported in part by the NIH/NHGRI Grant No. HG00772-05.
[2]The author's work was supported in part by the DOE SBIR Phase II Grant No. DE-FG02-95ER81906.

92. Data Visualization for Distributed Bioinformatics

Gregg Helt, Suzanna Lewis, Nomi Harris, and Gerald M. Rubin
Berkeley Drosophila Genome Project, University of California, Berkeley, California
gregg@fruitfly.berkeley.edu

A significant challenge for genome centers is to make the data being generated available to biologists in a succinct and meaningful way. We are addressing this problem by creating extensible, reusable graphical components specifically designed for developing genome visualization applications. With careful planning and design this toolkit enhances the ability

for others and ourselves to rapidly develop genome visualization applications for the Internet and as editing applications .

The visualization toolkit is written in the Java programming language. Our applications are being designed to read XML files. We will describe our component based approach and demonstrate a variety of visually distinct applications that are all based upon the same underlying components. These range from whole genome views of chromosomes to multiple alignment views of sequence data. The different views of the data are interconnected via shared, common data models that underlie the various displays. The views are also linked to external databases to retrieve and display textual data on selected features. Different types of analysis can be dynamically performed on the data and the results displayed on the maps. This analysis code is dynamically loaded when requested, to minimize the initial loading time.

Other groups can reuse this work in various ways: genome centers can reuse large parts of the genome browser with minor modifications, bioinformatics groups working on sequence analysis can reuse components to build front ends for analysis programs, and biology labs can reuse components to publish results as dynamic Web documents.

The BDGP has established a collaborative agreement with a small startup, Neomorphic, to develop improved versions of the widgets initially begun at the BDGP. This will provide the additional resources necessary to allow us to provide commercial-grade, thoroughly documented products. Under the terms of this collaboration, all the products of the collaboration will be made available to academic and government institutions for a nominal fee.

93. A Figure of Merit for DNA Sequence Data

Mark O. Mundt, Allon G. Percus, and **David C. Torney**
Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, New Mexico
dct@lanl.gov

We have implemented a new measure of the quality of sequence data. Given a sample of sequence data and its phred scores, our figure is the predicted net error rate for finished sequence that would be generated from a given coverage of sequences of comparable quality. It is reasonable to use our figure of merit for assessing the quality of batches of sequence data for continuous quality control in sequencing factories.

This figure of merit avoids the complexities of fragment assembly. It assumes that the sequence reads occur at random positions, uniformly across the target sequence, and with each orientation being equally likely. The figure of merit is then the expected composite rate of erroneous basecalls, for a given coverage in sequences comparable in quality to those of the sample dataset.

Thus, an average is taken over the different ways in which the bases and their associated phred scores "align" on the different base positions.

We implemented this figure of merit in an executable Java computer program, available by anonymous ftp from cell.lanl.gov., in the directory pub/fom. The inputs to the program are the standard phred output for the sample of sequences whose quality is to be assessed, and also the desired coverage to be used. There are essentially no restrictions upon the size of the dataset: it could consist of the phred scores for one or multiple reads. This program can trim off specified sequences, such as vector sequences. We illustrate the result of using different coverage parameters for three datasets, one of which has noticeably lower quality. It is surprising that expected net error rates manifest no trace of a dependence on the parity of the number of times sequenced, arising from "majority rule" statistics.

94. Probabilistic Basecalling

Terry Speed, Lei Li, Dave Nelson, and **Simon Cawley**
University of California, Berkeley
scawley@stat.berkeley.edu

Basecalling is the process of converting raw data from automated DNA sequencing machines to a sequence of bases. The process is typically subdivided into the tasks of color separation, mobility shift correction, deconvolution and decoding. A probabilistic model of the process is presented, at center of which lies an Hidden Markov Model (HMM). The class of HMMs is chosen for its flexibility and for the availability of efficient algorithms for training and decoding. Performance of this approach to basecalling is compared with the standard available basecalling algorithms.

95. The FAKtory Sequence Assembly System

Susan J. Miller, Eugene W. Myers, Kedarnath A. Dubhashi, and Daniel E. Garrison
University of Arizona, Tucson, Arizona
susanjo@cs.arizona.edu

The FAKtory system facilitates sophisticated prescreening, assembly, contig layout manipulation and finishing for DNA sequencing projects. The system allows specialized database configuration, definition of operations performed on sequences, constraint-based assemblies, and convenient linking to post-assembly analysis software. Each user can configure various FAKtory displays and can select the degree of control desired over each of the operations. Retaining our original design goals of customizability and sound ergonomics, we have continued to add features to FAKtory.

Among our recent additions to the system are allowing input of pre-trimmed sequence data and input of SCF format data. We have made a dramatic speed improvement and have increased the sensitivity of overlap detection in the underlying FAKII assembly kernel. A report generator has been added to FAKtory and we have developed a separate graphical viewer for comparing CAF or ace format assemblies. The finishing editor now has the capability of running the GENSCAN program and displaying any exons found in the six reading frames. Optionally, the consensus sequence can be filtered through the CENSOR or RepeatMasker programs before searching for exons.

Currently under development are improved overlap detection using quality values and enhancements to the Finishing editor. In addition, fragment prescreening based on quality numbers will be incorporated into the system. We also plan to add the capability of importing GAF format assemblies into FAKtory and to integrate the graphical comparison tool for more convenient comparisons of alternate assemblies generated by FAKII or by other assemblers.

96. Hidden Markov Models in Biosequence Analysis: Recent Results and New Methods

Christian Barrett, Mark Diekhans, Richard Hughey, Tommi Jaakkola, Kevin Karplus, David Kulp, Stephen Winters-Hilt, and **David Haussler**
Computer Science Department, University of California, Santa Cruz, CA 95064
haussler@cse.ucsc.edu

Currently there is an acute need for effective methods for locating genes in DNA sequences, along with their splice sites and regulatory binding sites, and for classifying new proteins by their predicted structure or function. Hidden Markov Models (HMMs) have proven to be useful tools for these tasks. We have recently extended the HMM-based genefinding system Genie so that it can simultaneously incorporate protein homology and EST information to improve gene finding. We have also built a new

library of HMMs for protein families and tested our methods against other methods for the detection of remote homologies between proteins in a large scale experiment conducted at the Laboratory for Molecular Biology in Cambridge. Results showed the method to be superior to other methods, including PSI-BLAST, the nearest competitor. Finally, we have developed a new method of biosequence classification called the Fisher kernel method. Here an HMM (or any parametric generative model for a family of biosequences) is used to embed the sequences into a linear space with a natural inner product defined using the Fisher information matrix. One can then employ a variety of classification methods to discriminate members of the family from nonmembers, for example, support vector machines. We present experiments for the protein superfamily classification problem that show the Fisher kernel method is superior to existing HMM approaches, and to simpler methods such as BLAST. In particular, the method is better at finding remote homologs in nearly all the 33 protein families we tested, including G proteins, retroviral proteases, interferons, and many others.

97. Java Based Restriction Map Display

Mark C. Wagner, Jan-Fang Cheng, Steve Lowry, Robert Sutherland, Norman Doggett, Laurie A. Gordon, and Anne S. Olsen
Joint Genome Institute, Lawrence Livermore National Laboratory, 7000 East Avenue, L-452, Livermore, CA 94550
wagner5@llnl.gov

The Clone Resources Task of the Joint Genome Institute generates large quantities of physical mapping data. These data require the use of a graphical display for ease of understanding. The restriction mapping data generated by the JGI (chromosome 5 data by Steve Lowry and Jan-Fang Cheng, chromosome 16 data by Robert Sutherland and Norman Doggett, and chromosome 19 data by Laurie Gordon and Anne Olsen) has been made available through the use of a Java based graphical interface.

This display enables both the biologists of the JGI and the public at large to view the current restriction maps of the JGI. We have recently upgraded the display to permit users to select a specific area of the chromosome, a gene/marker of interest, a particular map of interest, or a particular clone. These methods permit the user to get directly to the area to be studied. We have also added the display of selected genes and markers to the map, so that relationships between various biological entities are more readily apparent. We have made our sequencing information available directly from the mapping display.

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

98. Mapping Data

Lixin Tang, Jeremy Boulton[1], Benjamin Liau[1], Hui Zhang[2], Wei Qin[3], Sung Ha Huh[1], Yicheng Cao, **Robert Xuequn Xu**, Glen George[1], and Ung-Jin Kim
Division of Biology, [1]Computer Science, [2]Atmospheric Chemistry, and [3]Electrical Engineering, California Institute of Technology, Pasadena, CA 91125
xux@cco.caltech.edu

In building the physical contig maps of human and other chromosomes, we found that a high degree of automation as well as a high degree of flexibility is usually desired at the same time. Despite remarkable progresses in physical mapping projects and biocomputing, a computer software tool that allows both is not yet available albeit there are some mapping tools that heavily rely on automation and allow limited human intervention.

AceDraw is a graphical software tool that we have developed in order to facilitate drawing, updating, and database entry of physical mapping data. It is capable of reading the content of ACeDB database, allows graphic display and freehand editing of the physical maps, and dumps the physical map in a file that can be parsed by ACeDB. The program was written in C++ and used a freely available relational database, MySQL, as a backend. It helps to manage

the mapping data in a way that can be organized and viewed in which the location and order of clones and landmarks and overlaps between clones are displayed along the length of chromosomal regions. It is similar to ACeDB in this respect, and it can actually take in the ".ace" files and draw more presentable and intuitive contig maps using a large number of colors, with each assigned color associated with certain traits of associated data, such as "sequenced", "fingerprinted", and/or "end sequenced", etc. It can save the mapping data back into ACEDB format after drawing, modification, and updating, thus allowing using of the special functionalities provided in ACEDB.

An important feature of AceDraw is that it is a freehand drawing tool that enables easy human intervention to resolve conflicts in the contig map by direct manipulations on clones and markers such as clone searching, moving, resizing and color changing. AceDraw also allows easy creation, modification and deletion of clones and landmarkers without low-level editing of database files, and thus facilitates map construction. Since AceDraw is associated with relational database model, querying the database for an object of interest can be done easily. Moreover, AceDraw supports map output into high resolution postscript files for the printing of hardcopy maps.

99. A Relational Database and Web/CGI Approach in the Analysis and Data Presentation of Large-Scale BAC-EST Hybridization Screens

Robert Xuequn Xu, Chang-Su Lim, Bum-Chan Park, Mei Wang, Jonghyeob Lee, Aaron Rosin, Eunpyo Moon, Melvin Simon, and Ung-Jin Kim
Division of Biology, Caltech, Pasadena, CA 91125
xux@cco.caltech.edu

Large-scale hybridization, in which both probes and targets are in huge numbers, are frequently used in the genome projects to mass-produce positive screening results. In our project of "Construction of a genome-wide human BAC-Unigene resource", probes (ESTs) are pooled in groups of 20 according to a pre-designed 20x20 matrix, and the pooled and labeled probes are applied to BAC library filters in hybridization.

We have developed a complete data management, analysis and presentation system with a relational database tool, combined with a web server and several Perl scripts, for the deconvolution results of our massive BAC-EST screening. It features screen-by-screen progress reporting, detailed description of each probe, automatic statistics report generation, and some quality control functions (http://www.tree.caltech.edu/lib_D_Unigene.html). The methodology can be easily adapted to any other large-scale hybridization projects using similar probe pooling strategy. With the help of this tool, the probe pooling matrix is virtually unlimited, and therefore the efficiency of hybridization screening can be greatly improved. For example, for 10,000 probes, if they are organized into 100x100 matrix, only 200 hybridizations are needed (100 row pools and 100 column pools), instead of 10,000 individual hybridizations.

The initial hybridization results are collections of BACs that are positive to particular probe pools – row and column pools. Since there are 20 row pools and 20 column pools, 40 hybridizations are performed. In order to resolve the individual probe-BAC relationship, The pooled screening results are fed into a relational database scheme: First, a probe table PROBE_RC is constructed which includes fields PROBE, ROW, and COL (PROBE is the IMAGE clone ID; ROW is the row number of the probe in the 20x20 matrix; and COL is the probe's column number). In a 20x20 matrix, ROW ranges from 1-20, and COL ranges from 21-40. Then, when the pool-wise hybridization results are available, those results are entered into the positive BAC table as BAC, POS in which BAC is the positive BAC address identified from the filter screen and POS is the pool number of probes that that BAC is positive to. Apparently, the POS's value will range from 1 to

40. Finally, after data entry, a relational database tool is started to create views that split the positive BAC table into BAC_ROW view and BAC_COL view, and these two views are joined at the BAC field (i.e. BAC_ROW.BAC= BAC_COL.BAC) to create a new view BAC_ROW_COL which contains BACs that appear on both row-pool screens and column-pool screens.

The resulting view BAC_ROW_COL (which consists of BAC, BAC_ROW.POS and BAC_COL.POS fields) is joined further with the probe table, with the join condition "PROBE_RC.ROW= BAC_ROW.POS AND PROBE_RC.COL= BAC_COL.POS", and the final "deconvolution" table is generated as selecting BAC, ROW, COL, PROBE. Any possible individual positive BAC-probe relations are revealed in this table, and it can be grouped, sorted and reported through the relational database tool's internal reporting function and publish to the Web; or through custom designed Perl/CGI scripts.

100. A Distributed Object System for Automated Processing and Tracking of Fluorescence Based DNA Sequence Data

Michael C. Giddings, **Jessica M. Severin**, Michael Westphall, and Lloyd M. Smith
University of Wisconsin, Madison, Wisconsin
jseverin@chem.wisc.edu

We have been developing a system which allows for rapid and easy construction of data handling and analysis for DNA sequencing through the use of modular componentware which can be assembled into a working analysis system through an easy to use graphical interface. This system utilizes network-distributed object communications for data transfer and process synchronization allowing the integration of processing programs and components located on different hardware and software platforms. Our analysis system focuses on automated processing of raw data collected on four color fluorescence based DNA sequencing instruments to the point of base calling. This system employs a distributed object framework which allows easy integration of new components into the analysis system. A new component can even be an existing analysis program available as a precompiled command-line tool such as Phred or Phrap. We currently have 5 individual component servers implemented using this frame work. These correspond to the steps of input of new gel data into the system, automatic lanetracking, manual checking of lane tracks, lane extraction into trace data, and finally lane trace preprocessing and basecalling by BaseFinder. The last element is the "System Controller" which handles setup of system, data flow, information tracking, configuration of individual processing steps, handling and recovery of server failures, and parallel distributed load balancing to facilitate the use of multiple servers of the same type distributed across multiple machines. Parallel distributed load balancing will allow the system to scale with sequence processing demands by the simple addition of more processing computers and duplicate servers.

101. Arraydb: CGH-Array Tracking Database

Donn Davy, Daniel Pinkel, Donna Albertson, Steve Clark, Joel Palmer, Don Uber, Arthur Jones, Joe Gray, and Manfred Zorn
Lawrence Berkeley National Laboratory, Berkeley, Calfornia
dfdavy@lbl.gov

In collaboration with the UCSF Cancer Center, we have developed a database to track data from all stages of the production and use of the CGH (Comparitive Genome Hybridization) expression array slides produced robotically by separate LBNL-Cancer Center collaboration. The system tracks clones or DNA and their sources as they are selected for use in the array, the DNA preparation, the microtiter plates, print-run specifications, slide-printing runs, and slides printed. Researchers then log experiments performed on slides, with the resulting slide-images and analyses written back to the database, providing results which link back to the original clone or DNA and its source.

The system is implemented in an Oracle 8 database, 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 can provide a fine-grained control of security and accessibility.

102. BCM Search Launcher — Analysis of the Genome Sequence

Kim C. Worley and Pamela A. Culpepper
Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030
kworley@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://gc.bcm.tmc.edu:8088/search-launcher/launcher.html) 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 WWW 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. Our collaboration with the Genome Annotation consortium (http://compbio.ornl.gov/tools/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).

103. Profile Search

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

Protein Sequence & Profile Search Software Product

The ultimate goal of this project is to produce a software product that will facilitate the ability to query many protein sequence and profile databases. Platform portability has been tested to include Solaris, Linux, and Windows 95 platforms. This software builds an optimized database from the protein and sequence database files for maximum performance.
The current features of this software system include:

- Prosite database search for profile and pattern matches on a given sequence query.
- Command line & Gui Interface
- Search Parameters:
  - From / To Match Ratio
  - Inclusive / Exclusive match ratio selection
  - Cut off score maximum
  - Number of acceptable matches

♦ Full profile documentation display
- Search results are saved to a file to interface with other products. A visual display is also presented for the gui user interface. The detail of the search results are grouped under the accession id of the matching profile and grouped by match location on the query sequence.
- Java GUI and command line interface.

The scheduled features include:

- Query protein sequence databases for a given pattern.
- Incorporate the facility to accept varied formats of the query sequence.
- Interactive graphical display of search results.

104. Computer Analysis of DNA Sequence Data to Locate SECIS Elements

Michael Giddings, Olga Gurvich, Marla Berry, John Atkins, and Raymond Gesteland
University of Utah, Department of Human Genetics, 15 N. 2030 E. rm. 6160, Salt Lake City, UT 84112-5330
giddings@genetics.utah.edu

The selenocysteine insertion sequence (SECIS) has been found in a number of organisms. In Eukaryotes it consists of a particular structure in the 3' untranslated region which modifies the behavior of ribosomal translation within the coding region, causing selenocysteine insertion at a UGA codon instead of the translational stop which would normally occur. SECIS elements vary in structure, but all contain several common elements. Known examples contain a core group of 4-5 nucleotides that do not conform to usual Watson-Crick pairing rules, as well as a stem-loop structure which contains a group of 2-3 adenosine nucleotides as either a bulge or member of the upper loop.

Locating new SECIS elements through pattern searches in DNA databases poses some significant challenges. We are using a three-pronged approach to this problem. The first phase utilizes a fast, rough scan of large databases with Ross Overbeek's "patscan" software to narrow the field of possibilities. The second phase, currently being implemented, performs a refined analysis of the candidates, scoring and ranking them with a new algorithm in which neural network and fuzzy logic approaches are being explored. The third phase then performs visualization of the candidates, based on an algorithm that utilizes rules regarding the formation of SECIS elements to fold and display them for human analysis.

We present the implementation details of this system in its current form, as well as initial results for scans of several genomic databases using this system.

We would like to acknowledge the following supporters of this work:
DOE Grant #DE-FG03-94ER61817, "Advanced Sequencing Technology"
DOE Grant (no # assigned yet), "Genomic Analysis of the Multiplicity of Protein Products from Genes"
NIH Genome Training Grant #T32HG00042

105. Sequence Landscapes

Gary D. Stormo, Samuel Levy, and Fugen Li
MCD Biology, University of Colorado, Boulder, CO 80309-0347
stormo@colorado.edu

Sequence Landscapes are a graphical display of the word frequencies from a database (DB) for every word of every length in a target sequence (TS) [see Levy et al. Bioinformatics 14: 74-80, 1998]. If the TS and the DB are the same sequence this is a convenient method to detect all of the repeated sequences, of any length. However, we have been exploring the use of this approach for classifying regions of DNA sequence into functional domains, such as exons, introns, promoters, etc. Using DB from each class, the landscapes can be used to derive likelihoods that every region of the sequence belongs to each possible class. We think information can be combined with other types of information to help provide improved recognition algorithms. We are especially interested now in improving methods for determining promoter regions and transcription initiation sites. The information in the landscape can

also be very useful for determining the best oligos to use on DNA chips. One of the criteria to be used in choosing the best oligos are those that are most specific for the gene being assayed. Therefore one would like to pick, for each, the oligo which has the most mismatches to the most similar other sites in the genome. This can be accomplished easily and efficiently with the landscape information. We return a list of candidate oligos which can then be ranked by other criteria, including hybridization energy and TM.

106. Protein Fold Prediction in the Context of Fine-Grained Classifications

Inna Dubchak, Chris Mayor, Sylvia Spengler, and Manfred Zorn
E. O. Lawrence Berkeley National Laboratory, Berkeley, CA 94720
ildubchak@lbl.gov

Predicting a protein fold and implied function from the amino acid sequence is a problem of great interest. We have developed a neural networks (NN) based expert system which, given a classification of protein folds, can assign a protein to a folding class using primary sequence data. It addresses the inverse protein folding problem from a taxonometric rather than threading perspective. Recent classifications suggest the existence of ~300-500 different folds. The occurrence of several representatives for each fold allows extraction of the common features of its members. Our method (i) provides a global description of a protein sequence in terms of the biochemical and structural properties of the constituent amino acids, (ii) combines the descriptors using NNs allowing discrimination of members of a given folding class from members of all other folding classes and (iii) uses a voting procedure among predictions based on different descriptors to decide on the final assignment. The level of generalization in this method is higher than in the direct sequence-sequence and sequence-structure comparison approaches. Two sequences belonging to the same folding class can differ significantly at the amino acid level but the vectors of their global descriptors will be located very close in parameter space. Thus, utilizing these aggregate properties for fold recognition has an advantage over using detailed sequence comparisons. The prediction procedure is simple, efficient, and incorporated into easy-to-use-software. It was applied to the fold predictions in the context of fine-grained classifications 3D_ALI[1] and the Structural Classification of Proteins, SCOP[2]. In attempt to simplify the fold recognition problem and to increase the reliability of predictions, we also approached a reduced fold recognition problem, when the choice is limited to two folds. Our prediction scheme demonstrated high accuracy in extensive testing on the independent sets of proteins.

A WWW page for predicting protein folds is available at URL http://cbcg.nersc.gov

[1]Pascarella, S., Argos, P. (1992). Prot. Engng., 5: 121-137
[2]Murzin, A. G., S. E. Brenner, T. Hubbard and C. Chothia. (1995). J. Molec. Biol., 247: 536-540.

107. Comparative Analyses of Syntenic Blocks

Jonathan E. Moore and **James A. Lake**
Molecular Biology Institute, University of California, Los Angeles, CA 90095
Lake@mbi.ucla.edu

The comparative analysis of syntenic blocks common to the genomes of closely related organisms, such as those found in humans and mice, appears to have enormous potential to aid in the identification of gene boundaries, open reading frames (ORFs), and the interpretation of gene organization. Recently, pattern filtering, a new genome analysis tool has been developed. Pattern filtering methods appear to be able to obtain an optimal signal to noise ratio when used to search for ORFs and also simplify the analysis of

codon periodicities. In initial studies, it appears to be a sensitive and robust indicator of ORFs, and of gene structure and organization. Our major goal is to develop rapid, simple, and effective methods for analyzing syntenic blocks from human, mouse, *Drosophila*, and *Caenorhabditis* genomes using pattern filtering to optimally determine rates of evolution and thereby map ORFs, gene boundaries, regulatory regions, and introns. Preliminary experiments with syntenic blocks in human chromosome 12p13 and the corresponding region in mouse, and also experiments with mammalian mitochondrial DNAs, will be used to illustrate the potential of the method.

108. Sensitive Detection of Distant Protein Relationships Using Hidden Markov Model Alignment

Xiaobing Shi and **David J. States**
Washington University in St. Louis, St. Louis, Missouri
states@ibc.wustl.edu

Hidden Markov models are statistical models of the primary structure of a sequence family. In this poster, an algorithm to align hidden Markov models (HMMs) of protein sequences is presented along with the software implementation. Aligning HMMs provides a way to compare sequence families. Compared to pair-wise sequence alignment, HMM alignment is more sensitive to identify relationships between sequence families and requires less computation. Our algorithm uses dynamic programming to identify similarities between two HMMs. Two scoring algorithms are used: the local alignment algorithm, which identify the most similar segments from two HMMs, and the "glocal" alignment algorithm, which aligns the entire length of one HMM to a similar segment of the other model.

We have developed software to perform the alignment and set up a website allowing users to perform the alignment on the internet. Besides allowing users to input or upload HMMs, the website can build HMMs form user-inputted raw sequences or multiple alignments. All HMMs in the Pfam database are also available for aligning on that website. We also provided a method to generate and then align two random HMMs, the score of which can be used to determine the significance of a HMM alignment score.

We have used this software to align all pairs of HMMs in the Pfam database, and the result has revealed some interesting relationships between existing protein families that have not previously been recognized. For example, the high HMM local alignment score of the Sodium:solute symporter family (SSF) and the Amino acid permease family suggests that these two families are closely related. Other examples include the Tropomyosin family and the Filament family, the GerE family and the sigma70 family.

109. Multiple Sequence Alignment with Confidence Estimates

David J. States
Institute for Biomedical Computing, Washington University in St. Louis, St. Louis, Missouri
states@ibc.wustl.edu

Multiple sequence alignment (MSA) is the basis for many aspects of molecular sequence analysis including phylogenetics, motif detection and molecular modeling. Because the space of possible multiple sequence alignments is very large and the information accessible through sequence data is limited, there are often regions of a multiple sequence alignment that are not well determined. Here we develop a theory for assessing the confidence of multiple sequence alignment, describes software that implements this algorithm, and discusses the application of these methods.

A hierarchical approach to MSA is used in which each constituent sequence is related to the full alignment as a leaf in a tree of nearest neighbor relationships. The algorithm uses a progressive strategy for building the multiple alignment. Hidden Markov Models (HMM) are used to describe each sequence or collection of sequences. At each phase in the alignment calculation, all current models are compared with each other using a dynamic programming calculation to calculate the maximum

scoring local alignment. A new HMM is derived from the pair of models with the highest alignment score, and this new model replaces both of the previous models. The iteration is repeated until only a single HMM remains. A site specific confidence estimate, C, for pairwise alignemnts is calculated by comparing the likelihood for the optimal alignment passing through a pair of residues with the sum of the likelihoods for all alternative pairings of either the query or target residue.

$$C = \frac{\exp(\Theta_{ij})}{\sum_{k+j} \exp(\Theta'_{ik}) + \sum_{k+i} \exp(\Theta'_{kj})}$$

where Θ_{ij} is the optimal score for an alignment passing through any pair of residues i and j calculated using a forward and back dynamic programing algorithm [Vingron and Argos, Bishop and Thompson]. Note that the alternatives, Θ', include the possibility that the site is deleted or inserted as well as being a matched pair of residues. C has the form of a probability and is bounded by $0 < C < 1$. The overall confidence for a site in the multiple sequence alignment is calculated as the product of the confidence in the all of the pairwise alignments making up the full MSA.

The algorithm provides an efficient way to build HMMs for large families of unaligned sequences. A web site provide access to this tool is available at http://www.ibc.wustl.edu/service/msa

110. Improved Specificity and Sensitivity in Sequence Similarity Search Through the Use of Suboptimal Alignment Based Score Filtering

Lisa Gu and **David States**
Washington University in St. Louis, St. Louis, Missouri
states@ibc.wustl.edu

The specificity of molecular sequence similarity search is often limited by the presence of repetitive elements present in biological sequences. Both repeat filtering and biased content filtering methods have been proposed to alleviate these problems, however these methods can mask off large portions of some query sequences limiting the utility of subsequent searches. We have examined the use of suboptimal alignment to automatically identify robust regions of sequence similarity and use this indirectly to filter out the repetitive regions whose alignment is not definite. In this algorithm, the alignment confidence is assessed by comparing the score of the optimal alignment in a pair of residues are aligned with the highest score for an alignment in which the two residues are not paired. Varying degrees of stringency can be applied by raising the threshold for accepting an aligned pair. A "confidently aligned residues" (CAR) score is obtained by performing an optimal Smith-Waterman optimal alignment and subtracting the pairwise score for those residues pairs in that alignment that can not be confidently aligned.

Protein families rich in repetitive sequence were examined and members within the same family were aligned with each other. The results CAR scores were compared to those obtained using the XNU filter as a masking technique and WU-BLASTP (2.0) as the search algorithm. For the collagen family, whose members have extensive and highly repetitive regions, CAR based scoring is uniformly more sensitive in the detection of family members compared with XNU + BLAST. Alignments are missed by XNU + BLASTP as a result of excessive masking by XNU, but large numbers of false positive alignments are seen if

BLAST is run without XNU. On the other hand, XNU + BLASTP is, in some cases, able to detect regions of similarity in the myosin heavy chain family, which has some members with a minimal amount of repetitive region. For non-collagen, non-myosin repetitive sequence proteins, CAR scores detected a significant number of similarities missed by XNU + BLAST and in no case was a similarity detected by XNU + BLAST missed with CAR scoring. Our results can be explained by the fact that suboptimal alignment algorithm imposes a more stringent constraint on the alignment between two sequences than BLASTP. Moreover, since the members have minimal repetitive regions, masking by XNU does not cause a tremendous loss of information. CAR scores appear to be a useful tool for enhancing the performance of sequence similarity search in the face of repetitive sequence regions.

111. Screening for Large-Scale Variations in Human Genome Structure

S. MacMillan[1], C. Hott[1], D. Anderson[1], E. C. Rouchka[2], B. D. Dunford-Shore, B. Brownstein[1], R. Mazzarella[2], V. Nowotny[2], and **D. J. States**[2]
Washington University in St. Louis, St. Louis, Missouri
states@ibc.wustl.edu

The human genome is polymorphic at all scales ranging from single nucleotide polymorphism to cytogenetically visible translocation spanning tens of megabases, but it remains difficult to characterize variation between these scales. We have proposed a method for screening for the presence of large-scale structural variants in the human genome. To demonstrate the feasibility of our strategy, STS markers derived from regions of finished genomic sequence are used to screen BAC and PAC libraries derived from 9 individuals with coverage in excess of 20 fold using a hierarchical multiplex hybridization and PCR approach. Recovered clones are subjected to both end-sequence analysis and four-enzyme restriction (RE) digest fingerprinting. End-sequence reads are aligned with the reference genomic sequence and their separation is compared with the molecular size of the clone as determined by the sum of the RE fragments sizes. The set of restriction digests predicted from the region spanned by the end-sequence alignments is compared with the experimental digests. Our method is validated by applying them towards verification of three BAC sequencing projects from the Chen laboratory, demonstrating a fingerprint sizing accuracy of better than 1% for bands with molecular weight between 1.2 and 15 kb. Successful fingerprints and end-sequence were generated for all clones. No false positive or false negative calls were identified in 302 bands scored for comparison. To date, 61 markers spanning 2 megabase of sequence (color vision, BRCA2, and TCR beta) have been screened retrieving 249 clones. 15 sites have been identified where multiple clone demonstrate a consistent pattern of deviation from the predicted digest pattern, including the presence of novel bands as well as the absence of predicted bands. To validate variations, a second tier of RE has been implement to further characterize these variants and PCR assays are being developed to test for the presence of these variants in uncloned genomic DNA.

112. Probabilistic Physical Map Assembly

David J. States, Thomas W. Blackwell, John McCrow, and Volker Nowotny
Washington University in St. Louis, St. Louis, Missouri
states@ibc.wustl.edu

Physical map assembly is the inference of genome structure from experimental data derived on clones and markers, and map assembly is central to genome analysis. Map assembly depends on the integration of diverse data including sequence tagged site (STS) marker content, clone sizing, and restriction digest fingerprints (RDF). Like any experimental data, these data are uncertain and error prone. Physical map assembly from error free data is algorithmically straightforward and can be accomplished in linear time in the number of clones. However, the assembly of an optimal map from error prone data is an NP-hard problem [Turner, Shamir]. In this abstract we present an approach to physical map assembly that is based on a probabilistic view of the data and seeks to identify those features of the map that can be

reliably inferred from the available data. Based on our alternative approach, we achieve several goals. These include the use of multiple data sources, appropriate representation of uncertainties in the underlying data, the use of clone length information in fingerprint map assembly, and the use of higher order information in map assembly. By higher order information, we mean relationships that are not expressible in terms of neighboring clone relationships. These include triplet and higher order constraints (a+, b, c+ => b likely to be +), the uniqueness of STS position, and fingerprint marker locations. Probabilistic descriptions of the map provide an alternative approach to the problem of physical mapping. In this view, we assert that it is impossible to know which of the many possible map assemblies is correct. We can only state which assemblies are more likely than others given the available experimental observations. Parameters of interest are then derived as likelihood weighted averages over map assemblies. Ideally these averages should be sums or integrals over all possible map assemblies, but computationally this is not feasible for real-world map assembly problems. Instead, Gibbs sampling is used to asymptotically approach the desired parameters. Software implementing our probabilistic approach to mapping has been written. Assembly of mixed RDF and STS maps containing up to 60 clones can be accomplished on a desktop PC with run times under an hour. A JAVA based physical map viewing tool has also been written to display the results of these calculations.

113. Multi-Resolution Molecular Sequence Classification

David J. States, Zhengyan Kan, and Brian Dunford-Shore
Washington University in St. Louis, St. Louis, Missouri
states@ibc.wustl.edu

Classification is the most reliable and widely used basis for inferring macromolecular function from primary sequence. Beginning with the pioneering work of Margaret Dayhof, a number of sequence classification algorithms have been proposed based including sequence signatures (Prosite), profiles (blocks), HMMs (pfam), and transitive closure relationships (HHS and others). There are intrinsically conflicting constraints on domain classifications that makes it difficult to achieve satisfactory performance in all applications all of the time. Classes must be general enough to represent all of the members of a class, but this generality limits the information content of any single pattern and reduces the sensitivity with which members can be detected. Further, the stochastic nature of mutations may result in domain detection in some sequences and failure to detect domains in other closely related sequences. In transitive closure methods where we are attempting to infer domain structure from similarity relationships, variations in the extent of sequence covered by sequence alignments may further confuse matters and result in the failure to consistently recognize a domain. Instead the algorithm defines several related domains with overlapping membership and sequence extents.

Here we present a novel approach to molecular sequence classification that addresses some of these problems. A multi-resolution approach is employed in which sequences are first classified into transitive closure groups (TCGs) on the basis of high scoring global sequence alignments. These TCGs are then grouped into superfamilies based on inferred domain content and local sequence similarity relationships. All of the members of a TCG are assumed to have identical domain structure providing more redundancy in the data available for domain definition and avoiding inconsistent domain annotation between closely related sequences. To date, 14,227 transitive closure groups with more than two members have been defined in a classification of non-redundant protein sequences derived from SwissProt, PIR, OWL, TREMBL, and GenBank. Work on HMM representations for TCG and the grouping of TCGs into superfamilies is on-going. Relating the annotation and literature reference accessible through primary sequence classification

with the structure-based classification being developed at SDSC is proposed as a goal for the Molecular Sciences Thrust.

114. PQ Edit—A Web-Based Database Table Editor and the Relational Database Abstraction Layer

Brian H. Dunford-Shore and **David J. States**
Washington University in St. Louis, St. Louis, Missouri
states@ibc.wustl.edu

High throughput genome sequencing necessitates the production and use of large amounts of information. To make such information usable, it must be easy to enter, edit, search, and manipulate and the information systems must evolve with changes in experimental design and formula. Relational database servers and tools such as form generators or Microsoft Access™ provide tools to implement part of the solution but does not offer client-independent, flexible, instantly useable data entry and editing. The PQ Edit program and the RDBAL (Relational Database Abstraction Layer) Perl 5 modules were written to fill this gap. PQ Edit is a Perl CGI script that provides general purpose, client-independent, web-based database table editing for relational database tables using automatically generated CGI forms. PQ Edit allows the editing of any database table as it currently exists despite any changes made to the definition (schema) of the tables. PQ Edit provides a reasonable entry form so that the (re)writing of data entry forms for relational databases is unnecessary most of the time. PQ Edit is based on the RDBAL Schema Object—a general purpose Perl library for retrieving database definitions and for searching or manipulating data. RDBAL is an abstraction layer for relational (SQL) databases, which allows middleware independent SQL execution and database schema (catalog) information retrieval. RDBAL tries to 'hide' details of implementation for scripts so that they need no changes to run on different platforms such as Linux, Solaris, or Windows NT and Sybase/MS SQL server or Oracle. The RDBAL Perl library uses Perl 5 objects to make it easy to retrieve information about a particular database's schema. The database connection is cached in the schema object. Database entities (tables, views, and procedures), their field's properties and their index information are retrieved when the schema object is created. Table primary and foreign key relationship information is also retrieved for all tables in a database. Currently PQ Edit and RDBAL support Transact-SQL (Sybase and MS SQL) and Oracle relational database servers via SybaseDBlib, ODBC, or DBI/DBD drivers. Other types of data sources, such as AceDB, are possible. PQ Edit and RDBAL have been tested and used on the Apache and MS IIS web servers and on Solaris and Windows NT.

115. Allele Frequency Estimation from Sequence Trace Data

David G. Politte, David R. Maffitt, and **David J. States**
Washington University in St. Louis, St. Louis, Missouri
states@ibc.wustl.edu

Parametric model fitting of unprocessed sequencing-gel trace data and a least squares optimization algorithm provide a method for accurately determining allele frequencies of a single nucleotide polymorphism in a population. The method uses trace data from one or two homozygous individuals as a reference to estimate allele frequencies present in DNA derived from a pooled population. A parametric model is fit to each of the traces to estimate the amount of each of the four fluorescent dyes that is present at each site. The parameters estimated from each trace are then normalized to account for scalar variations due to differences in the amount of template or sample loaded. The parameters estimated from the trace of the heterozygous individual or from the mixture are viewed as a weighted sum of the parameters estimated from the traces of the homozygous individuals. The weights, or allele frequencies, are estimated by minimizing the sum of squared errors between the linear combination of homozygous traces and the mixed trace. Comparison of allele frequencies estimated by our method to known frequencies at polymorphic sites in three pools of CEPH individuals show that our method is accurate to ~10% even when

only a single homozygous reference is available. The allele frequency estimator is accessed via a portable Java based interface that reads ABD or SCF format trace files and allows the user to interactively select sites of interest. When a site has been identified, allele frequency estimation calculations are performed remotely using HTTP mediated requests. Our method is automatic and much less labor intensive than previous approaches. Software is available at http://www.ibc.wustl.edu/software/allele-estimation .

116. Improved Detection of Single Nucleotide Polymorphisms (SNPs)

Scott L. Taylor, Natali Kolker, and Deborah A. Nickerson
Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, WA 98195
stay@u.washington.edu

Single nucleotide substitutions and unique base insertions and deletions are the most common form of polymorphism and disease-causing mutation. Based on the natural frequency of these variants, they are likely to be the underlying cause of most phenotypic differences among humans. Because of their functional importance, their frequency, and amenability to automated genotyping, large mapping of single nucleotide polymorphisms (SNPs) are now underway for the human genome. We have developed a computer program known as PolyPhred which together with Phred, Phrap, and Consed automatically identifies single nucleotide substitutions using fluorescence-based sequencing. Over the past year, we have evaluated several approaches to increase the accuracy and selectivity of PolyPhred. We will present information on a binning process that greatly improves SNP identification by PolyPhred and that speeds the analysis of sequence diversity in human genes.

117. The Genome Sequence DataBase (GSDB): Advances in Data Access, Analysis, and Quality

C.A. Harger, M. Booker, A. Farmer, W. Huang, J. Inman, D. Kipart, C Kodira, S. Root, F. Schilkey, J. Schwertfeger, A. Siepel, M.P. Skupski, D. Stamper, N. Thayer, R. Thompson, J. Wortman, J.J. Zhuang, and M.M. Harpold
National Center for Genome Resources, 1800 Old Pecos Trail, Suite A, Santa Fe, NM 87505
cah@ncgr.org

Two primary foci of GSDB (http://www.ncgr.org/gsdb) located at the National Center for Genome Resources (NCGR), in Santa Fe, NM, are to expand the data access and analysis capabilities that are provided to researchers and to continue to improve and automate data quality assurance procedures. Substantial progress in both of these areas has been made during the last 18 months.

Recently NCGR has launched two data utilization tools which provide significant enhancements in data access and analysis capabilities. First, NCGR has begun implementation of sequence similarity searching by making the BLAST suite of algorithms available for researchers to search sequences in GSDB. The addition of sequence similarity searching complements the gene localization capabilities, e.g., MarFinder, already provided by NCGR. NCGR is planning to expand this analysis capability by making Frame Search, Clustalw, and Smith-Waterman publicly available.

Second, NCGR has introduced Sequence Viewer, a platform-independent graphical viewer for sequence data in GSDB. This tool provides easy visualization of sequence and associated annotation together with simple text presentation of non-graphical data. The benefits of Sequence Viewer are augmented by its integration with other GSDB data access tools, such as Maestro, a web-based database query tool. The availability of Sequence Viewer provides a significant improvement in the ability to retrieve and

review sequences and associated annotation from GSDB.

During the last year NCGR has also made important advances in data quality assurance procedures. First, NCGR has improved the suite of programs that automatically acquire data from the International Nucleotide Sequence Database Collaboration (IC) databases. These improvements have resulted in a significant reduction of the amount of manual curation necessary to ensure quality and completeness of data acquired from the IC. Second, NCGR has implemented daily curatation of several database fields, including source molecule, chromosome, and the taxonomic information. The increased data consistency resulting from these efforts allows NCGR to provide researchers with flexibility in selecting BLAST search sets. For example, these search sets could range from the entire database to a variety of taxonomic-based subsets or to individual human chromosome sets.

These enhancements and improvements are designed to make GSDB more accessible to researchers, extend the rich searching capability already present in GSDB, and to facilitate the integration of sequence data with additional types of biological data.

118. Analysis of Ribosomal RNA Sequences by Combinatorial Clustering

Poe Xing, Casimir Kulikowski, Ilya Muchnik, Inna Dubchak, Sylvia Spengler, Manfred Zorn, and Denise Wolf
DIMACS and CS Department, Rutgers University, New Jersey; Lawrence Berkeley National Laboratory, Berkeley, Calfornia
xingpoe@cs.rutgers.edu

In our present study, multi-aligned sequences of eukaryotic and procaryotic small subunit rRNA were analyzed using a novel clustering procedure in an attempt to extract subsets of sequences sharing common features. This procedure includes two new models - data segmentation and a core separation and consists of the following four steps: a) sequence segmentation and identification of likely conserved segments according to some specific criterion (i.e. gap frequency); b) clustering of sequences based on each of these segments; c) intersection of clustering results from all the conserved segments; d) comparison of the results of the steps a)-c) with a phylogenetic tree.

Segmentation is a result of global optimization of a new objective function that finds the most homologous consequent partition of a given set of aligned sequences. It was developed as a very efficient and simple dynamic programming procedure. Segmentation was performed on the multi-alignment of 409 eucaryotic rRNA sequences and, independently, on the multialignment of 6205 procaryotic rRNA sequences. In both cases we tested different levels of granularity of segmentation by changing total number of segments. The position and the length of the conserved segments in the multi-alignment were relatively stable. Segment-specific score function discriminated sequence segments mostly composed of gaps from those less frequently interrupted by gaps. Among eucaryotes we found seven conserved segments with less than 20% gaps in the segment, and among procaryotes - nine conserved segments with less than 40% of gaps.

Using the novel clustering procedure, we examined these, minimally interrupted by gaps, segments of the multi-alignment. Every segment was analyzed individually by the clustering procedure, which extracted optimal (exact and unique) subset of 'correlated elements' among all aligned sequences. From each segment we obtained one core cluster and one complementary tail cluster. In the core cluster, all sequences were close to each other and also similar to the consensus sequence of the corresponding segment. For this reason, we call the core cluster a 'homogeneous group', and the tail cluster a 'heterogeneous group'. The sizes of the homogeneous groups derived from each segment in eucaryotes were 284, 344, 361, 343, 366, 335, 317 sequences, respectively. From this result, we can see that rRNA sequences are indeed highly conserved in eukaryotic organisms since among 409 analyzed sequences, a majority belongs to the homologous groups. In procaryotes homogeneous groups derived from each segment contained 3838, 3343, 2378, 2447, 4312, 2641, 1491, 837, 3179 sequences, respectively.

Although a relative fraction of sequences in the homologous groups is lower than in eucaryotes, it is still significant and reached 69 % for one of the segments.

Clusters resulting from different conserved segments are fairly consistent. We performed the intersection of all clustering results on all segments by labeling each sequence with an occurrence label. Although there are 27, or 128 types of occurrence patterns possible among seven conserved segments of eucaryotes, only 33 patterns were observed, which indicatesd a significant deviation from a random sequence classification. Furthermore, of the 33 patterns, only 4 patterns could be considered significant because they were shared by a large enough number of sequences. To integrate clustering information from all conserved segments, we ranked each sequence according to its occurrence label, and aggregated them based on the rank. We found that 249 of the 409 rRNA sequences fell into the group with the highest rank: 7, which means they are homologous as determined by clustering of all seven conserved sequence segments. In procaryotes distribution of patterns is also non-random, although clusters resulting from 9 different conserved segments are not very consistent. Among 29, or 512 types of occurrence patterns, 320 patterns were observed, and among those only 11 combinations were represented by more than 100 sequences and 249 by less than 20. 59 Sequences fell into the group with the highest rank: 9, which means they were homologous as determined by clustering of all the nine conserved sequence segments. There were 415, 705, and 940 sequences in the clusters of rank 8, 7 and 6 respectively, which also suggests a substantial homology among the sequences. There are 470 sequences in the cluster of rank 0, meaning that these sequences share little similarity among all nine conserved segments.

Prevalence of the homologous sequences in all segments indicates that using only conserved sequence segments greatly reduces the effect of random information from non-conserved or nonessential sequence fragments on the evaluation of relationship between sequences. Comparison of the phylogenetic classification of the rRNA sequences with our clustering results showed that each phylum usually corresponds to one or two major clusters that are adjacently ranked in our analysis. The advantage of presented algorithm is that: (1) We avoid the interference of frequent gaps that exist in the multi-aligned sequences, and base our clustering only on uninterrupted sequence segments potentially corresponding to essential functional units of rRNA molecules. (2) By identifying these conserved segments, in future we will be able to develop new procedures to cluster unaligned sequences. (3) The algorithm provides the means to apply a polynomial clustering procedure of O(n2) by using the special properties of the objective function defined on the conserved segments.

Since our clustering is based on an objective criterion defined by specific statistical properties of the sequences, and uses no prior knowledge of the biological relevance of the sequences being analyzed, the consistency of our clustering result with an independently derived phylogenetic organization of the associated organisms suggests that it is feasible to apply such an objective and stable clustering method to discover phylogenetic correlations among large number of biological sequences. It can serve as a framework to organize these sequences in an efficient and easily searchable manner.

119. Ribosomal RNA Alignment Using Stochastic Context Free Grammars

Michael P.S. Brown
University of California at Santa Cruz, Santa Cruz, California
mpbrown@cse.ucsc.edu

I present a method for aligning ribosomal RNA using a well principled probabilistic method that models pairwise interactions in a computationally efficient manner, Stochastic Context-Free Grammars

(SCFG's). I show this method has superior performance characteristics in relation to several other alignment methods. This method has applications in areas such as phylogenetic tree reconstruction. A webserver is located at http://www.cse.ucsc.edu/research/compbio/ssurrna.html.

SCFG's have been used previously for modeling structures such as tRNA (Sakakibara94, Eddy+Durbin94) and have been demonstrated to have the highest specificity of any method (Lowe97). This performance comes from SCFG's pairwise modeling ability as well as it's probabilistic foundations that allow specific estimations of parameters such as gap and mutation costs. Unfortunately SCFG's require a relatively high computational cost, O(n^3), where n is the length of the sequence. Previous work to reduce this cost has been done by preprocessing databases with a fast approximate method and presenting only likely strings to the SCFG for further processing (Lowe97). I extend this idea in a new direction using Hidden Markov Models (HMM's).

HMM's are used not only to preprocess the database but to also constrain the SCFG computation in a principled way using posterior decodings. These constraints allow the analysis of large molecules such as rRNA to be done using the full power of complex SCFG models in a reasonable amount of time. I analyze several methods for RNA structure prediction and show that SCFG's have the highest specificity and generalization capabilities using the Ribosomal Database Project alignment of small subunit rRNA as a gauge (Maidak97).

Alignment of ribosomal RNA is important for several reasons. Historically, rRNA was used by Carl Woese to relate all organisms and reconstruct the tree of life (Woese77). Recently, Norman Pace pointed to an opportunity for an environmental genome survey in which rRNA is gathered from the environment to provide a sequence based snapshot of the microbial biodiversity (Pace97).

In order to relate organisms based on their biosequence identity, a multiple sequence alignment is necessary. Indeed, alignment is a very important process in correct phylogenetic tree reconstruction (Morrison97). Current methods of computing this alignment involve a combination of computer alignment with human fine tuning (O'Brien98). This leads to a computational bottleneck as evidenced by the large number of unaligned rRNA sequences in the Ribosomal Database Project. Full analysis of widescale environmental biodiversity projects will exacerbate this problem.

Stochastic Context-Free Grammars are an automatic method of determining RNA alignment using a well principled probabilistic model that accounts for pairwise interactions in a computationally efficient manner. SCFG's have superior performance properties in relation to other methods and have several important application areas including phylogenetic tree reconstruction.

(Sakakibara94) Y.Sakakibara et. al. Nucleic Acids Research. (22)5112-5120. (1994).
(Eddy+Durbin94) S.R.Eddy and R.Durbin. Nucleic Acids Research. (22)2049-2088. (1994).
(Lowe97) T.Lowe and S.Eddy. Nucleic Acids Research. (25)955-964. (1997).
(Maidak97) B.L.Maidak et al. Nucleic Acids Research. (25)109-111. (1997).
(Woese77) C.R. Woese and G.E. Fox. Proc. Natl. Acad. Sci. USA. (74)5088. (1977).
(Pace97) N.R. Pace. Science. (276)734-740. (1997).
(Morrison97) D. Morrison and J. Ellis. Mol. Biol. Evol (14)428-441. (1997).
(OBrien98) E. O'Brien et. al. Bioinformatics. (14)332-341. (1998).

120. Ribosomal Database Project II

James R. Cole, B. Maidak, T.G. Lilburn, B. Li, C.T. Parker, S. Pramanik, G.M. Garrity, T.M. Schmidt, and Jim Tiedje
Center for Microbial Ecology, Michigan State University, East Lansing, Michigan
colej@pilot.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 have great potential value for functional genomics. In addititon they are widely used in molecular phylogeny and evolutionary biology, microbial ecology, organism identification, characterizing microbial populations, and in understanding the diversity of life. RDP-II is a value-added database that offers aligned and annotated rRNA sequence data, analysis services, and phylogenetic inferences derived from these data. These services are available to the research community through the RDP-II website (http://cme.msu.edu/RDP).

In December 1997, the RDP officially moved to The Center for Microbial Ecology at Michigan State University from its previous home at The University of Illinois. A new, greatly enhanced website, and a major data update (version 7) were released on July 31, 1998. The new data release, the first since June '97, contains 9835 aligned sequences, an increase of 66% over the previous release. In addition, this is the first release to be generated from a new custom dbms. Generating the release from the dbms provides the user with better, more consistent formatting of the data within sequence records, and consistent formatting of shared data (eg. reference data) between records.

The new RDP-II website offers a significant improvement over the older website. It exhibits a new, clean, easy to understand user interface. Most of the functions have been enhanced with easier user data input, and improved, more informative output. In addition, we offer several new functions, including a similarity matrix generator, a T-RFLP analyzer, and a java based phylogenetic tree browser. In the first full month of operation (August '98) the website handled 23,032 requests from 1399 distinct hosts in 40 different countries.

We are currently focused on reducing the delay between the time rRNA sequence data becomes available in the primary sequence repository (GenBank) and the time these sequences are available in annotated and aligned format through RDP-II. To that end, we are working on further automation of the sequence harvesting, alignment, and annotation procedures. In addition, we are working on procedures to enhance our phylogenetic tree building capability and to simplify user sequence submission. Our goal is to have data available in RDP-II within three months of its GenBank release.

Functional Genomics

121. The Regulatory Network of a Eukaryote

Matthew N. Ashby, Tod Flak, and Darren H. Wong
Acacia Biosciences, Inc., 4136 Lakeside Drive, Richmond, CA 94806
ashbym@acaciabio.com

Eukaryotic cells possess the ability to orchestrate the expression of thousands of genes in response to a changing environment. While numerous genome sequencing projects of eukaryotic model organisms are currently under way, only that of the yeast *Saccharomyces cerevisiae* has been completed. The modest size of the yeast genome, approximately 6000 hypothetical open reading frames, represents a significant opportunity to study the organization and inter-relationships of the regulation of gene expression on a genomic scale. The Genome Reporter Matrix (GRM) consists of a high density array of yeast colonies each harboring one of over 6000 yeast promoter-reporter fusions. The GRM can measure patterns of gene expression in living cells in response to external stimuli or mutations. The response of yeast exposed to an extensive panel of environmentally important compounds as well as exposure to ionizing radiation will be examined at the level of changes in gene expression. Compensatory changes in gene regulation will also be examined in response to a collection of mutations. Analyses of the 1300 expression profiles of a set of 864 reporters in response to pharmaceutical agents revealed the presence of 26 unique regulons. These analyses will be extended to over 6000 reporters in response to the proposed environmental stimuli. The generality of the regulons identified from these experiments will be assessed by a series of directed experiments in human cells in tissue culture. These experiments will provide a map or framework for the regulatory circuitry within a eukaryote and help determine the extent of the evolutionary conservation between yeast and human cells.

122. Genomic Hot Spots for Homologous Recombination

Jerzy Jurka, Jiong Ma, and Sun-Yu Ng
Genetic Information Research Institute, 1170 Morse Ave., Sunnyvale, CA 94089
jurka@charon.girinst.org

Non-LTR retrotransposons, or retroposons integrate at short, consensus-defined DNA targets in mammals in a process mediated by L1 element[1,2]. These targets appear to be hot spots for homologous recombination. We have determined that significant recombination occurs only in cells lacking p53 tumor suppressor protein, such as C33A cell line. Co-transfection of p53 gene to C33A inhibited the recombination. We have also studied recombinogenic effects of different mutations within the targets. The results will be presented. We will discuss implications of our research for understanding genomic instability in cancer and germ line cells as well as its potential applications in gene therapy.

[1]Jurka, J. Proc. Natl. Acad. Sci. U.S.A. 94: 1872-1877 (1997).
[2]Feng, Q., Moran, J.V., Kazazian, H.H. & Boeke, J.D. Cell 87: 905-916 (1996)

123. Development and Application of Subtractive Hybridization-Based Approaches to Facilitate Gene Discovery

Maria de Fatima Bonaldo, Brian Berger, and **Marcelo Bento Soares**
The University of Iowa, 451 Eckstein Medical Research Building, Iowa City, IA 52242
bento-soares@uiowa.edu

It is widely recognized that the generation of Expressed Sequence Tags (ESTs) from 3' terminal exons of cDNA clones randomly picked from libraries constitutes an efficient strategy to identify genes (Adams et al. 1992; Adams et al. 1991; Adams et al. 1995; Adams et al. 1993; Houlgatte et al. 1995; Khan et al. 1992; Matsubara and Okubo 1993; Okubo et al. 1992). However, it is important to acknowledge that despite its advantages, there are several problems associated with the EST approach. One of the problems commonly observed in large scale EST programs is the redundant generation of ESTs corresponding to the most common RNAs (i.e. mRNAs of the super-prevalent and intermediate frequency classes, mitochondrial RNAs, and rRNAs). This is a problem that can significantly impair the overall efficiency of a gene discovery program that relies solely on the generation of ESTs from cDNA clones randomly picked from standard libraries. The use of normalized cDNA libraries has been shown to expedite gene discovery in large scale EST programs (Berry et al. 1995; Hillier et al. 1996). Because in a typical normalized cDNA library the frequency of all clones is within an order of magnitude range (Soares et al. 1994), redundant identification of the most common RNAs is greatly minimized. Normalized libraries can be generated by a number of reassociation-kinetics based approaches (Bonaldo et al. 1996; Soares and Bonaldo 1996; Soares et al. 1994). It is noteworthy, however, that the process of normalization only contributes to minimize redundancies within (not across) libraries. Redundant identification of ESTs derived from mRNAs that are expressed in multiple tissues and therefore are represented in multiple libraries constitutes a major problem at advanced phases of gene discovery programs. The use of normalized libraries cannot help to solve this problem. Hence, we have argued that this problem can be more effectively addressed by the use of subtractive libraries that are progressively enriched for novel ESTs (Bonaldo et al. 1996). With this support from the U.S. Department of Energy, we have developed a subtractive hybridization-based gene discovery strategy, which we named "Serial Subtraction of Normalized Libraries", which involves the generation of ESTs from subtracted libraries enriched for novel cDNAs. Serial Subtraction of Normalized Libraries is an iterative approach whereby all arrayed cDNA clones from a library (which have been or will be used for generation of ESTs) are pooled and used as a driver in a subtractive hybridization with the library from which they originated. Since the representation of the driver population is significantly reduced in the resulting subtracted library, redundant generation of ESTs, regardless of abundance, is significantly minimized. Hence, every new library of a series is enriched for novel ESTs. Most importantly, however, this process enhances the proportional representation of rare transcripts rather significantly. It should be emphasized that such transcripts are likely to be missed in more random sampling approaches, unless very large numbers of ESTs are generated from a library, which inevitably ends up becoming costly and inefficient due to the very high redundancy levels that are reached. This strategy has been successfully applied in the rat gene discovery program that we are conducting at The University of Iowa with NIH support. We have been able to minimize redundancies rather significantly and thus maintain a high frequency of identification of novel ESTs (62 % overall average) after a total of approximately 32,000 ESTs submitted to GenBank since February 1998. Most importantly, we were able to identify approximately 20,000 unique clusters from a total of 32,000 3' ESTs, a gene discovery efficiency that is unprecedented in any EST program described to date.

124. Generation of Large-Insert Mouse cDNA Libraries

Lisa Stubbs[1], Jimmy Spearow[2], and Xiaojia Ren[1]
[1]DOE Joint Genome Institute and Human Genome Center, Lawrence Livermore National Laboratory, Livermore, CA 94550 and [2]Section on Neurobiology, Physiology and Behavior, University of California, Davis, CA 95616
stubbs5@llnl.gov

We have developed straightforward, reliable and efficient protocols for generating large-insert clone libraries from size-selected cDNA. We have found enzyme combinations and RNA preparation protocols that routinely produce double-stranded cDNA products with excellent representation of very large cDNA fragments (5-15 kb). After cDNA synthesis, the products are fractionated on sucrose gradients, and each size fraction is cloned and plated as a separate sub-library. Size fractions are chosen for screening according to information obtained from Northern blot analysis or from PCR screening of pooled sublibrary clones. Our most recent efforts, funded as part of the JGI functional genomics pilot program, have focused upon improvements in methods of library production and screening; we have also begun to experiment with new methods to normalize large insert pools before cloning. Using these improved protocols, we have recently created a series of new mouse cDNA libraries representing brain, thymus, ovary, and other mouse tissues. We will work with I.M.A.G.E. to share these resource libraries with the Genome research community as widely as possible, to serve as a resource for isolation of full-length cDNA clones.

125. The DOTS Resource for Gene Expression Analysis and Genome Annotation

Chris Overton, Brian Brunk, Jonathan Crabtree, Philip Le, and Jules Milgram
Center for Bioinformatics, University of Pennsylvania, Philadelphia, Pennsylvania
coverton@pcbi.upenn.edu

We have created a linking database that integrates a wide range of high quality, carefully analyzed information on eukaryotic transcribed sequences. This new resource, termed DOTS (Database of Transcribed Sequences, builds upon and substantially expands previous work on creating LENS, a database linking information on ESTs generated in the IMAGE/WashU/Merck project (http://agave.humgen.upenn.edu/lens). The DOTS resource supports gene expression analyses ongoing at Penn and large-scale genome annotation as part of the DOE sponsored Genome Annotation Collaboratory. The central organizing concept of the database is a representation for mature messenger and structural RNAs and their predicted sequences accompanied by links to genomic sequences and proteins, and associated information, e.g., gene expression arrays and gene expression experiments. Construction of DOTS requires ongoing computational analyses to identify putative transcribed sequences as determined from databases of experimentally identified mRNAs, ESTs and genomic sequences. In the long term, however, most of the effort in building and maintaining DOTS involves integration of data from across multiple online resources (and to some extent directly from the scientific literature). For example, DOTS incorporates keywords and functional taxonomies from GenBank, OMIM, SwissProt, and EGAD among others, enabling complex queries such as "Display all transcription factors with a greater than 4-fold difference in mRNA abundance level at day 11 of erythropoiesis in adult and cord blood." The integration process is facilitated through the K2 system, developed at Penn, for integration of information in distributed, heterogenous databases.

Presentation of data is through the bioWidgets visualization toolkit also developed Penn.

126. Web Based Quality Reporting of Completed DNA Sequencing

Robert D. Sutherland
Los Alamos National Laboratory, Los Alamos, New Mexico
rds@lanl.gov

I have created a Web based mechanism to report quality statistics on completed DNA sequencing projects. The motivation for this project was to streamline the processing of sequence and phrap quality data to the Web in automated manner for access by the public. This function crosses over facility boundaries and provides a single point of access for all of the JGI sequencing data. Currently, the JGI is required to (1) run quality codes to create the sequencing statistics, (2) transfer summary statistics to an Excel spreadsheet, (3) convert the spreadsheets to HTML and post to the Web. This is costly in time and effort, and the data from all sites cannot be viewed in a single report.

This new process automates the above steps in a single Web application that will meet the increased growth of sequencing within the JGI.

The Web implementation of this project is in three parts. The first part connects Perl scripts to the Web to run quality numbers on sequencing projects. This is all done locally at each facility. Once the quality numbers are acceptable, the Web part gathers more information about the project and posts all the data to one common database. The third Web part is the reporting mechanism which can give standard reports or create limited ad hoc queries to see only a portion of the sequencing projects.

This application utilizes several integrated technologies including the WWW, CGI, and Database engines using HTML, Perl, SQL, and C-shell.

The reporting part of the application can be accessed at http://jgi.doe.gov. This work is funded by the United States Department of Energy.

127. IMAGEne II: EST Clustering and Ranking of I.M.A.G.E. cDNA Clones Corresponding to Known and Unknown Genes

Peg Folta, Tom Kuczmarski, and Christa Prange
Lawrance Livermore National Laboratory, Livermore, California
pfolta@llnl.gov

With just under 2 million entries in dbEST, the ability to select the best cDNA clone(s) to conduct costly research is becoming increasingly difficult. The I.M.A.G.E. Consortium has developed the IMAGEne II product to increase the value of its cDNA collection by organizing its corresponding dbEST information at a gene level. IMAGEne II first clusters I.M.A.G.E. clones to both known and "unknown" genes. It then ranks the clones within a cluster as to their ability to represent the gene. A java-based display allows users to query against this database of information and view the alignment of the clusters at the nucleotide level. While the current product deals with the human species only, it will soon be extended to include other species and multi-species clusters.

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

128. Screening for Mutant Phenotypes in Mice at ORNL

D.K. Johnson, K.C. Goss, G.S. Sega, J.C. Schryver, M.J. Paulus, M.N. Ericson, and L.S. Webb
Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8077
lyy@ornl.gov

The Life Sciences, Instrumentation and Controls, and Chemical and Analytical Sciences Divisions at Oak Ridge National Laboratory have launched a broad-based, high-throughput primary screening program designed to recover mouse mutations exhibiting subtle phenotypes. We are validating screening tools for behavioral, biochemical, morphological, and physiological changes induced by experimental mutagenesis by screening about 100 test-class mice per week.

Our behavior-testing set currently includes the Porsolt forced swim test, rotorod, Poly-Track open-field activity system, and PhotobeaM activity monitor. We are introducing modifications into our cued and contextual fear conditioning test for learning and memory deficits and in our startle response tests, which have not proved adequate for reliable mutant identification so far. Biochemical tests include gas chromatography/mass spectrometry analysis of fatty acids, organic acids, and neurotransmitters in blood and tissue, as well as standard package analysis on an Abbott Cell-dyne 3500 Hematology Analyzer. For urine, we perform standard dipstick and specific gravity tests.

Tool development includes a microCT scanner with image analysis software for mice, and a subdermal microbiosensor for the measurement of activity patterns, heart rate, body temperature, and, eventually, blood pressure. We are organizing joint screening programs with clinical and academic institutions across the state of Tennessee in order to broaden our screening and greatly enhance our expertise. Our goal is to maximize the number of whole-organism mutant phenotypes that we can detect in a high-throughput, broad-based, and cost-effective primary screening effort at ORNL.

[Research sponsored jointly by the Office of Health and Environmental Research, USDOE, under contract DE-AC05-960R22464 with Lockheed Martin Energy Systems, Inc., and by the National Center for Human Genome Research (HG 00370).]

129. Using Overlapping Deletions in the Analysis of Recessive Phenotypes

Yun You, Hanna Chao, Sarah Mentzer, Rebecca Bergstrom[1], and John Schimenti[1]
Life Sciences Division, Oak Ridge National Laboratory, PO Box 2009, Oak Ridge, Tennessee 37831-8077
[1]The Jackson Laboratory, Bar Harbor, Maine 04609
lyy@ornl.gov

Chromosomal deletions have been exploited to perform a systematic characterization of functional units in *Drosophila melanogaster*. The Human Genome Project will generate nucleotide sequences of 10^9 base pairs, an estimated 80,000 to 100,000 genes in human, and only a small percentage of them has a known role. As a model system, the mouse is an indispensable tool to decipher mammalian gene function. A high throughput method has recently been developed to induce chromosomal deletions at any region of the mouse genome by radiation in embryonic stem (ES) cells. Lines of mutant mice carrying deletions around the *D17Aus9* locus have been generated by this strategy. Deletion analysis of mutant mice called $D17Aus9^{dfl0J}$ carrying a small deletion showed that an early lethal gene is located near the *D17Aus9* locus. Early lethality renders further deletion analysis of this region difficult. In our deletion analysis this problem was easily avoided by using $Del(17)T^{7J}$, another mutant line carrying a deletion, which does not encompass the *D17Aus9* locus, but overlap with the deleted region found in $D17Aus9^{dfl0J}$. By crossing $D17Aus9^{dfl0J}$ /+ to $Del(17)T^{7J}$ /+, the heterozygous compound deletions

unveiled a late action recessive lethal locus. The deletion analysis data and initial characterization of this lethal mutant will be presented.

Above results illustrate the importance to generate sets of overlapping deletion complexes in the mouse chromosome 15 mutagenesis project at Oak Ridge National Laboratory (see abstract of E. Rinchik, et. al). The deletions will be used as mapping tools to locate the ENU-induced point mutations, and will also serve as reagents to identify functional units and clone genes important for mouse development along chromosome 15.

[Research currently sponsored by USDOE, under contract DE-AC05-960R22464 with Lockheed Martin Energy Research, Inc.]

130. Germline Deletion Complexes in Embryonic Stem Cells for Mapping Gene Function in Mouse-Human Homology Regions

Edward J. Michaud, Irina Khrebtukova, Carmen M. Foster, and Tuan Vo-Dinh
Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8077
michaudej@bio.ornl.gov

Rapid progress has been made by the human genome sciences community in the last several years in generating nearly complete physical maps of several human chromosomes. In the very near future, the map positions and DNA sequence of the estimated 100,000 genes that make up a healthy individual will also be known. Sequence information alone, however, is often insufficient to ascertain the biological roles that genes play in normal human development and health. In order to determine the organismal function of every human gene and to understand how specific DNA mutations in genes result in birth defects and disease, strategies will need to be employed that are cost effective, scaleable to the entire genome, and that complement the mapping and sequencing data.

One powerful approach for mapping the biological functions of many human genes that reside along large segments of chromosomes is to generate nested sets of chromosomal deletions in the homologous regions in mice. Deletion complexes at defined loci on mouse chromosomes permit fine-structure gene-function maps to be constructed, based on heritable mutations with specific phenotypes, that are then correlated with the available physical maps. Unfortunately, deletion complexes are currently available for only about 14% of the mouse genome. However, a new method was recently described (You et al., Nature Genet. 15:285-288, 1997; Thomas et al., Proc. Natl. Acad. Sci. USA 95:1114-1119, 1998) that permits deletion complexes to be generated anywhere in the mouse genome in F1 hybrid embryonic stem (ES) cells. The method is rapid and cost effective because the deletions are generated at defined locations in the genome and selected for in the ES cells. Additionally, many different deletions can be generated in one experiment and the extent of the deletion breakpoints can be mapped with available polymorphic markers before producing lines of mice.

The objective of this project is to develop ES-cell reagents to facilitate the generation of functional maps in gene-rich regions that are homologous to portions of human chromosomes being mapped and sequenced by the Joint Genome Institute. The initial focus of this project will be to generate nested sets of chromosomal deletions in ES cells for the proximal 23 cM of mouse Chr 7 (human 19q homology) and a 16 cM region of proximal mouse Chr 11 (human 5q homology). ES-cell clones containing chromosomal deletions are the first reagents that will be generated and made available to the scientific community. As the project progresses, the goal will be to generate lines of mice harboring these chromosomal deletions and to archive these mutations in the form of cryopreserved embryos and spermatozoa. The reagents generated during this project will be advertised on the Joint Genome Institute Functional Genomics web site.

This work is supported by the U.S. Department of Energy FWP ERKP293, in collaboration with the Joint Genome Institute.

131. Mouse Genetics and Mutagenesis for Functional Genomics: The Chromosome 7 and 15 Mutagenesis Programs at the Oak Ridge National Laboratory

E. M. Rinchik, D. A. Carpenter, E. J. Michaud, Y. You, P. R. Hunsicker, and D. K. Johnson
Life Sciences Division, Oak Ridge National Laboratory, PO Box 2009, Oak Ridge, Tennessee 37831-8077
rinchikem@ornl.gov

The development of detailed mutation maps of regions of the mouse genome provides new resources for the study of mammalian biology and serves as an important functional complement to the human genome program. Mouse-human linkage homologies permit a type of "surrogate genetics" to be developed for regions of the human genome that is based on analyzing the molecular and organismal consequences of mutations mapping within the corresponding mouse genomic segment. One of the major goals of the mouse genetics program at ORNL is to apply our experience in chemical germ-cell mutagenesis and mutation recovery and propagation, as well as recently developed and evolving broad-based phenotype screening, for creating a large, user-friendly mouse-mutation resource for use by the functional-genomics and wider biological communities.

For a number of years, we have been molecularly characterizing regions of mouse Chromosome (Chr) 7 while recovering, in parallel, N-ethyl-N-nitrosourea (ENU)-induced, recessive single-gene mutations mapping within those regions by two-generation hemizygosity screens with radiation-induced deletions. Mutagenesis of one 6- to 11-cM region surrounding the albino (c; Tyr) locus has been completed, yielding 31 mutations representing ten complementation groups. An on-going screen of another ~4- to 5-cM Chr-7 region, proximal to the pink-eyed dilution (p) locus (human 11p and 15q homologies), has so far yielded 19 new mutations, representing 8 complementation groups, from a screen of just 1218 gametes. Both of these screens have greatly increased the fine-structure genetic and functional maps of the corresponding regions. In addition to these hemizygosity screens, we shall also describe new work about to get underway that involves three-generation, homozygosity strategies to induce mutations in proximal Chr 7 (human 19q homology), mid Chr 7 (human 15q homology), and mid-to-distal Chr 15 (human 8q, 22q, and 12q homologies), which are large, multi-megabase regions that are currently not covered by complexes of deletions. We shall discuss our emphasis on up-front investment in developing genetic reagents so that any mutation created can be maintained and used by a wide variety of investigators with no molecular genotyping. In addition, we shall discuss the potential value of "parallel processing" in regional mutagenesis of the mouse genome, in which chromosomally "pre-mapped" mutations are recovered by three-generation screens with inversions in parallel to (not following) the development of deletions in embryonic stem cells for use as finer-mapping and gene-identification reagents. Mutant stocks generated in any of our screens will be advertised on the Web (http://lsd.ornl.gov/htmouse/mmdmain.htm) and made available to the scientific community.

[Research currently sponsored by the Office of Biological and Environmental Research, US DOE, under contract DE-AC05-960R22464 with Lockheed Martin Energy Research, Inc., and in the past by US DOE and the National Human Genome Research Institute (HG 00370).]

132. Comparative Analysis of Structure and Function in an Imprinted Region of Proximal Mouse Chromosome 7 and the Related Region of Human Chromosome 19q13.4

Joomyeong Kim, Anne Bergmann, Xiaochen Lu, Anne Olsen, Jane Lamerdin, and Lisa Stubbs
Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551 and DOE Joint Genome Institute
kim16@llnl.gov

Our group is interested in coupling mouse genetics and biology to comparative sequence that is being generated by JGI teams, with the aim of generating in-depth functional annotation the sequenced human regions. One special target for these studies has been a 2 Mb-region of human chromosome 19q13.4 (H19q13.4) and the syntenically homologous region of mouse chromosome 7 (Mmu7). The sequence of the human region is nearly completed by the JGI genome sequencing team, and sequence analysis of the homologous murine has recently been initiated. Since the murine region is known to be parentally imprinted and imprinting is a generally conserved in mammalian species, the structure and function of this region are of special biological interest.

Our group's efforts begin as DNA sequence and basic annotation of specific DNA segments are completed. Our goal is to characterize genes and conserved regulatory sequences predicted to exist in the sequenced regions, especially those that may contribute to parent-of-origin specific functions in humans and mice . The human region is especially rich in clustered zinc finger containing genes (ZNFs); about 90% of the genes found in 19q13.4 appear to be actively expressed Kruppel-type ZNF loci. We have also identified a number of other types of genes in the 2 Mb human interval, including genes encoding an *Aurora*-related serine/threonine kinase (*STK13*), a sulfotransferase2 (*ST2*), and one anonymous gene homologous to a yeast hypothetical protein P38334 (HYP). The mouse region is similar in terms of gene content, but physical mapping studies have also revealed several chromosomal changes that are unique to the mouse. For example, there are at least 5 copies of the *STK13*-related genes in mouse, and these copies appear to have been duplicated in tandem as part of a large unit that also contains ZNF-related gene sequences. This tandem duplication appears to have occurred in very recent evolutionary history. One gene (*Cln4-2*), whose homolog is located in the pseudoautsomal region of the human X chromosome and which had previously been mapped to Mmu7, also appears to have been very recently transposed into the mouse zinc-finger gene cluster region. Other orthologs of human 19q13.4 genes, including *ST2,* and *HYP* are present in the related mouse region, although their relative positions within the mouse and human regions have not been strictly conserved. The intronless nature of many of these genes suggests that they were duplicated by retroposition and inserted into this region after the zinc-finger gene clusters were elaborated.

Detailed functional studies have been focused primarily on the imprinted genes that are located in this region. Only one imprinted gene, *paternally expressed gene 3* (*Peg3*) had been identified at the outset of this study, but we have recently identified several new genes located near *Peg3.* At least one of the new genes, called *Zim1,* is also imprinted in mouse. *Peg3* and *Zim1* are located next to each other in both species, and the two genes are reciprocally-imprinted: *Peg3* is paternally expressed whereas *Zim1* is expressed only from the maternal allele and specifically in embryonic tissues. As has been found in other well-known imprinted domains, such as Prader-Willi/ Angelman syndrome region of H15q11-q13/Mmu7 and Beckwith-Wiedemann region of H11p15.5/Mmu7, the H19q13.4/proximal Mmu7 imprinted domain is expected to contain a number of additional imprinted genes. A comprehensive update on our recent studies, including gene identification, gene expression, and functional analysis of this gene-rich, imprinted region will be presented.

133. Differential Expansion of Homologous Zinc-Finger Gene Families in Human Chromosome 19q13.2 and Mouse Chromosome 7

Mark Shannon[1], Elbert Branscomb[1], Loren Hauser[2], Anne Olsen[1], Laurie Gordon[1], Linda K. Ashworth[1], and Lisa Stubbs[1]
[1]Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, P.O. Box 808, L-452, Livermore, CA 94550 and [2]Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831
shannon8@llnl.gov

Mapping studies indicate that many of the 600-1000 mammalian zinc-finger (ZNF)-containing genes reside within familial clusters, particularly those genes encoding Kruppel-associated box (KRAB) motifs. However, little is known about family content, organization, or evolutionary conservation. In previous studies, we identified and characterized homologous KRAB-containing ZNF gene families located in human chromosome 19q13.2 and mouse chromosome 7. Here we present details of the construction and characterization of contigs that completely span these families. The human cluster spans 700kb and is comprised of 16 members that are arrayed in tandem. By contrast, the mouse family spans approximately 400kb and contains just 10 genes. We have also identified cDNA clones corresponding to each family member and have analyzed their sequences. The KRAB A domains encoded by the human and mouse genes are highly similar in sequence, but other portions of the predicted proteins encoded by the clustered paralogs may be more divergent in structure. To predict the evolutionary relationships between genes within and between the families, ZNF-containing regions were compared using computational methods. These studies uncovered three pairs of putative orthologs, but also provided evidence for the continued evolution of the families in both species after their divergence from a common ancestor. Recent evolutionary events include intragenic ZNF repeat alterations as well as complete gene duplications. These studies therefore expose complex, yet discernible, histories of sequence duplication and divergence and pave the way for studies of the evolution of gene function within the related families.

134. YAC-ES (Y-ES) Cell Libraries for In Vivo Analysis of JGI Sequences

Yiwen Zhu, Veena Afzal, Jan-Fang Cheng, and Edward Rubin
Department of Genome Sciences, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
yzhu@lbl.gov

Libraries of the human genome, propagated in bacteria and somatic cells, have been an invaluable tool in the identification of genes based on in vitro assays. We have expanded upon this concept by creating a >10 Mb "in vivo" library of regions of the human genome sequenced by the JGI in the form of human YACs propagated in totipotent mouse embryonic stem (ES) cells.

Megabase human YACs from JGI sequenced regions were first characterized for integrity by Southern hybridization and STS content mapping. Appropriate YACs were then retrofitted with selectable markers and introduced individually into germ line transmitting ES cells by yeast spheroplast - ES cell fusion. The content and integrity of the human YACs in the ES cells were assayed and clones with intact human transgenes without detectable rearrangements have been cryopreserved to serve as publicly available reagents to explore the function of genes contained within the human sequences (for more information, visit our Web site at http://grail.lsd.ornl.gov/projects/jgi/fung.shtml). In our initial experiments, we have fused six 5q31 YACs and two 19q13.4 YACs into ES cells. The 19q13.4 Y-ES clones are being injected into mouse blastocysts to test their capacity to contribute to the germline. We have obtained good chimeras from one of the 19q13.4

Y-ES clones tested. We are now in the process of characterizing another 50 megaYACs in 5q region.

Possible uses of the Y-ES clones to biological researchers include: 1) ready made reagents for the investigation of expression/function of a human gene of interest; either in tissue culture or in transgenic mice derived from Y-ES clones; 2) as a reagent for fine mapping of mouse mutations based on functional in vivo complementation of the mutant mouse phenotype by YAC transgenes; and 3) for sifting through large candidate regions of the genome identified by human complex trait mapping studies using gene expression patterns or functional assays in mice propagating these regions.

135. Comparative Functional Genomics

George M. Church, Pam Ralston, Martha Bulyk, Abby McGuire, Rob Mitra, Saeed Tavazoie, and Jason Hughes
Department of Genetics, Harvard Medical School, Boston, Massachusetts
church@arep.med.harvard.edu

We have developed technologies for annotating genome sequences, including intergenic regions and regulon/operon comparisons. Performing enzymatic reactions on oligonucleotide chips or microarrays (of kbp-sized DNA) allowing us to replicate DNA chips using microcontact printing. RNA quantitations from chip, microarray and SAGE can be merged, clustered, and the motifs mechanistically responsible for the clusters of coregulated RNAs can be determined. Methods for measuring and modeling in vivo concentrations of protein, RNA, metabolite, protein interactions and mutant growth rates in response to diverse environments provide the foundations for a genome sequence function database.

Nature Biotech. 16:566-571; Nature Biotech. 16: 939-945; J. Molec. Biol. 284: 241-254. (http://arep.med.harvard.edu)

136. A Targeted 450 Kb Deletion in Mouse Chromosome 11 Identifies a Novel Gene Dramatically Impacting on VLDL Triglyceride Production

Yiwen Zhu, Miek Jong, Elaine Gong, Kelly Frazer, Jan-Fang Cheng, and Eddy Rubin
Department of Genome Sciences, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
Yzhu@lbl.gov

In order to multiplex the examination of the function of genes present within JGI sequenced regions of the human genome the Cre Lox system was employed to delete several genes at a time within the human 5q31 / mouse chromosome 11 syntenic region. A 450 Kb stretch between IRF1 and CSF-GM gene containing nine genes, all of no known function, were deleted in ES cells. Mice homozygous for the deletion though prenatal viable demonstrated premature morbidity with approximately 75% dying before 100 days of age. The other significant finding in these animals was massive enlargement of the liver owing to the engorgement of hepatocytes with triglycerides. Plasma triglyceride levels were approximately ten-fold greater than control animals.

Due to the importance of triglyceride metabolism as an atherosclerotic risk factor we investigated the mechanism underlying the hypertriglyceridemia in these mice. Of the three major factors effecting plasma triglyceride levels (synthesis, lipolysis in the periphery and clearance via hepatic uptake) abnormalities were only present with regard to synthesis. Homozygous animals exhibited a four-fold increase in hepatic VLDL triglyceride production while animals heterozygous for the deletion had a two-fold increase in triglyceride production compared to control mice. These deletion mice represent a unique model of enhanced hepatic triglyceride production coupled with increased hepatic fat accumulation.

To identify the gene responsible for this phenotype through in vivo complementation the 450 Kb deletion mice were crossed with mice containing human YAC and mouse BAC transgenes covering the entire deleted region. Animals homozygous for the deletion

and hemizygous for the transgenes were analyzed with regard to the phenotypes associated with the deletion. A human YAC containing approximately 120 Kb of the deleted region successfully corrected the hepatic fat accumulation, hypertriglyceridemia and premature lethality associated with the homozygous deletion. Three potential candidate genes are present in YAC: two identified as EST hits with no homology with known genes and one with homology to a rat liver specific transporter-like protein.

137. Identification and Functional Analysis of Evolutionarily Conserved Non-Coding Sequences in the Human 5q31 Cytokine Cluster Region

Gabriela Cretu, Webb Miller, Catherine M. Brion, Jan-Fang Cheng, Christopher H. Martin, William Kimberly, Edward M. Rubin, and **Kelly A. Frazer**
Genome Sciences Department, Lawrence Berkeley National Laboratory, 1 Cyclotron Road MS 84-171, Berkeley, CA 94720
kelly@mhgc.lbl.gov

The human 5q31 region chosen by the JGI for large scale sequencing is biologically interesting because it harbors a family of cytokine genes which are important regulators of the immune response. We previously annotated the 1 Mb Cytokine Gene Cluster Region on human 5q31 computationally and biologically resulting in the identification of 23 genes and the determination of their expression patterns. To further annotate this 1 Mb region we are currently identifying evolutionarily conserved non-coding sequences and attempting to determine their biological function. Since the interleukin loci in this region are likely to have arisen by ancient duplications of an ancestral gene, several intervals of this region may be paralogous with each other. To identify evolutionarily conserved non-coding elements we compared the potential human paralogous sequences in this 1 Mb region to each other as well as with their orthologous mouse chromosome 11 sequences. Several highly conserved non-coding sequences that potentially have biological function were identified. We have started functionally analyzing two of these non-coding elements: an 86 bp element located 5' of the human interleukin 13 (IL 13) gene which is 91% identical with another non-coding element in the human 5q31 region, and a 400 bp element located between the IL 4 and IL 13 genes which is 85% identical in humans and mice.

To investigate the biological function of the 86 bp element upstream of IL 13 we have deleted it in a human 5q31 YAC and have used this manipulated YAC to generate transgenic mice. Production of human IL 4 and IL 13 is markedly reduced in mice harboring the mutated YAC lacking the 86 bp conserved element compared with the production of these human proteins in mice harboring a wild type YAC. These data suggest that the 86 bp conserved element is involved in regulating the expression of the human IL 4 and IL 13 genes.

The biological role of the non-coding 400 bp element located between IL 4 and IL 13 is being investigated by several strategies. We determined by PCR amplification and sequence analysis that this 400 bp element is also highly conserved in the cow, dog, rabbit, pig, and rat which provides additional evidence indicating that this human-mouse conserved non-coding elememt is likely to be functionally important. To determine the biological role of this conserved element we are in the process of homozygously deleting it in mice and will determine how its absence affects the expression of the murine IL 4 and IL 13 genes. We are also performing comparative studies of human IL 4 and IL 13 expression in mice harboring a 5q31 YAC lacking the 400 bp conserved element with mice harboring a wild type 5q31 YAC. To be able to assay for small changes in the expression of the human IL 4 and IL 13 genes it is important to eliminate variation in their expression due to integration site of the mutated and wild type YAC in the mouse genome. To accomplish this we have surrounded the 400 bp conserved element on the human YAC with lox P sites and are

using this YAC to generate transgenic mice. The founder mice will be mated with wild type mice and with mice expressing CRE recombinase. In this manner we plan to generate lines of mice that harbor at the same site of integration the human YAC lacking and containing the 400 bp element. By assaying for changes in the expression of the human IL 4 and IL 13 genes in these mice we hope to gain insight into the function of this highly conserved 400 bp non-coding element.

138. Discovering the Genes Affected by Schizophrenia Using DNA Micro-Array

Yang Qiu, Edward M. Rubin, and Jan-Fang Cheng
Genome Science Department, Lawrence Berkeley National Lab, 1 Cyclotron Road MS 84-171, Berkeley, CA 94720
yqiu@lbl.gov

Schizophrenia is a devastating psychiatric disorder that affects 1% of the population. Genetic factors make important contributions to the etiologies of this disease. It is highly likely that multiple genes and environmental factors are involved. Chromosome 6p has been shown to have linkage with schizophrenia in several independent studies. The current drugs treating schizophrenia including clozapine, risperidone and olanzapine are all far from perfect with substantial side effects. It is thus important to be able to identify the genes affected by schizophrenia, which would greatly enhance the drug discovery leading to a better treatment.

We are taking advantage of the technology of DNA micro-array at Lawrence Berkeley National Lab which can hold thousands of genes on one single glass slide and the development of the human and mouse Unigen set (uniquely expressed sequences) through the effort of genome community. The expression of thousands of genes at different physiological condition can be analyzed in parallel. New genes can be identified and biological functions of the genes can be further studied. The DNAs to be spotted on the DNA micro-array are as following:

(1) 10,000 human Unigen clones representing ~20% of expressed human genes.
(2) 309 BAC clones available from the physical mapping project covering 90% of the schizophrenia candidate region at chromosome 6p.
(3) 269 genes singly selected through thorough literature search which include neurotransmitter receptor (dopamine receptor, glutamate receptor, serotonin receptor, acetylcholine receptor, etc), brain function related genes and other possible genes involved in schizophrenia.
(4) ~ 40 clones identified to be differentially expressed in neuropsychiatric disorders by Stanley Neurovirology Laboratory at the Johns Hopkins University, School of Medicine.

The postmortem brain tissue from individuals with schizophrenia and normal controls will be obtained from Standley Foundation Neuropathology Consortium. Total RNAs are to be extracted from the different brain tissues and hybridized with the DNA micro-array. The genes that are affected in schizophrenia can be identified when using a large sample sets to minimize the individual variations in the gene expression.

Meanwhile, a mouse model for schizophrenia is underway for this study. It has been established that the mice treated with psychotic drug PCP (angel dust) mimic some symptoms of schizophrenia in which the prepulse inhibition is diminished in schizophrenia patients. We are treating mice with PCP as well as antipsychotic drugs including clozapine and risperidone. The gene expression patterns in the mouse brain will be followed at different times after each treatment using the DNA micro-array. The genes that are affected by drugs can be identified as the candidate genes for schizophrenia.

139. Gene Expression in Cardiac Hypertrophy as Measured by cDNA Microarrays

Carl Friddle, Teiichiro Koga, James Bristow, and Edward M. Rubin
Lawrence Berkeley National Laboratory, Berkeley, California
cjfriddle@lbl.gov

The mouse heart is an ideal model because it offers the benefits of a whole animal system in an organ comprised of relatively few cell types. We are studying the changes in heart gene expression that correspond to the onset and progression of cardiac hypertrophy. We wish to know the normal distribution of gene transcripts in the heart chambers in contrast to the distribution found in the hypertrophic state. Much has been learned about the expression profile of specific genes that play a role in cardiac hypertrophy (e.g.: ANF, MLC-2, c-Fos, c-Jun, Egr1). However, the discovery of new pathways involved in hypertrophy, and even the rapid identification of genes in known pathways, would benefit from an approach that analyzes thousands of genes in parallel. One could then perform a more detailed analysis of those genes that show a change in expression levels in conjunction with the onset of cardiac hypertrophy.

We have chosen to apply cDNA microarray technology to these questions. A array of over 3000 mouse EST clones was generated from the sequenced libraries of the IMAGE consortium. Included are ESTs from both heart, embryonic, liver and brain libraries. This array allows us to generate an expression profile for both the normal mouse heart and for hypertrophic tissues.

Hypertrophy was induced in vivo by treating mice with Isoproterenol. Heart weight was increased by 50% over the course of a week. Mice were sacrificed daily to generate a time course of hypertrophy induction. We then monitored the expression of the genes represented by our 3000 clones and used this information to identify classes of genes that are regulated in coordination with the onset and progression of cardiac hypertrophy.

140. Genetic Factors Affecting Globin Switching

Sluan D. Lin, Phil Cooper, Mary E. Stevens, and Edward M. Rubin
Genome Science Department, Lawrence Berkeley National Laboratory, One Cyclotron Rd., MS 84-171, Berkeley, CA 94720
slin@mhgc.lbl.gov

Low level expression of fetal gamma globin inhibits red cell sicklying and its pathological consequence in individuals homozygous for the Beta-S alleles. However, fetal gamma globin switches adult beta globin shortly after birth. To furthering our understanding of the genetic factors affecting the sickle cell disease, we are studying the "transacting" modifier gene(s) that impact on the switching of human gamma to beta globin in transgenic mice.

Creation of transgenic mice that persistently express human gamma globin: We have made transgenic mice using a YAC containing the entire human beta-cluster with a -117 Agm mutation. The mutation causes the Greek form of hereditary persistence of fetal hemoglobin (HPFH) in human. We found that the HPFH mutation can also causes the human gamma chains to be expressed postnatally in the transgenic mice. This feature has greatly facilitated the study of the globin switching parameters and the level of gamma globin expression after birth.

Different genetic backgrounds affect the globin switching parameter among the F1 animals: The heterozygous FVB transgenic animals were crossed with different inbred strains including DBA/2N, Balb/C, 129/SvJ and SWR/J. The blood of YAC-positive animals screened by PCR were collected on 10, 15, 30 and 60 days after birth. The human gamma/beta globin ratio of the

DBA/2N-derived F1 hybrid shows consistent highest level throughout the sampling period. Comparing DBA/2N-derived F1 with the transgenic FVB, the p-values are 1x10-7 and 3x10-10 on day 30 and day 60, respectively. Thus, we've verified the hypothesis that different genetic background of F1 hybrid mice, derived from crossing the transgenic with other inbred strains, can affect the level of gamma globin expression.

Backcross suggests more than one genetic loci contribute in regulating gamma globin expression: We first generated 70 backcross transgenic animals as a pilot study of the possible number of genetic factors involved in up-regulating the gamma expression. The lack of a bimodal distribution of the backcross animals suggesting that there are more than one genetic loci that contribute in regulating the expression of human gamma globin in the transgenic mice. By applying the classical formula of Wright (1968), we estimate the number of QTLs controlling the gamma expression is 2.4 and the genetic contribution to the phenotypic variance is 48%. By comparing to others publication of similar situation, We estimate that we would need approximately 200 backcross transgenic animals in total to map the QTLs. To test the polymorphism between the FVB and DBA/2N genomic DNA, we are screening for 83 SSLP markers across the mouse genome at an average genetic interval of 17 CM. Using these markers, we are performing a genome scan starting with the 20% backcross animals at the two phenotypic extremes. Once a significant lod score of 3.3 is achieved, we will pick more markers around the particular locus to fine map the modifier gene.

141. Resources for Functional Genomics in *Drosophila*

Gerald Rubin, Suzanna Lewis, Ling Hong, Damon Harvey, E. Jay Rehm, Amy Beaton, Peter Brokstein, Guochun Liao, Erwin Frise, and Allan Spradling
University of California, Berkeley, California
gerry@fruitfly.berkeley.edu

A major goal of the *Drosophila* Genome Project is to biologically annotate the DNA sequence of the *Drosophila melanogaster* genome as it emerges and to provide community resources for functional genomics. To this end, we are carrying out several related projects: (1) insertional mutagenesis using P transposable elements that have been engineered to allow controlled misexpression, in addition to insertional inactivation, of the gene at the site of insertion; (2) generation of a "unigene" set of arrayed, full-length *Drosophila* cDNAs; (3) highly accurate DNA sequencing of a selected subset of these cDNAs; (4) single-pass sequencing of the remainder of the cDNAs as the corresponding genomic sequence becomes available to generate a transcript map of the genome; and (5) determination of the expression pattern of individual genes by tissue in situ hybridization to embryos at a variety of developmental stages and by hybridization to gene microarrays.

142. Isolation of *Drosophila* DNA Repair Genes

R. Scott Hawley, Kenneth C. Burtis, and Gerald M. Rubin[1]
University of California, Davis, California and
[1]University of California, Berkeley, California
kcburtis@ucdavis.edu

The ultimate goal of this project is to complete the identification and mapping of all of the genes involved in repair of genome damage in *Drosophila melanogaster* and to initiate their functional characterization. The substantial quantity of *Drosophila* genomic and cDNA sequences obtained to date, in combination with the genetic information available for this organism, provide a powerful base from which to begin a comprehensive description of the DNA repair genes operating in *Drosophila*. We have now initiated a multi-faceted approach to complete this process, using genetic, molecular and bioinformatic approaches.

We have initiated genetic screens to extend previous, non-saturating screens for mutagen-sensitive mutations. These genetic screens have only just begun, and no new mutagen-sensitive loci have yet been isolated. However, using a combined molecular and genetic approach, we have made some progress

in identifying two repair genes previously uncharacterized in *Drosophila*; the *Drosophila* homolog of the yeast RAD10 gene (now designated mei-10), and the *Drosophila* homolog of XPG. We have shown that the fly MEI-10 protein physically interacts with the MEI-9 (dm Rad1) protein by yeast two-hybrid studies. We are now in the process of creating mutants in the mei-10 gene. Preliminary mapping suggests that the MEI-10 gene may correspond to the previously identified mus210 locus. We have also obtained strong evidence that the *Drosophila* XPG homolog corresponds to the mus201 locus. The two extant mus201 alleles display phenotypes expected for an NER defect, but no discernable meiotic defect.

We are also continuing genetic studies of several known repair-deficient loci. Most notably we have identified a null allele of the repair/checkpoint gene mei-41 and demonstrated intermediate levels of repair competency and checkpoint function in heterozygotes for this mutation. We have used this and other such mutants as substrates in screens for dominant enhancer or suppressor mutations. These screens are allowing us to identify mutations that might be lethal or sterile when homozygous and thus be missed in more convential screens for mutagen-sensitive mutations.

A second approach to identifying genes involved in the *Drosophila* response to genome damage will involve the use of DNA microarrays. We have recently completed construction of an arraying robot using the design developed by Pat Brown's lab at Stanford, and have initiated the production of arrayed collections of EST-characterized *Drosophila* cDNAs produced by the Berkeley *Drosophila* Genome Project. These arrays will be used to identify transcription units whose expression is regulated in response to DNA damage. The genes thus identified will be further characterized molecularly and genetically, and correlated to the extent possible with repair genes identified by other means.

Finally, we will present an up-to-date summary of the *Drosophila* DNA repair genes identified through genomic and EST sequences obtained to date by the Berkeley *Drosophila* Genome Project.

143. Ribozyme Gene Delivery for Gene Target Discovery and Functional Validation

Xinqiang Li, Peter J. Welch, Mark C. Leavitt, Flossie Wong-Staal, and **Jack R. Barber**
Immusol, Inc., 3050 Science Park Road, Second Floor, San Diego, CA 92121
barber@immusol.com

Ribozymes (Rzs) are RNA molecules that can be engineered to cleave and inactivate other RNA molecules in a sequence-specific fashion. Thus, Rzs can be designed to selectively inactivate the expression of any target gene ("gene knockdown") and its corresponding protein. We have used Rz genes, delivered with viral vectors, as a tool for gene functional validation and discovery. We have used hairpin ribozyme gene delivery to rapidly and effectively inhibit expression of a number of viral and cellular genes.

To expedite the process of associating genes with cellular function, we have developed Rz gene vector libraries. Retroviral and Adenoassociated viral vectors have been generated that efficiently deliver and express Rz genes whose target recognition sequences have been randomized, generating a library of Rz genes capable of recognizing a total of more than 1x107 possible gene target sequences. The library of Rz genes is delivered into large numbers of tissue culture cells (one Rz gene per cell for each Rz gene in the library), followed by selection for individual cells that have lost a particular function. The sequence of the Rz target recognition domain thus identified allows the identification and cloning of genes that are necessary for a given cellular function. As an example of the power of the technology, we will present data demonstrating its use to identify a

novel tumor suppressor gene, without prior sequence information.

144. Microfabricated Microfluidic Devices for Proteome Mapping

R.S. Ramsey, R.S. Foote, R.D. Rocklin, M.I. Lazar, Y. Liu, and J. M. Ramsey
Chemical and Analytical Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6142
RamseyJM@ornl.gov

Miniaturized chemical instruments, "Lab-on-a-Chip" technologies, are being developed for rapid, comprehensive analysis of cellular proteins, as an alternative to the slow and labor-intensive 2D gel methods currently used for protein mapping. The microfabricated devices will integrate on a single structure, elements that enable multidimensional separations of protein mixtures and electrospray ionization of the analytes for direct, on-line interfacing with mass spectrometry. The platform exploits the many advantages of Lab-on-a-Chip devices, including small size, inexpensive fabrication, high speed, low volume materials consumption, high throughput, and automated operation. Potential applications of the technology include quantification of gene product levels in specific cell types, comparative analysis of patterns of gene expression in different tissues at different stages of development, analysis of structural and/or expression level changes resulting from mutagenesis or genetic disease, and identification of specific protein markers of disease.

The conventional 2D PAGE method for resolving cellular proteins is not only laborious but also has poor reproducibility, sensitivity, and sample recovery. Individual spots may be identified off-line using mass spectrometry (MS) but sample extraction and transfer processes are inefficient. Column liquid chromatography or capillary electrophoresis (CE), which are more easily coupled with MS using electrospray ionization (ES), in general, lack the resolution required for the analysis of complex biological samples. Two-dimensional separations greatly increase the resolving power, provided the individual methods are orthogonal, and when combined with MS result in a powerful technique, given the multiplicative effect of joining different separation mechanisms. We have designed and demonstrated an integrated device combining micellar electrokinetic chromatography and high-speed free zone electrophoresis. The orthogonality of these techniques, an important factor for maximizing peak capacity or resolution elements, was verified by examining each technique independently for peptide separations. The two dimensional separation strategy was found to greatly increase the resolving power over that obtained for either dimension alone. The integrated device operates by rapidly sampling and analyzing effluent in the second dimension from the first dimension. Second dimension analyses are performed and completed every few seconds. Total analysis times are less than 10 min and the peak capacity has been estimated to be in the 500 to 1000 range. The operation of the device is completely automated. Microchips have also been interfaced to a time-of-flight mass spectrometer that has acquisition rates necessary to capture mass spectra from rapidly eluting components. The electrospray element will eventually be integrated with the two dimensional separations to allow on-chip MS analysis for ultra-high throughput protein mapping. Increases in sample throughput are anticipated to be greater than two orders of magnitude as compared to 2D PAGE.

145. Using Phage Display in Functional Genomics

Peter Pavlik, Rob Segal, Daniele Sblattero, Vittorio Verzillo, Roberto Marzari, and **Andrew Bradbury**
Los Alamos National Laboratory, Los Alamos, New Mexico and Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy
bradbury@icgeb.trieste.it

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. A new method to make phage antibody libraries has been developed and a new large library using this method is presently under construction. A library provided by Jim Marks, UCSF, has been used to select antibodies against a number of cell cycle and DNA repair proteins. We have succeeded in miniaturising selection on proteins to a 96 pin format. Should gene products be available this is a very efficient way to select antibodies in a high throughput format. We have also used scanned peptides (180 in total) derived from five different proteins (ubiquitin, cdk2, human serum albumin, cyclin D, transglutaminase) to select antibodies. Some of the antibodies selected are able to recognise the native protein. We are attempting to derive rules, based on physicochemical characteristics and other predictive algorithms, which predict which peptide sequences will select antibodies recognising the full length protein.

146. One Gene - How Many Proteins?

Raymond F. Gesteland, Chad Nelson, Mike Giddings, Norma Wills, Jiadong Zhou, Barry Moore, Mike Howard, and John Atkins
University of Utah, Department of Human Genetics, Salt Lake City, UT 84112-5330
ray.gesteland@genetics.utah.edu

This is a new pilot project to use mass spectrometry methods to determine the multiplicity and character of proteins coming from individual mRNAs. Many processes contribute to the complexity of gene products that come from one gene. In addition to alternate splicing and RNA editing that increase mRNA complexity, protein modification and alternative translation can all expand the population of proteins that come from one mRNA species. Although we know a good deal about protein modifications on a protein by protein basis, we know little in a genome-wide sense. We know even less about the frequency of occurrence of unusual translation events. Alternative translations, or recoding, include programmed frameshifts, bypassing of mRNA regions, and redefinition of stop codons to encode one of the twenty amino acids or selenocysteine the 21st amino acid.

We are developing technology to ask how many different protein products come from each mRNA species. We are using Electrospray Liquid Chromatography Mass Spectrometry. With a genome of known sequence, such as yeast, we can fractionate proteins and by accurately determining their masses, see if they can be accounted for by predicted molecular weights from the known open reading frames. Identification with genes can be verified by mass analysis of tryptic digests. If initial molecular weights do not conform to any known ORF, alternate origins must be considered. Protein modifications will add predictable masses up to a few hundred Daltons, and again confirmation can be obtained by analysis of tryptic peptides. Recoding events such as frameshifting or bypassing will often result in more drastic changes in mass. Tryptic peptide analysis will

identify the genome origin from which a limited number of possible masses due to recoding events can be predicted. Again, analysis of tryptic peptides should allow identification of the specific recoding event.

We are initially analyzing mitochondria of the yeast *Saccharomyces cerevisiae* since this will limit complexity to a fraction of the whole yeast genome - perhaps 500 genes out of 6,500. We are also pursuing tagging methods that are suited for examining one gene at a time and that will be more suited for analysis of the complexity of proteins coming from human genes. From this approach we hope to define the real complexity of the genome products.

147. ASDB: Database of Alternatively Spliced Genes

M. S. Gelfand, **I. Dubchak**, I. Dralyuk, and M. Zorn
Institute of Protein Research, Russian Academy of Sciences, Pushchino, 142292, Russia and National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
ildubchak@lbl.gov

Alternative splicing is an important regulatory mechanism in higher eukaryotes[1]. By recent estimates, at least 30% of human genes are spliced alternatively (Mironov, A.A. and Gelfand, M.S. Proc. 1st Int. Conf. on Bioinformatics of Genome Regulation, 1998. v. 2, p. 249). Alternative splicing plays a major role in sex determination in *Drosophila*, antibody response in humans and other tissue or developmental stage specific processes[2-5]. The database of alternatively spliced genes can be of potential use for molecular biologists studying splicing, developmental biologists, geneticists, and cell biologists. Version 1.1 of ASDB contains information about protein products of alternatively spliced genes. Selecting all SwissProt entries containing the words "alternative splicing" has generated 1663 proteins. Then clusters of proteins that could arise by alternative splicing of the same gene were created. Two proteins from the same species belong to a cluster if they have common fragments not shorter than 20 amino acids. Each cluster is represented in the database by the multiple global alignment of its members, allowing for easy identification of regions produced by alternative splicing. The database contains 241 clusters with more than one member. The database can be searched using Medline, SwissProt, and GenBank identifiers and accession numbers. Standard context search can be performed over SwissProt keyword, description, taxonomy, and comment fields and feature tables. ASDB contains internal links between entries and/or clusters, as well as external links to Medline, GenBank and SwissProt entries. Next steps of ASDB development will be incorporation of DNA data, classification of main types of alternative splicing, incorporation of data on aberrant splicing and splicing mutations. Automated processing of existing databases with minimum manual curation produced the current version of the database. In future we plan to add manual curation of the database, including addition of splicing variants described in the literature but not annotated in GenBank.

AVAILABILITY

ASDB is currently available at the URL http://cbcg.nersc.gov/asdb. The administrator of the database can be contacted by Email: asdb@lbl.gov.

REFERENCES

[1]Sharp, P.A. (1994) Cell. 77, 805-815
[2]Stamm, S.; Zhang, M.Q.; Marr, T.G. and Helfman, D.M. (1994) Nucleic Acids Res. 22,1515-26.
[3]Chabot, B. (1996) Trends Genet. 12, 472-478
[4]Breitbart, R.E., Andreadis, A. and Nadal-Ginard, B. (1987) Annu. Rev. Biochem. 56, 467-495
[5]Smith, C.W., Patton, J.G., and Nadal-Ginard, B. (1989) Annu. Rev. Genet. 23, 527-577

148. Prediction of Protein Structural Domains

Robert Miller[1], Winston A. Hide[1], and **David C. Torney**[2]
[1]South African National Bioinformatics Institute, University of the Western Cape and [2]Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, New Mexico
dct@lanl.gov

New challenges of sequence analysis have arisen with the advent of functional genomics. In particular, there is a premium on being able to make good use of small collections of example sequences, of known function, for classifying and predicting the functions of new sequences. Established techniques of classification have thus far not performed as well as needed, even with relatively abundant data, as in the case of exon prediction. We have therefore developed new example-based Bayesian statistical techniques for classification. These approaches can use conserved sequence motifs when these are present, but such overt similarities are not required because our techniques capture and employ all the statistical properties exhibited by a collection of example sequences. Thus, the likelihood for any sequence being a member of a given functional class is derived based on examples from the class. As many classes of structurally or functionally related biological sequences have only a relatively small number of examples, the prior specification of "what the statistical properties of a class might comprise" is critical. Our techniques include judicious choices for this prior, using insights about the statistical and physical properties of the sequences. One promising application of our techniques is the development of automatic clustering methods for use with a class of sequences. This will enable the discovery of heterogeneity within a class, improving the ability to predict class membership and deriving new classes.

To establish and refine our techniques, as well as provide the basis for predicting structural and functional aspects of new protein sequences, we created datasets of sequence-dissimilar examples of known secondary structures, using DSSP applied to Brookhaven PDB files. We obtained 64,775 residues of alpha-helix, 47,304 residues of beta-sheet, and 45,549 residues of coil, exhibiting recognized structural features such as helix capping mechanisms. The application of our techniques classifies regions of novel protein sequences into these three categories. We will report the details of the implementation and performance, making comparisons with established approaches. Data may be submitted for analysis by our methods via the World Wide Web (http://www.sanbi.ac.za/karoo).
Supported by the U.S. D.O.E. Office of Biological and Environmental Research under contract W-7405-ENG-36.

149. Rapid and Sensitive Characterization of Proteomes; an Adjunct to the Genome

Richard D. Smith, Ljiljana Pasa Tolic, Mary S. Lipton, Pamela K. Jensen, Gordon A. Anderson, and James E. Bruce
Environmental Molecular Sciences Laboratory, Mail Stop: K8-98, Pacific Northwest National Laboratory, Richland, WA 99352
dick.smith@pnl.gov

In contrast to an organism's virtually static and well defined genome, the proteome continually changes in response to external and internal events. 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, and the speed and sensitivity of subsequent mass spectrometric analyses for identification.

Our objective is to circumvent the limitations of this approach by directly characterizing the cell's polypeptide constituents by combining the speed of capillary isoelectric focusing (CIEF) and the mass accuracy and sensitivity obtainable with Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. CIEF-FTICR MS studies require orders of magnitude smaller sample sizes than required by 2-D PAGE technology, and initial efforts have demonstrated sensitivities well into the attomole range, as well as the potential for further significant improvements. A key attraction of FTICR is the enhanced facility for protein identification based upon the use of genome sequence data. Isotopically depleted growth media allow highly accurate molecular mass determinations for larger proteins than otherwise possible, and further improves achievable sensitivity and detection limits. We describe our efforts aimed at developing on-line CIEF-FTICR techniques, their comparison with conventional methodologies, and their initial application to several prokaryotes for which complete genome sequences are available. We will also describe new approaches for the determination of precise expression levels for large numbers of proteins in the same measurement.

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.

Microbial Genome Program

150. Archaeal Proteomics

Carol S. Giometti[1], Sandra L. Tollaksen[1], Xiaoli Liang[1], Michael W. W. Adams[2], James F. Holden[2], Angeli Menon[2], Gerti Schut[2], Claudia I. Reich[3], Gary J. Olsen[3], and John Yates, III[4]
[1]Center for Mechanistic Biology and Biotechnology, Argonne National Laboratory, Argonne, IL 60439; [2]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602-7229; [3]Department of Microbiology, University of Illinois, Urbana, IL 61801; and [4]Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195-7730
csgiometti@anl.gov

The genomes of several Archaea are either partially or completely sequenced, revealing the presumptive sequences of encoded proteins. However, the functions of these proteins can only be inferred from sequence similarity with known proteins, and the mechanisms by which the expression and function of most of the proteins are regulated remain unknown. The goal of the Archaeal Proteomics Project is to identify archaeal proteins and regulatory pathways relevant to bioremediation and energy technology, processes of interest to the U.S. Department of Energy. We are using two-dimensional gel electrophoresis (2DE) to purify and quantitate proteins expressed in Archaea grown under a variety of conditions designed to modulate specific metabolic pathways. The compartmentalization of Archaeal proteins is being determined by 2DE of subcellular fractions. Proteins are identified on the basis of similarities between observed peptide masses for tryptic digests generated from proteins in the 2DE gels and calculated peptide masses for the proteins encoded in the genome sequences. We are obtaining peptide masses by using matrix-assisted laser desorption ionization mass spectrometry. Initial work is focused on the proteomes of *Pyrococcus furiosus* and *Methanococcus jannaschii*, both hyperthermophilic Archaea with growth temperatures near 100 °C and enzymatic capabilities that promise to be of value in bioremediation reactions, energy conversion, and chemical processing systems. Whereas many of the enzymatic activities associated with primary metabolic pathways have been characterized in *P. furiosus*, the metabolic capabilities of *M. jannaschii* have only been inferred from gene sequence information. Thus far, the most abundant proteins in the 2DE patterns of *M. jannaschii* and *P. furiosus* lysates have been identified using peptide mass searches, membrane and cytosolic proteins from *P. furiosus* have been compared, and quantitative changes in *M. jannaschii* proteins under several different growth conditions have been analyzed. These preliminary results are the foundation for the *M. jannaschii and P. furiosus* proteome databases. This work is support by the U.S. Department of Energy, Office of Biological and Environmental Research, under Contract No. W-31-109-Eng-38.

151. Microbial Genome Sequencing and Analysis at TIGR

William C. Nierman, Tamara Feldblyum, Rebecca A. Clayton, Robert D. Fleischmann, Owen White, Claire M. Fraser, and J. Craig Venter
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850
wnierman@tigr.org

Advances in automated DNA sequence analyses and a whole genome shotgun strategy pioneered by The Institute for Genomic Research resulted in the first complete genome sequence for a free living bacterium, *Haemophilus influenzae* in 1995. Since then the TIGR microbial sequencing program has expanded to include 8 completed bacterial genomes representing 12.3 Mb of sequence data and another 16 genomes in progress. Data analyses of the sequenced organisms indicates that on the average about 46% of the genes have no assigned biological function and 26% of those genes are unique to a particular species. The number of genes with unknown function and the number of genes unique to an organism indicate the wide variety and adaptability of the bacterial world, their ability to adjust to extreme living conditions, their ability to metabolize a variety of chemical compounds as sources of energy, and our rudimentary knowledge of the scope of the physiology and metabolism of earth's microbes.

The availability of complete microbial genome sequence data opens new ways of investigating these organisms. Analysis of TIGR sequenced microbial genomes has provided new and exciting insights into the phylogenetic relatedness of organisms, novel metabolic pathways, biochemical strategies of pathogenic microbes, the functional identification of genes, and the minimal gene content of free living organisms.

152. Genomics and Engineering of a Radioresistant Bacterium

Kenneth W. Minton, Kira S. Makarova, Michael J. Daly, Eugene V. Koonin, Hassan Brim, L. Aravind, and Ajay Sharma
Uniformed Services University of the Health Sciences, Bethesda, MD 20814-4799
kminton@usuhs.mil

The eubacterium *Deinococcus radiodurans* is the most DNA damage-resistant organism discovered to date. It is therefore of intrinsic interest to study its DNA repair mechanisms, and towards this end the full genomic sequence of this organism has recently been obtained by TIGR. We have fully annotated this sequence with special attention to properties that might render this organism radioresistant. Features noted to date include a novel enzyme, combining potential repair domains from three independent repair proteins. This gene is currently being knocked out of the deinococcal genome and properties of the null mutant will be reported. Similarly, desiccation-resistance proteins similar to those seen in plants have also been discovered in the *Deinococcus radiodurans* genome. This is of particular significance, as there is a known positive correlation between deinococcal desiccation-resistance and radioresistance. The properties of knock out mutants will be reported.

Finally, an expansion of several protein families, including phosphatases, proteases, acyl transferases, the mutT family of pyrophosphatases, and thioredoxins have been noted. *Deinococcus radiodurans*' genome is extraordinarily rich in repeated sequences, suggesting a mechanism of repair that will be presented. In addition, it is the first bacterium to be sequenced that has multiple chromosomes (three). Engineering of this versatile organism for organopollutant degradation in radioactive mixed waste environments and engineering of heavy metal resistance in this organism will be described. Finally, current attempts to acquire large amounts and crystallize Deinococcus' extraordinary and highly toxic RecA protein will be described.

153. Functional Analysis of *Deinococcus radiodurans* Genomes by Targeted Mutagenesis

Kwong-Kwok Wong, William B. Chrisler, Lye Meng Markillie, and Richard D. Smith
Pacific Northwest National Laboratory, Molecular Biosciences, MS P7-56, P.O. Box 999, Richland, WA 99352
kk.wong@pnl.gov

D. radiodurans, previously known as *Micrococcus radiodurans*, strains R1, has extreme resistance to genotoxic chemicals, oxidative damage, high levels of ionizing and UV radiation, and desiccation. The ability to survive such extreme environments is attributed in part to a unique DNA repair system in combination with its chromosome copy number and structure, as well as factors affecting the survival of other cellular components. There is evidence suggesting that the carotenoids which cause red pigmentation in *D. radiodurans* may act as free radical scavengers, thus increasing resistance to DNA damage by hydroxyl radicals. High levels of two oxygen toxicity defense enzymes, superoxide dismutase and catalase, are also found in *D. radiodurans*. In addition, the Deinococcal outer membrane lipids are complex and distinct from those found in the rest of the bacterial world and it has been suggested that they, together with the plasma membrane, may also be involved in stress resistance. However, the genetic basis for these stress resistance is still not clear. With the genomic sequence information of *D. radiodurans* R1, we have developed a simple and general targeted mutagenesis method to perform a genome-wide analysis of putative genes involved in the stress resistance. We have generated mutations in katA (catalase) and sodA (superoxide dismutase). Both katA and sodA mutants are shown to be required for the extreme ionizing radiation resistance. Several other mutations have been generated and are being analyzed for their roles in stress resistance.

154. Complete Genome Sequence of *Deinococcus radiodurans*

Owen White, John Heidelburg, Claire Fraser, and J. Craig Venter
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20855
owhite@tigr.org

Deinococcus radiodurans is a non-pathogenic, non-sporulating, red-pigmented Gram+ bacterium. *D. radiodurans* was originally found in radiation sterilized food that under went spoilage. It is remarkable in that it is the most radioresistant organism to have ever been isolated (Moseley, 1983). An important component of this resistance is the ability to repair damage to chromosomal DNA. *D. radiodurans* cultures exposed to 1.5 Mrad of radiation displayed reduction in size of genomic DNA fragments corresponding to approximately 100 double stranded breaks (DSBs) per genome. (Typically, most prokaryotic and eukaryotic organisms cannot tolerate more than 5 double stranded breaks per genome without reduced survival.) Remarkably, within eight to ten hours after exposure, *D. radiodurans* genomic fragment lengths are restored to size ranges seen in non-treated cells. During this repair time, cellular replication of *D. radiodurans* is arrested (Daly et al., 1994); however, after this eight to ten hour interval, the cells display 100% survival with no detectable mutagenesis of their completely restored genomes. The genome sequence of *Deinococcus* is complete and we have determined the genome is composed of 3 chromosomes and a small plasmid; a number of unique sequence elements have been identified. The content of the genome, along experimental results will be discussed in context of this organism's unique ability to withstand gamma radiation.

155. Complete Genome Sequencing of *Shewanella putrefaciens*

Rebecca A. Clayton, John Heidelberg, Kenneth Nealson, Eric Gaidos, Alexandre I. Tsapin, James Scott, J. Craig Venter, and Claire M. Fraser
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850 ; NASA Jet Propulsion Lab, 4800 Oak Grove Drive, Pasadena, CA 91109
rclayton@tigr.org

We are midway through the closure process in the complete genome sequencing of *Shewanella putrefaciens*. Random sequencing was completed in July, 1998, and closure began in August, 1998. We present preliminary annotation of the genome, with whole genome comparisons with other completed microbial genomes. *Shewanella putrefaciens* has high 16S rRNA sequence similarity to *Escherichia coli*, and also to *Vibrio cholerae*, a genome in the final stages of closure at TIGR. approximately half its open reading frames have high sequence similarity to *E. coli*. About 15% of the *S.putrefaciens* genome has high sequence similarity with *V. cholerae* but not with *E. coli* K12. Analysis of the assemblies suggests that the completed genome size will be approximately 5 Mb.

156. Whole Genome Sequence and Structural Proteomics of *Pyrobaculum aerophilum*

Sorel Fitz-Gibbon[1], Ung-Jin Kim[2], Heidi Ladner[1], Elizabeth Conzevoy[1], Gigi Park[2], Karl Stetter[3], Jeffrey H. Miller[1], and Melvin I. Simon[2]
[1]Department of Microbiology and Molecular Biology Institute, University of California, Los Angeles, California; [2]Biology, California Institute of Technology, Pasadena, California; and [3]University of Regensberg, Germany
sorel@mbi.ucla.edu

Pyrobaculum aerophilum is a hyperthermophilic archaeon, isolated from a boiling marine water hole, that is capable of growth at 104°C. This microorganism can grow microaerobically, unlike most of its thermophilic relatives, making it amenable to a variety of experimental manipulations and a candidate as a model organism for studying archaeal and thermophilic microbiology. We have sequenced the entire genome using a random shotgun approach (3.5X genomic coverage) followed by oligonucleotide primer directed sequencing, guided by our fosmid map. The 2.2 Mb genome codes for more than 2000 proteins, 30% of which have been identified by their sequence similarities to proteins of known function. Only 15% of the *Pyrobaculum aerophilum* proteins have related high resolution structures. In collaboration with the DOE/UCLA Laboratories and Los Alamos National Laboratories we have initiated a project to express and purify proteins for structure determination by NMR or xray diffraction. The three dimensional structures of the *Pyrobaculum aerophilum* proteins will give one the power to understand and manipulate protein function and are crucial to fully exploiting the information in the genome. At this time several proteins have been cloned, expressed in *E.coli*, purified and crystals which diffract to high resolution have been obtained.

157. The Genome Sequence of a Hyperthermophilic Archaeon: *Pyrococcus furiosus*

Robert B. Weiss[1], Diane Dunn[1], Mark Stump[1], Raymond Yeh[1], Joshua Cherry[1], and Frank T. Robb[2]
[1]Dept. of Human Genetics, University of Utah, Salt Lake City, Utah and [2]Center of Marine Biotechnology, University of Maryland, Baltimore, Maryland
bob.weiss@genetics.utah.edu

Pyrococcus furiosus is a strictly anaerobic archaeon that grows optimally at 100°C by a fermentative-type metabolism in which complex peptide mixtures such as yeast extract and Tryptone, and also certain sugars, are oxidized to organic acids, H_2 and CO_2. The organism was isolated from geothermal marine sediment in shallow waters off Vulcano Island, Italy. We have determined the complete sequence of this organism's genome. It is 1,908,253 base pairs in length, with a GC-content of 40.8%. Recently, the

complete sequence of a distantly-related species, *Pyrococcus horikoshii*, has been determined by a group in Japan (www.bio.nite.go.jp). This species was isolated from a hydrothermal vent at a depth of 1395 meters in the Sea of Japan. Comparative analysis is revealing complex gene re-arrangements and changes in gene content between these two *Pyrococcus* species.

The genome content and organization reveals many potential operons, one rRNA operon, 46 tRNAs, 22 insertion elements, 7 SR elements, and 14 inteins. 50% of the ORFs are of unknown function, and of this class 19% are in common between the two *Pyrococcus* species. The 22 putative insertion elements in the genome of *P. furiosus* are not found in the *P. horikoshii* genome.

The genome of *P. furiosus* is 170 kb larger than *P. horikosihii*, with about 70% DNA identity conserved within open reading frames. Genome to genome dot plot alignment reveals the remnant of a conserved diagonal. The longest co-linear segment between the two genomes is 70 kb, and much of the remnant diagonal is interspersed with inversions, deletions and insertions. The list of major gene clusters present in *P. furiosus* but not in *P. horikoshii* include: maltose/trehalose transport, phosphate uptake system, major parts of the urea and TCA cycle, and amino acid metabolism of tryptophan, aromatics, arginine, and isoleucine/valine. The maltose/trehalose transport operon is within a 17 kb segment flanked by putative insertion elements. This segment is also found in *Thermococcus litoralis*, another isolate from a Mediterrean marine geothermal location. The high identity (>99%) between these two segments suggests a recent lateral transfer event between *T. litoralis* and *P. furiosus*.

158. The *Chlorobium tepidum* Genome Sequencing Program at TIGR

Karen A. Ketchum, Matthew D. Cotton, Cheryl Bowman, M. Brook Craven, Tanya Mason, Terrence Shea, William Nierman, and Claire M. Fraser
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850
kketchum@tigr.org

The genus *Chlorobium* is placed in the taxonomic group of green sulfur bacteria (Chlorobiaceae). They are formally classified as Gram-negative organisms. Members of this genus are photoautotrophs that can generate chemical energy through an electron transport chain in the cytoplasmic membrane that is associated with a light-harvesting complex housed in a specialized organelle called the chlorosome. The components of this light-harvesting apparatus and some of its organizational structure are reminiscent of photosystems found in plant chloroplasts and, therefore, the evolutionary relationship of these prokaryotes to eukaryotic organelles is of interest. *Chlorobium* species can also fix CO_2, although the biochemical pathway used by these prokaryotes is distinct from the Calvin cycle found in higher plants.

C. tepidum was initially identified from a hot spring in New Zealand (Wahlund et al. 1991). This species is thermophilic with and optimum growth temperature of @ 47°C. It has a genome size of 2.1 Mb (Naterstad et al., 1995) with a G + C content of 56.5 mol%. *C. tepidum* was nominated for sequencing by the DOE because it has a prominent role in global carbon cycling and an interesting phylogenetic position in the Eubacterial kingdom.

The *C. tepidum* genome project was initiated in March of 1998. Genomic DNA was generously supplied by Dr. Donald A. Bryant, Earnest C. Pollard Professor of Biotechnology and Professor of Biochemistry and Molecular Biology at The Pennsylvania State University. Random sequencing of a small insert (1.6 - 2.5 kb) plasmid library began

in May and is now complete. We obtained 32,246 sequencing reads for 8X coverage of the genome. The overall success rate was 82%. We are now sequencing a large insert lambda library which will provide linking information for our contigs. The current genome assembly has 41 groups with 38 sequencing gaps and 80 physical gaps. A progress report on the *C. tepidum* project will be presented.

159. Searching for Synteny: A Whole-Genome Comparison of *Caenorhabditis elegans* with *Saccharomyces cerevisiae*

Karen L. Diemer and **Kelly A. Frazer**
Genome Sciences Department, Lawrence Berkeley National Laboratory, 1 Cyclotron Road MS 84-171, Berkeley, CA 94720
kelly@mhgc.lbl.gov

Characterizing the syntenic relationships of genes in different species has been a valuable tool for deciphering a variety of biological phenomenons. The completely sequenced yeast *Saccharomyces cerevisiae* and the extensively sequenced worm *Caenorhabditis elegans* genomes provide us with the opportunity to search on a whole-genome wide basis for conservation of gene order between these distantly related eukaryotic organisms. The yeast and worm genomes diverged approximately 965 million years ago (Doolittle et al. 1996), therefore any conservation of gene order is likely due to biological forces dictating genome organization rather than a lack of shuffling of genes that were neighbors in the last common ancestor. We compared protein translations of the 6221 yeast ORFs to the available worm sequence data (85% of total) to determine whether any paired genes, loci that are consecutive (neighbors) in both organisms, exist. Ten pairs of adjacent yeast ORFs were identified that have significant matches (TBLASTN expect values <1e-21) adjacent to each other in the worm genome. Four of these paired ORFs consist of genes encoding for different core histones, three consist of genes that encode for proteins of no known related function, and three consist of ORFs that are part of the same gene in yeast but had not yet been identified as such. These data indicate that the study of conserved gene pairs in distantly related eukaryotes may provide insights into the selective pressures governing the clustering of certain genes as well as serve to facilitate the assignment of putative ORFs into protein encoding units.

160. Microbial Genome Sequencing and Comparative Analysis

D.R. Smith, M. Ayers, R. Bashirzadeh, H. Bochner, M. Boivin, G. Breton, S. Bross, A. Caron, A. Caruso, R. Cook, P. Daggett, L. Doucette-Stamm, J. Dubois, J. Egan, D. Ellston, J. Ezedi, T. Ho, K. Holtham, P. Joseph, M. LaPlante, H-M. Lee, R. Gibson, K. Gilbert, J. Guerin, D. Harrison, J. Hitti, P. Keagle, J. Kozlovsky, G. LeBlanc, W. Lumm, P. Mank, A. Majeski, J. Nölling, D. Patwell, J. Phillips, B. Pothier, S. Prabhakar, D. Qiu, J.N. Reeve[1], M. Rossetti, M. Sachdeva, P. Snell, [2]P. Soucaille, L. Spitzer, R. Vicaire, K. Wall, Y. Wang, L. Wong, A. Wonsey, K. Weinstock, Q. Xu, and L. Zhang
[1]Dept.of Microbiology, The Ohio State University, Columbus, Ohio and [2]INSA, Toulose, France; Genome Therapeutics Corp., Waltham, Massachusetts
doug.smith@genomecorp.com

This project is applying automated sequencing technology and bioinformatics tools to the analysis of microbial genomes with potential applications in energy production and bioremediation., Efforts have focused on two genomes in particular, those of *Methanobacterium thermoautotrophicum* strain ΔH, and *Clostridum acetobutylicum* strain ATCC 824.

Methanobacterium thermoautotrophicum strain ΔH is a thermophilic archaeon that grows at temperatures from 40-70° C, and was isolated in 1971 from sewage sludge. The complete 1,751,377 bp sequence of the genome of *M. thermoautotrophicum* was determined by a whole genome shotgun sequencing approach. The results of extensive comparative and functional analysis work were published last year in the Journal of Bacteriology, Volume 179, 7135-7155.

C. acetobutylicum strain ATCC 824 has a 4.2 Mb, AT-rich genome, and is one of the best-studied solventogenic clostridia (it has been used commercially to produce acetone, butanol and ethanol). The shotgun sequencing phase has been completed, with 4.9 Mb of multiplex and 21.3 Mb of ABI raw sequence reads (6.3 fold total redundancy) that produced 551 contigs spanning 4,030,725 bases when assembled using PHRAP with quality scores. The genome has been finished to 27 ordered contigs, with quality enhancement, at the time of this writing.

Physical mapping of the *C. acetobutylicum* genome by P. Soucaille and coworkers (INSA, Toulose) has shown that this strain harbors a large plasmid, designated pSOL1, of about 210 kb in size. Further studies by the same group revealed that loss of this plasmid coincides with the loss of the capacity to produce acetone and butanol and that the genes involved in solvent formation reside on pSOL1. We now have the complete 203 kb sequence of this plasmid.

C. acetobutylicum contains a variety of genes involved in the utilization of polysaccharides such as starch, hemicelluose and cellulose. The potential to degrade cellulose, indicated by the presence of an entire set of genes predicted to code for a cellulose-hydolysing mutlienzyme complex termed cellulosome, is surprising as cellulolytic activity is unknown for *C. acetobutylicum*. In addition, a gene similar to the toxin A encoding gene from the pathogenic clostridium *C. difficile* is present in the non-pathogenic *C. acetobutylicum*, coding for a polypeptide of ~ 2800 residues the majority of which is organized in ~125 repeats of 20 amino acids each. The genome of *C. acetobutylicum* ATCC 824 seems to be nearly void of mobile genetic elements. Only a single copy of a transposase gene, belonging to the Tn3 family and located on plasmid pSOL1, could be identified. Two gene clusters of four genes each show similarity to bacteriophage-like elements. There are 11 ribosomal operons. The data are available in GenBank and on our Web page (http://www.genomecorp.com).

161. Genome Sequencing and Analysis

G. J. Olsen, C. I. Reich, N. C. Kyrpides, J. H. Badger, D. E. Graham, P. J. Haney, L. K. McNeil, G. M. Colón González, A. A. Best, B. P. Kaine, and **C. R. Woese**
Department of Microbiology, University of Illinois, Urbana, IL 61801
carl@ninja.life.uiuc.edu

Our work is directed toward the sequencing and interpretation of selected microbial genomes, and has several components.

To study the sequence basis of thermal adaptation, we have been comparing the proteins encoded in the genome of *Methanococcus jannaschii* (an extreme thermophile) with those of related mesophiles. We have documented specific amino acid changes correlated with the difference in organismal growth temperatures, as well as systematic changes in amino acid properties. These trends are recurring themes; they are observed in 82-93% of all compete protein sequences analyzed. To generate more data for this comparative analysis, we have prepared sequencing quality genomic DNA libraries from *Methanococcus maripaludis*, and have started partially sequencing clones from this library (this sequencing is supported by NASA).

To ensure the availability of data from key (diverse) eukaryotic microorganisms, we have prepared sequencing quality genomic DNA libraries for *Giardia lamblia*. These libraries are being used by Mitchell L. Sogin (Marine Biology Laboratory) to generate a nearly complete genome sequence for this organism (this sequencing is supported by the NIH). These data will be critical to understanding the origins of eukaryotes and their unique cellular organization.

The sequence data resulting from our participation in the Microbial Genome Initiative has stimulated additional research in our laboratories. Specifically:

1. We have continued to make new gene identifications in the sequenced genomes;

2. We have begun an experimental verification of the function of some novel RNA methylase genes;

3. We have cloned and expressed RNA polymerase genes and transcription initiation factors from the Archaea and have experimentally identified new protein-protein interactions in the transcription apparatus; and

4. We have entered into a collaboration with Carol Giometti (Argonne to National Laboratory) and Michael Adams (University of Georgia, Athens) study the proteomes of *Methanococcus jannaschii* and *Pyrococcus furiosus*.

162. Use of Suppressive Subtractive Hybridization to Identify Genomic Differences among Enteropathogenic Strains of *Yersinia enterocolitica* and *Yersinia pseudotuberculosis*

Lyndsay Radnedge, Peter Agron, Lisa Glover, and Gary Andersen
Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California
andersen2@llnl.gov

A comparison of genomic sequences among closely related species is likely to reveal unique DNA regions that define the genetic basis for the underlying differences in their phenotypic variation. An example of two closely related human pathogens that differ in their ability to colonize animal hosts as well as their persistence in the environment are the enteropathogenic bacteria, *Yersinia enterocolitica* and *Y. pseudotuberculosis*. Of the two pathogens, *Y. enterocolitica* is more often associated with human infection, especially in day-care centers where the disease is transmitted through infected food, water or soil. Although less frequently diagnosed, infection with *Y. pseudotuberculosis* is most commonly transmitted through contact with infected birds or mammals. A large percentage of *Y. pseudotuberculosis* infections are subclinical with no observable symptoms in the exposed individuals. Unlike *Y. enterocolitica*, *Y. pseudotuberculosis* may enter the bloodstream of predisposed individuals, causing a lethal septicemia.

We used a PCR-based subtractive hybridization method termed suppressive subtractive hybridization (SSH) (Diatchenko et al., 1996. Proc. Natl. Acad. Sci. 93:6025-6030) to define the differences between the genomes of *Y. enterocolitica* and *Y. pseudotuberculosis*. This technique uses PCR amplification to enrich for unique segments of restricted DNA and simultaneously limits non-target amplification by suppression PCR. The bacterial genome of interest in this comparison is called the tester DNA and the comparison genome is called driver DNA. Using pair-wise comparisons among four strains of *Y. enterocolitica* and four strains of *Y. pseudotuberculosis* our initial aim was to identify tester-specific sequences in the type-strains of both species. Control subtractions yielded no PCR products, indicating that the protocol effectively subtracts identical DNA sequences. We have optimized the reaction conditions for the subtraction experiments. Subtracted DNAs were successfully cloned into pGEMT-Easy plasmid vector (Promega); almost 100% of resulting white colonies contained an insert. The clones so far characterized contain inserts that range in size from 200 bp to 1700 bp. The band size distribution of the cloned products represents the distribution of the amplified subtractive hybridization products. Plasmids containing an insert were sequenced on an ABI 377 using dye terminator chemistry.

BLAST searches of tester-specific DNA sequences reveal homologies to known bacterial genes (including genes involved in pathogenicity) and eleven novel DNA sequences. Included in the regions unique to the *Y. enterocolitica* type-strain is a difference product with homology to the response-regulator phoP, which has been associated with virulence in *Salmonella typhimurium*. 92% of oligonucleotide probes designed using these tester-specific DNA sequences distinguished genomic

DNA isolated from *Y. enterocolitica*, while 100% of such probes were specific for *Y. pseudotuberculosis*. Furthermore, SSH has proved to be sensitive enough to design probes against tester-specific DNA sequence that have been shown to discriminate between genomic DNA isolated from two strains of *Y. enterocolitica* (58% of the probes successfully discriminate tester DNA).

Streamlining, automating the steps, and increasing the throughput of this technique should enable large-scale genomic comparison among closely related strains and the generation of strain-specific oligonucleotide probes for molecular epidemiology studies.

163. Exploring Whole Genome Sequence Information for Defining the Functions of Unknown Genes and Regulatory Networks in Dissimilatory Metal Reduction Pathways

Jizhong Zhou[1], Douglas Lies[2], Gary Li[1], Rebecca Clayton[3], Kenneth H. Nealson[2], Claire Fraser[3], James M. Tiedje[4]
[1]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831; [2]Department of Geology and Planetary Sciences, Jet Propulsion Laboratory and California Institute of Technology, Pasadena, CA 91109; [3]The Institute of Genomic Research, Rockville, MD 20850; and [4]Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824-1325
zhouj@ornl.gov

The goal of this project is to explore whole genome sequence information to understand the genetic structure, functions, regulatory networks and mechanisms of dissimilatory metal reduction pathways. The following objectives will be pursued: (1) To identify the genes involved in dissimilatory metal reduction pathways in MR-1; (2) To generate and characterize deletion mutants for defining the functions of the unknown genes expressed under metal reducing conditions; (3) To understand the metabolic and genetic control of gene expression at the genome level under iron reducing conditions; and (4) To explore genetic diversity of the dissimilatory iron reduction pathways in selected thermophilic and psychrophilic iron-reducing bacteria. To achieve these objectives, we will construct microarrays consisting of all ORFs from MR-1, and use them to monitor gene expression patterns under different growth conditions for identifying the genes involved in dissimilatory metal reduction and for defining the putative functions of unknown ORFs. We will also use them to compare the gene expression patterns when MR-1 is shifted from aerobic to anaerobic iron-reducing conditions, and to compare the gene expression patterns between wild type and specific regulatory mutants for understanding the metabolic control and regulatory networks of iron reduction pathways. In addition, we will generate and characterize specific deletion mutants for defining the functions of unknown ORFs. Finally, we will use the microarrays to assay the genetic diversity of iron reduction pathways at the genomic level among representative thermophilic and psychrophilic iron-reducing bacteria.

To optimize the conditions for microarray hybridization, we are constructing prototype microarrays containing genes involved in anaerobic metabolisms to understand how these genes are regulated under anaerobic conditions. As a part of this project, nine psychrophilic iron-reducing bacteria have also been isolated from Siberia and Alaska permafrost soils, deep marine sediments and Hawaii deep sea water. These bacteria are also able to reduce cobalt, chromium at low temperature. Phylogenetical analysis showed that they are closely related to Shewanella and Vibrio species. In addition, we are using sequences of genes known from mutational studies to be involved in metal reduction (mirAB) as hybridization probes to search for homologues in additional Shewanella species and other metal-reducing bacteria.

164. Identification, Isolation, and Genome Amplification of Abundant Non-Cultured Bacteria from Novel Phylogenetic Kingdoms in Two Extreme Surface Environments

Cheryl R. Kuske, Susan M. Barns, John D. Dunbar, Jody A. Davis, and Greg Fisher
Environmental Molecular Biology Group, LS-7, M888, Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
Kuske@lanl.gov

Microbial genome sequencing projects have produced a wealth of information on microbial genetics, biochemistry, and evolution with important medical, environmental, agricultural and industrial applications, but have focused primarily on species we can easily culture. Cultured bacteria are only a small fraction of the total bacterial diversity present in the environment. Non-cultured organisms of considerable genetic and biochemical diversity are present in arid and extreme surface environments. 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 initiated a project to (1) identify novel bacterial kingdoms in 16S ribosomal RNA gene (rDNA) libraries from two extreme surface environments to expand our current understanding of the scope of environmental bacterial diversity, (2) determine the abundance and activity of these novel organisms in the environment by using rRNA-targeted fluorescent probes, and (3) collect cells of novel phylogenetic groups that are abundant and active in the environment by flow cytometry and isolate or amplify DNA from the pools of bacterial cells.

Our preliminary results demonstrate the extensive diversity of bacteria present in two extreme terrestrial surface environments. Sequence analysis of 60 clones from two 16S rDNA libraries indicated that bacterial diversity in these two environments is extensive. RFLP fingerprint analysis of 800 clones in the two libraries demonstrated that diversity in the remaining clones was also great and that so far only 5% of them have fingerprints similar to those clones already sequenced. In fact, almost every clone produced a unique pattern, and there was very little overlap in patterns between the two environments. Preliminary RFLP and 16S rDNA sequence analysis indicate that most of the bacterial species represented in the clone libraries fall outside the known, previously-described bacterial taxa. Thus these libraries are a rich resource for identifying and isolating novel bacterial species with novel genes. We are developing fluorescently tagged oligonucleotide probes for detection and collection of some of these bacterial groups from environmental samples.

The pooled DNA of isolated, 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. Identification and sequence analysis of genes from abundant bacterial species will greatly enhance our understanding of their functional roles in the environment and will significantly expand the set of unique genes and proteins available for DOE missions, as well as for medical, industrial and agricultural applications.

165. WIT System: Advantages of Parallel Analysis of Multiple Genomes

Ross Overbeek, Mark D'Souza, Gordon Pusch, **Natalia Maltsev**, and Evgeni Selkov
Argonne National Laboratory, Argonne, Illinois
maltsev@mcs.anl.gov

The WIT system (http://wit.mcs.anl.gov/WIT2/wit.html) was designed and implemented to support genetic sequence and comparative analysis of sequenced genomes and metabolic reconstructions from the sequence data. It now contains data from 34 distinct genomes, although a few of the genomes are quite incomplete. It provides access to thoroughly annotated genomes within a framework of metabolic reconstructions, connected to the sequence data; data on regulatory patterns, protein alignments and phylogenetic trees; as well as data on gene clusters 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 major physiological designs. We recently developed a method for using conserved clusters of genes on the chromosome from numerous genomes to predict functional coupling between genes in those genomes[1]. The results obtained by applying this method to analysis of 34 genomes in WIT collection were very encouraging. We were able to predict major portions of most of the pathways of central metabolism (e.g. glycolysis, purine, pyrimidine biosynthesis, signal transduction, transmembrane transport pathways, etc). Our results agree well with the functional connections between the genes previously described in the literature. We believe that the precision of prediction and the amount of accessible functional coupling increases dramatically as more genomes are included in the analysis. As the number of genomes increases, this class of data may well become one of the significant resources in the effort to establish the function of the hypothetical proteins, better understanding of the functions of the paralogous genes and reconstruction of the functional connections in the higher level functional subsystems.

[1]Ross Overbeek, Michael Fonstein, Mark D'Souza, Gordon D. Pusch and Natalia Maltsev. Use of Contiguity on the Chromosome to Predict Functional Coupling. (June, 1998) In Silico Biol. 1, 0009

166. Microbial Protein and Regulatory Function Analysis and Database Program

Temple F. Smith
BioMolecular Engineering Research Center, Boston University, Boston, Massachusetts
http://bmerc-www.bu.edu/

We will have completed the first of three planned years on this project in December 1998. We have already made significant progress. Our first goal was to construct two preliminary profile databases. The first has been generated as part of the functional analyses of the various bacterial and archaeal genes (ORFs) that showed sequence similarity to probable Yeast mitochondria genes. We have generated profile-defining set sequences from a broad set of functional families. We have carefully studied the set of S. cerevisiae mitochondrial proteins and their homologs with the completed genomes and used them to create 367 profiles in which we have confidence that cover a broad set of biological functions. We have created 855 profiles from the Pfam protein family database by reviewing protein sequences in SWISS-PROT using a set of Hidden Markov Model-derived similarity families. There has been an effort to automate the generation of profile-defining sets from the blast-defined similar ORF families from the complete genomes and an initial set of profiles has been derived. This has produced a set of 807 profiles. Current efforts are centered on creating a method for automatically determining a set of disjoint profiles, that is, a set of profiles that are not redundant. We have also begun to investigate, in collaboration with Julio Collado-Vides (CIFN, Mexico), the potential of coordinate regulation among genes that are in neighbors in various biochemical pathways. Here we began with sets of genes in *E. coli* or some other bacteria or archaea that are organized in operons. Next, each of the operon sets are being examined in Yeast and *C. elegans* for shared regulatory words. The initial work here led to the identification of two different types of

eukaryotic operon equivalent organizations in Yeast, and led to our recent publication (Zhang and Smith, Microb. & Comp. Genomics, 1998). As part of our microbial genome comparisons, we developed a set of integral functions to examine nucleotide base composition along the entire length of the genome. These "excess plots" describe the abundance of a nucleotide property. Instead of examining base composition over a fixed window, excess plot values are calculated cumulatively at each position in the genome , adding one if the next nucleotide shares the property of interest, and subtracting one for each nucleotide with the converse property. These "excess plots" describe the abundance of a nucleotide property over its opposite property. The minima of the Purine Excess plots correlate with the origins of replication for seven bacterial genomes (*Escherichia coli*, *Bacillus subtilis*, *Mycoplasma pneumoniae*, *Mycoplasma genitalium*, *Helicobacter pylori*, *Haemophilus influenzae*, and *Synechocystis* PCC6803), while the maxima of the Purine Excess plots track with the three known replication termini (from *E. coli*, *B. subtilis* and *H. influenzae*). Keto Excess minima and maxima track the same replicative features in four of the nine bacterial genomes available at the time of study. Additionally, there is a strong correlation between purine excess and coding strand excess, evidenced most remarkably by *E. coli* and *Methanococcus jannaschii*, an archaebacterium. It is an ongoing effort to track and analyze excess plots for each new microbial genome, as well as the multiple chromosomes of the available eukaryotic genomes.

Citations:

Freeman, J.M., Plasterer, T.N., Smith, T.F. and Mohr, S.C. (1998). Patterns of genome organization in bacteria. Science 279, 1827.

Zhang, Xiaolin and Smith, Temple F. (1998). Yeast "operons". Microbial and Comparative Genomics 3(2), 133-140.

167. Annotation of Microbial Genomes

Frank Larimer, Richard Mural, Morey Parang, Manesh Shah, Victor Olman, Inna Vokler, Jay Snoddy, and Edward C. Uberbacher
Computational Biosciences, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee
fwl@ornl.gov
http://compbio.ornl.gov

Because of their completeness, sequenced microbial genomes present a number of challenges and opportunities not yet fully addressed by genomics. Conventional annotation is inherently single gene–protein centered, yet the operon and regulon organization of microbial genomes immediately accentuates the incompleteness of this simple gene-protein model. Additionally, few attempts have been made to represent regulatory features. Complete genomes require that regulatory and coding elements as well as global and local structural detail be addressed. Although less than a third of the major bacterial taxa have been sampled, a lack of comprehensive tools for representing evolutionary relationships and the richness of microbial diversity is already evident. Finally, the rapid proliferation of completed genomes emphasizes the need for regular updates to annotation.

We are developing microbial annotation systems to address these needs within the context of the Genome Channel and the Genome Annotation Consortium. In cooperation with The Institute for Genomic Research, we currently have views of the various complete microbial genomes sequenced by TIGR in the Genome Channel. Other complete genomes will be added shortly, and views of genomes in progress will be developed. Among the features being implemented are the following:

- A visual, integrated contextual browser for viewing genomic relationships from the chromosome to the nucleotide level, within and between genomes;
- Improved and consistent gene calling, with emphasis on accurate prediction of translation

start as well as accurate calling of short genes (<300 nt);
- 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; and
- Linkage and integration of the gene/protein/function catalog to phylogenetic, structural, and metabolic relationships.

The rapidly growing microbial genome database poses significant challenges in both analysis and presentation, particularly in multigene, multiple-genome comparisons. The exploration of microbial diversity and an understanding of the nature and origin of evolutionary change requires depth of analysis as well as breadth of sampling. Comprehensive annotation tools will be critical to access the richness of genomic complexity.

168. Insights into Evolution from the *Thermotoga maritima* Genome

K.E. Nelson, R.A. Clayton, O. White, J.C. Venter, and C.M. Fraser
The Institute for Genomic Research, Rockville, Maryland
kenelson@tigr.org

Thermotoga maritima is the most extreme thermophilic organotrophic bacterium known, and one of the earliest branching Eubacteria. This obligate anaerobe is capable of utilizing various carbohydrates, including glucose, maltose, starch, cellulose and xylan as energy sources. In an attempt to further understand *T. maritima*, a whole genome random shotgun sequencing project was initiated at The Institute for Genomic Research (TIGR). The 1,860,725 bp *T.maritima* genome contains 1872 predicted coding regions, 54% (1005) of which have functional assignments, and 46% (867) of which are of unknown function. Of the sequenced Eubacteria, *T. maritima* has the highest percentage (24%) of genes that are most similar to archaeal genes. Eighty-one of these genes are clustered in regions of the genome that range in size from 4 - 20 kb. Five of these regions have a composition substantially different from the rest of the genome, suggesting that lateral gene transfer has occurred between the thermophilic Archaea and Eubacteria. In addition to repeat structures which can be identified only in thermophiles, there are 108 genes on the *T. maritima* genome that have orthologues only in the genomes of other thermophilic Eubacteria and Archaea. Along with a range of pathways for the degradation of both simple and complex carbohydrates, the *T. maritima* genome is revealing genes whose thermostable products may be useful for industrial processes. The genome sequence is also revealing similarities between the thermophilic Archaea and Eubacteria, and allowing us to address existing theories on evolution. The findings from an analysis of the complete genome sequence will be presented.

Ethical, Legal, and Social Issues

169. Genetics Adjudication Resource Project

Franklin M. Zweig
Einstein Institute for Science, Health and the Courts, Bethesda, Maryland
einshac@intr.net

During the first 19 months of project operation beginning in March, 1997, the Genetics Adjudication Resource Project (GARP) has operated from the center of the judicial system. It has provided basic science and ELSI orientation to 884 judges in attendance at nine genetics in the courtroom conferences. Approximately 60 federal and state judicial faculty members experienced several conferences and now comprise a sophisticated teacher core within the Judicial Branch. Another 25 judges attended the first and last conference in the series, experiencing both the basic orientation and an advanced course that extended the basics to biomarker testing in the toxic tort context. The GARP has designed and implemented a unique and effective judicial educational technology. The GARP has created a collaborative relationship with the Lawrence Livermore and Lawrence Berkeley National Laboratories and has mobilized the participation of 250 neutral, independent scientists, many of whom, by virtue of their participation, have increased their legal system knowledge, thereby building bridges across the deep institutional chasm separating science and the courts. The GARP has collected the most thorough genetics-related legal literature to date, covering case law and exact statutes in 20 legal categories. This collection was published in August, 1998 as an Adjudication Source Book in conjunction with a 22 judge, 16 scientist Working Conversation on Genes as Property in the American Law Tradition. GARP has published a primer for courts as a dedicated theme issue ("Genetics in the Courtroom") in the American Bar Association's main line, judicial division magazine, The Judge's Journal (Vol. 36., No. 3, 1997.) One article (Walsh, Admissibility) was selected as the best of the ABA for 1997. The GARP has produced a durable hypothetical case library and workbooks as initial archives for judges' use in early case assignment. These accomplishments provide a platform for ELSI/genetics-related Judicial System

leadership as the Human Genome Project's research moves toward establishing a reference map and sequence.

GARP objectives for the next three years include: (1) Conduct of a 1999 national neurogenetics in the courtroom conference; and California, Southeast, and Arizona/Southwest regional conferences on genetics in the courtroom for 500 additional federal and state, trial and appellate judges; (2) Conduct, in conjunction with Lawrence Livermore National Laboratory, of hands-on training for an expanded judicial faculty, dedicated leadership conference for the Chief Justices of our state courts, and a joint egislative/judicial conference; (3) Development of a model and procedural guidelines for the real time teleconferenced and videotaped testimony of court appointed expert witnesses in genetics-related cases; (4) an Alternative Dispute Resolution Techniques Guide Book with respect to ELSI/Genetics issues, a means for courts to manage genetics-related conflicts without subjecting the parties to the stress of formal trials; (5) Creation of a plan for international Judicial dialogues on gene testing and therapy issues; (6) Completion of a video ELSI-genetics curriculum for courts, to accompany the volume entitled "Genetics in the Courtroom: Judge's Handbook" to be printed for distribution in January, 1999; (7) Completion of a background paper for courts on the state of genetic property doctrines and issues; (8) Completion of a law school moot court initiative to orient future judges and law faculty to the management of ELSI-genetics related cases; (9) conduct of a working conversation on the legal status of genetic counselors; (10) institution of an ELSI/Genetics teaching courtroom in the District of Columbia as part of the EINSHAC's newly-established Law and Science Academy in the Courts of the District of Columbia; (11) development of a case tracking and GARP impact evaluation system.

These resources will promote judicial leadership on behalf of our society. Given the patchwork, inconsistent nature of state legislation concerning genetic property, privacy, and discrimination, the courts will be the bulwark of our ability to integrate new science with our framework of emerging as well as established rights. Given the paralysis of Congressional action, the Judicial Branch will be pushed front and center to interpret and adapt the existing law to rapidly spiraling new technological domains. It is timely and feasible to provide a durable and full toolbox for a judiciary that will be tested as never before at the margins of changing ethical precepts, social aspirations and high velocity science.

170. Measuring the Effects of a Unique Law Limiting Employee Medical Records to Job-Related Matters

Mark Rothstein
University of Houston Law Center, Houston, Texas
mrothstein@uh.edu

The grant attempted to measure the effects of a Minnesota law enacted in 1983, which prohibits employers from obtaining any employee medical information that is not strictly related to the ability to perform the job. The investigators reviewed the cases filed with the Minnesota Department of Human Rights, conducted interviews with Minnesota employment lawyers, surveyed occupational physicians and human resource managers, and assessed Minnesota economic data. The study concluded that the effects were not ascertainable because there was very little knowlege of the existence of the law by any of the groups. Nevertheless, the investigators determined that the Minnesota approach remains a more attractive alternative to preventing genetic discrimination in employment than genetic-specific laws recently enacted in 18 states. The findings are being published as: MA Rothstein, BD Gelb, & SG Craig, "Protecting Genetic Privacy by Permitting Employer Access Only to Job-Related Employee Medical Information: Analysis of a Unique Minnesota Law," *American Journal of Law and Medicine* 24 (1998): 399-417.

171. *TRUTH & JUSTICE*: Science and Its Appeals

Noel Schwerin
Backbone Media, 1327 Church St, San Francisco, CA 94114
schwerin@backbonemedia.org

TRUTH & JUSTICE is a three-part, three-hour documentary Special for national broadcast on PBS. Produced by Backbone Media (a public benefit, nonprofit charitable corporation*) in association with Oregon Public Broadcasting, *TRUTH & JUSTICE* will stimulate the public to think critically about the real strengths and important limits of science both in framing and in resolving social conflict. In three parts, the program will profile individuals - judges and scientists, lay people and lawyers - as they grapple with questions of science and law in a handful of actual cases. In Part One, Novel Cases will demonstrate how new technologies - particularly genetic technologies - create unexpected, unprecedented legal conflicts which challenge fundamental legal, ethical and social principles. In Part Two, Judging Science will look at what happens when new laws oblige the courts to distinguish between "good" and "bad" science. In Part Three, Due Process will examine science as one "way of knowing the world," often in conflict or competition with other ways of knowing in the courts.

In the style of *A Question of Genes*, the PI's recent award-winning, DOE-funded PBS program, each hour will closely observe two or three pairs of people 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. By profiling people at the center of actual conflicts, it will use compelling, accessible human drama as its vehicle.

Use of the DOE funds has focused on four goals; half has been spent on the first three goals, and the rest has been reserved for the fourth goal: (1) development of the conceptual framework and specific story content and treatment for the program; (2) development of a substantial body of institutional and individual support for the project, including an active and distinguished Board of Advisors, a community of academic and professional support and advice, an experienced production and promotion team, a distribution plan, and direct and ongoing relationships with story participants; (3) development toward other funding, including extensive foundation research and the submission of proposals to targeted foundations and federal agencies; (4) the production of a story from the program to aid in fund-raising and to launch the project.

*Formerly NoelEye Documentaries

172. *The DNA Files: Unraveling the Mysteries of Genetics*
A Nationally Syndicated Series of Radio Programs on the Social Implications of Human Genome Research and its Applications

Bari Scott and Jude Thilman
SoundVision Productions, 2991 Shattuck Ave., Ste. 304, Berkeley, CA 94705
bariscot@aol.com

The DNA Files is a series of nationally distributed public radio programs furthering public education on developments in genetic science. The series began broadcast on, at this writing, over 140 stations in November 1998. The producers anticipate an ultimate carriage of approximately 200 stations. *The DNA Files* is hosted by John Hockenberry and is distributed by National Public Radio. Program content is guided by a distinguished body of advisors and includes the voices of prominent genetic researchers, people affected by the clinical application of genetic medicine, members of the biotech industry, and others from related fields. They

provide real-life examples of the complex social and ethical issues associated with new discoveries in genetics. In addition to the general public radio audience, the series targets educators, scientists, and involved professionals. Ancillary materials are distributed in digital form through the project's web site, which also features ethical scenarios with which the visitor can interact. The site address is http://www.dnafiles.org. Tapes and transcripts are available on request by calling 303.823.3000.

With information linking major diseases such as breast cancer, colon cancer, and arteriosclerosis to genetic factors, new dangers in public perception emerge. Many people who hear about them could mistakenly conclude that these diseases can now be easily diagnosed and even cured. On the other end of the public perception spectrum, unfounded fears of extreme, and highly unlikely, consequences also appear. Will society now genetically engineer whole generations of people with "designer genes" offering more "desirable physical qualities"? *The DNA Files* will ground public understanding of these issues in reality. The programs in the series are documentary in format, featuring on location interviews, as well as radio theater and other techniques for conveying information about genetics and the social issues. An overview program, entitled "The Human Genome Project: Mapping the Future" consists of a hosted panel discussion with experts, with questions from the lay public interspersed into the discussion.

These nine programs describe the basic science of DNA, genes and heredity, while illustrating the accompanying social and ethical issues. "Law and the Genetics of Identity," for example, reviews the scientific methodology of genetic fingerprinting and explains the accuracy and use of this tool in criminal cases, as well as for establishing the identity of missing persons. "Gene Therapy: Medicine for Our Genes" follows the case of one man with mesothelioma, who is being treated with experimental gene therapy. His case provides a realistic illustration of the promise, as well as the current limits, of gene therapy. Other shows include "Genetics and Biotechnology: DNA in the Marketplace," "Prenatal Genetic Testing: Do You Really Want to Know Your Baby's Future?" and "The Genetics of Human Evolution: Where Did We Come From? Where Did We Go?"

Supported by ELSI grant DE-FG03-95ER62003 from the Office of Health and Environmental Research of the U.S. Department of Energy.

173. The Science and Issues of Human DNA Polymorphisms

David Micklos, John Kruper, Scott Bronson, and Matthew Christensen
DNA Learning Center, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724
micklos@cshl.org

The DOE Program introduces high school biology faculty to a laboratory-based unit on human DNA polymorphisms, which provides a uniquely personal perspective on the science and ELSI aspects of the Human Genome Project. Thus far, 170 high school faculty have participated in eight three-day workshops held in Atlanta, Boston, Denver, Los Angeles, New York, Richmond, Salt Lake City, and San Francisco. Participants learn simplified lab techniques for amplifying three types of DNA polymorphisms: an Alu insertion, a VNTR repeat, and point mutations (SNPs) in the mitochondrial (mt) control region. These polymorphisms illustrate the use of DNA variations in disease diagnosis, forensic biology, and identity testing - and provide a starting point for discussion of the uses and potential abuses of genetic technology.

During the past year, we developed ready-to-use teaching kits to support the three human PCR experiments introduced in the DOE workshop. The kits, which are now available through Carolina Biological Supply Company, incorporate a three-part PCR chemistry that greatly simplifies reaction set-up and improves reproducibility. Template DNA (obtained from cheek or hair cells) is added to primer mix and a freeze-dried reagent pellet (containing Taq polymerase, deoxynucleotide triphosphates, and buffer). Loading dye is incorporated in the primer

mix, allowing amplified reactions to be immediately loaded for gel electrophoresis.

We also provided proof of concept for a Sequencing Service to process mt DNA samples submitted by teachers around the country. During each DOE workshop, participants amplified the mt control region, and the samples were returned to the CSHL Genome Sequencing Center for cycle sequencing. The completed sequences were then posted at the DNALC's WWW site (http://vector.cshl.org) in a Student Sequence Database, which currently contains 350 sequences. This process was replicated with 59 student samples submitted, by mail, from teachers in New York, Maryland, Utah, and Virginia. A dedicated DNA sequencer to support student sequencing is included in the capital budget for a 6,000 square foot BioMedia addition to the DNALC, on which construction will begin next year.

We have continued to develop step-by-step Internet templates that allow students to use their own polymorphism data to solve cases illustrating key principles of genomic biology. In one case, the Student Allele Database facility is used to compare students' Alu insertion data with data from world populations. In another case, multiple sequence alignment is used to compare a student and reference mt DNA sequences to determine whether Neanderthal hominids were our direct ancestors. We are currently developing Bioforms that further simplify data submission/presentation - allowing students to focus on the biological question at hand, rather than being overwhelmed navigating complex forms at Internet genome servers. The initial Bioform allows students to analyze mt DNA sequences to identify the remains of the Romanov family and determine if Anna Anderson was the missing princess Anastasia.

174. Medical Confidentiality in the Market Driven Managed Care Setting: Does the Law Protect Against Misuse of DNA-Based Tests?

J.S. Kotval, Anthony Clarizia, and Patricia Salkin
The University at Albany, State University of New York, Albany, New York
jsk03@health.state.ny.us

The rise of managed care as the primary method of health care delivery raises new concerns regarding the misuse of confidential medical information -- especially the misuse of DNA-based tests that predict the likelihood of late-onset, high-cost disorders. We have surveyed state and federal law related to protection of medical confidentiality, and find several gaps in the legal framework intended to protect individuals from misuse of their genetic information in the managed care setting. The occurrence of these gaps stems primarily from reliance on a patchwork of state laws, regulations and case law to protect confidential medical information through professional licensure laws, medical malpractice laws and regulations, and by direct protection of an ill-defined entity: the medical record. These measures have become inadequate to protect medical confidentiality within the context of many managed care contracts. In addition, large numbers of persons insured through managed care plans that fall under ERISA are not touched by the state policy framework and no laws directly address the mis-use of genetic information within managed care plans that have the effect of discriminating against individuals with high cost illnesses. Furthermore, the absence of uniformity among states with regards to law and public policy creates confusion for both the managed care organizations that operate in multiple jurisdictions and for the consumers that seek care in more than one state.

In the absence of federal or state law to protect individuals from the unconsented dissemination of genetic information, there are few legal grounds for a

successful cause of action against a managed care organization. Legal theories that could serve as basis for successful causes of action include vicarious liability through respondeat superior and ostensible agency theories. In assessing the merits of a plaintiff's case claiming discrimination, the courts are likely to consider issues including the structure, administration and internal procedures implemented by the managed care organization; the representations and disclosures made in contracts, brochures, and advertising; the financial incentives, cost containment procedures and utilization review procedures used by the managed care organization and, of course the selection, control and contracts between the managed care organization and the physician. Even if the plaintiff does have 'the law on his side' it would take him or her years to resolve the matter, not to mention the expenses involved, thus creating a disincentive to seek redress and creating a hostile environment for advocating consumer rights. Recommendations regarding appropriate federal and state law are being explored.

175. *Geneletter*: An Internet Newsletter on Ethical, Legal, and Social Issues in Genetics

Dorothy C. Wertz and Philip R. Reilly
The Shriver Center, 200 Trapelo Road, Waltham, MA 02452
dwertz@shriver.org

Geneletter (http://www.geneletter.org) is among the few informative and readable sites on ELSI topics aimed at the nonspecialist. It has received over 1 million hits and 200,000 site visits since September 1996. Currently it has about 450 users (2500 hits) a day, with an average session length of 9 minutes. Judging from our emails, many users are students, ranging from sixth grade to postgraduate, but some are college deans, science fiction writers, state policymakers, epidemiologists, and law professors. At least 14% are international, in Canada, Australia, UK, Germany, Singapore, Sweden, France, Japan, Malaysia, New Zealand, Netherlands, Italy, Norway, Israel, Philippines, and Brazil. In North America, major user areas are VA, CA, OH, ONT, WA, MA, FL, PA, IL, and MD. Topics receiving greatest visitor attention are cloning (by far the most popular), genetics of homosexuality, genetic "Adam and Eve," overview of genetic screening, the calico cat, false paternity, teratogens, and eugenics. *Geneletter* has also been used in a course on Genetics and Identity at a four-year technical college, with evaluation of ethical views before and after use.

176. Competition Between Public & Private Research Funding in Genomics

Rebecca S. Eisenberg
University of Michigan Law School, Hutchins Hall, 625 S. State St., Ann Arbor, MI 48109
rse@umich.edu

The field of genomics exhibits considerable overlap in the type of research that is supported by public and private funding. Recent announcements from two private firms that they plan to complete the DNA sequence of the human genome ahead of the publicly funded HGP is the most recent, and perhaps the most dramatic, example of this phenomenon. Sometimes public and private institutions commingle funds and work together collaboratively, but sometimes they are openly competitive. These interactions shed an interesting light on the relationship between public and private funding of scientific research. A number of features of public-private interactions that have been highly salient in this particular setting play little role in standard accounts of the relationship between public and private research funding, including scientific rivalry for priority in achieving overlapping if not identical goals, a tortoise-hare struggle over the relative virtues of speed and thoroughness, and recurring controversies over intellectual property and terms of access to data and discoveries.

A standard account of the relationship between public and private research pictures publicly-funded research as promoting activities that are entirely distinct from the sort of research that private firms are likely to pursue with their own funds. In his classic argument for continued government funding of research following the conclusion of World War II, Vanevar Bush called for government funding of

"basic research" to compensate for inadequate commercial incentives to invest in the pursuit of fundamental knowledge, as distinguished from "applied technology." Economists have similarly suggested that government research funding should compensate for "market failure" that limits private motivation to invest in research, despite the high social value of such investments, because of uncertainty as to who will profit from research results and difficulty in appropriating results as intellectual property. An alternative justification for public funding of research that finds little support among economists but nonetheless enjoys considerable political popularity emphasizes promoting technological innovation by U.S. firms. In this account government-funded research is pictured as an advanced scouting mission to identify promising opportunity for short-sighted or risk-averse firms. None of these accounts contemplates public funding of research that is competitive with private sector research.

When does it make sense to allocate taxpayer dollars to funding research that that resembles work being pursued in the private sector? Are there circumstances in which public funding of research may be justified as a means of forestalling private appropriation of research results as intellectual property? When are judgments about the wisdom of patenting certain types of discoveries best left to the patent system in its determinations of what may be patented, and when are such judgments appropriately made by funding agencies in deciding what sort of research to fund and in limiting the rights of grantees to pursue patent rights? What are the proper roles of the patent system and research funding agencies in mediating the boundaries between public and private in research science, and what sorts of judgments on this issue are within the competence of the institutions that manage these systems?

177. Microbial Literacy Collaborative: *Intimate Strangers: Unseen Life on Earth*

Cynthia A. Needham and Susan E. Kee
American Society for Microbiology, Washington, District of Columbia
skee@asmusa.org

A Report from the Microbial Literacy Collaborative -- DE-FG02-97ER62339
Capturing the Imagination to Capture the Mind: Using the Power of Informal Learning to Advance Science Literacy.

The Microbial Literacy Collaborative (MLC), a partnership of organizations committed to advancing scientific literacy through a focus on the microbial world, will report on four components of the initiative: (1) the science documentary for public television, entitled *Intimate Strangers: Unseen Life on Earth* (2) the set of 17 hands on community based microbial activities entitled *Microbe Mania*, and (3) youth leadership training for pre-college students from traditionally under-represented communities and (4) a 12 part telecourse for undergraduate use. The organizations that comprise the MLC include The American Society for Microbiology, The National Association of Biology Teachers, Oregon Public Broadcasting, and Baker & Simon, Associates, an independent production company. Other organizations include The Association of Science-Technology Centers, Inc. and The American Association for the Advancement of Science.

The MLC is funded by The Department of Energy through the Human Genome Project, The National Science Foundation, The American Society for Microbiology, the Annenberg/CPB Project, The Corporation for Public Broadcasting, and The Archer Vining Davis Foundations.

Intimate Strangers: Unseen Life on Earth has been completed and is expected to air on PBS in the fall of 1999. The four hours of the series include: (1) "The

Tree of Life," (2) "Dangerous Friends and Friendly Enemies," (3) "Keepers of the Biosphere and (4) "Creators of the Future." "The Tree of Life" delves into our evolutionary past. The key message of this hour is that all living things today evolved from microbes and share fundamental biologic properties with them. "Dangerous Friends and Friendly Enemies" examines our ancient rivalry with the microbial world. "Keepers of the Biosphere" explores the central role that microbes play in sustaining the earth's ecosystems. "Creators of the Future" examines our present and future use of microbial technologies to solve long standing problems that affect the way we live. We will have one hour of the series, "The Tree of Life," available for viewing at the conference.

Microbe Mania is a collection of 17 hands on activities designed for use in both informal and formal learning environments. The activities complement the major themes within the television documentary. They require little or no knowledge of microbiology and little or no specialized equipment to conduct. The activities will support open ended experimental design and help to address elements of the National Science Education Standards. We will have an example of the activities available for demonstration.

Microbe Mania Youth Leadership Training is a week long experience 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. We will report on the first of two summits, which was held in August 1998, on the St. Paul campus of the University of Minnesota. The training experience was organized with The Association of Science-Technology Centers, Inc. and their Youth Alive! Program. Participants represented 12 science museums from around the country, with youth leaders drawn primarily from challenged home environments.

The Unseen Universe is a 12 part telecourse for use in both undergraduate and pre-college classrooms. Each 30 minute film focuses on a different aspect of the microbial world. The telecourse was designed to address the curriculum standards endorsed by the American Society for Microbiology and will be accompanied by teacher and student guides. The telecourse will support a full distance learning course in microbiology or serve as supporting materials for traditional classroom environments. We will have a 30 minute segment of the telecourse available for viewing.

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

Barbara Handelin and Susan Katz
Public Responsibility in Medicine and Research (PRIM&R), Boston, Massachusetts
bhandelin@compuserve.com

IRBs are under extreme stress to provide adequate review of all manner of protocols. A central assumption that underlies the IRB's charge to protect the rights and welfare of human subjects involved in research, is that each individual IRB will possess or develop the requisite expertise to accomplish this mission adequately. The increasingly complex ethical, regulatory and scientific issues presented to IRBs in reviewing genetic research protocols challenge the validity of this assumption. Individual IRBs have inadequate time and resources to develop the necessary genetics expertise and facility to deal with this new challenge. Thus, our project has been developed to solicit specific needs from IRBs so that specific working "tools" can be created to address those needs. We will report on our progress toward that end. But that is not all....as we are also addressing the concomitant increased pressure on biotech and genomics companies to conform to a standard of practice in conducting research studies and in developing marketing plans for gene based products and services. As such companies become engaged in clinical studies involving human subjects or tissues it has become apparent that they may need help in effecting quality, IRB-like review. In this project we are seeking to exploit the synergistic needs and expertise found in these two types of organizations: the ethics oversight capabilities and systems of IRBs and the genome expertise in

industrial R&D shops. We will report on the dynamic interplay and relative perspectives that IRBs and the biotech communities have of one another and how we are proposing to weave common solutions to critical issues in the safe and ethical participation of human subjects in genetic research.

179. Your World/Our World - Exploring the Human Genome

Jeff Alan Davidson
Alliance for Science Education
73150.1623@Compuserve.com

The Pennsylvania Biotechnology Association (PBA) in cooperation with the Alliance for Science Education (ASE) publishes the biotechnology science magazine YOUR WORLD/OUR WORLD to introduce middle and high school students to the underlying science and the social issues raised by modern biological research and technology. In the Spring of 1996, with partial DOE funding, a special enlarged issue of YOUR WORLD/OUR WORLD dealing with the underlying science of genomics, and the ethical, legal, and social issues raised by the Human Genome Project (HGP) was published.

PBA and ASE are developing additional instructional materials for use by middle and high school students to facilitate a more extensive presentation of the subjects covered in the special issue. These materials are being built in two phases. First, by developing new materials by PBA to create a comprehensive supplemental materials package that provides resources in several different media. Second, by running a national contest for science teachers and students to encourage classroom development of new and original approaches to teaching the material. Materials from both phases will each in turn be packaged and made available to the 45,000 middle school and high school biology teachers in the United States over the next 24 months.

Phase I Materials

Eight multimedia lectures designed for teacher use and richly annotated with teacher notes have been developed that can be presented directly from a CD-ROM or from color or laser printed overheads that can be printed from the disk. The lectures make extensive use of three-dimensional animations to explain the science clearly and interestingly. The materials also include an extensive glossary and directory of internet resources.

The lecture topics are summarized below:

- Inheritance and Uniqueness - The Role of DNA
- DNA Structure and Function, DNA Replication (Highlights from Animations and Your World), DNA Transcription and Translation (Highlights from Animations and Your World)
- The Human Genome Program and Gene Mapping (Animation & Lecture)
- Genetic Testing
- Molecular Evolution
- A Glimpse Into the Future of Gene Discovery and Application
- Careers in Fields Related to Gene Discovery and Research
- Introducing a Gene Teaching Material Development Contest to the Students

The three dimensional animations Featured on the CD - ROM include:

- DNA Structure and DNA Replication
- DNA Transcription and Translation - An Overview
- DNA Transcription and Translation - A More Detailed Exploration
- Searching an Online Database for a Matching Gene
- Three Levels of Maps - (Genetic Linkage Map, Physical Map, Sequence Map)

Phase II Materials are expected to include:

- Additional Teaching Plans
- Graphics, Art, Photographs
- Additional Science and ELSI Text & Animations

- Additional Experiments, Demonstrations, & Activities
- Multimedia Presentations and Graphics, Web Sites, Games
- Role Playing Exercises
- Science Fair or Research Project Suggestions
- Plays or Literature - written or performed and videotaped
- Articles for the General Press or Radio or Television

180. Human Genome Teacher Networking Project

Debra L. Collins
University of Kansas Medical Center, Kansas City, Kansas
dcollins@kumc.edu

Over the past few years, the Human Genome Project has increased our knowledge about human genetics dramatically. However, it is difficult to keep up with all the new technological advances. Secondary science and biology teachers have difficulty determining which new genetic advances need to be incorporated into their curriculum. As well, families, health care providers, and the general public need accurate human genetics information, and a background information to help them interprete all the new information in newspapers, television, and other media.

To help bridge the gap between the general public's background genetic knowledge and new genome technological advances, we designed a national education program for secondary science and biology teachers. Since approximately 95% of high school students take a biology class, these teachers' classrooms may provide the last formal science course before their students become parents, voters, legislators, policy makers, journalists, or others needing accurate genome information.

Over 5 years, 177 secondary teachers attended a series of human genome workshops focused on applications of human genome project technology, including ethical, legal, and social implications (ELSI). The project required a two year committment of each teacher, who attended summer workshops, used new materials with students, conducted peer and community education programs, and contacted genetic and ELSI experts to enhance classroom experiences. Ongoing networks between teachers, liaisons with genetic professionals, and on-line computer communications continue to help educators and their students obtain current genome information.

Teachers participating in the project became more prepared and confident to teach complex genome technology and applications than their peers ($p<.05$). They expanded their knowledge of human genetics, and integrated more information into existing science curricula, increasing time devoted to teaching genetics. Teachers developed new school genetic courses, and advised district curricula development committees to increase human genetics course content.

Participants became better prepared to help students understand the ramifications of HGP discoveries and readily access information on many aspects of the Human Genome Project, including decisions regarding genetic testing. Their students scored significantly higher ($p<.05$) on a survey of knowledge, compared to students whose teachers did not attend the workshop.

Participants presented genetic programs to over 10,000 peers. Through this dissemination, more than 1,000,000 students were exposed to new genome information, resources, and applications.

Continued and increased support for teacher education workshops is needed to increase literacy on human genetic topics not available in current published textbooks.

Genetic resource materials, lesson plans, the mentor network, and career information are on the web site: http://www.kumc.edu/gec
[DOE #DE-FG02-92ER61392]

181. Electronic Scholarly Publishing: Foundations of Genetics

Robert J. Robbins
Fred Hutchinson Cancer Research Center, Seattle, Washington
rrobbins@fhcrc.org

As the Human Genome Project (HGP) moves toward its successful completion, more and more people are becoming interested in understanding this project and its results. Since the HGP has significant ethical, legal, and social implications for all citizens, the number of individuals who do, or should wish to become familiar with the project is very high. 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 the results of HGP research, however, requires a familiarity with the notions of basic genetics, and this is often not available to most individuals. We have created an 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. Together, this collection should be of great use to those wishing to appreciate and understand genetics and genome research.

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.

The site is intended to be useful not only to students, teachers, scholars, but also to general readers. Indeed, we consider the general public to be our primary audience. Data currently available suggest that we are succeeding in reaching our target audience. The bulk of our users are accessing the site from clients that use a dial-up Internet service provider. Since scholars and scientists usually have full Internet access from their university facilities (that all have *.EDU domains), the data suggest that the bulk of our users are from the general public. We specifically have logged visits from more than 100 high-school sites and we know of several high-school web projects that have established links to the ESP 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. By January 1999, we will move our site to a different physical server that will allow us much more control over the functionality that we can deliver. This will allow us to offer custom services to the user, including personalized searches and file storage, as well as custom annotated versions of classic texts.

182. The Community College Initiative

Sylvia J. Spengler and Laurel Egenberger
Life Sciences Division, Center for Science and Engineering Education, Ernest Orlando Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720
sjspengler@lbl.gov

The Community College Initiative prepared community college students for careers in biotechnology. Lawrence Berkeley National Laboratory (LBNL) collaborated with California

Community Colleges in developing mechanisms that encourage students to pursue science studies, to participate in forefront laboratory research, and to gain work experience. The initiative was structured to upgrade the skills of students and their instructors through Summer Student Workshops.

The Summer Student Workshops provided a four-week summer residential program for students who had completed the first year of the biotechnology academic program. Ethical, legal and social concerns were integrated into the laboratory exercises. Students learned to identify commonly shared values of the scientific community as well as increase their understanding of issues of personal and public concern. In the three-year period of the grant, we involved over thirty-five students. Students in the second and third summers were awarded laboratory internships.

183. Genes, Environment, and Human Behavior

Michael J. Dougherty and Joseph D. McInerney
Biological Sciences Curriculum Study (BSCS), 5415 Mark Dabling Blvd., Colorado Springs, Colorado 80918-3842
jdougherty@bscs.org and jmcinerney@bscs.org

The Biological Sciences Curriculum Study (BSCS) is developing an instructional module titled Genes, Environment, and Human Behavior for use in high school biology classes. The module will rectify the deficient treatment of the biology of behavior in the current curriculum and will help to dispel misconceptions about genes and human behavior that often pervade media reports of research in this area. The materials also will address some of the ethical, legal, and social issues generated by research into the biological basis of behavior and will help to change traditional assumptions about the teaching of genetics at the high school level. The draft instructional activities are designed to help students move through the following major concepts:

1. variation in behavior exists in populations;
2. there are complex genetic and environmental components to this behavioral variation;
3. scientists have methods for investigating the source of differences in human behavior;
4. these methods have strengths and limitations; and
5. there are ethical, social, and legal implications to understanding that genes influence behavior.

The project employs the process of curriculum development that BSCS has refined continually since the inception of the organization in 1958. That process involves advisory meetings, writing conferences, pilot and field testing of draft materials, and periodic reviews of progress by members of the education committees of the National Society of Genetic Counselors, the American Society of Human Genetics, the Council of Regional Networks for Genetic Services, and other independent experts in genetics. The project also is drawing upon the experience BSCS acquired during the development, distribution, and implementation of three genome-related instructional modules between 1991 and 1996.

In May of 1998, BSCS field tested a complete draft of the module with 20 teachers and over 1200 students in 13 states. Students completed pre- and post-tests to determine common misconceptions about behavioral genetics and to assess changes in student learning after using the module. Statistical analysis of the data showed that students improved significantly on every learning outcome. In addition, mean pre-test scores of 56% correct improved to 71% correct on the post-tests ($p < 0.001$).

After analyzing the field-test data, BSCS hosted the second meeting of the project's advisory committee, which made a number of recommendations for improving the module. In late July 1998, BSCS completed its second and final writing conference, during which experts in behavioral and medical genetics, ethics, and high school biology teaching produced drafts of a revised module based on the recommendations of the advisory committee. BSCS staff are currently revising those drafts to produce final materials, which will comprise student activities, support materials for the teacher, and extensive background information for teachers. Following production and printing, BSCS will

distribute the module free of charge to 20,000 interested biology teachers.

184. Hispanic Role Model and Science Education Outreach Project: Human Genome Project Education & Outreach Component

Clay Dillingham
Institute of Genetics Education, 1611 Don Gaspar Ave. Santa Fe, NM 87505; the Self-Reliance Foundation; and the Hispanic Radio Network, 518 C St. NE, Washington, DC 20002
cd@ncgr.org

Currently, Hispanics make up between 11-15 percent of the U.S. population; about 30 million people. Hispanics are also the fastest growing minority in the U.S. The dominant language of 64 percent of all Hispanics living in the United States is Spanish. Furthermore, Hispanics feel discriminated against by the U.S. health care system. According to the American Journal of Health Promotion (Vol.9, No. 4 1994/95):

- 78.3 % of Hispanics report they are most comfortable speaking Spanish at home
- 27% believe they face discrimination in the quality of health care to which they have access
- 30% believe they are not treated with respect at clinics
- 28% believe they do not have the same opportunities as others in obtaining health care information

Hispanics are Outside the Mainstream of Information about Health, Science, the Human Genome Project (HGP), and its Ethical, Legal, and Social Implications (ELSI)

Hispanics are largely "out of the loop" of the public health information mainstream because of the substantial linguistic preferences of Hispanic residents. For example, disseminating information related to the rapid advances in health and science technology, like the HGP-ELSI, and information about how and where to access health care services, may not be targeted to or successfully reach Spanish speaking residents.

This DOE-funded project is currently completing its second year of providing Spanish radio programming and outreach services that focus on the HGP and its scientific, medical ELSI implications. The purpose of this project is to help inform the Spanish-speaking population in the U.S. about the HGP and its ELSI implications, and motivate them to access the resources available for further education and information on these issues.

185. The Hispanic Educational Genome Project

Margaret C. Jefferson, Mary Ann Sesma, and Patricia Ordonez
California State University, Los Angeles, California
mjeffer@flytrap.calstatela.edu

The primary objectives of this grant were to develop, implement, and distribute culturally competent, linguistically appropriate, and relevant curricula that lead to Hispanic student and family interactions regarding the science, ethical, legal, and social issues of the Human Genome Project.

Two curricula were developed: (a) that designed for students and (b) that designed for parents. The student component consists of available materials (e.g., the BSCS HGP-ELSI curricula; laboratory projects; University of Washington High School Human Genome Program exercises; Virtual FlyLab; and more) and newly developed materials (e.g., teacher-developed activities in four major units of biology; student-developed surveys; and more). The parent component consists of newsletters written by students available in both English and Spanish and

parent focus groups which discuss issues related to genetics and health. Discussions have been in both English and Spanish with translators available. Information on materials are available via our WWW home page:

http://vflylab.calstatela.edu/hgp/

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under grant # DE-FG03-94-ER61797.

186. The High School Human Genome Program

Maureen Munn and Leroy Hood
Department of Molecular Biotechnology, University of Washington, Box 352145, Seattle, WA 98195
mmunn@u.washington.edu

The discovery that DNA is the information molecule of living organisms is one of the most significant scientific breakthroughs of the 20th Century and is critical to our understanding of inheritance, development, biodiversity and evolution. Advances in genetics, molecular biology and biotechnology have revolutionized biological research, medicine, agriculture and manufacturing, and will continue to do so in the 21st Century. Along with many benefits, genetic research and biotechnology evoke complex ethical and legal issues that impact individuals and society as a whole. Pre-college genetics education should do the following:

- present genetics as a unifying concept of biology;
- provide knowledge and experience with experimental approaches used in genetics and biotechnology;
- inform students about career choices in genetics and related fields;
- challenge students to consider related ethical issues so that they can develop the critical thinking skills needed to understand and evaluate them.

The High School Human Genome Program (HSHGP) encourages high school students to think constructively about the scientific and ethical issues of genetic research by enabling them to participate in both. This program was developed through a partnership of scientists at the University of Washington (UW) and several Seattle area high school teachers. Students sequence a portion of human DNA as part of an ongoing research project being carried out at the University of Washington. Our current project is focused on understanding the molecular basis for nicotine addiction by sequencing one subunit of the nicotine-binding receptor in the brain. This project is being carried out in collaboration with Dr. Carl Ton, an acting assistant professor in the Department of Medicine, Division of Medical Genetics at the University of Washington. Sequencing is carried out either in classrooms, using a manual technique, or at the UW Genome Center, using fluorescent sequencing. Each student group sequences a small portion of the overall segment of DNA being studied, and then their data is merged using the DNA assembly program, Sequencher. Students and their teachers can access Sequencher through a tutorial called "virtual DNA sequencing" that is available on our web site (http://hshgp.genome.washington.edu).

The Ethics unit focuses on presymptomatic genetic testing. This module was developed by Sharon Durfy and Robert Hansen from the UW Department of Medical History and Ethics. The module utilizes a role-playing scenario to involve students in the complex issue of whether, as a character in the scenario, they would choose to be tested presymptomatically for Huntington's disease (HD). Materials provided include background information on the genetics, molecular biology and clinical aspects of HD, directions for constructing a pedigree and analyzing the laboratory data used to determine whether someone carries the HD gene, a tool for assessing student learning, and a teacher's guide. Students use a decision making model to assist them in making a justifiable ethical decision.

Our program offers professional development for teachers from Washington state and other locations in the US and Canada during a one-week summer workshop. Activities include completion of the DNA sequencing and ethics modules, presentations by guest speakers, and informal discussions about classroom implementation and student assessment. During the academic year, local teachers are provided

with the necessary equipment and reagents to carry out the experiment in their classrooms. Teachers from outside the Seattle area can borrow equipment through the loan program of the Howard Hughes Program at Washington State University (WSU), while teachers in the Vancouver area are supported by a partner site at WSU Vancouver. Scientist volunteers from UW and local biotechnology companies assist during classroom experiments. This program is currently serving over 50 high school and college teachers in Washington State, as well as 20 teachers outside the state.

Molecular Biotechnology's Education Outreach contributes to K through 12 science education through a variety of outreach efforts. These outreach programs share several important features, including a strong emphasis on presenting science as inquiry and the development of partnerships between teachers and scientists. Genetics is an integral part of many of these outreach programs. For example, in conjunction with the Seattle Partnership for Inquiry-Based Science, Education Outreach presented one-week workshops on genetics and biodiversity for Seattle elementary teachers in the summer of 1998. The Integrated Science Partners, a Howard Hughes-funded program focused on the development of curriculum for middle school science teaching, has developed a module on genetics. We are currently coordinating a project called the Genetics Education Partnership, in conjunction with teachers and genetics professionals from around the state. The purpose of this project is to examine genetics teaching in grades K through 12, identify useful materials for teaching genetics at different grade levels and foster the development of genetics learning communities throughout the state.

Recent Publications

Munn, M. M., O'Neill Skinner, P., Conn, L., Horzma, G. and Gregory, P. "The Involvement of Genome Researchers in High School Science Education". Review submitted to Genome Research.

Internet-based Publications and Projects:
Genetics Education Database (http://genetics-education.mbt.washington.edu/database)
Web-site for the High School Human Genome Program (http://hshgp.genome.washington.edu)
Web-site for the Genetics Education Partnership (http://genetics-education-partnership.mbt.washington.edu)

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

Sara L. Tobin and Ann Boughton
Stanford University, Stanford, California, and Thumbnail 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 individuals do not understand the limitations or the promise of the current explosion in knowledge of the human genome. This is also true of physicians, most of whom completed their medical training prior to the application of recombinant DNA technology to medical diagnosis and treatment. This lack of training prevents physicians from appreciating many of the recent advances in molecular genetics and may delay their acceptance of new treatment regimens. In particular, physicians practicing in rural communities are often limited in their access to resources that would bring them into the mainstream of current molecular developments. This project is designed to fill two important functions: first, to provide solid training for physicians in the field of 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 engaged in the development of a flexible, user-friendly, interactive multimedia CD-ROM designed for continuing education of physicians in applications of molecular medical genetics. We have the designed the navigational system, completed a prototype, carried out a preliminary evaluation of the prototype by physicians, and continued to create content. 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 the human genome project, technical advances, and disease diagnosis and prognosis; and (4) Societal implications, focusing on issues such as privacy and impact on the family. The CD format permits the use of animation, video, and audio, in addition to graphic illustrations and photographs. A hypertext glossary, user notes, practice tests, and customized settings will be utilized to tailor the CD to the needs of the user. Brief, multiple-choice examinations will be evaluated for continuing medical education credits by the Stanford Office of Postgraduate Medical Education. The CD will function as a 'hybrid' product, capable of seamless interaction with Internet resources. This capability permits continuous updating of the course content.

The development of the CD is supervised by a Board of Advisors, and the completed courseware will be evaluated by physician focus groups. Commercial distribution will be arranged through the Stanford Office of Technology Licensing. The courseware is designed to provide a powerful tool for the education of physicians and the public about the potential of the Human Genome Project to benefit human health.

188. Individualizing Medicine Through Genomics: Medical and Social Implications

Henry T. Greely, Barbara A. Koenig, and Laura L. McConnell
Stanford Center for Biomedical Ethics, Stanford Law School, Stanford, CA 94305-8610
hgreely@stanford.edu

This grant partially supported a process that led to a conference at Stanford on October 17, 1998 on the implications of the increasing use of genetic variation in medicine. Scientists, physicians, and industry increasingly are recognizing the potential medical importance of such variation. The genetic variation involved can be that of a pathogen, a tumor, or healthy human tissue; the medical implications may be in prevention, treatment, or prognosis. In all these respects, the potential for applying individualized genomic information to medicine is an extension of existing knowledge based on the growing availability of inexpensive and convenient determination of what DNA sequence in known genes. Among the specific topics examined were

- Ethical issues in the research necessary for greater medical use of genetic variation, such as research correlating vast amounts of phenotypic data to vast amounts of genotypic data
- The consequences of individual genomic variation for public health initiatives.
- The possible uses of individualized genetic information by managed care to make treatment decisions.
- The effects of greatly expanded use of individual genetic data in medicine on privacy.
- The implications of studies of medically relevant genetic variation for public and medical views of "race."
- Problems raised by the commercialization of treatments involving genetic variation, including intellectual property and FDA issues.

The conference, which was videotaped, included presentations from, among others, Drs. Anthony Carrano, Francis Collins, Paul Berg, Barbara Koenig, Laurie Zoloff-Dorfman, and Robert

Cook-Deegan and Law Professors Rebecca Eisenberg, Alta Charo, and Henry Greely. At least one publication, a summary article, will be forthcoming from the conference; a broader set of articles may also result.

189. AAAS Congressional Fellowship Program

Elaine Strass
The American Society of Human Genetics, Bethesda, MD 20814-3998
society@genetics.faseb.org

Few individuals in the genetics community are conversant with federal mechanisms for developing and implementing policy on human genetics research. In 1995 the American Society of Human Genetics (ASHG), in conjunction with DOE, initiated an American Association for the Advancement of Science (AAAS) Congressional Fellowship Program to strengthen the dialogue between the professional genetics community and federal policymakers. The fellowship will allow genetics professionals to spend a year as special legislative assistants on the staff of members of Congress or on congressional committees. Directed toward productive scientists, the program is intended to attract independent investigators.

In addition to educating the scientific community about the public policy process, the fellowship is expected to demonstrate the value of science-government interactions and make practical contributions to the effective use of scientific and technical knowledge in government. The program includes an orientation to legislative and executive operations and a year-long weekly seminar on issues involving science and public policy.

Unlike similar government programs, this fellowship is aimed primarily at scientists outside government. It emphasizes policy-oriented public service rather than observational learning and designates its fellows as free agents rather than representatives of their sponsoring societies.

One of the goals of DOE and ASHG is to develop a group of nongovernmental professionals who will be equipped to deal with issues concerning human genetics policy development and implementation, particularly in the current environment of health-care reform and managed care. Graduates of this program will serve as a resource for consultation in the development of public-health policy concerning genetic disease.

Fellowship candidates must demonstrate exceptional basic understanding of and competence in human genetics; hold an earned degree in genetics, biology, life sciences, or a similar field; have a well-grounded and appropriately documented scientific and technical background; have a broad professional background in the practice of human genetics as demonstrated by national or international reputation; be cognizant of related nonscientific matters that impact on human genetics; exhibit sensitivity toward political and social issues; have a strong interest and some experience in applying personal knowledge toward the solution of social problems; be a member of ASHG; be articulate, literate, adaptable, and interested in working on long-range public policy problems; be able to work with a variety of people of diverse professional backgrounds; and function well during periods of intense pressure.

DOE Grant No. DE-FG02-95ER61974.

Infrastructure

190. 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 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 Barbara Dorsey at Science and Engineering Education Programs, ORISE, MS 36, P.O. Box 117, Oak Ridge, TN 37831-0117 (423) 576-9975; Fax (423) 241-5220.

191. 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 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; (423) 576-6669; Fax: (423) 574-9888
bkq@ornl.gov
General: http://www.ornl.gov/hgmis
Research: http://www.ornl.gov/hgmis/research.html

The Human Genome Management Information System (HGMIS), begun in 1989, helps the Task Group of the DOE Human Genome Program (HGP) fulfill its commitment to informing scientists, policymakers, and the public about the program's goals, funded research, and applications. HGMIS products, including the Web sites and a newsletter, have won technical and electronic communication awards and have been reviewed and featured in well-known publications.

The HGP requires contributions from many disciplines to accomplish its goals and to make sure its outcomes are used to their greatest beneficial potential. Through its scientific communication role, HGMIS seeks to (1) help foster such collaborations and (2) make HGP science, resources, and societal

implications accessible to nongenome researchers who are using these new tools and data to solve basic research problems traditional to their fields. Additional targeted groups are medical and legal personnel; bioethicists; educators and other professionals who are being impacted by genomics; and the public.

Through its communication of scientific and societal issues to nonresearch audiences, HGMIS seeks to increase public literacy in genetics, thus laying a foundation for more informed personal decision making and policy development. The hope is to maximize HGP benefits while simultaneously protecting against misuse of personal genetic information. To meet the coming flood of court cases involving genetic evidence, since 1995 HGMIS has been participating in a project to educate judges on the basics of genetics and gene testing. Recently, HGMIS has established genome Web pages for the medical community to help them prepare for the new era of molecular medicine.

Print and Electronic Information Resources

Publications. A forum for the wide exchange of knowledge, *Human Genome News (HGN)* uniquely presents a spectrum of genome-related topics not found in any other single resource. More than 75% of *HGN*'s diverse body of 15,000 domestic and foreign subscribers are non-HGP scientists who would not find this information in their discipline-specific publications. HGMIS also produces the DOE *Primer on Molecular Genetics*, progress reports on the DOE Human Genome Program, contractor-grantee workshop proceedings, one-page topical handouts, and other related resource material.

Document Distribution. In addition to *HGN*, HGMIS has distributed more than 175,000 copies of publications requested by subscribers, meeting attendees, and managers of genetics meetings and educational events. About 120 such requests are processed each month.

Electronic Communication. Since November 1994, HGMIS has produced a comprehensive, text-based Web server called "Human Genome Project Information." Through its newly created "Research in Progress" site, the HGMIS server is devoted to topics relating to the science and societal issues surrounding the genome project. The HGMIS Web sites contain more than 1800 text files that are accessed about 3 million times a year. Each month, around 15,000 host computer domains connect to the HGMIS server directly or through more than 2400 other Web sites. HGMIS also maintains Web pages on the human and microbial genome programs, bacterial artificial chromosomes, cDNA full-length sequencing, and genomics meetings for DOE as well as the Genetics section of the Virtual Library from CERN in Switzerland. HGMIS moderates the BioSci Human Genome Newsgroup.

Direct Information Source. Staff members answer individual questions and supply other information about genetics and the Human Genome Project. Around a hundred such queries are received each month via the Website, fax, and telephone. HGMIS reaches diverse scientific and educational groups when the DOE Human Genome Project traveling exhibit and posters are displayed at conferences and public meetings and when staff members make presentations to educational, judicial, medical, and other groups. As more people become aware of the HGP's impact, HGMIS is striving to strengthen the content relevancy of its services to meet the growing and varied demands for information. Comments and suggestions 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.

192. Human Genome Program Coordination Activities

Sylvia J. Spengler and Janice L. Mann
Human Genome Program, Mailstop 84-171, Ernest Orlando Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720
sjspengler@lbl.gov

The DOE Human 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 is now the Biotechnology Forum. The Forum is chaired by Dr. A Patrinos, Associate Director, OBER. Members of the Human Genome Program Management Task group and of the Biological and Environmental Research Advisory Committee's subcommittee on the Human Genome are ex-officio members of the Biotechnology Forum. Responsibilities of the Forum include assisting OBER in overall coordination of DOE-funded genome research; facilitating the development and dissemination of novel genome technologies; recommending establishment of ad hoc task groups in specific areas, such as informatics, technologies, and model organisms; and evaluation of progress and consideration of long-term goals. Members also serve on the Joint DOE-NIH subcommittee on the Human Genome, for interagency coordination. The coordination group also participates in interface programs with other facilities and provides scientific support for development of other OBER goals, as requested.

This work was supported by the Director, Office of Energy Research, Office of Biological and Environmental Research, Human Genome Program, of the US Department of Energy under Contract No. DE-AC03-76SF00098.

193. The JASON Study of the Human Genome Project

Gerry Joyce, S. Block, J. Cornwall, F. Dyson, S. Koonin, N. Lewis, and R. Schwitters
The Scripps Research Institute, La Jolla, California
gjoyce@scripps.edu

In 1997, the JASON organization conducted a DOE-sponsored study of the human genome project with special emphasis in the areas of sequencing technology, quality assurance and quality control, and informatics. A summary of this report was published in Science magazine (Science 1998 January 2; 279: 36-37). In 1998, the study was continued and expanded to include a consideration of what role DOE might play in the "post genomic" era, following the acquisition of the complete human genome sequence.

The JASON study recommended that DOE should (1) help ensure the development of a full length cDNA clone resource, (2) expand its efforts in comparative genomic sequencing of model organisms, (3) work to establish community-wide standards for software operation and the quality of data entered into public databases, including the development and operation of functional genomics databases, and (4) foster progress in genome-wide technologies for functional genomics.

Appendices

Appendix A: Author Index

Presenting authors are in **bold**.

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z

Appendix B: National Laboratory Index

U.S. Department of Energy Laboratories

Human Genome Program work at the national laboratories is described in the following abstracts.