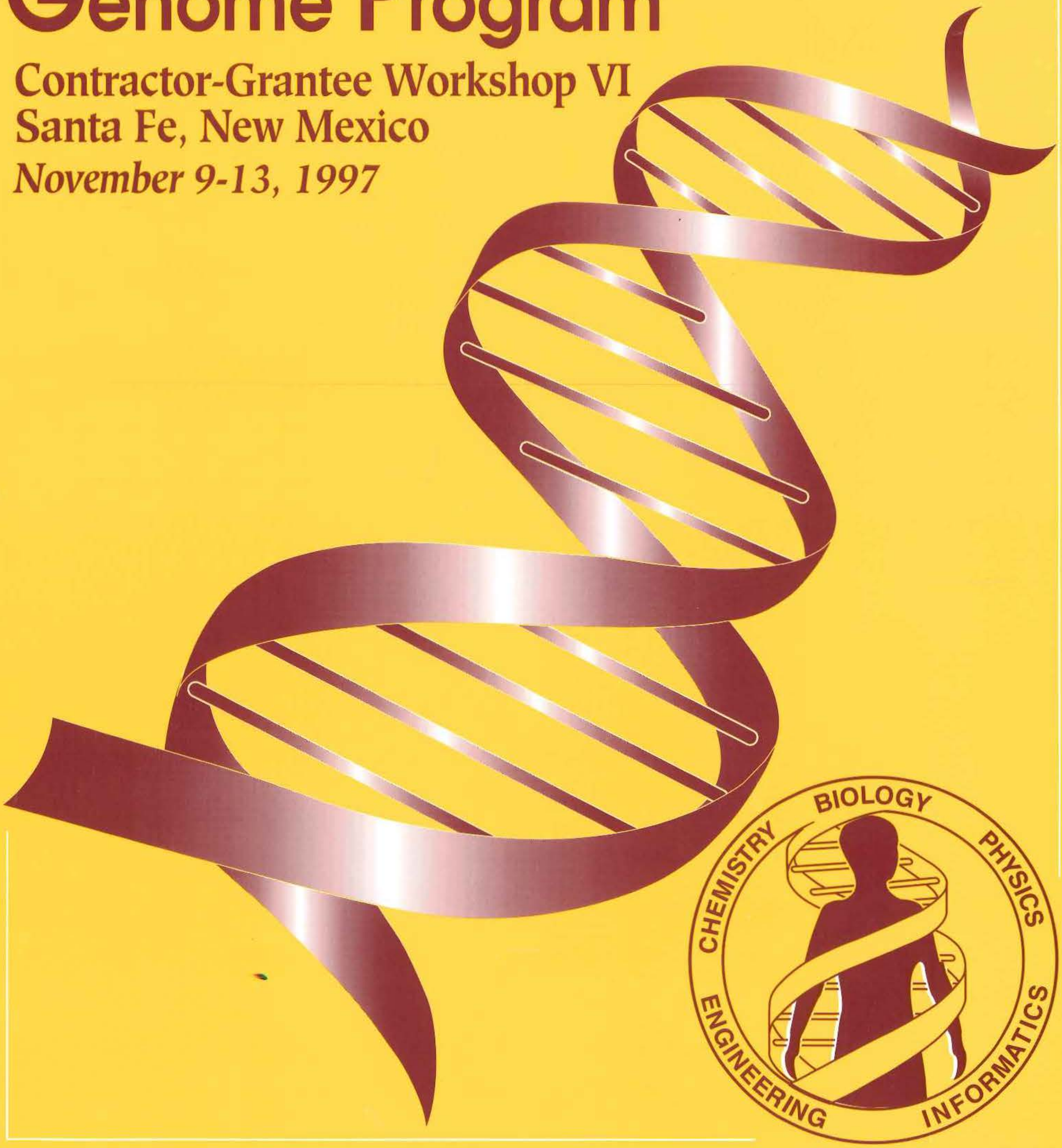
CONF-971146

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

DOE Human Genome Program

Contractor-Grantee Workshop VI
Santa Fe, New Mexico

November 9-13, 1997



BIOLOGY · PHYSICS · INFORMATICS · ENGINEERING · CHEMISTRY

Contact for queries about this publication:

Human Genome Program
U.S. Department of Energy
Office of Biological and Environmental Research
ER-72 GTN
Washington, DC 20585
301/903-6488, Fax: 301/903-8521
E-mail: *genome@er.doe.gov*

A limited number of print copies are available. Contact:

Betty K. Mansfield
Oak Ridge National Laboratory
1060 Commerce Park, MS 6480
Oak Ridge, TN 37830
423/576-6669, Fax: 423/574-9888
E-mail: *bkq@ornl.gov*

An electronic version of this document will be available, after the November 1997 meeting, at the Human Genome Project Information Web site under Publications (*http://www.ornl.gov/hgmis*).

This report has been reproduced directly from the best obtainable copy.

Available to DOE and DOE contractors from the Office of Scientific and Technical Information; P.O. Box 62; Oak Ridge, TN 37831. Price information: 423/576-8401.

Available to the public from the National Technical Information Service; U.S. Department of Commerce; 5285 Port Royal Road; Springfield, VA 22161.

CONF-971146

DOE Human Genome Program
Contractor-Grantee Workshop VI
November 9–13, 1997
Santa Fe, New Mexico

Date Published: October 1997

Prepared for the
U.S. Department of Energy
Office of Energy Research
Office of Biological and Environmental Research
Washington, D.C. 20585
under budget and reporting code KP 0404000

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

Sequencing

Poster
Number Page

Sequencing Technologies

Poster
Number Page

Mapping and Resources

Poster
Number Page

Informatics

Ethical, Legal, and Social Issues

Poster
Number Page

Infrastructure

Appendices

Introduction to Contractor-Grantee Workshop VI

Welcome to the Sixth 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 gives genome scientists and administrative staff 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 always-beautiful surroundings in Santa Fe.

The 158 abstracts in this booklet describe the most recent activities and accomplishments of grantees and contractors funded by DOE's human and microbial genome programs, as well as the research of a few invited guests. All projects will be represented at the poster sessions, so you will have an opportunity to meet with researchers. Plenary sessions will be held at the Eldorado Hotel and poster sessions at the La Fonda Hotel. New informatics resources also will be demonstrated during the poster sessions.

The main challenge facing the genome program today is high-throughput sequencing. DOE is addressing this challenge with the formation of the Joint Genome Institute (JGI) under the direction of Elbert Branscomb. JGI will take advantage of 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. A major step in JGI's inauguration is the creation of a DNA sequencing factory in Walnut Creek, California, located halfway between the Lawrence Berkeley and Lawrence Livermore National Laboratories. Factory output will be high-accuracy human DNA sequences that will be deposited into public databases in accordance with "Bermuda principles."

The U.S. Human Genome Project is nearing the end of its second 5-year plan, which was produced in 1993 by DOE and the National Institutes of Health following rapid progress toward the goals of the original 5-year plan of 1990. We look forward to defining and meeting the goals of the project's next 5 years, now being developed with input from the broad genome scientific community. Although many challenges lie ahead, we are optimistic about the success of this grand project and await its many contributions to science and society.

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 the promise of the genome program.

Sincerely,

Ari Patrinos, Associate Director
Office of Biological and Environmental Research
U.S. Department of Energy

Sequencing

Large Scale Genomic Sequencing at Washington University

Stephanie L. Chissoe and the Genome Sequencing Center
Washington University School of Medicine, St. Louis, MO, USA 63108
schissoe@watson.wustl.edu

Large scale genomic sequencing is ongoing at the Washington University Genome Sequencing Center, focusing on human and *C. elegans* DNA. The *C. elegans* project is in its final year with a total of 71 Mb finished sequence (in collaboration with The Sanger Centre, Hinxton, UK). The human project has been maturing since funding began last year, and we now have 15 Mb finished human genomic sequence. The goals for each project are similar; to achieve contiguous, base-perfect sequence. For *C. elegans*, a clone-based sequence-ready map has provided the majority of sequencing substrates, including cosmids and YACs. A minimal tiling path of cosmids were chosen, although recently we have selected YACs for direct sequencing. Additionally, direct probing of a *C. elegans* fosmid library has provided bacterial clones for regions previously spanned only by YACs, facilitating sequence contiguity. We have incorporated an up-front, sequence-ready mapping paradigm for human genomic DNA. Chromosome 7 STS information is being translated into sequence-ready bacterial contigs. As the sequencing progresses, we are actively working to close gaps in the bacterial clone map by re-screening the large insert libraries with probes based on end-sequence data, additional markers, or overlapping YACs. Candidate clones are chosen for sequencing after evaluating the fingerprint data to verify clone fidelity and overlap through a region.

Sequencing is performed using a mixed shotgun strategy, which includes initial sequencing of random subclones followed by directed approaches for gap closure and ambiguity resolution. The quality of the assembly and sequence editing are monitored by reassembly and comparison to the finished sequence. Additionally, the restriction digest sizes are compared to that based on the finished sequence. During analysis and annotation of the finished sequence, potential exons are identified by similarity to EST data and known protein sequences, and by gene prediction programs.

An overview of the projects will be presented, addressing our strategies, progress, and methods for quality control at each stage of the process. Additionally, recent software tools and technology developments (see abstract by E. Mardis) which have increased our throughput of high quality data will be included.

The UTSW High Throughput DNA Project

Glen A. Evans, Maria Athanasiou, Lisa Hahner, Sherri Osborne-Lawrence, Terry Franklin, Nina Federova, Cynthia English, Shelly Hinson-Cooper, Joel Dunn, James McFarland, Juan Davie, Travis Ward, Paul Card, Parul Patel, Margaret Gordon, Jackie Newton, Danny Valenzuela, Jeff Schageman, Jeff Harris, Garrett Gotway, Mathenna Syed, Ken Kufer, Peter Schilling, Vanetta Gee, Mujeeb Basit, Stafford Brignac, Odell Grant, Ron Burmeister, Kevin O'Brien, Skip Garner, and Roger Schultz
Genome Science and Technology Center,
University of Texas Southwestern Medical Center at Dallas, Texas 75235-8591

In order to complete the DNA sequence of the human genome, collaboration of a group of high throughput sequencing centers will be necessary. The UTSW Genome Sequencing Center has as its goals: 1) development of 100 mb/year sequencing capacity, 2) complete sequencing of human chromosomes 11 and 15 and 3) developing exportable technology and robotics to support high throughput genomic sequencing. Aspects of this project include high throughput mapping to generate sequence ready PAC and BAC template sets, pilot projects to sequence megabase-sized contigs on chromosome 11p15.5, 11p14, 11q23, 15q26 and 15q11, and technology development projects in robotics and informatics. The strategy for integrated map construction, sequencing and data annotation involves assembly of a precise sequence-ready map using STSs, ESTs, regionally mapped genes and other probes screened against a 10X total human BAC library which meets ethical and legal standards of informed consent and anonymity. BAC clones are isolated by array hybridization with radiolabeled pooled oligonucleotides, identity confirmed by PCR with specific markers and FISH, and clones fingerprinted. BAC/PAC end-sequencing is widely utilized for gap filling, contig establishment and sequence annotation. BAC clones are sequenced by M13 shotgun sequencing and fragments assembled into initial contigs using PHRED and PHRAP. Gap filling and quality improvement to an estimated sequence accuracy of 99.99% are accomplished by oligonucleotide-directed sequencing from BAC and PAC templates. Sequence processing, assembly, quality control and annotation and implemented through informatics tools developed in the center. The effort is augmented by the robotics and instrumentation developed in the center including MERMADE high throughput oligonucleotide synthesizers, the Sequencing Support System, a 3 m Sagian rail robot developed for automated sequencing, and a high capacity DNA sequencer under development. The center has produced 4.2 mb of DNA sequence and analysis and annotation of 750 kb of contiguous sequence for 11p15.5 region have been completed with an anticipated 2 mb of continuous sequence by the end of 1997.

Progress Towards Sequencing Selected Regions of Human Chromosome 19

Jane Lamerdin, Aaron Adamson, Karolyn Burkhart-Schultz, Linda Danganan, Jeff Garnes, Ami Kyle, Melissa Ramirez, Stephanie Stilwagen, Glenda Quan, Pat Poundstone, Robert Bruce, Evan Skowronski, Arthur Kobayashi, David Ow, Anthony V. Carrano, and Paula McCready
Joint Genome Institute, Lawrence Livermore National Laboratory, Livermore, CA, 94550
lamerdin1@llnl.gov

Chromosome 19 is the most GC-rich of the human chromosomes as determined by flow cytometry. It is predicted to be very gene rich, with an estimated 2000 genes contained within ~60 Mb of euchromatin. A high resolution physical map of human chromosome 19, constructed largely in bacterial-based clones, serves as a resource for targeted genomic sequencing in regions of high biological interest. One interesting feature of chromosome 19 is the high density of clustered gene families such as zinc finger genes (ZNFs), olfactory receptors (OLFRs) and cytochrome P-450s (CYPs). In order to understand their evolution and subsequent functional diversification, several of these clusters are current sequencing targets. We are also interested in genes involved in DNA repair, and have performed genomic analyses of 6 such loci, many in both human and mouse. To date we have completed over 2 Mb of genomic sequence from chromosome 19 and other human and rodent targets, using a shotgun strategy. Our largest completed sequence contig is ~1 Mb and is located in 19q13.1, flanked by the genetic markers D19S208 and COX7A1. Preliminary analysis of this contig indicates a relative gene density of ~1.7 per 40 kb, and an average Alu density of 1.1 Alu/kb (~31%), which is comparable to other previously sequenced regions of chromosome 19. Of the 43 putative genes identified, two appear to be pseudogenes, 7 encode putative cell surface proteins (e.g. glycoproteins), and 14 are completely novel. Progress towards completion of an >800 kb contig in 19p12 and regions containing clustered OLFR and ZNF gene family members will also be presented.

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.

Analysis of Nine Tandem Zinc Finger-Containing Genes Located within a 339kb Region in Human Chromosome 19q13.2

Mark Shannon, Linda Ashworth, Laurie Gordon, Anne Olsen, and Lisa Stubbs
Human Genome Center, Lawrence Livermore National Laboratory, Livermore, CA 94551
Mark_Shannon@quickmail.llnl.gov

Genetic and physical mapping studies indicate that hundreds, if not thousands, of zinc finger

(ZNF)-containing genes populate the human genome, and that many of these genes, including those with Kruppel-associated box (KRAB) motifs, are arranged in familial clusters. In a previous study, we identified a KRAB-containing ZNF gene family located near the XRCC1 gene in human chromosome 19q13.2 (H19q13.2). Preliminary characterization of this family indicated that the genes are arranged in a 'head-to-tail' tandem array with an average intergenic spacing of 20-30kb. Current estimates based upon physical mapping data suggest that this family is comprised of at least 15 members, which span a minimum of 600kb in 19q13.2. Restriction mapping studies, in combination with Southern blot analyses using ZNF consensus and KRAB sequence probes, have allowed a preliminary determination of this family's content and organization. In depth studies of the largest cosmid contig from the region have demonstrated that nine genes are present within 339kb, encompassing the proximal one-half to two-thirds of the gene family. A number of methods, including Southern blot analysis, PCR analysis and DNA sequencing, have been used to localize expressed sequences within an EcoRI restriction map (RMAP) of this region. Several sequences present in the Genbank database (ZNF45, ZNF155, and HZF4 as well as ESTs 429289, 20113, and 28165) have been localized to specific sites within the RMAP. The region also contains three previously unknown gene sequences. Sequence analysis of cDNA clones for eight of the genes indicates that while the KRAB A domains of the sibling genes are highly similar in structure, other portions, including the KRAB B and ZNF domains, are highly divergent in sequence. These observations suggest that this gene family may have evolved to encode a collection of transcription factors that bind to different target DNA sequences as a result of divergent ZNF arrays, while participating in common transcription control complexes due to their highly similar KRAB A domains. Interestingly, the genes are expressed widely in adult tissues and are co-expressed in many sites. However, tissue-specific variations in levels of transcript between the genes are also evident. Such overlapping, but not identical, expression patterns are consistent with the idea that, like coding sequences, the duplicated cis-acting regulatory regions of the sibling genes have diverged over evolutionary time as they acquired new biological functions. Comparative mapping studies have demonstrated that a homologous Xrcc1-linked gene family is present in the mouse genome. Preliminary studies indicate that three members of the murine gene cluster are orthologous to genes located in H19q13.2. Future studies will address whether the historically homologous mouse and human genes are subject to the same upstream regulation and whether they regulate the same downstream genes.

This work was supported by the U.S. Department of Energy under contract Nos. DE-AC05-96ORO224674 and W-7405-ENG-48.

Genomic Sequencing of 16p13.3

D. C. Bruce, D. O. Ricke, J. L. Longmire, P. S. White, J. M. Buckingham, L. A. Chasteen, D. L. Robinson, M. D. Jones, A. C. Munk, J. D. Cohn, A. L. Williams, M. O. Mundt, L. L. Deaven, and **N. A. Doggett**
Los Alamos National Laboratory, Life Sciences Division and Center for Human Genome Studies, Los Alamos, New Mexico 87545
doggett@gnome.lanl.gov

We have recently begun full genomic sequencing of a 3.0 Mb cosmid/P1 contig of the human chromosome region in 16p13.3 extending from the polycystic kidney disease 1 (PKD1) 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 encompasses the recently cloned Familial Mediterranean Fever gene, and the syntenic breakpoint between mouse chromosomes 16 and 17. The average overlap between clones in the contig is 25%. Sample sequencing of this region has 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 involves nebulization to randomly break DNA, size selection of 3 kb fragments, double adapter cloning into bluescript KS+ plasmid, and sequencing of both ends to 5 X random sequencing coverage. Assembly of sequence contigs is constrained by the inherent relationship of the end sequences being approximately 3 kb apart. Closure is achieved by a combination of primer walking, longer reads, and alternate chemistry reactions. To date we have sequenced to approximately 2 X coverage for the first Mb of this contig and approximately 1 X coverage for the remaining 2 Mb. Supported by the US DOE, OBER under contract W-7405-ENG-36.

The Beginning of the End: Mapping and Sequencing Human Telomeric DNA

Robert K. Moyzis, Han-Chang Chi, and Deborah L. Grady
Department of Biological Chemistry, University of California, Irvine, Irvine, CA 92697
rmoyzis@uci.edu

The Human Genome Project is undergoing a rapid transition from an emphasis on generating physical maps to the large-scale finished sequencing of human DNA. Current technology will allow a large fraction of human DNA to be sequenced in the next 5-10 years by highly automated, high-throughput sequencing "factories". A significant fraction of the human genome, however, will be difficult to sequence to completion by such "factory" approaches. These are regions that: 1) contain a high percentage of repetitive DNA sequences, 2) contain internal tandem duplications, including multigene families, and/or 3) are unstable in all current sequencing vectors. This would be irrelevant if such regions were rare, or contained little of intrinsic informational value. Such is not the case. The first five years of the Human Genome Project mapping efforts have indicated that such regions represent approximately 20% of human DNA. This includes such critical regions as centromeres and telomeres, as well as a greater abundance of low-copy repeats and multigene families than previously anticipated. Producing quality DNA sequence of these regions, which faithfully represent genomic DNA, will be a continuing challenge.

We propose that a large-scale, yet distributed "boutique" approach to mapping and sequencing such regions is warranted, where individual laboratories specialize in genomic regions they have special expertise in investigating. Such efforts would complement and integrate with the few truly large-scale sequencing centers that are likely to evolve during the next few years, such as the proposed DOE Joint Genome Institute Sequencing Center. Our initial iboutiquei target is telomeric regions, which exhibit high levels of repetitive DNA composition, cloning instability, and population heterogeneity. Numerous investigations have implicated genes near telomeres as likely targets for alterations during aging and cancer progression. Through the efforts of a number of laboratories, most notably Dr. Harold Riethmanis at the Wistar Institute, nearly all human telomeres have now been cloned by functional complementation in yeast. My laboratory has finished the 0.23Mb 7q telomere sequence (Chi et.al., this meeting), the first RARE cleavage confirmed human telomere region to be sequenced directly up to the terminal (TTAGGG)n repeat. An important QC/QA aspect of this project was the extensive confirmation of the sequence against genomic DNA by PCR-sequencing (White et.al., this meeting). Greater than 3Mb of confirmed telomeres are now available for sequencing, with another 6-12Mb currently being confirmed. These represent 43 of the 46 unique human chromosome ends. The finished and annotated sequence of these clones will "cap" the world-wide genome sequencing effort, and identify numerous important genes and polymorphic markers.

7q Telomere: Complete Sequencing

Han-Chang Chi[1,2], Elizabeth H. Saunders[1], Judy M. Buckingham[1], Darrell O. Ricke[1], A. Christine Munk[1], Rebecca Lobb[1], Samantha Y.-J. Ueng[1], Mark O. Mundt[1], P. Scott White[1], Owatha L. Tatum[1], and Robert K. Moyzis[1,2]
[1] Center for Human Genome Studies and Genomics Group, LS-3, Los Alamos National Laboratory, Los Alamos, NM 87545 and [2] Department of Biological Chemistry, College of Medicine, University of California, Irvine, Irvine, CA 92697

225,432 bp of DNA immediately internal to the (TTAGGG)n telomere repeat of human chromosome 7q was sequenced. The telomeric end of chromosome 7q, unlike most human chromosomes, contains only a few (<1.5 kb nucleotides in total) subtelomeric repetitive DNA. The lack of subtelomeric repeats, therefore, suggested that this region would likely contain subtelomeric genes whose expression would be affected by telomere alterations accompanying aging and cancer progression (Moyzis et. al., this meeting). 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. Sequence validation was performed on PCR amplified human genomic DNA from at least 6 individuals including the cell line used to construct HTY146. Primer pairs (spaced 300-900 bp) were randomly picked from 25 sites that are almost evenly distributed along the entire 7q terminal region. The QC/QA results confirmed that the cloned YAC/*E. coli* sequences were a faithful representation of genomic DNA, containing less than one error in 10,000 bases (White et. al., this meeting). Computer analysis uncovered numerous open reading frames, expressed sequence tags (ESTs), and potential exons dispersed along the entire 226 kb region, as well as 19 variable number of tandem repeats (VNTRs) and 20 microsatellite repeats. Approximately 192 kb internal to the (TTAGGG)n terminal repeat the first and second exons for the human vasoactive intestinal peptide receptor 2 (VIPR2) gene were located. This gene is involved in a diverse set of physiological functions including smooth muscle relaxation, electrolyte secretion, and vasodilation. SASE-parallel primer walking is efficient for finished sequencing and gene lacalization of a long range genomic target, especially a idifficulti region like telomeres.

Finished Sequence of 7q Telomere Region: Features, Validation and Polymorphism Detection

P. Scott White, Han Chi, Larry L. Deaven, Darrell O. Ricke, Elizabeth Saunders, Owatha L. Tatum, and Robert K. Moyzis

Los Alamos National Laboratory, Life Sciences Division and Center for Human Genome Studies, Los Alamos, New Mexico 87545
swhite@telomere.lanl.gov

The 7q telomere region was sequenced to examine if this region contained expressed genes close to the telomere or served as a buffer area between the telomere and expressed genes in the advent of telomere shortening. A CpG island next to the few exons of the vasoactive intestinal polypeptide receptor 2 precursor gene were discovered (remain exons presumed to be centromeric of the sequenced region). This 225,432 bp sequence contains multiple EST clusters and evidence for additional candidate genes as well as a large array of repetitive sequences and one type I phosphatidylinositol-4-phosphate 5-kinase pseudogene. Interesting, a second candidate gene exists in the form of 7 close regions of homology with chromosome 4 cosmid L191F1. Evidence for a third candidate gene can be pieced together from EST clones with homologies to both the 5' and 3' ends. These paired EST end sequences link together multiple candidate exons in what may be an alternatively spliced gene. An update of the sequence analysis of this region will be presented.

As part of the sequence verification for the 7q region, we have PCR amplified and sequenced ~300 bp from each of 19 sites along this region, from at least 6 individuals representing 4 diverse ethnic groups. In addition, we sequenced the genomic DNA used to build the 7q YAC library, and the supporting YAC clone, HTY-146. For each site, PCR primers were selected in regions free of known repeats. All PCR products were sequenced and compared to detect errors and polymorphisms. Sequence comparisons among individuals give us an idea of the amount of natural variation, as well as validate the clone sequence to the genomic DNA from which the clone originated. Out of 5722 nucleotides resequenced (2.5 % of total) there was one heterozygous site found in the individual from which the library was constructed, and 9 single nucleotide polymorphisms (SNPs) among the other individuals. All other sequences were identical between the clone sequence and that of the individual from which the library was made. These results are in agreement with previous estimates that natural variation at the primary sequence level is at least 1 in 1000, which needs to be considered when performing sequence validation aimed at detecting a one in 10,000 error frequency.

Large-Scale High-Quality Genomic Sequencing of Human Chromosome 7

Shawn Iadonato, Jun Yu, Gane K.-S. Wong, Charles Magness, Phil Green, and Maynard Olson
Human Genome Center, Department of Medicine, University of Washington, Seattle, WA 98195.

We are implementing an approach to large-scale sequencing of human genomic DNA that emphasizes high-quality at a reasonable cost. Quality targets are: 1) a single-base-pair error rate of better than 0.01%, 2) no gaps, in either the shotgun sequence or the physical map, across mega-base sized regions, and 3) validation of the sequenced large-insert clones to an average resolution of 200-bp. This is accomplished by the systematic introduction of objective quality measures throughout the data production process, from the physical mapping through to the shotgun sequencing.

Detailed sequence-ready restriction-maps are produced by the multiple-complete-digest (MCD) method. These maps have proven to be very accurate, with average sizing errors better than 1%. They allow us to both validate the large-insert clones to a 200-bp average resolution, and to confirm the correctness of the subsequent sequence assemblies. The sequence production is distinguished by an emphasis on long read-lengths instead of maximum machine utilization. Data and quality analysis are performed with the Phred/Phrap/Consed system and average Phrap-alignable read lengths are 735-bp. Overlapping large-insert clones are finished independently and the sequence overlaps are used to estimate the single-base-pair error rate, which has been better than 1 in 100,000-bp. We will present data from a CONTIGUOUS 2-Mbp region on human chromosome-7 (near 7q31.3). Evidence will be presented to support all of the above quality assertions.

Sequencing Progress at Lawrence Berkeley National Lab and the Formation of JGI

Chris Martin and Mohan Narla
Human Genome Sequencing Department, Life Sciences Division, Lawrence Berkeley National Laboratory and the DOE Joint Genome Institute, Berkeley, CA

The Genome Centers at DOE's Los Alamos, Berkeley, and Livermore National Laboratories have merged into the Joint Genome Institute (JGI). This reorganization is a significant undertaking that is aimed at achieving economies of scale in our genome efforts while also leveraging off of the expertise available at the three sites. In the area of genomic sequencing, the large majority of the JGI's effort will be moving the a new facility, located in Walnut Creek, California, in early to mid 1998. This will provide a custom designed space sufficient for the scale up of the JGI's production sequencing effort in a factory like setting, termed the PSF (Production Sequencing Facility). Until this time, the production sequencing efforts at the three laboratories will be scaled up in place, while also working towards the goal of finalizing a uniform production process for use within the PSF.

The sequencing approach at Berkeley has recently been altered by the increase in the emphasis on the up-front shotgun phase of the process, which utilizes double end plasmid subclone sequencing. Additionally, we have adopted the use of additional bacterial strains for all of our subclone libraries that seem to alleviate under-representation of genomic regions due to cloning biases. These changes are helping us to significantly reduce the time required for the complete sequence of a given physical mapping clone to be determined. Data will be presented on the current status of a set of 16 human and mouse P1's, pac's and bac's that are now in progress using this new sequencing process.

Comparative Analysis of Human and Mouse Orthologous DNA to Identify Conserved Regulatory Regions

Kelly A. Frazer, Gabriela G. Cretz, Christopher H. Martin, Jan-Fang Cheng, and Edward M. Rubin
Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
kelly@mhgc.lbl.gov

Human chromosome 5q31 was chosen by the JGI for large-scale sequencing because it harbors a family of interleukin genes which are important regulators of the immune response. Interleukin-3 (IL-3), IL-4, IL-5, IL-13 and granulocyte-macrophage colony-stimulating factor (GM-CSF) are clustered within 600 kilobases of each other on human 5q31. Previously, we have computationally and biologically analyzed the available genomic sequence data in the 1 Mb region of human 5q31 containing the interleukin family cluster and identified 16 new genes, as well as 7 previously known genes. We also performed comparative mapping studies and demonstrated that 13 of the human genes in the 5q31 region are located in the syntenic region of mouse chromosome 11.

Numerous experimental studies have indicated that noncoding regulatory sequences controlling gene expression are evolutionarily conserved in mice and humans. Comparative analysis of human and mouse orthologous sequences is a powerful tool for identifying these conserved noncoding regulatory elements. To facilitate the annotation of coding regions and permit noncoding conserved regulatory elements to be readily identified in the human 5q31 sequence data, the JGI has chosen to sequence the 1 Mb region containing the interleukin growth factor cluster on mouse chromosome 11.

To date, we have isolated and begun sequencing a 150 kb mouse BAC which contains 5 complete genes: IL-4 and IL-13 – genes with coordinated expression in T-Cells; septin – a ubiquitously expressed gene with complex alternative splicing; and cyclin-like and KIF3 – genes which are both predominantly expressed in the brain and are physically near each other. Importantly, the expression patterns of these five genes suggest that human-mouse conserved noncoding regulatory elements are likely to be in this 150 kb region and identified by comparative sequence analysis.

Comparative Genomic Sequencing of the Human and Mouse Fibroblast Growth Factor Receptor Type III Gene

Ana V. Perez-Castro, Julie Wilson, Cleo Naranjo, and **Michael R. Altherr**
Los Alamos National Laboratory, Genomics Group, Los Alamos, New Mexico 87545
altherr@telomere.lanl.gov

We have sequenced the human and mouse genomic segments encoding the fibroblast growth factor receptor 3 gene (FGFR3). FGFR3 is a member of the receptor tyrosine kinase superfamily and a developmentally regulated transmembrane protein. Mutations in FGFR3 contribute to a number of significant human maladies. The human and mouse genes exhibit significant similarity in terms of their structural domains and genomic organization. In both species, the gene consists of 19 exons and 18 introns spanning greater than 15 kbp of sequence. The coding regions across the entire gene are 84% and 92% identical at the nucleic acid and amino acid levels respectively. The alternatively spliced exon 8 is 95% similar at the nucleic acid level and 93% similar at the amino acid level. While the sequence similarity of the introns is (on average) less than 50%, the size of the individual introns is very similar. The 5' flanking regions of the gene in both species shows a similar high degree of conservation. In general, the analysis of a 400 bp segment preceding the initiator ATG shows only a 60% similarity. However, several common transcriptional regulatory sequences are present in both. Consensus binding sites for Sp1, AP2, Krox24, and IgHC.4 are located in this region. It is also worth noting that the position and spacing between these sites is conserved in these species. These data suggest that mouse comparative genomic sequencing can be used to identify and annotate significant functional domains in the human genome.

Support: DOE Contract W-7405-ENG-36 and LANL LDRD funds.
E-mail Contact: ALTHERR@LANL.GOV
Day Phone: (505)665-6144

Comparative Genomic Sequencing of the Murine Syntenic Segments Corresponding to Human Chromosome 16p13.3

Darrell O. Ricke, Norman A. Doggett and **Michael R. Altherr**
Los Alamos National Laboratory, Genetics Group, Los Alamos, New Mexico 87545
altherr@telomere.lanl.gov

The terminal short arm band of human chromosome 16 (16p13.3) is syntenic with three different mouse chromosomes (11, 16, and 17). At least 8 loci define the syntenic overlap between human 16 and mouse 17. Less than 1 Mbp from the proximal boundary of this segment is another cluster of eight loci that comprise 21 centimorgans of mouse chromosome 16. Our human genomic sequencing targets include this 1 Mbp segment of chromosome 16 where the syntenic breakpoint between mouse chromosomes 16 and 17 occurs. In addition, this segment overlaps at least one unidentified locus of medical significance, CATM, and contains the recently identified gene for familial Mediterranean fever. The CATM locus is a gene that, when defective, causes congenital cataract and microphthalimia. It is clear from previous comparative sequencing studies that both structural and regulatory regions of genes can be identified by similarity comparisons of human and mouse genes. In fact, our preliminary sequencing efforts of this region of 16p13.3 have already identified 33 mouse EST clusters with greater than 80% similarity to the human sequence. These ESTs are being used to develop primer pairs to identify mouse BAC clones. These mouse BAC clones will be subcloned and sequenced at low redundancy to identify and annotate potentially interested regions of the human genome. This effort will be closely coordinated with higher redundancy efforts at LLNL and LBNL in an effort to optimize the depth of sequence coverage required to identify conserved sequence segments and presumably functionally important domains.

Support: DOE Contract W7405-ENG-36 and LANL LDRD funds.
E-mail Contact: ALTHERR@LANL.GOV
Day Phone: (505)665-6144

High-Throughput Sequencing of Eubacterial and Archeal Genomes

Owen White, Mark D. Adams, Rebecca Clayton, Hans-Peter Klenk, Anthony R. Kerlavage, Karen E. Nelson, J. Craig Venter
The Institute for Genomic Research
owhite@tigr.org

The Institute for Genomic Research (TIGR) has completed the genomic sequences for the DOE-funded projects of the eubacteria *Mycoplasma genitalium*, and two archeal genomes *Methanococcus jannaschii* and *Archaeoglobus fulgidus*. The eubacterial genomes *Deinococcus radiodurans* and *Thermotoga maritima* are now in closure, while a fifth genome project, *Shewanella putrefaciens* is currently in the random sequencing phase at TIGR. These projects result in over 11.5 Mb of finished sequence that contain approximately 10,500 genes. Data relating to each of these bacterial projects are located in our local database, disk-based files used for searching, an external web server, the public sequence archives, and other web servers (e.g. Argonne's biochemical pathways for *Haeomophilus influenzae*). One challenge of the distributed archetecture of bacterial information is to effectively propagate changes from one data location to another, preferably in an automated manner. We will report on a system that is under development that encodes system-wide data dependencies necessary for update information in a large scale sequencing facility. We have also begun implementation of enhanced high-throughput data analysis that adds intergenomic and intragenomic gene families, hydrophobicity plots, block and motif searching, medline and web-derived information to our previous database searching methodology. We will report on our enhanced annotation techniques and will discuss them in context of the recently completed genomes.

High-Throughput Transposon Mapping and Sequencing

Robert B.Weiss[1], Mark Stump[1], Joshua Cherry[1], Cindy Hamil[1], Frank Robb[2] and Diane Dunn[1]
[1] Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, [2]Center of Marine Biotechnology, University of Maryland, Baltimore, MD 21202.
bob.weiss@genetics.utah.edu

The ability to sequence moderate-size plasmid inserts (10-15 kb) using mapped transposons is being tested on both microbial and human sub-clone libraries. The transposon serves as the priming site for initiating a set of bi-directional di-deoxy ladders within the plasmid insert. The priming site locations are mapped by automated Southern analysis using a restriction digest that releases the insert from the vector, while cutting in the center of the transposon. The inserts propagate on a vector designed to maintain the plasmid at a few copies per cell, while enabling efficient DNA purification after temperature-induced runaway plasmid replication. Sub-clone libraries are built in a multiplex family of this vector backbone, where each vector attaches unique sequence tags to the insert for use in mapping. Automated Hybridization and Imaging Instruments (AHII) are used to sequentially probe for the mapping tags. These instruments are fully automated devices for detecting enzyme-linked fluorescence from DNA hybrids on nylon membranes, and are used in the mapping and sequencing phase of the project

We are in the process of using the transposon mapping and sequencing technique to complete the 2.1 Mb genomic sequence of *Pyrococcus furiosus* (DSM 3638), a hyperthermophilic heterotrophic member of the Archaea. This organism, isolated from hot marine sediments, grows vigorously at temperatures near 100oC. Its lifestyle is anaerobic and it derives energy by fermenting peptides and carbohydrates to organic acids and CO2. *P. furiosus* plasmid and cosmid libraries were built in a family of 21 multiplex vectors. These vectors provide multiplex tags for both the end sequencing and transposon mapping phases of the process. End sequences were determined on 2875 plasmid inserts and 400 cosmid inserts. The cosmid inserts provide a global scaffold of the genome, while the plasmid inserts are used in the primary process of transposon mapping and sequencing. Overlaps between plasmid inserts derive from the matching of end sequence onto transposon sequence contigs, and contigs are grown from re-feeding the mapping process with inserts that extend or bridge the sequence contigs.

Common growth and DNA prep formats feed both the mapping and sequencing process. Minimal spanning sets of clones are predicted from the mapping phase and fed to the sequencing process, where cycle sequencing is performed on double-stranded templates. A single hybridization instrument is capable of imaging 1728 lanes of mapping data in each eight hour cycle; with a 20-fold deep multiplex set, these instruments acquire mapping data on 34,560 clones over the course of a single, unattended run. The use of this technique on the *Pyrococcus* genome is demonstrating the robustness and cost-effectiveness of using this technique. The use of moderate-size plasmid insert libraries providing a reduced clone feed and finishing information results in a complementary approach to small-size M13 insert strategies. This project is an alpha test of methods, instrumentation and software under development at the Utah Center for Human Genome Research.

This work is funded by DOE grant DE-FG03-94ER-61950 (R.B. Weiss, P.I.).

Complete Sequencing of the 2.3Mbp Genome of the Hyperthermophilic Archaeon *Pyrobaculum aerophilum*

Sorel Fitz-Gibbon[1], **Ung-Jin Kim**[2], Heidi Ladner[1], Yongweo Cao[2], Gony H. Kim[2], Barbara Perry[2], Enrique Colayco[2], Ronald V. Swanson[3], Terry Gaasterland[4], Jeffrey H. Miller[1], and Melvin I. Simon[2]
[1] Department of Microbiology and Molecular Genetics, and Molecular Biology Institute, University of California, Los Angeles
[2] Biology, California Institute of Technology
[3] Now at Diversa Corp., San Diego

[4] Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL; Department of Computer Science, University of Chicago, Chicago, IL

Pyrobaculum aerophilum is a hyperthermophilic archaeon discovered from a boiling marine water hole at Maronti Beach, Italy that is capable of growth at 104°C. This microorganism can grow aerobically, unlike most of it's thermophilic relatives, making it amenable to a variety of experimental manipulation and a potential candidate for a model organism for studying archaeal microbiology and thermophilism. Sequencing the entire genome of this organism provides a wealth of information on the evolutionary and phylogenetic relationship between archaea and other organisms as well as the basis of thermophilic nature of this organism. We have constructed a physical map that covers estimated 2.3 Megabase pair genome using a 10X fosmid library. The map currently consists of 96 overlapping fosmid clones. We have completed sequencing the entire genome using a random shotgun approach with the supplement of oligonucleotide primer directed sequencing. A total of 16,098 random sequences corresponding to approximately 3.5X genomic coverage were obtained by sequencing from both ends of 2-3 kbp genomic DNA fragments cloned into pUC18/19 vectors using vector-specific primers. These fragments were assembled into several hundred contigs using the Phrap program developed by Dr. Phil Green at University of Washington, Seattle. Gaps and regions of low quality base calls have been resolved using primers specifically designed to extend the problem regions. We have successfully closed all remaining gaps after a total of 2,300 directed sequencing reactions and reassembly. Our current full length genomic sequence still suffers from low data quality: only approximately 99% of the nucleotide sequences are accurate. This is mainly due to the low redundancy (3.5 fold) in random sequencing. We plan to perform 2-3,000 more directed sequencing reactions to polish the sequence to 99.99% accuracy. We are currently running MAGPIE, a WEB based system for sequence annotation, in our UltraSparc server (date.tree.caltech.edu) and plan to analyze and completely annotate the genome.

Microbial Genome Sequencing and Comparative Analysis

D.R. Smith, T. Aldredge, R. Bashirzadeh, H. Bochner, M. Boivin, S. Bross, D. Bush, A. Caron, A. Caruso, G. Church*, R. Cook, C.J. Daniels#, C. Deloughery, L. Doucette-Stamm, J. Dubois, J. Egan, D. Ellston, J. Ezedi, T. Ho, K. Holtham, P. Joseph, M. LaPlante, H-M. Lee, D. Blakely, R. Cook, R. Gibson, K. Gilbert, A. Goyal, J. Guerin, D. Harrison, J. Hitti, L. Hoang, N. Jiwani, P. Keagle, J. Kozlovsky, W. Lumm, J. Mao, P. Mank, A. Majeski, S. McDougall, J. Nölling, D. Patwell, J. Phillips, S. Pietrokovski@, B. Pothier, S. Prabhakar, D. Qiu, J.N. Reeve#, P. Rice, P. Richterich, M. Rossetti, M. Rubenfield, M. Sachdeva, H. Safer, G. Shimer, P. Snell, R. Spadafora, L. Spitzer, H-U. Thomann, R. Vicaire, Y.Wang, L.Wong, K. Weinstock, J. Wierzbowski, Q. Xu, L. Zhang
Genome Therapeutics Corp., Waltham, MA
*HHMI, Dept. of Genetics, Harvard Medical School, Boston, MA
@Fred Hutchinson Cancer Research Center, Seattle, WA
#Dept. of Microbiology, The Ohio State University, Columbus OH
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 delta H and *Clostridum acetobutylicum* ATCC 824.

Methanobacterium thermoautotrophicum 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. Analysis of the sequence predicted 1,855 polypeptide-encoding ORFs, 807 (44%) of which could be classified according to function. The putative gene products were compared with sequences from *Methanococcus jannaschii*, as well as eucaryal, bacterial and archaeal specific databases. These analyses indicated that most ORFs are most similar to sequences described previously in other Archaea, but that there has been extensive divergence between the two sequenced methanogen genomes. Most gene products predicted to be involved in cofactor and small molecule biosyntheses, intermediary metabolism, transport, nitrogen fixation, regulatory functions and interactions with the environment are more similar to bacterial than eucaryal sequences, whereas the converse was true for most proteins predicted to be involved in DNA metabolism, transcription, and translation. There are 24 polypetides that could form two-component sensor kinase-response regulator systems, homologs of bacterial DnaK and DnaJ, homologs of eucaryal DNA replication initiation Cdc6 proteins, an X-family repair-type DNA polymerase and an unusual archaeal B-type DNA polymerase. There are 39 tRNA genes, two rRNA gene clusters, one intein containing gene, several repeated regions, two large clusters of short repetitive elements, and numerous other interesting features.

The Clostridia are a diverse group of gram-positive, rod-shaped, spore forming anaerobes that include several toxin-producing pathogens and a large number of terrestrial species. The latter have been used extensively for industrial solvent production (acetone, butanol and ethanol) by fermentation of starches and sugars. *C. acetobutylicum* strain ATCC 824 has a 4.1 Mb, AT-rich genome and is one of the best-studied solventogenic clostridia. 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. A total of 4018 putative polypeptide encoding ORFs were identified and searched against public databases to provide preliminary annotation. The finishing phase of the project is currently underway utilizing a quality-based finishing paradigm and a set of integrated bioinformatics tools. The data are available at http://www.cric.com.

Sequencing the *Borrelia burgdorferi* Genome

John J. Dunn, Laura-Li Butler-Loffredo, Ting Chen, William C. Crockett, Jan Kieleczawa, Jeremy Medalle, Sean McCorkle, Keith H. Thompson, Jeanne R. Wysocki, Shiping Zhang and **F. William Studier**
Biology Department, Brookhaven National Laboratory, Upton, New York 11973
jdunn@bnl.gov, studier@bnl.gov

The ~900 kbp linear chromosome of *Borrelia burgdorferi*, the bacterium that causes Lyme disease, is being sequenced to spur the development of an integrated system for accurate, high-throughput, low-cost genome sequencing with minimal human involvement. We are testing a modified whole-genome shotgun approach, using random 1st-end and directed 2nd-end sequencing on a clone set with good physical coverage. A network of linked clones is generated at relatively low sequence redundancy, and primer walking on linking clones is used to close the gaps and complete the sequence of both strands. This strategy can achieve highly accurate sequence at an overall redundancy of 4-fold or even less.

Two commercial fluorescent sequencers have been used in this development phase, with the intention of scaling up capacity with a capillary sequencing system currently under development. Data management and analysis capabilities tailored to this sequencing strategy have been developed and implemented, including software for managing the sequence process, selecting 2nd-ends for sequencing, assembling the sequence, and selecting walking primers. The routine biochemical protocols needed for Mbp-scale sequencing have also been implemented. Primers for walking are generated from a library of all 4096 hexamers, by ligating hexamers on hexamer templates to form 12-mers for cycle sequencing.

The *Borrelia* sequence is almost finished. As of August 21, approximately 5400 end sequences and 2700 primer-walking sequences had been obtained on approximately 2700 plasmid clones (average insert length of 2.1 kbp) and 106 fesmid clones (average insert length of 35 kbp), representing about 4x sequence coverage and 10x physical coverage of the linear chromosome. The sequence redundancy was higher than necessary because of extra sequences obtained during protocol development. A single contig of approximately 900 kbp aligns with the published restriction map and contains all but an estimated few kbp at each end. This sequence is posted on our web site at www.bio.bnl.gov and will be updated periodically as the sequence of both strands is completed. As of August 21, about 92% had been sequenced on both strands and approximately 200 single-strand gaps remained to be filled. Our sequence agrees well with the sequence of the chromosome of the same *Borrelia* strain determined independently at The Institute for Genomic Research (TIGR).

Software and database support for this sequencing strategy continues to be refined, with the aim of removing as much human decision making as possible. Automation of sample selection and data entry will greatly reduce sources of human error, which cause more problems in primer walking than in shotgun sequencing. A newly developed single-copy amplifiable vector in which nested deletions are easily produced should allow the use of clone libraries of longer average fragment length, improve sequencing efficiency, and help in resolving repeated sequences.

Sequence Analysis of the Plasmids of *Bacillus anthracis*

R. T. Okinaka, K. G. Cloud, O. A. Hampton, K. K. Hill, P. Keim, S. Kumano, D. Manter, J. J. Renouard, D. O. Ricke, and P. J. Jackson
Los Alamos National Laboratory, Life Sciences Division, Los Alamos, New Mexico 87545 and Northern Arizona University, Flagstaff, Arizona
okinaka@telomere.lanl.gov

Virulent strains of *Bacillus anthracis* contain two large plasmids, pX01 and pX02. These plasmids are known to carry the toxin genes (lethal factor, edema factor, and protective antigen) and the three capsule genes that enhance the virulence of the organism (Cap A, B and C). The vast majority of the genes on pX01 and pX02, however, have not yet been characterized. The two plasmids combined contain sufficient sequence information to code for an additional 200-250 microbial sized open reading frames (ORF), i.e., 185 kb and 85 kb of DNA in pX01 and pX02, respectively. As an initial step to identify and characterize potential ORFs we have begun to sequence, assemble and analyze the entire DNA sequence of each plasmid. The pX01 and pX02 genomes are being analyzed by random cloning of sheared or restriction digested 2-4 kb fragments followed by high throughput DNA sequence analysis on automated Applied Biosystems sequencing machines. We will report on our progress and sequence analysis of the assembled shotgun sequence contigs for both pX01 and pX02.

This work is sponsored by the U.S. Department of Energy.

Sequencing Technologies

Rhodobacter capsulatus Genome Sequencing Project: DENS Technology Testing Ground

Mugasimangalam Raja, Vincent Molloy, Lev Lvovsky, James Akowski, Jonathan Schisler, Yakov Kogan#, Michael Fonstein#, Robert Haselkorn# and Levy Ulanovsky
Argonne National Laboratory, Argonne, IL
#University of Chicago, Chicago, IL

We are using DENS (Differential Extension with Nucleotide Subsets, see Ref.1 and the accompanying abstract by Raja et al.) for sequencing the genome of *Rhodobacter capsulatus* by primer walking without custom primer synthesis. A cosmid library of *R. capsulatus* was constructed and mapped with high resolution (2). Of the total 3.5 Mb genome, 1.4 Mb has already been sequenced by limited shotgun sequencing followed by conventional primer walking. Forty-eight plasmid subclones containing 5 kb inserts were isolated from each cosmid and both ends sequenced. The resulting sequences were then assembled into contigs. In conventional primer walking, approximately 100 custom synthesized primers per cosmid are required to close the sequencing gaps and generate the second strand. Around 10 additional primers are needed to fill the physical gaps between subclones by direct primer walking on the cosmid.
In primer walking by DENS, the conventional custom synthesized walking primers are replaced with DENS octamers (containing two degenerate positions each) from our presynthesized library of 2,048 octamers (50% of all possible sequences). At the current DENS success rate of 62% on dsDNA templates, the use of DENS is cheaper than primer walking using conventional custom-made primers, even though the latter yields an 85% success rate. Future closed end automation of DENS primer walking, made possible by the instant availability of primers, should reduce sequencing cost by more than an order of magnitude. The results of the *R. capsulatus* sequencing show that DENS is a viable option. Updated results on this pilot project and implications of the DENS technique will be discussed.

1. Raja et al. (1997) Nucleic Acids Res. 25: 800-805.
2. Fonstein et al. (1995) EMBO J. 14: 1827-1841.

Cosmid Sized Templates for DENS Sequencing Technique

Mugasimangalam Raja*, Vincent Molloy*, Lev Lvovsky# and Levy Ulanovsky*#
*CMB, Argonne National Laboratory, Argonne, IL
#Weizmann Institute of Science, Rehovot, Israel

DENS (Differential Extension with Nucleotide Subsets) is a technique used for template directed enzymatic synthesis of unique primers, avoiding the chemical synthesis step in primer walking (1). DENS works by selectively extending a short primer and making it a long one at the intended site only. The procedure starts with an initial extension of the primer (at 20-30°C) in the presence of only two out of the four possible dNTPs. The primer is extended by 5 bases or longer at the intended priming site, which is deliberately selected, (as is the 2 dNTP set), to maximize the extension length. The subsequent termination reaction at 60-65°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.

Until recently we were getting about 70% success rate on ssDNA plasmid and asymmetric PCR products and 62% on ds plasmids. Now, we have found that the usable template size for DENS is not limited to several kb (typically plasmids), as we thought previously. It turns out that DENS sequencing can be performed on cosmid-sized templates on the condition that the sequence of most of the template is known (as in filling physical gaps). Then the computer excludes candidate priming sites for whom the chosen octamer primers have alternative priming sites in the known part of the template if at any alternative site the extension crosses the failure threshold (about 4-5 bases). Of 17 DENS sequencing reactions performed using Lambda phage as the template, 10 worked well.

In addition, we have developed a novel version of DENS that uses only one out of the 4 possible dNTPs in the differential extension (either dCTP or dGTP extending the octamer by 3 or more bases). One-dNTP DENS allows primer walking on such a template as a whole cosmid, even if most of its sequence is completely unknown. What makes it possible is that the probability of the occurrence of alternative sites with long differential extensions using a single dNTP is dramatically lower than that of extensions using a two-dNTP subset. Of five octamer primers tested on Lambda using one-dNTP DENS, four produced good sequences at the intended unique sites. The occurrence of suitable priming sites for the one-dNTP DENS is lower than for the two-dNTP DENS but is still high enough to perform primer walking if all 4,096 octamers are used.

1. Raja et al (1997) Nucleic Acids Res. 25, 800-805.

Sequencing of Human Telomeric Region DNA by Differential Extension with Nucleotide Subsets (DENS)

Dina Zevin-Sonkin#, Anahit Ghochikyan#, Arthur Liberzon#, Lev Lvovsky# and Levy Ulanovsky*#
# Dept. of Structural Biology, Weizmann Institute of Science, Rehovot, ISRAEL
* CMB, Argonne National Laboratory, Argonne, IL

DENS allows primer walking without custom primer synthesis and each walk involves a two step procedure. It starts with a limited initial extension of an 8-mer primer (degenerate in 2 positions) at 20-30°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 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 (see Ref. 1). Both steps involve thermocycling.

Here we show an example of DENS primer walking on three human genomic DNA subclones of the 3-4 kbp length each from telomeric region of human chromosome 7, kindly provided by Robert Moysis group, LANL, Los Alamos, NM. The full length sequences (3024, 3473 and 3707 bp) were obtained from both strands of each subclone using dye- terminators and the ABI-373 sequencer. For these clones we tested an easy method for ss template preparation avoiding plasmid prep of any kind. The entire insert of a plasmid clone was amplified by PCR in which one of the two primers was 5'-phosphorylated. One of the two strands was digested with lambda exonuclease (Boehringher Mannheim, cat#1666908) which selectively digests 5'-phosphorylated strand only (Ref.2). Using this method ssDNA template is produced by PCR using vector-specific primers directly from overnight bacterial culture.

For the DENS reaction, we have optimized parameters related to the stability of the extended primers and used them for the selection of DENS priming sites. Upon this optimization, the success rate of DENS primer walking using the ss template preparation by PCR was 70%.

1. Raja et al. (1997) Nucleic Acids Res. 25, 800-805.
2. Little et al. (1967) J. Biol. Chem. 242, 672-678.

Primer Walking Through Alu Repeats Using DENS Sequencing Technique

Dina Zevin-Sonkin#, Anahit Ghochikyan#, Arthur Liberzon#, Lev Lvovsky#, and Levy Ulanovsky*#
#Weizmann Institute of Science, Rehovot, Israel
*CMB, Argonne National Laboratory, Argonne, IL

Primer walking using conventional primers is believed to be problematic when tandem Alu repeats are present in the template. In contrast to conventional 18-20mer primers, DENS (Differential Extension with Nucleotide Subsets, see Ref. 1 and accompanying abstracts by Raja et al.) uses octamers degenerate in 2 positions. We have found that DENS has an advantage over conventional primer walking in sequencing through tandem Alu repeats. A single mismatch in the octamer/template complex prevents priming, enabling discrimination between nearly identical repeats. It is possible to walk through repeat-rich regions by selecting hypervariable sites for DENS priming within the Alu consensus sequence. Additional mismatch discrimination is provided during the differential extension stage, since the extension will be shorter if a template base is non-complementary to either of the two dNTPs provided (e.g. if dATP and dGTP were the only nucleotides present, and a "G" were encountered in the template). Finally, the extended primer is also destabilized by a single mismatch at the higher annealing temperature of the cycle sequencing stage. All of these factors cause DENS to be very effective at distinguishing similar, but not identical, priming sites within a DNA template, whereas a conventional long primer is less able to provide such discrimination.

1. Raja et al. (1997) Nucleic Acids Res. 25, 800-805.

Structural Insights into the Properties of DNA Polymerases Important for DNA Sequencing

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

Current methods for DNA sequencing require a DNA polymerase to extend a primer with each of the four natural nucleotides, as well as a variety of analogs such as fluorescent and chain-terminating dideoxynucleotides. We have been characterizing the structure and function of DNA polymerases to better understand and to modify those properties important for DNA sequencing, in particular the incorporation of nucleotide analogs and the processivity of DNA synthesis. Our work has focused on the DNA polymerases belonging to the Pol I family, that includes T7 DNA polymerase, Taq DNA polymerase, and *E. coli* DNA polymerase I. An important advance in our understanding of these enzymes has resulted from our recent determination of the 2.2 Å 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]. The incoming dNTP fits snugly into a pocket formed by the fingers of the polymerase closing on its palm and the 3'-terminus of the primer. Numerous interactions of the bound nucleotide with the template base, the polymerase, and two metal ions specify the correct base-pair in the active site, and provide insight into the mechanism of discrimination against analogs with modifications in the sugar moiety (e.g. dideoxynucleotides, ribonucleotides and 3' fluoro derivatives) as well as bases containing bulky fluorescent substituents. We are making use of this structure to construct mutant polymerases that incorporate various nucleotide analogs more efficiently that are modified in either their sugar, base or triphosphate groups.

The processivity factor for T7 DNA polymerase, *E. coli* thioredoxin, is located at the tip of the thumb of the polymerase in a position poised to prevent dissociation of the DNA from the polymerase. It binds to a 74 residue domain in T7 DNA polymerase that is attached to the thumb by a flexible tether. While this domain is unique to T7 DNA polymerase, it is a modular in that it can be transferred to other homologous polymerases by gene fusion to generate hybrid enzymes that have dramatically increased processivity for DNA synthesis [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 critical basic residues in T7 DNA polymerase to interact electrostatically with the DNA backbone and thus prevent its dissociation. We are carrying out extensive mutagenesis of this region in order to further define the critical structural features and to engineer new DNA polymerases that have increased processivity.

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, in press.

[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).

Semiautomated Mutagenesis and Screening of T7 RNA Polymerases

Mark W. Knuth, Scott A. Lesley, and Heath E. Klock
Promega Corporation 2800 Woods Hollow Road, Madison, WI 53597
mknuth@promega.com

The objective of this grant is to develop an engineered RNA/DNA polymerase for primerless DNA sequencing, among other applications. In order to develop this enzyme, systems have been developed for facile semiautomated mutagenesis, expression, micropurification, and screening of mutants of T7 RNA polymerase.

Two publications have reported T7 RNAP mutations that confer some ability to incorporate dNTPs; we have further characterized their performance in parameters important to DNA sequencing and other applications, and have created screens to test for improved performance.

To date, clones representing all possible amino acid substitutions at 96 sites have been created, using a novel mutagenesis kit developed at Promega. These mutant enzymes are being expressed, purified and tested for increased dNTP incorporation. We anticipate that at full capacity, around 12 positions will be screened per week, using techniques and equipment that can be generalized to other enzymes requiring some purification before screening.

Vectors and Biochemistry for Sequencing by Nested Deletions

John J. Dunn and Matthew Randesi
Biology Department, Brookhaven National Laboratory, Upton, New York 11973
jdunn@bnl.gov

An ordered set of nested deletions whose ends are separated by ~400 bp allows rapid sequencing across one strand of a cloned fragment, using a universal primer. Any gaps remaining after this process can be closed by primer walking on the original clone. Even highly repeated DNA can easily be assembled correctly, knowing the relative locations of the sequences obtained. We have developed vectors and protocols that allow simple and reliable production of nested deletions suitable for such a sequencing strategy, from cloned fragments at least as large as 17 kbp and potentially 40 kbp.

We have made two single-copy, amplifiable vectors suitable for this strategy, and have validated and made reliable a previously described method for generating nested deletions enzymatically. In both vectors, clones are stably maintained at low copy number by the F replication and partitioning functions and can be amplified from an IPTG-inducible P1 lytic replicon to prepare DNA. A synthetic version of the phage f1 origin of replication is located a short distance upstream of the multiple cloning site. Vector pND-1 is used primarily for obtaining clones by transformation or electroporation; pND-2 has phage lambda cos sites that allow efficient cloning of 30-40 kbp fragments in a lambda packaging system.

Reaction conditions have been defined where purified f1 gene 2 protein efficiently introduces a strand-specific single nick in the f1 origin sequence with very little rejoining. Large amounts of stable gene 2 protein are easily obtained from a clone by a rapid purification procedure we developed. To create the nested deletions, the nick is expanded unidirectionally into the cloned fragment by 3' to 5' digestion with *E. coli* Exo III, the resulting single-stranded regions are digested with S1 nuclease, and the ends are repaired and ligated with T4 DNA polymerase and ligase. The Exo III digestion is highly synchronous and processive, and the deletion lengths are proportional to incubation time. To prevent undeleted DNA from giving rise to clones, the treated DNA is digested with one of several restriction enzymes whose 8-base recognition sequences lie between the fl origin and the cloning site. Nested deletion clones are then obtained by electroporation.

Pooling samples from several different times of Exo III digestion before subsequent treatment generates a good distribution of deletion clones. Growth and amplification of randomly selected clones in 1 ml of medium in 96-well format followed by a simple DNA preparation protocol provides ample DNA for analyzing deletion length by gel electrophoresis and for DNA sequencing reactions. Imaging and sizing software is now being tested for automated selection of an appropriate set of deletions for sequencing.

Exploratory work has demonstrated the effectiveness of this nested deletion strategy for sequencing fragments at least as large as 17 kbp cloned from a human BAC.

Trapping of DNA in Non-Uniform Oscillating Electric Fields

Charles L. Asbury, **Ger van den Engh**
Department of Molecular Biotechnology,
University of Washington, Seattle, WA 98105

We present a method, analogous to optical trapping, which allows manipulation of DNA molecules in aqueous solution. Due to the induction of an electric dipole, DNA molecules are pulled by a gradient force to regions of high electric field strength. With the use of very thin gold films and oscillating fields, molecules can be trapped individually or locally concentrated. The molecules do not become permanently attached to the gold. Spatial control over the trapped molecules is achieved because they are confined to a width of ~5 mm perpendicular to the edges of the gold films. It is possible to mix static and oscillating electric fields in order to move trapped molecules from one edge to another, or to make them follow very precise trajectories along the edges. This phenomenon may be useful in microdevices for manipulation of small quantities or single molecules of DNA.

Energy Transfer Fluorescent Primers Optimized for DNA Sequencing and Diagnostics

Su-Chun Hung*, Yiwen Wang**, Richard A. Mathies** and **Alexander N. Glazer**
Departments of Molecular and Cell Biology* and Chemistry**, University of California, Berkeley, CA 94720

Energy transfer (ET) primers are markedly superior to single dye-labeled primers in DNA sequencing, and in multicomponent genetic analyses, such as forensic identification, and genetic typing of short tandem repeats.[1,2] We describe here improvements in the properties of ET fluorescent primers that enhance their value in all of these applications. In designing improved ET primers, we focused on the spectroscopic characteristics most important for DNA sequencing and PCR product analysis: the relative acceptor fluorescence emission intensity and the amount of residual donor fluorescence emission, and on improving the match in electrophoretic mobilities of the DNA fragments extended from ET primers. Hung et al.[3] showed that 3-(e-carboxypentyl)-3'-ethyl- 5,5'-dimethyloxa-carbocyanine (CYA or C), a dye with a high absorption cross-section but a low fluorescence quantum yield, is superior to FAM (F) as a donor in ET primers and that 6-carboxyrhodamine-6G (R6G or G) and 5&6-carboxyrhodamine-11) (R110) can serve as alternative acceptors to JOE and FAM, respectively, in ET primers. CYA-labeled ET primers have stronger acceptor emission and higher spectral purity than ET primers with a FAM donor. The ET primer set C10R110, C10G, C10T, and C10R with only rhodamine derivatives (R110, R6G, TAMRA, and ROX) as the four acceptors has superior spectroscopic properties and gives results in DNA sequencing with near-perfect match in the electrophoretic mobility of single-base extension DNA fragments both in capillary and slab gel electrophoresis.[3-5] ET primers with CYA as donor and a donor-acceptor spacing of 4-6 nucleotides offer excellent acceptor emission intensities coupled with negligible donor emissions. In multiplex separations, this allows precise quantitation of the ratio of the signals from DNA fragments labeled with one or another of two different primers. We have applied the two-color ET primer sets C4G/C4R and C6G/C6R to bladder cancer diagnosis based on electrophoretic analyses of polymerase chain reaction-amplified short tandem repeats (STRs) where the diagnosis depends on the detection of loss of heterozygosity at particular loci. The success of the analysis depends on the accurate multiplex quantitation of the amplified DNA fragments from two different samples, normal and tumor cell DNA. Multiplex analyses with the two-color primer sets C4G/C4R and C6G/C6R allowed quantitative determination of allelic ratios with a precision of ±10%.[6] We performed a quantitative comparison of sets of primers differing in the nature of the donor-acceptor combinations, CYA-ROX, FAM-ROX, and BODIPY503/512-BODIPY581/591. Variables examined included the length of the 5'-amino linker arm, the number of base pairs between the donor and acceptor, and the excitation wavelength (488 or 514 nm). Of the primers examined, CYA-ROX primers offer the best combination of acceptor fluorescence emission intensity and spectral purity.[7]

Supported by the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under Contract DE-FG-91ER61125. Financial support from the Amersham Life Science Inc. is also gratefully acknowledged.

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. S-C. Hung, J. Ju, R.A. Mathies, and A.N. Glazer Anal. Biochem. 243, 15-27 (1997).
4. S-C. Hung, J. Ju, R.A. Mathies, and A.N. Glazer Anal. Biochem. 238, 165-170 (1996)
5. S-C. Hung, R.A. Mathies, and A.N. Glazer Anal. Biochem. 251, (1997) - in press.
6. Y. Wang, S-C. Hung, J.F. Linn, G. Steiner, A.N. Glazer, D. Sidransky, and R.A. Mathies Electrophoresis 18, (1997) - in press.
7. S-C. Hung, R.A. Mathies, and A.N. Glazer Anal. Biochem. - in press.

Direct PCR Sequencing with Boronated Nucleotides

Kenneth W. Porter, Dima Sergueev, Ahmad Hasan, J. David Briley and **Barbara Ramsay Shaw**
Department of Chemistry, Duke University, Durham, NC 27708

DNA can be simultaneously amplified and sequenced using a new class of nucleotides containing boron. During the polymerase chain reaction, boron-modified nucleotides, i.e. 2'-deoxynucleoside 5'-alpha-[P-borano]-triphosphates, are incor-porated into the product DNA. The boranophosphate linkages are resistant to nucleases and thus the positions of the boranophosphates can be revealed by exonuclease digestion, thereby generating a set of fragments that terminate in a boranophosphate linkage and define the DNA sequence. The boranophosphate method offers an alternative to current PCR sequencing methods. Single-sided primer extension with dideoxynucleotide chain terminators is avoided, with the consequence that the sequencing fragments are derived directly from the original PCR products. Boranophosphate sequencing has been demonstrated with the Pharmacia and the Applied Biosystems 373A automatic sequencers, producing data that is comparable to cycle sequencing.

The method has been improved recently by streamlining sample preparation and by employing modified boranophosphate nucleotides. Sample preparation is streamlined by implementing direct digestion of the PCR products immediately after amplification. The base-specific PCR products are combined and digested by direct addition of Exonuclease III and Phosphodiesterase I. Subsequently, the digestion products are purified by spin column chromatography, concentrated by isopropanol precipitation, and separated by gel electrophoresis. The uniformity of the digestion products is increased by the addition of modified boranophosphates to the PCR amplification and by the addition of Phosphodiesterase I to the digestion mixture.

Additionally, we have synthesized a 2'-deoxyadenosine alpha-borano-triphosphate with a fluorescent tag attached at the C-8 position of adenine via an alkylamine. Experiments are in progress to determine its ability to be incorporated during amplification and its nuclease resistance, and to develop a screen that will allow for rapid evaluation of other naturally occurring or mutant exonucleases.

The boron method 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 should be skewed to long fragments. Direct sequencing of PCR products simplifies bidirectional sequencing and provides a simple, direct, and complementary method to cycle sequencing.

(Supported by DOE grant 94-ER61882 and continuing grant 97ER-62376 to B.R.S.)

Advances in DNA Analysis Using Capillary Array Electrophoresis

Indu Kheterpal, James R. Scherer, William W. Ja, Yiwen Wang and **Richard A. Mathies**
Department of Chemistry, University of California, Berkeley, CA 94720

Recent advancements in DNA analysis with capillary array electrophoresis (CAE) have included (i) the optimization of sample preparation, sample loading and separation matrices for DNA sequencing, (ii) the development of a scanner for detecting separations in large numbers of capillaries (~1000) in parallel, (iii) the development and evaluation of new 3-color extended binary coding strategies, and (iv) the development of rapid and sensitive methods for high-throughput cancer screening:

(1) Four-color confocal fluorescence CAE scanners using our standard flat bed design[1] are being used along with energy transfer (ET) primers for sequencing of mitochondrial (mt) DNA, *Chlamydia*, and *Anabaena*. Twelve motifs of the hypervariable region I of human mt DNA from a Sierra Leone population have been sequenced with an average accuracy of 99.7%.[2] Over 40 kilobases of *Chlamydia* clones have been resequenced and a 8 kilobase fragment has been assembled with 99.63 % accuracy. Several genes from *Anabaena* have also been sequenced as a part of an undergraduate research project and fragments of up to 8 kilobases have been assembled.

(2) A capillary array scanner (CAE) capable of acquiring four-color data at the rate of 4 Hz from over 1000 capillaries has been constructed. The scanner has a rotating objective which excites and collects fluorescence from one to 1088 capillaries which are positioned in precisely machined grooves (spaced at 260 microns) in a cylindrical objective housing. Four-color data from four photomultipliers is obtained simultaneously with four independent ADCs. The acquired data is stripped of non-sample gaps, averaged across each capillary and also averaged across a variable number of successive rotations as it is acquired. The instrument has been designed to use replaceable matrices introduced by pressure filling. This scanner will be evaluated through sequencing of *Chlamydia* in collaboration with Ron Davis' group at Stanford.

(3) New three-color extended binary coding strategies for multiplex DNA sequencing have been developed using ET primers.[3] These three-color methods are found to be nearly as good as traditional four-color coding methods with sequencing accuracy rates of 99.6% and 99.9%, respectively. Methods have been developed to deconvolve the three-color data into the four base concentrations. This three-color approach is the first step towards the development of higher order multiplex coding schemes for DNA sequencing and other analyses.

(4) Finally, short tandem repeat (STR)-based bladder cancer diagnosis methods have been developed using two-color labeling with ET primers and CAE electrophoresis.[4] Rapid (< 35

min.) separations are achieved on capillary arrays using replaceable separation matrices and the allelic ratios are quantitatively determined with a precision of 10%. These methods provide a significant improvement in the speed, ease and precision of STR analyses.

Supported by the Director, Office of Energy Research, Office of Health and Environmental Research of the U. S. Department of Energy under Contract DE-FG-91ER61125.

1. R. A. Mathies, X. C. Huang, Nature (London), 359, 167-169, 1992.
2. I. Kheterpal, J. R. Scherer, S. M. Clark, A. Radhakrishnan, J. Ju, C. L. Ginther, G. F. Sensabaugh, R. A. Mathies, Electrophoresis, 17, 1852-1859, 1996.
3. I. Kheterpal, L. Li, T. P. Speed, R. A. Mathies, Anal. Chem., submitted
4. Y. Wang, S. -H. Hung, J. F. Linn, G. Steiner, A. N. Glazer, D. Sidransky and R. A. Mathies, Electrophoresis 18, in press, 1997.

Development of a Low Cost, High Throughput, Four Color DNA Sequencer

Michael S. Westphall, David R. Rank and Lloyd M. Smith
University of Wisconsin-Madison, Department of Chemistry, 1101 University Ave., Madison, WI 53706
mwestpha@whitewater.chem.wisc.edu

It has been recognized from the beginning of the Human Genome project that in order to successfully complete this tremendous undertaking the development of new and improved technologies for DNA sequencing would be required. We have focused on several aspects of this technology development from automated Front End sample preparation to base calling of collected data. Central to the development of this and any other automated sequencing system is the electrophoresis platform. To meet the electrophoresis throughput requirements of our automated Front End DNA sample preparation system, we have developed a 4 color fluorescent sequencing instrument emphasizing long read lengths, dense sample loading, and low construction cost.

The system employs a 250 micron thick cross-linked polyacrylamide gel, 10 inches wide x 24 inches long which is temperature regulated on both sides. A scanning four color detection system is employed to collect the emitted fluorescence. Data is collected bi-directionally with one of four bandpass filters in position per scan. Image processing and base calling is performed using GelImager and BaseFinder software packages developed in our lab.

The system provides sufficient resolution to base-call DNA fragments beyond 1000 bases in length (albeit requiring high quality template preparations and sequencing chemistry) and to routinely deliver runs with 750 bases of useable sequence (98% accuracy). The system currently process 88 samples in parallel and can be built at a cost of $23,000. Instrument design details will be presented along with performance characteristics (in combination with our analysis software) as obtained through in-house sequencing projects.

Multiplexed Integrated On-line System for DNA Sequencing by Capillary Electrophoresis: From Template to Called Bases

Edward S. Yeung* and Hongdong Tan
Ames Laboratory-USDOE and Department of Chemistry, Iowa State University, Ames, IA 50011
yeung@ameslab.gov

DNA sequencing as practiced today involves a series of steps starting from isolating DNA from the biological sample, cutting these into fragments with convenient sizes, amplifying the fragments biochemically, introducing a label for detection while generating a nested set of ordered fragments, separating the ordered set of fragments and identifying the nucleotide sequence, and reassembling the short sequence data into a continuous sequence. Recent developments of capillary electrophoresis, especially in multiplexed arrays, show great promise for substantially increasing the speed and throughput of the separation and identification steps. The issue of cost when combined with the other steps in the whole sequencing process still remains. Our project is aimed towards the development of novel front-end strategies, whereby the speed and throughput of sample preparation can be significantly increased while the amount of manual operation and total cost can be significantly reduced. We will present our latest results on multiplexed sample preparation in small volumes without the use of robotics. This promises to reduce the cost of reagents and at the same time provide high-speed high- throughput operation.

An integrated on-line prototype for coupling a microreactor to capillary electrophoresis for DNA sequencing has been demonstrated. A dye-labeled terminator cycle-sequencing reaction is performed in a fused-silica capillary. Subsequently, the sequencing ladder is directly injected into a size-exclusion chromatographic column operated at ~95 °C for purification. On-line injection to a capillary for electrophoresis is accomplished at a junction set at ~70 °C. High temperature at the purification column and injection junction prevents the renaturation of DNA fragments during on-line transfer without affecting the separation. The high solubility of DNA in and the relatively low ionic strength of 1x TE buffer permit both effective purification and electrokinetic injection of the DNA sample. The system is compatible with highly efficient separations by a replaceable poly(ethylene oxide) polymer solution in uncoated capillary tubes.

We will present data from an 8-capillary system, where individual templates are simultaneously injected from standard microtiter wells. After injection, a completely computer controlled system takes the samples through terminator-labeled cycle sequencing, purification, introduction to 8 parallel electrophoresis capillaries for separation and detection, and base calling. Scaling up to 100 simultaneous channels is straightforward. Future research should allow starting from single bacterial colonies injected into each capillary and the nucleotide sequence identified in a fully on-line and automated system, perhaps multiplexed to 1000 at a time.

Production Sequencing Evaluation of a Nearly Automated 96-Capillary Array DNA Sequencer Developed at LBNL

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

As the Human Genome Project moves towards a scale-up of its sequencing phase, it is apparent that gel electrophoresis will remain the main technology for DNA sequencing. Although slab gel electrophoresis is currently the dominant technology used in production-level sequencing, recent rapid progress in capillary electrophoresis technology suggests that it soon will have the capability of replacing slab gel sequencer in production work. At LBNL we have developed a beta-test version of a 96-capillary system capable of production level sequencing at increased rates and more importantly, improved automation. The system is based on an adaptation of the best available technology currently being developed by several laboratories. In particular, we employ a sheath-flow excitation/detection geometry derived from earlier work by Norman Dovichi at the University of Alberta. Our system, fully integrated from the assembling of the capillary array to automated DNA sequencing, to base-calling, has demonstrated the effectiveness of the sheath-flow approach using 96-channel array, with a separation speed of 700 bases/hour/channel. The instrument's operation has been fully computerized with automatic control of the laser, high voltage power supply, data acquisition and processing. The gel replacement and sample loading were nearly automated, and the sequencing cycle time was less than 2 hours. A standardized baseline protocol for capillary coating and filling and sample preparation has also been developed. Currently, this system has undergone a series of beta-testing runs in the production-sequencing environment at LBNL's sequencing Center. By loading our system with the sample remainders left by production sequencing at LBNL and directly comparing our sequencing results with those generated by the

ABI-377 machines, we have found that a comparable sequencing quality has been achieved by our instrument. Details of results will be represented.

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

DNA Sequencing by Capillary Electrophoresis from Sample Preparation to Resulting Sequence: A Robust High Throughput Procedure with Long Read Length

Salas-Solano, O., Carrilho, E., Ruiz-Martinez[1], M.C., Goetzinger[2], W., Kotler, L., Sosic, Z., and **Karger, B.L.**
Barnett Institute and Department of Chemistry, Northeastern University, Boston, MA, 02115
bakarger@lynx.dac.neu.edu

Capillary electrophoresis (CE) is being developed in our lab and elsewhere for high speed automated DNA sequencing. However, while several multichannel sequencing instruments are currently under development by different groups, the significant issues of robustness and high overall throughput of the DNA sequencing analysis have not been adequately addressed. We are developing a fully automated multicapillary DNA sequencing system, starting from sample preparation to DNA sequence generation. To reduce the failure rate to a minimum and, hence, increase the throughput of such a system, proper care is necessary for every step of the process. We have developed a robust protocol for thorough purification of DNA samples that includes both desalting and template removal. This procedure also resulted in 10-fold increase in the amount of DNA sequencing fragments compared to conventional desalting by ethanol precipitation or gel filtration columns. The purification procedure is economical, very reproducible and compatible with different sequencing chemistries. Conditions for electrokinetic injection of the purified samples have also been optimized. With respect to the column, we previously found that long read length sequencing runs may be achieved using 2% w/w linear polyacrylamide (LPA) with a molecular weight close to 9 MDa. A new, convenient method for high molecular weight LPA preparation using inverse emulsion polymerization has been developed. With this procedure, LPA forms a white powder of unlimited shelf life and allows fast and reproducible preparation of working polymer solutions. Additionally, base calling software has been improved to increase the accuracy of extended read length (see a separate abstract of A.W. Miller and B.L. Karger).

We are currently working on a fully automated multicapillary array DNA sequencing system, which will include a robotic sample preparation and purification system, separation matrix replacement and sample injection automated devices. We will present our latest results in this area. The described sequencing technology advances will greatly increase overall throughput of the automated multi-channel capillary DNA electrophoresis system.

This work is being supported by DOE grant #DE-FG02-90ER60895.

[1] Present address: Curagen Corp., Branford, CT 06504
[2] Present address: Arqule Inc., Medford, MA 02155

A Fully Automated 96-Capillary DNA Sequencer

Qingbo Li, Thomas E. Kane, Changsheng Liu, John Kernan, and James R. Hoyland
Premier American Technologies Corp.
qbli@aol.com

A high throughput DNA sequencer is constructed, where the instrument operation (e.g., sample introduction, DNA separation, instrument reconditioning, data processing) is carried out and tightly controlled by the instrument computer. The complete instrument consists of two units. The main unit houses the optical detection system, the 96-capillary cartridge, the mechanical system for sample introduction, and the associated electronic controllers. The other unit is the liquid handling module dedicated to gel delivering and capillary reconditioning. The two units interface through a liquid conducting tubing. In the detection system, an air-cooled argon ion laser is efficiently coupled with the optics to excite fluorescence from all 96 capillaries. The instrument is portable. In the sample introduction system, a carrousel assembly allows automatic processing of seven 96-well sample trays without human intervention.

The high throughput arises primarily from four advantages: (1) the instrument design that allows complete automatic operation; (2) the use of 96 discrete capillaries, with which 96 separate DNA samples can be analyzed simultaneously; (3) the use of a high speed CCD camera that allows simultaneous monitoring of fast separation in all 96 capillaries; (4) the use of dilute replaceable gel matrix that minimizes the time required for gel filling and capillary reconditioning. Preliminary sample runs of 400 - 500 DNA bases per capillary have been achieved, yielding 38,400 - 48,000 bases read per instrument run. Including the time for sample introduction, separation, and capillary reconditioning, the PATCO prototype is capable of one complete run within two hours. This sample analysis throughput is comparable to the demands of the Human Genome Project.

The 96-capillary prototype will be available by the end of the year. In the next stage, the instrument will be scaled up to 384 capillaries for 4X improvement in the instrument throughput.

Development of a Microchannel Based DNA Sequencer

Courtney Davidson, Joseph Balch, Larry Brewer, Joe Kimbrough, Steve Swierkowski, David Nelson, Ramkrishna Madabhushi, Ron Pastrone, Ann Lee, Paula McCready, Aaron Adamson, Bob Bruce, Ray Mariella, and Anthony Carrano
Lawrence Livermore National Laboratory, Human Genome Center
davidson4@llnl.gov

We are developing instrumentation for DNA sequencing based on the use of an array of microchannels fabricated on a glass substrate. Arrays with up to 101 microchannels, 48 cm long have been fabricated in plates of borosilicate float glass. Since last reporting we have further improved our microfabrication process to etch channels of arbitrary width and depth. We have etched substrates with 12, 24, and 101 channels per plate varying from 150 - 200 um wide and 30 - 60 um deep. The microchannels are constructed with two plates of glass (nominally 7.5 cm x 58 cm) which are fusion bonded. The channel plate is bonded to a top plate that has the input and output ports for sample introduction and buffer reservoir interconnects. Channels of various cross section sizes have been built and tested using low viscosity solutions of linear poly(dimethylacrylamide). This sieving media dynamically coats the walls of the glass microchannels to significantly reduce the electro-osmotic flow so that sequencing separations can be done in uncoated channels. Also, the relatively low viscosity of the sieving media allows it to be pumped into the channels via a simple syringe pump. This syringe pump is coupled directly to a common output port of the microchannel plate so that all channels can be filled with sieving media in one simple pump operation. A linear scanning confocal PMT-based detection system is used to detect the laser induced four color fluorescence from the DNA fragments. The data acquisition and control system is based on a Pentium class personal computer running LabVIEW. Data compression and transfer, signal analysis, and basecalling routines have been developed in S-PLUS and C on a Sun Microsystems Ultra Enterprise 2 server. Recent experimental results of microchannels 60 um deep by 250 um wide and having a 38 cm load-to-read length resulted in an electrophoretic resolution greater than 400 bases in about 90 min. for a 160 V/cm separation field. A 24 microchannel plate is presently operational providing electrophoretic resolution of 450 to 500 bases. Currently we are building and assembling a 96 channel system based on this technology into an "alpha-phase" DNA sequencing instrument which will be networked with the Sun for data analysis and base calling. We will report on-going results obtained from it for high throughput DNA sequencing.

Work is supported by a grant from the National Center for Human Genome Research, National Institutes of Health and by The Department of Energy Human Genome Program. Work was performed by Lawrence Livermore National Laboratory under auspices of the U. S. Department of Energy under contract no. W-7405-ENG-48.

Microchannel Process Development and Fabrication for DNA Sequencing

Steve Swierkowski, Joseph Balch, Courtney Davidson, and Lisa Tarte
Lawrence Livermore National Laboratory, Human Genome Center
swierkowski1@llnl.gov

We have developed a process for the production of microchannel arrays on single glass substrates as an alternative electrophoresis technology to arrays of discrete capillaries for DNA sequencing. This technology approach provides a number of advantages for building large arrays of electrophoresis microchannels for DNA sequencing. By fabricating the array of microchannels on a single glass substrate, the arrays of microchannels are very robust mechanically and can be handled without any special care. By means of photolithography and chemical etching techniques the dimensions of rectangular cross-section channels can be optimized by making the channel depth thin to minimize the thermal dispersion of DNA bands while at the same time the channel width can be made large to increase the amount of dye-labeled DNA available for strong fluorescence signal generation and detection. The detection of the fluorescence signal is also made easier by having a flat optical window over the channels through which laser excitation of fluorescence occurs with less scattered light of the primary laser beam to contribute to the overall noise level.

Microchannel arrays for electrophoresis with up to 101 channels, 48 cm long have been fabricated in plates of borosilicate float glass. The channels are constructed with two plates of glass that are 7.6 cm wide and 58 cm long and are fusion bonded at 650°C. The channel plate is 5 mm thick and typically has 12, 24, or 101 channels per plate; the channels are 150 - 200 um wide and about 30 - 60 um deep. The channel plate is bonded to a top plate that has the input and output ports in it. Two different types of input ports have been tested. For a 5 mm thick top plate, input ports 1 mm in diameter have been ultrasonically milled through the top plate and registered with the channels before the bonding process. For a 1.2 mm thick top plate, input ports as small as 150 um have been fabricated(about the same as the channel diameter) and these have been registered to within 20 um accuracy to the channels before bonding.

The patterning of these plates employs simple contact printing with flat panel display industry type photomasks onto standard photoresists. Special apparatus was constructed to coat the plates with photoresist and also to expose them with a simple contact printing method. A critical procedure was developed to eliminate microscopic damage to the glass before processing begins and to clean the glass at the beginning of the processing. This special procedure was essential to reduce the microchannel etching defects by many orders of magnitude that would have otherwise rendered the plates useless for high resolution genome sequencing. Extensions of this fabrication technology should enable very high(400 - 500) channel count plates to be made that would, in turn, greatly increase the throughput and efficiency of sequencing instruments.

Work is supported by a grant from the National Center for Human Genome Research, National Institutes of Health and by The Department of Energy Human Genome Program. Work was performed by Lawrence Livermore National Laboratory under auspices of the U. S. Department of Energy under contract no. W-7405-ENG-48.

An Optical Trigger for Locating Microchannel Position

Laurence R. Brewer, Joseph Kimbrough, Courtney Davidson, and Joseph Balch
Lawrence Livermore National Laboratory, Human Genome Center
brewer1@llnl.gov

We have developed a technique for automatically aligning a microchannel plate with a scanning fluorescence detector in a high throughput DNA sequencer. An optical signal from each microchannel can be used to dramatically reduce the amount of data collected while further eliminating the effects of hysteresis and velocity variation of the scanning motor. While the system can be run in a high resolution mode (~2800 points per scan) for diagnostic purposes such as evaluating band shape, the described technique can be adapted for the collection of a single data point per channel per scan. Such a reduction is prerequisite for practical application to high throughput DNA sequencing. The technique makes use of the difference in reflection of the air-glass, gel-glass, and bonded glass-glass interfaces present in the microchannel plate. A thin piece of glass is used to collect back reflected light from laser illuminated microchannels and sent to a photodiode for detection. This signal delineates the position of the microchannels with a high signal to noise ratio and is used as an electronic trigger for data collection.

Work is supported by a grant from the National Center for Human Genome Research, National Institutes of Health and by The Department of Energy Human Genome Program. Work was performed by Lawrence Livermore National Laboratory under auspices of the U. S. Department of Energy under contract no. W-7405-ENG-48.

Multiple Capillary DNA Sequencer Illuminated by a Waveguide

Mark A. Quesada, Harbans S. Dhadwal, David J. Fisk, Janine S. Graves and F. William Studier
Biology Department, Brookhaven National Laboratory, Upton, New York 11973
quesada@genome1.bio.bnl.gov

Capillary electrophoresis through a replaceable polymer matrix has great promise for improving the speed and efficiency of DNA sequencing if many different capillaries can be analyzed simultaneously. However, illumination of the interiors of multiple capillaries and collection of the emitted fluorescence from each is complicated by the cylindrical shape and small radii of curvature of the capillaries. We have solved this problem by using the capillaries themselves as optical elements in a waveguide.

Refraction of light at the surfaces of a capillary depends on the radii of curvature of the capillary walls and the change in refractive index in crossing each surface. With appropriate dimensions and refractive indices, refractive effects can confine a beam of light to pass through the interior of each successive capillary in a parallel array. The illuminating beam must be in the plane of the array and normal to the first capillary, have minimal divergence, and have a radius comparable to or smaller than the internal radius of the capillary. This condition is readily achieved by delivering the beam through an integrated fiber optic transmitter.

Losses of light due to reflection at each successive capillary surface can be minimized by reducing the differences in refractive index across the surfaces while still satisfying the conditions for beam confinement. Illuminating with coaxial beams from opposite ends of the array also improves the uniformity of illumination. Using commercially available materials, it would be feasible to make 96-capillary waveguide sequencers that would be expected to have only a few percent variation in the intensity of illumination of each capillary across the entire array. The fluorescence can be collected by an array of optical fibers whose spacing is identical to that of the capillaries and whose ends are positioned normal to the capillaries. This configuration allows simultaneous alignment of all of the collection fibers.

A 12-capillary prototype sequencing apparatus was fabricated and tested, validating the key elements of this design. The capillaries are illuminated efficiently with only 30 mWatts of laser power, and the fluorescence is efficiently collected by the matched optical fibers with little cross-talk between channels. The collected light is delivered to a spectrograph and the full fluorescence spectrum of all the capillaries in parallel is displayed on the surface of a CCD and read into a computer about 3 times per sec. Replacement of capillaries and alignment of the system is very simple.

Samples are electrokinetically injected simultaneously into all 12 capillaries from a single row of a microtiter plate. A dimethylpolyacrylamide polymer matrix is used in uncoated capillaries. Our current preparations can be used for about 30 sequencing runs per capillary before performance begins to degrade. We are developing base calling software based on Bayesian analytical methods. The system currently reads almost 500 bases per capillary when run at room temperature, with a cycle time of about 2 hr.

Technology Development and Informatics Tools for Genome Sequencing

E.R. Mardis, L. Hillier, A. Chinwalla, M. Cook, M. Holman, G. Marth, R. McGrane, D.A. Panussis, D.C. Peluso, L.L. Rifkin, J.E. Snider, J. Strong, E. Stuebe, R.H. Waterston, M. Wendl, R.K. Wilson
Washington University School of Medicine, Genome Sequencing Center, St. Louis, MO 63108
emardis@alu.wustl.edu

Efficient performance of high throughput, large-scale genome sequencing projects depends upon improvements to techniques and devices for their performance that streamline data production, as well as computer tools to complement these improvements. Thus, the ability to study processes and apply the appropriate modifications, devices and/or informatics tools is critical to our ability to increase productivity and efficiency. Several examples of technology development and informatics contributions to our processes will be highlighted. These include efforts to increase sample capacity on DNA sequencers, improve sample loading processes, and streamline liquid transfer steps, as well as scripts and algorithms to reduce time spent on gel retracking, to facilitate sequence data entry and to examine assembled projects.

Single Molecule DNA Sequencing

Hong Cai, Peter M. Goodwin, James H. Jett, **Richard A. Keller**, Nicholas P. Machara, Susan L. Riley, and David J. Semin
Los Alamos National Laboratory, Los Alamos, New Mexico 87545
keller@lanl.gov

Our flow-cytometry based approach to DNA sequencing involves: (1) labeling DNA fragments with base-specific fluorescent tags; (2) suspending a single labeled fragment in a flow stream; (3) digesting with an exonuclease that sequentially cleaves the end nucleotide and releases it into the flow stream; (4) detecting and identifying the individual cleaved nucleotides as they pass in order of cleavage, through a focused laser beam.

We have implemented an optical trap to suspend the DNA laden microsphere upstream from the detection laser. This resulted in: reduced background; improved detection efficiency; and simplified sample handling. With this system, our detection efficiency of labeled nucleotides is ~ 90% and false positive signals from the background are a few per second. Enzymatic cleavage rates with Exo III are ~ 5 per second at 37 °C on fluorescently labeled substrates. Progress towards a two color sequencing demonstration will be described.

On the Road to Rapid Exonuclease Screening for DNA Sequencing

Hong Cai, # Susan L. Riley, # Kristina Kommander, * John Nolan, * **Richard A. Keller#**
Los Alamos National Laboratory, Chemical Science and Technology# and Life Sciences*

The limitation of current automated sequencers is 1800 base pairs/day, which results in the need for 4600 machine years to create the first finished sequence of one human genome (~3 x 109 bases pairs). A rapid laser-based technique for sequencing of 10 kb or larger fragments of DNA at a rate of 100 to 1000 bases per second is being developed in our laboratory. Successful completion of this would greatly reduce the time and effort needed in the sequencing of the human genome and other genomes. This new method relies on the attachment of fluorescent labeled DNA to a microsphere, introduction of this microsphere into a flowing sample stream, and detection of the individual labeled nucleotides as they are cleaved from the DNA fragment by an exonuclease. In order to increase analysis rates, an exonuclease with a fast digestion rate is required.

Extensive testing of commercially available exonucleases in our laboratory has not revealed a suitable exonuclease capable of rapidly cleaving fluorescently-labeled DNA. This has lead us to search for either mutant forms of exonucleases or exonucleases derived from other bacterial strains. In order to screen the numerous exonucleases, a rapid screening assay based on flow cytometry has been developed. Compared to conventional techniques, this new assay is sensitive, rapid, and requires no radiation labeling or separation. This will allow the screening of hundreds of samples per day.

Sizing of Individual DNA Fragments

Zhengping Huang, **Yongseong Kim**, Jonathan L. Longmire, Nancy C. Brown, James H. Jett and Richard A. Keller
Los Alamos National Laboratory, Los Alamos, New Mexico 87545
kimys@lanl.gov

Our flow-cytometry based approach to DNA fragment sizing involves: (1) staining a restriction digest of DNA with a dye that intercalates stoichiometrically with the fragments such that the number of incorporated dye molecules is proportional to the fragment length; (2) diluting the sample to ~ 10-13 M; passing the sample through our single molecule detection apparatus and (3) measuring the magnitude of the fluorescence from individual, stained fragments. A histogram of the fluorescence intensities gives the size distribution of the DNA fragments, i.e. a DNA fingerprint. Samples containing less than a femtogram of DNA are sized in minutes with an accuracy of ~ 98%. We have demonstrated the applicability of this technique for sizing DNA fragments as small as 212 bp and as large as 340 kbp. In comparison with pulsed-field gel electrophoresis for the sizing of large DNA fragments, this approach is more accurate, much faster, requires much less DNA, and is independent of the DNA conformation. Applications to the characterization of PAC and BAC clones for DNA library construction and identification of bacteria strains by their DNA fingerprint will be described.

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

We have shown that single-molecule fluorescence burst counting is a highly sensitive method for detecting electrophoretic separations of ds-DNA fragments.[1] In previous work, methods for optimizing dye labeling, laser power and data analysis were developed, which enabled detection of single DNA fragments as small as 100 bp in capillary electrophoresis separations.[2] These separations were detectable when only 50-100 molecules passed through the probe volume. To further enhance the applicability of this method to low level pathogen and mutation detection, we have now successfully performed single molecule detection of DNA separations in microfabricated glass capillary electrophoresis (CE) chips. By fabricating CE chips with a 280 mm thick top cover plate and by using a 40X NA 1.3 immersion microscope objective, the S/N ratio for single molecule detection was enhanced by more than two-fold over conventional capillaries. By constricting the sample in the detection region to a ~10 mm wide by ~5 mm deep cross section, approximately 10% of the molecules could be probed by the ~2 mm wide focused laser beam. Cross channel and separation channel dimensions were systematically varied to optimize the injection of the DNA sample. We have now achieved a detection limit of 500 molecules on-column or 200 attograms/ml for 500 bp DNA fragments.[3] This accomplishment is important because DNA-based methods are becoming increasingly important in environmental monitoring and in health care diagnostics to detect trace levels of pathogen contamination or DNA mutation.

Supported by the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under Contract DE-FG-91ER61125.

[1] B. B. Haab and R. A. Mathies, "Single molecule fluorescence burst detection of DNA fragments separated by capillary electrophoresis," Anal. Chem. 34, 3253-3260 (1995).
[2] B. B. Haab and R. A. Mathies, "Optimization of single molecule fluorescence burst detection of ds-DNA: Application to capillary electrophoresis separations of 100-1000 bp fragments," Appl. Spec., 51, N10 (1997).
[3] B. B. Haab and R. A. Mathies, in preparation.

Sizing Megadalton DNA by Mass Spectrometry

W. Henry Benner
LBNL, 1 Cyclotron Road, Berkeley, CA 94720
whbenner@lbl.gov

Mass spectrometry has important potential applications in the measurement of molecular masses relevant to the goals of the Human Genome Project. Thin gels, capillary gels and DNA chip technology appear to offer improvements in DNA analysis rates over slab gels but "mass spectrometry has perhaps demonstrated the greatest near-term potential [to increase through-put]." To the extent that direct instrumental measurements of molecular size could replace gel electrophoresis as a routine tool for DNA-sequencing separations and general molecular biology experimentation, a significant saving in time could be effected. Electrophoresis gels take hours to run but mass spectra are acquired in seconds to a few minutes. Additionally, mass spectrometers produce a mass measurement compared to gel electrophoresis which separates ions according to mobility, a relative measurement.

Electrospray mass spectrometry is an important type of mass spectrometry applicable to DNA analysis because it does not break apart DNA molecules. For relatively small electrospray DNA ions, the different charge states provide a way to calculate mass but for large ions with numerous charge states the calculation of ion mass is precluded because the charge states are not resolved in most mass spectrometers. We recently demonstrated the direct detection of charge on large electrospray ions as a way to solve this problem so that the mass of large DNA ions could be measured.

A patent application has been submitted for the invention of charge-detection-mass-spectrometry. More recently, we have significantly improved the measurement capability of this technique by implementing this detection system in a new type of ion trap. The gated electrostatic trap consists of a detector tube mounted between two sets of ion mirrors. The mirrors define symmetrically-opposing potential valleys which guide axially-injected ions to cycle back and forth through the charge-detection tube. A low noise charge-sensitive amplifier, connected to the tube, reproduces the image charge of individual ions as they pass through the detector tube. Ion mass is calculated from measurement of ion charge and velocity following each passage through the detector. Individual ions carrying more than 250 charges at an energy of 200 eV/charge have been trapped for 10 ms corresponding to 450 cycles through the detector tube. At this level of trapping time, a theoretical precision for charge measurement as small as 2 electrons RMS can be achieved. The operation of the system was demonstrated by trapping a 4.3 kilobase long circular DNA molecule of bacterial plasmid pBR322. The sodium form of this molecule has a molecular weight of 2.88 MDa. A mass value of 2.79 +/- 0.09 (ave +/- s.d.) MDa was determined. The accuracy of the mass measurements and the speed of this technique suggest that this measurement approach could be applied to the routine sizing of cloning vectors for the purpose of quality control of the cloning process.

Evaluation of New Membrane Surfaces for Chemiluminescent Assays

Christopher S. Martin, Jing Ying Lee, Betty Liu, Jeffrey Shumway, John C. Voyta and Irena Bronstein
Tropix, Inc., 47 Wiggins Ave., Bedford, MA 01730
chrisma@tropix.com

Nylon membrane is the preferred support for a multitude of molecular biology applications due to its robustness and retension of high levels of bioanalytes. Chemiluminescent detection with 1,2-dioxetane substrates on nylon membrane is facilitated by the enhancement properties of nylon, but limited by high levels of non-specific binding of enzyme labeled reagents. We have developed polymers of quaternary amines for enhancement of chemiluminescence intensity from 1,2-dioxetanes in solution. These polymers also improve membrane assays when added to the substrate buffer. Poly(benzyltributyl)ammonium chloride(TBQ)(SapphireAE II) and Poly(benzyldimethylvinylbenzyl)ammonium chloride (BDMQ) (SapphireAE I) are current Tropix chemiluminescent enhancer products. Various quaternary amine polymers were screened to determine if membranes coated with such polymers would exhibit superior performance compared to commercially available nylon, PVDF, or nitrocellulose supports. Due to the superior chemiluminescent enhancement properties of THQ (Polyvinylbenzyltrihexylammonium chloride), this polymer was used to overcoat different membrane supports. After overcoating with THQ, biotinylated DNA was detected with a chemiluminescent dot blot assay. Polyethersulfone membranes exhibited the best performance in these assays. Recently, we have collaborated with an outside membrane manufacturer to bench cast membranes with THQ polymer. These membranes were subsequently tested by performing chemiluminescent detection of biotin or fluorescein labeled DNA in dot blot and Southern assays. The results indicate that sensitive detection and low background signal are attained on these membranes. Development of a superior membrane for chemiluminescent assays is of great benefit, enabling more rapid imaging of signals with x-ray film and electronic imaging devices (i.e. CCD cameras).

This work was funded by the DOE Genome Program.
Contract No. DE FG05 92ER81389

Advances in Microfabricated Integrated DNA Analysis Systems

Adam T. Woolley, Peter C. Simpson, Shaorong Liu, Kaiqin Lao, Stephen J. Williams, Mary X. Tang, Lester Hutt, Alexander N. Glazer and **Richard A. Mathies**
Department of Chemistry and Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720

The microfabrication of DNA sample preparation, electrophoretic analysis and detection devices is making possible a new generation of high-speed, high-throughput DNA analysis systems. Our early work showed that high-quality fragment sizing as well as DNA sequencing separations could be performed on microfabricated capillary electrophoresis (CE) chips.[1,2] We also demonstrated that PCR amplification could be directly performed on our CE chips to make the first integrated DNA sample preparation and analysis devices.[3] It was also possible to increase the throughput of these microdevices by making capillary array electrophoresis (CAE) chips that could carry out parallel genotyping separations of up to 12 samples on a single chip in under 160 seconds.[4] Recent advances in the use of replaceable denaturing separation matrices, in injection methodology, and in channel fabrication now enable DNA sequencing separations on chips with single base resolution to =500 bases in only 12 minutes. Improvements in fabrication permit the construction of larger defect-free devices on 10 cm diameter glass wafers.[5] We have also developed (i) novel injection modules for serially introducing multiple (2-4) samples onto the same capillary, (ii) elastomer sample well arrays for the facile loading of up to 96 samples, and (iii) electrode arrays for addressing up to 96 samples. These improvements have led to the development of a CAE chip that can separate 96 DNA fragment samples in less than 8 minutes using 48 parallel capillaries, each capable of analyzing two different samples.[6] These analyses have all been performed by using high-sensitivity, laser-excited confocal fluorescence detection. However, to produce truly portable high-speed microdevices it is desirable to eliminate expensive and bulky optical components and laser systems. Toward this end we have been working on the development of integrated electrochemical detection systems for CE chips.[7] In these devices, the working electrode, formed by RF sputter deposition of Pt (2500 Å) on a 200 Å Ti adhesion layer, was photolithographically placed =30 um outside the end of the separation channel to avoid interference from the separation field. Electrophoretic separations of neurotransmitters were performed in under 100 seconds demonstrating the speed, resolution and attomole detection sensitivity of these devices. Indirect electrochemical detection of DNA fragment separations was performed by using the redox-active intercalator Fe(1,10-phenanthroline)32+ in the separation buffer; transient dips in the constant background current from free intercalator indicated the migration of DNA-intercalator complexes through the detection region. On-chip electrochemical detection provided excellent sensitivity (=103 molecules) for rapid (=200 s) and high-quality separations of DNA restriction fragments and PCR products. This work is the harbinger of a paradigm shift in the application of CE chips to genomic sequencing and analysis.

Supported by the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under Contract DE-FG-91ER61125.

[1] Woolley, A. T.; Mathies, R. A. Proc. Natl. Acad. Sci., USA 91, 11348-11352 (1994).
[2] Woolley, A. T.; Mathies, R. A. Anal. Chem. 67, 3676-3680 (1995).
[3] Woolley, A. T.; Hadley, D.; Landre, P.; deMello, A. J.; Mathies, R. A.; Northrup, M. A. Anal. Chem. 68, 4081-4086 (1996).
[4] Woolley, A. T.; Sensabaugh, G. F.; Mathies, R. A. Anal. Chem. 69, 2181-2186 (1997).
[5] Simpson, P. C.; Woolley, A. T.; Mathies, R. A. BioMEMS 1, in press (1997).
[6] Simpson, P. C.; Roach, D.; Thorsen, T.; Johnston, R.; Sensabaugh, G. F.; Mathies, R. A. manuscript in preparation (1997).
[7] Woolley, A. T.; Lao, K. Glazer, A. N.; Mathies, R. A. submitted for publication (1997).

The Flowthrough Genosensor Program at Oak Ridge National Laboratory

Kenneth Beattie, Mitchel Doktycz, Ming Zhan, Gabriel Betanzos, William Bryan, John Turner
Oak Ridge National Laboratory, Oak Ridge, TN 37831-6123
beattiekl@ornl.gov

A flowthrough genosensor instrument for ultrahigh throughput analysis of gene structure and function is being developed. The core of this microscale instrumentation is a microchannel hybridization array, containing a library of thousands of specific DNA sequences, immobilized within microporous cells in a thin layer of silicon or glass. A nucleic acid sample is passed through the microchannel genosensor at precisely controlled temperature and flow rate, and each porous hybridization cell binds nucleic acid sequences that are complementary to the immobilized DNA probe. The quantitative binding pattern reflects the base sequence of the nucleic acid strands present in the analyte and reveals the relative abundance of different sequences. The porous glass configuration has several important advantages over the flat surface DNA chip being developed by others: greatly improved hybridization kinetics, superior detection sensitivity, the ability to analyze dilute solutions of nucleic acids, and direct detection of heat-denatured PCR fragments without prior isolation of single strands.

The prototype genosensor system includes a temperature-controlled fluidics module and a CCD imaging system for quantitation of hybridized fluorescent-labeled strands. A key objective in the project is to develop important applications of the flowthrough genosensor for analysis of gene structure and function and DNA diagnostics. Feasibility studies for key genosensor applications are being pursued in collaboration with the mouse genetics program at ORNL and with several other organizations. In one application area, a series of miniature "genochips" containing arrays of genomic DNA fragments are being prepared for use in gene discovery and mapping. Another application area, being pursued via a CRADA with Gene Logic, Inc. (Columbia, MD) and in collaboration with Dr. Jeffrey Trent (NHGRI) employs flowthrough arrays of DNA probes for transcriptional profiling, facilitating the discovery of genes that function in specific biological processes. A third application of the flowthrough genosensor, led by Dr. Mitch Doktycz, involves model hybridization studies with defined nucleic acid sequences, aimed at providing a more complete understanding of the specificity of oligonucleotide hybridization, which will facilitate intelligent selection of probes and valid interpretation of hybridization patterns. A fourth application of the flowthrough genosensor, being pursued in collaboration with Dr. James Weber at the Marshfield Medical Research Foundation, is high throughput genotyping. In this work miniature flowthrough genosensors will be fabricated to simultaneously analyze thousands of biallelic single nucleotide polymorphisms.

In another approach, 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. These reactions will include: (1) Cycle sequencing reactions – Sequencing primers, nucleotides and Taq polymerase are flowed into each of the porous glass "microreaction cells," then the wafer will be sealed and placed into a thermocyler to carry out the cycle sequencing reactions. Products will be eluted and analyzed by Dr. Joe Balch at LLNL, using his parallel microcapillary array electrophoresis apparatus. This process will be carried out in numerous successive cycles, each time with a new set of primers, to acquire a large amount of sequence information from each BAC. The sequencing primers can be selected from oligonucleotide libraries as suggested by Ulanovsky and others, to achieve rapid primer walking along the entire set of BACs. (2) Mapping of expressed sequences in BAC libraries – Libraries of BACs immobilized in the channel glass array will be hybridized with individual cDNA clones or ESTs to localize each expressed sequence across the BAC array. This process will be repeated with numerous expressed sequences, to achieve rapid assignment of expressed sequences to their genomic clones.

Technical Aspects of Fabrication and Quantitative Analysis of DNA Micro-Arrays for Comparative Genomic Hybridization

Damir Sudar, Steve Clark, Ian Poole, Rick Segraves, Stephen Lockett, Arthur Jones, Donna Albertson, Joe Gray, Daniel Pinkel
Lawrence Berkeley National Laboratory
d_sudar@lbl.gov

We have developed a method of performing comparative genomic hybridization (CGH) to microarrays of genomic DNA clones (cosmids, P1s, BACs, etc.) that permits high resolution detection and mapping of DNA copy number variations in the human genome (Albertson et al., this meeting). In array CGH, spots of cloned DNA are arrayed onto a microscope slide and hybridized with total genomic DNA from a test specimen, labeled with one fluorochrome, and reference genomic DNA labeled with a spectrally different fluorochrome. The ratio of the fluorescence intensities on each target clone is proportional to the relative copy number of those sequences in the test and reference genomes. Efficient implementation of array CGH requires overcoming several major technical challenges including production of high density arrays, and rapid quantitative readout and analysis of the fluorescence signals. We have made considerable progress in both of these areas.

We are developing a robotic arraying system with a multi-pin tool to print DNA target solutions from 864 well plates onto fused silica slides mounted on a precision X-Y stage. The pins, each made from a segment of capillary electrophoresis tubing are spring mounted on 3 mm centers. The upper end of each pin is connected to a manifold by flexible tubing to permit pressurization for printing and suction for cleaning. We have demonstrated the feasibility of printing targets on 100 um centers using this approach, providing a density of 10,000 targets / cm2.

Our imaging system was designed to provide: (a) large field-of-view (1cm2, to match the print format), (b) sufficient resolution (10 pixels per target diameter), and (c) high sensitivity (accurate quantitation down to single-copy DNA quantities). A digital CCD camera and 2 high-speed lenses in a back-to-back configuration image the sample at 1x magnification. Fluorescence is excited using a mercury arc lamp with filter wheel for wavelength selection. A fiber-optic delivery system illuminates the sample at a 45 degree angle through the back side of the slide. A right-angle fused-silica prism is used in "total internal reflection" conditions to efficiently excite the fluorescent dyes without allowing excitation light to enter the imaging optics. A multi-band barrier filter was used in the emission path so it did not have to be changed when imaging different fluorochromes. We acquire a DNA counterstain (DAPI) image and the test and reference images with an exposure of several seconds or less. Software automatically identifies the spots using the DAPI counterstain and measures the background-corrected fluorescence intensities and intensity ratios of the spots.

We evaluated the performance of the system and analysis software using test samples made by spiking 200 ng of total human genomic DNA with 0, 1, 2, 20, 200, and 2000 pg of lambda DNA and a reference sample consisting of 200 ng of total genomic DNA spiked with 20 pg of lambda DNA. In this situation 3 pg of lambda DNA is equivalent to a single copy sequence. We found that the changes in the fluorescence ratios were detectable from below single copy equivalent level, and were quantitatively proportional to DNA sequence copy number over three orders of magnitude.

Partially supported by the Director, Office of Energy Research, Office of Health and Environmental Research of the U.S. Department of Energy under contract NO. DE-AC03-76SF00098.

Multilabel SERS Gene Probes for DNA Sequencing

T. Vo-Dinh[*], D.L. Stokes[1], G.D. Griffin[1], N. Isola[1], J.P. Alarie[1], U.J. Kim[2], M.I. Simon[1], T. Bunde[1]

[*] Corresponding Author; Oak Ridge National Laboratory, Oak Ridge, TN 37831-6101
[1] Advanced Monitoring Development Group, Life Sciences Div., Oak Ridge National Laboratory, Oak Ridge, TN 37831-6101
[2] Div. of Biology, Calif. Inst. of Technology, Pasadena, CA 91125
tvo@ornl.gov

We describe a new type of DNA gene probe based on surface-enhanced Raman scattering (SERS) label detection. Raman spectroscopy is a useful tool for chemical analysis due to its excellent capability of chemical group identification. One limitation of conventional Raman spectroscopy is its low sensitivity, often requiring the use of powerful and costly laser sources for excitation. However, a renewed interest has recently developed among analytical spectroscopists as a result of observation that Raman scattering efficiency can be enhanced by factors of up to 10(8) when a compound is adsorbed on or near special metal surfaces. The technique associated with this phenomenon is known as Surface-Enhanced Raman Scattering (SERS) spectroscopy. The surface-enhanced Raman gene (SERGen) probes described here do not require the use of radioactive labels and have great potential to provide both sensitivity and selectivity for DNA sequencing. The SERGen probes can be used to detect DNA biotargets (e.g., gene sequences, bacteria, viral DNA fragments) via hybridization to DNA sequences complementary to that probe. The use of stable clone resources containing large human DNA insets has opened new possibilities to contig building for the Human Genome Project. The objective of this research is to apply the SERS multi-label technique for use in DNA mapping and bacterial artificial chromosomes (BAC) colony hybridization. The technology is based on a system that will integrate several concepts including: i) multi-label SERS detection, ii) spectral multiplex mapping, and iii) BAC colony hybridization.

Emphasis is on detection techniques that minimize the time, expense and variability of preparing samples by combining the BAC mapping approach with SERS "label multiplex" detection. Large numbers of DNA samples can be simultaneously prepared by automated devices. With this device, multiple samples can be separated and directly analyzed using multiple SERS labels simultaneously. The results demonstrate the feasibility of the SERGen approach in the detection of two gene probes simultaneously.

MicroArray of Gel Immobilized Compounds

A. Mirzabekov
Joint Human Genome Program: Argonne National Laboratory, U.S.A., and Engelhardt Institute of Molecular Biology, Moscow

Technologies for robotic manufacturing of Microarrays of Gel-Immobilized Compounds on a chip (MAGIC™ chip) have been developed and the MAGIC™ chips are being tested for various applications. These microchips are polyacrylamide gel pads fixed and separated on a glass surface by hydrophobic spacers. The microchips can be used like an array of micro test tubes in which chemical and biochemical reactions can be carried out separately in each gel pad. Different oligonucleotides, DNA antibodies, and proteins have been immobilized in specified gel pads to produce oligonucleotide, DNA, and protein microchips. The usual size of the gel pads is 100x100x20 m. However, microchips with gel element sizes as small as 10x10x10 m can be produced by photopolymerization. Applications of oligonucleotide microchips have been demonstrated for detection of mutation, identification of microorganisms, HLA allotyping, DNA fractionation, DNA enzymatic phosphorylation and ligation on specified or all microchip elements, and thermodynamic analysis of DNA duplexes. Generic microchips containing all 4,096 possible 6 mers have been manufactured, and their use for de novo sequencing and DNA sequence analysis will be described.

Flow Cytometry-Based Polymorphism Detection and Analysis

John P. Nolan, Hong Cai, Kristina Kommander, and P. Scott White
Los Alamos National Laboratory, Life Sciences Division, Los Alamos, New Mexico 87545
nolan@telomere.lanl.gov

Single-base polymorphism analysis of the human genome on a large scale requires robust and sensitive screening methods which are amenable to automation and high throughput analysis. We are developing a suite of microsphere-based approaches employing fluorescence detection by flow cytometry to screen for and analyze single base polymorphisms. One approach being developed for polymorphism detection is to immobilize on microspheres proteins which recognize specific DNA structures. When a fluorescently labeled DNA molecule binds to the immobilized protein, the binding can be measured by flow cytometry. An example is the recognition of heteroduplex DNA by the mutS protein as a means to detect single base mismatches. To analyze nucleotide base frequencies at a polymorphic site, we are developing an approach based on minisequencing in which immobilized primers designed to interrogate a specific site are used to bind the region of interest in an unknown sample. The primers are then extended by polymerase using fluorescent ddNTPs and flow cytometry is used to read the frequency of each differently colored base. Alternately, the ligation of fluorescent oligonucleotides containing base variation at the site of interest is used to detect the base frequency at that site. Apart from the advantages of sensitivity and low sample consumption, the flow cytometric approaches have the advantages of the potential for multiplexed analysis using different color or size beads and automated sample handling. Multiplexed analysis could enable simultaneous analysis of base frequencies at dozens of different loci, which combined with automated sample handling could provide a powerful tool for high throughtput screening of single base polymorphisms. Supported by NIH, DOE, and LDRD.

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

David S. Wunschel, Ljiljana Pasa Tolic, David. C. Muddiman, James E. Bruce, Steve A. Hofstadler and **Richard D. Smith**
Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA 99352
rd_smith@pnl.gov

Mass spectrometry offers the potential for high speed DNA sequencing and other applications. In addition to the development of sequencing approaches, ongoing work in the laboratory is exploring applications using Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. These efforts include the characterization of polymerase chain reaction (PCR) products, enzymatically produced oligonucleotide mixtures, modified DNA and the development of methods for the analysis of DNA large fragments. The analysis time required is on the order of seconds, and is made possible by isotopic resolution of each component's charge states obtained using FTICR. High mass accuracy measurements for PCR products have been achieved for products up to 114 base pairs in length. As an example, the mass accuracy allowed single base substitutions to be detected with de novo identification of an unreported base substitution (1,2). This approach was extended to examine a multi-component reaction from a single primer pair where a base pair deletion was identified with the putative identification of inter- variability within a single bacterial strain (3). Recent efforts have focused on increasing the size of products amenable to analysis with a goal of providing comparable "read-lengths" to traditional sequencing methods, and have involved improvements to sample preparation methods and the exploitation of improved methods for dynamic range expansion. We are also exploring the use of collision induced dissociation methods with PCR products to provide sequence information. This would allow for direct selection and analysis of individual components from within mixtures that may share a high degree of similarity without cloning. This alternative would not only eliminate that time-consuming step, but also potentially allow identification of low abundance products without an intensive screening process.

In this presentation the recent advances at our laboratory will be described. These include the development of new interfacing methods for realizing greatly improved sensitivity and the implementation of new high performance FTICR that has been designed to achieve greater sensitivity as well as resolution and mass measurement accuracy.

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. "Length and Base Composition of PCR-Amplified Nucleic Acids Using Mass Measurements from Electrospray Ionization Mass Spectrometry", D. C. Muddiman, G. A. Anderson, S. A. Hofstadler and R. D. Smith, Anal. Chem. 69, 1543-1549 (1997)

3. "Inter-Operon Variability in B. cereus by Normal PCR using ESI-FTICR Mass Spectrometry", D. S. Wunschel, D. C. Muddiman, K. F. Fox, A. Fox and R. D. Smith, submitted.

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.

Improved Mass Spectrometric Resolution for PCR Product Size Measurement

Gregory B. Hurst, Kristal Weaver, and Michelle V. Buchanan
Chemical and Analytical Sciences Division, Oak Ridge National Laboratory, Oak Ridge TN 37831-6365
hurstgb@ornl.gov

While a wealth of biological and genetic information can be gleaned from properly-designed polymerase chain reaction (PCR) assays, currently-used technologies for analysis of the resulting oligonucleotides all suffer from limitations in speed, accuracy, safety, or convenience. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) offers considerable potential for rapid and accurate molecular mass determination of biopolymers, such as proteins and DNA. In order to achieve this potential, we are working to improve the utility of MALDI-MS for measurement of PCR product size.

The resolution (ability to distinguish products of similar molecular mass) of MALDI-MS is determined by both chemical and instrumental factors. The presence of reaction components necessary for polymerase activity, particularly metal ions, causes broadening of the observed peaks in MALDI mass spectra of PCR products. Post-PCR removal of these metal ions and other interferences can be performed efficiently using reverse-phase cartridges in syringe-mounted or microtiter plate formats. The wide energy distribution imparted to biomolecules by the laser desorption process also broadens mass spectrometric peaks. Delayed ion extraction has greatly improved the resolution of MALDI-MS by compensating for this energy spread. Combining these techniques, PCR products differing in length by a single base can be resolved up to a total length of 60 bases or more, and larger oligonucleotides can be detected if single- base resolution is not required. The design of PCR products that are shorter than typically used for gel electrophoresis is thus a high priority in improving the applicability of MALDI-MS.

Other important practical considerations are the reproducibility and throughput of MALDI-MS. Commercially-available instrumentation allows robotic loading onto multiple-sample plates followed by automated analysis. However, samples outside narrow constraints of concentration, size, and purity still require human intervention because of the inhomogeneity of the dried matrix/sample mixture. We are currently developing methods for combining oligonucleotides with the MALDI matrix to yield a homogenous preparation resulting in a uniform signal at any point interrogated by the desorption laser.

K.W. acknowledges support through an appointment to the Oak Ridge National Laboratory Postgraduate Research Program administered jointly by the Oak Ridge Institute for Science and Education and Oak Ridge National Laboratory. Research supported by the Environmental Management Science Program and 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.

Laser Desorption Mass Spectrometry for Quick DNA Sequencing and Analysis

C. H. Winston Chen, N. I. Taranenko, Y. F. Zhu, S. L. Allman, V. V. Golovlev and N. R. Isola
Life Science Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6378

During the past two years, we have used laser desorption mass spectrometry (LDMS) to obtain the following major achievements. They are (1) LDMS sequencing of ss-DNA longer than 100 nucleotides with DNA ladders (2) Direct DNA sequencing without ladders (3)LDMS for hybridization detection (4) STR detection for forensic applications and (5) Rapid disease diagnosis.

For conventional gel electrophoresis for DNA sequencing, major steps include (1) DNA ladders preparation (2) Separation of different sizes of DNAs and (3) detection by either autoradiogram or laser-induced fluorescence. Laser desorption mass spectrometry (LDMS) can be used to enhance DNA sequencing speed. One approach is to use a mass spectrometer as a detector only. In general multiplexing is used to increase the sequencing speed. However, gel electrophoresis and DNA ladders preparation are still required. Another approach is to use LDMS for both separation and detection. With this approach, gel electrophoresis is not needed but the preparation of DNA ladders is required. Recently, we succeeded in using LDMS in sequencing single-stranded DNA with the size longer than 100 nucleotides. With primers for both strands, a double-stranded DNA segment with the size up to 260 base pairs can be sequenced. Both cycle sequencing and standard Sanger's sequencing have been tried with successful results.

Since the preparation of DNA ladders is somewhat time consuming, it is very desirable to be able to sequence DNAs without the need of ladder preparation. We recently found that preferred bond cleavage can be obtained during the laser desorption process if adequate matrices and laser fluences are used. We took this approach and recently succeeded in sequencing an oligonucleotide with 35 bases. This direct sequencing by MALDI without the need to prepare DNA ladders can be conveniently used to sequence primers and short DNA probes.

In addition to the DNA sequencing, we alsc apply LDMS for the detection of DNA probes for hybridization. Preliminary results indicate that LDMS as detector for hybridization process can reduce the time and cost for DNA sequencing by hybridization (SBH). LDMS was also used to obtain short tandem repeats (STR) for forensic applications. Clinic applications for disease diagnosis such as cystic fibrosis due to the base deletion and point mutation have also been demonstrated. Different schemes for resolution and detection efficiency improvements will be pursued in the future to increase the sequencing and/or analysis speed. Experimental details will be presented in the meeting.

Research has been sponsored by the Office of Health and Environmental Research, U.S. Department of Energy under contract DE-AC05-84OR21400 with Lockheed Martin Energy System, Inc.

Genome Analysis Technologies

George Church, Martha Bulyk, Sonali Bose, Chris Harbison, Linxiao Xu, Poguang Wang, Laura Kutney, T. O'Keeffe, Dereth Phillips
Department of Genetics, Harvard Medical School, 200 Longwood Ave., Boston, MA 02115
church@salt2.med.harvard.edu

Together with Bruker Instruments Inc., Genome Therapeutics Corp., and Northeastern University, we have developed a laser-desorption electron capture mass spectrometry method to quantitate up to 400 "Mass-tags" every 10 to 200 milliseconds. Both DNA and proteins have been mass-tagged in ways analogous to fluorescent-tags. The tags are laser-heat-releasable electrophores. Advantages over fluorescence include more numerous cleanly separated spectral peaks, better detection, and higher throughput. Applications to capillary electrophoretic (CE) assays, DNA microarray chips, in situ hybridization, and in situ PCR imaging are under evaluation. The electrophoretic applications are designed (but not yet tested) to collect 100 sequence base pair equivalents per second. To assess the capacity of our CE (75 micron inner diameter), we have obtained dideoxy sequence data from a multiplex PCR sequence of a mixture of 47 ds-templates. Another feature of the Mass-tag is the 400 internal standards, which should enhance the ability to computationally align and quantitate 400 images for chip and in situ applications. For the new complete microbial genome sequences, we have developed chip and CE technologies for systematic analyses of DNA-protein interactions and competitive growth phenotypes.

L. Xu, N. Bian, Z. Wang, S. Abdel-Baky, S. Pillai, D. Magiera, V. Murugaiah, R.W. Giese, P. Wang, T. O'Keeffe, H. Abushamaa, L. Kutney, G.M. Church, S. Carson, D. Smith, M. Park, J. Wronka, F. Laukien. Electrophore Mass Tag

Dideoxy DNA Sequencing. Analytical Chemistry in press.
http://arep.med.harvard.edu/

2'-Fluoro Modified Nucleic Acids: Polymerase-Directed Synthesis, Properties, and Stability to Analysis by Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry

Tetsuyoshi Ono, Mark Scalf, Lloyd M. Smith
University of Wisconsin-Madison Chemistry Dept.
tetsu@whitewater.chem.wisc.edu

DNA fragmentation is a major factor limiting mass range and resolution in the analysis of oligonucleotides by Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry (MALDI-MS). Protonation of the nucleobase leads to base loss and backbone cleavage by a mechanism similar to the depurination reactions employed in the chemical degradation method of DNA sequencing. In a previous study, the stabilizing effect of substituting the 2' hydrogen with an electronegative group such as hydroxyl or fluorine was investigated. These 2' substitutions stabilized the N-glycosidic linkage, blocking base loss and subsequent backbone cleavage. For such chemical modifications to be of practical significance, it would be useful to be able to employ the corresponding 2'-modified nucleoside triphosphates in the polymerase-directed synthesis of DNA. This would provide an avenue to the preparation of 2'-modified PCR fragments and dideoxy sequencing ladders stablilized for MALDI analysis. In this paper methods are described for the polymerase-directed synthesis of 2'-fluoro modified DNA, using commercially available 2'-fluoronucleoside triphosphates. The ability of a number of DNA and RNA polymerases as well as reverse transcriptase to incorporate the 2'-fluoro analogs was tested. Four thermostable DNA polymerases (Pfu (exo-), Vent (exo-), Deep Vent (exo-) and UlTma) were found that were able to incorporate 2'-fluoronucleotides with reasonable efficiency. In order to perform Sanger sequencing reactions, the enzymes' ability to incorporate dideoxy terminators in conjunction with the 2'-fluoronucleotides was evaluated. UlTma DNA polymerase was found to be the best of the enzymes tested for this purpose. MALDI analysis of enzymatically produced 2'-fluoro modified DNA using the matrix 2,5-dihydroxy benzoic acid showed no base loss or backbone fragmentation, in contrast to the extensive fragmentation evident with unmodified DNA of the same sequence.

Mapping and Resources

Development and Application of Subtractive Hybridization Strategies to Facilitate Gene Discovery

Maria de Fatima Bonaldo, Greg Lennon and **Marcelo Bento Soares**
The University of Iowa, Iowa City, IA
bento-soares@uiowa.edu

The methods we originally developed to normalize directionally cloned cDNA libraries (Soares et al., 1994; Bonaldo, Lennon & Soares, 1996) have been successfully utilized to generate a number of human, mouse and rat cDNA libraries. All human and mouse libraries (and soon the rat libraries as well) have been contributed to the I.M.A.G.E. consortium and they have been extensively used for large scale generation of expressed sequence tags (ESTs). Both the ESTs and their respective clones are publicly available. Although the use of normalized libraries has proven most advantageous to minimize the redundant identification of the mRNAs of the super-prevalent and intermediate frequency classes within a particular tissue, it cannot prevent the redundant identification of mRNAs (of any frequency class) that are expressed in multiple tissues. In other words, normalization alone cannot avoid the redundant identification of ESTs that have been obtained previously from other libraries. This problem is becoming increasingly more relevant as we approach completion of the ongoing human and mouse gene discovery efforts. Hence, we proposed to take advantage of subtractive hybridization strategies that we developed, to generate libraries enriched for novel cDNAs. Briefly, pools of I.M.A.G.E. clones from which ESTs have been derived, are used as drivers in hybridizations with single or multiple normalized libraries thus generating subtracted libraries enriched for cDNAs not yet represented in public databases. Subtracted libraries are characterized by Southern hybridization to assess reduction in representation of clones of the driver population and then contributed to the I.M.A.G.E. consortium for large-scale arraying and sequencing. Sequence analysis of two subtracted libraries that we have generated indicated a four-fold reduction in representation of the driver population. It is anticipated that the use of subtracted libraries will become increasingly advantageous as we strive towards the ultimate goal of identifying all human and mouse genes. This project has two major goals: (1) to optimize the method for construction of subtracted libraries, and (2) to generate subtracted libraries to facilitate the ongoing human and mouse EST programs.

EURO-IMAGE: the European IMAGE Consortium for Integrated Molecular Analysis of Human Gene Transcripts

Auffray, C.[1], Devignes, M.D.[1], Pietu, G.[1], Ansorge, W.[2], Schwager, C.[2], Ballabio, A.[3], Borsani, G.[3], Banfi, S.[3], Estivill, X.[4], Lynch, M.[4], Gibson, K.[5], Mundy, C.[5], Lehrach, H.[6], Poustka, A.[7], Wiemann, S.[7], Korn, B.[7], O'Brien, J.[7], Uhlen, M.[8], Lundeberg, J.[8]
[1] CNRS UPR 420, BP 8, 94801 Villejuif cedex, France
[2] EMBL, Postfach 10.2209, 6900 Heidelberg, Germany
[3] TIGEM, Via Olgettina 58, 20132 Milano, Italy
[4] IRO, Hospital Duran I Reynals, Autovia de Castelldefels 2,7, Barcelona, Spain
[5] HGMP, Hinxton, CB10 1RQ, United Kingdom
[6] MPI, Ihnerstrasse 73, Dahlem, 14195 Berlin, Germany
[7] DKFZ, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[8] KTH, Teknikringen 34, 10044 Stockholm, Sweden
auffray@infobiogen.fr

The general objectives of the European IMAGE Consortium are:

- To generate a minimal set of non redundant cDNA clones for most human gene transcripts and a master set of unique full-length cDNA clones representing 3,000 transcripts based upon the IMAGE Consortium resources (arrays of cDNA and CpG islands clones).
- To characterize by DNA sequencing with high accuracy the complete sequence of the master set of 3,000 human gene transcripts (6 Mbases of finished sequence).

- To obtain high resolution and comparative functional mapping localization in man and model organisms of 1,000 genes represented in the master set.
- To develop the IMAGE Consortium Data Base to provide an easy access to an integrated view of the sequence, map and expression data generated.

The European IMAGE Consortium will devote 20% of the resources to the assembly of the physical resources (cDNA and CpG island arrays characterized by end sequencing and fingerprinting, minimal sets of clones selected after comprehensive sequence, clone and functional clustering, master set of 3,000 full-length clones). These resources will be available throughout the European Union for the user community in both academy and industry. The Consortium will serve the future needs of the scientific community in the systematic identification of all human genes and their regulatory sequences by deciphering in an efficient and economic manner 6 Mb of complete, finished sequence for 3,000 transcripts of average size 2 kb: 50% of the resources will be devoted to the sequencing of full-length cDNAs and CpG islands selected by Consortium on the basis of their map position, similarity to known families or expression profiles. In order to ensure that advances in basic genetic knowledge is used to further enhance human health, the Consortium will seek to contribute to the identification of genes involved in human biology and diseases by correlating precise map location and phenotypic expression data, exploiting various comparative approaches for 1,000 of the genes represented in the master set: 10% of the resources will be devoted to high-resolution and comparative functional mapping in close interaction with the mapping consortia in order to obtain the most precise and evolutionary relevant map location. Last but not least, 20% of the resources will be devoted to the IMAGE Consortium Data Base. This will provide the community with up to date, integrated sequence, mapping and expression data related to the IMAGE consortium arrays, as they are collected by IMAGE Consortium members in Europe, the United States and Japan, and will help in sharing and harmonizing such data.

Supported by the participating institutions and the European Union BIOMED 2 Program.

The Functional Genomics Initiative at Oak Ridge National Laboratory

Dabney Johnson, Monica Justice, Ken Beattie, Michelle Buchanan, Michael Ramsey, Rose Ramsey, Michael Paulus, Nance Ericson, David Allison, Reid Kress, Richard Mural, Ed Uberbacher, **Reinhold Mann**
Life Sciences, Instrumentation and Controls, Chemical Technology, and Robotics and Process Systems Divisions, Oak Ridge National Laboratory, Oak Ridge TN 37831
mannrc@ornl.gov

The Functional Genomics Initiative at the Oak Ridge National Laboratory integrates outstanding capabilities in mouse genetics, bioinformatics, and instrumentation. The 50 year investment by the DoE in mouse genetics/mutagenesis has created a one-of-a-kind resource for generating mutations and understanding their biological consequences. It is generally accepted that, through the mouse as a surrogate for human biology, we will come to understand the function of human genes. In addition to this world class program in mammalian genetics, ORNL has also been a world leader in developing bioinformatics tools for the analysis, management and visualization of genomic data. Combining this expertise with new instrumentation technologies will provide a unique capability to understand the consequences of mutations in the mouse at both the organism and molecular levels.

The goal of the Functional Genomics Initiative is to develop the technology and methodology necessary to understand gene function on a genomic scale and apply these technologies to megabase regions of the human genome. The effort is scoped so as to create an effective and powerful resource for functional genomics. ORNL is partnering with the Joint Genome Institute and other large scale sequencing centers to sequence several multimegabase regions of both human and mouse genomic DNA, to identify all the genes in these regions, and to conduct fundamental surveys to examine gene function at the molecular and organism level. The Initiative is designed to be a pilot for larger scale deployment in the post-genome era. Technologies will be applied to the examination of gene expression and regulation, metabolism, gene networks, physiology and development.

The initiative was launched in 1996 and is comprised of the following component efforts:

Directed High-Efficiency Mouse Mutagenesis (Monica Justice) - We have established a comprehensive high-efficiency mouse germline mutagenesis program to examine mammalian gene function. N-ethyl-N-nitrosourea (ENU) is currently the mutagen of choice because it induces point mutations that reflect single gene function. Large genomic regions can be scanned for mutations that may reflect loss of function, gain of function, or partial loss of function. In parallel, we are developing sequence-ready BAC contigs of a region that we are mutagenizing. Allelic series for loci will be obtained that will complement other mutagenesis approaches, such as gene disruptions. Our program incorporates many different methods of genetic screening to isolate mutations, and the mutations will serve as a resource to the mouse and human genome communities.

The Mouse Screenotyping Center (Dabney Johnson) - We have established a mouse phenotype-screening center to complement and extend current mouse mutational analyses by developing screening protocols for biochemical and behavioral abnormalities. High-throughput methodologies are being developed to efficiently screen for induced aberrations in behaviors, in locomotor and neuromuscular function, and in biochemical and hematological parameters. The screens are designed to detect mouse models of a variety of human diseases from among large populations of potentially mutant mice.

Comprehensive Cellular Protein Analysis using Microfluidic Devices (Rose Ramsey) - ORNL's unique lab-on-a-chip technology is being combined with mass spectrometric methods to provide uniquely powerful methods for comprehensive cellular protein analysis using microfluidic devices. The microchips will integrate on a single structure, elements that enable multidimensional separations of protein mixtures with electrospray ionization of the analytes for direct, on-line interfacing with mass spectrometry. This system will provide unparalleled throughput and sensitivity, allowing robust protein identification in small samples.

Gene Expression Studies using Genosensor Microchip Technology (Ken Beattie) - Porous glass flowthrough Genosensor chip technology provides unique hybridization surface area and sensitivity for use in large-scale gene expression studies, mapping, and mutational surveillance.

Genome Lab-on-a-Chip (Michael Ramsey) - Microfluidic and microinstrumentation capabilities are being applied to labor intensive assays used in gene mapping and analysis of genetic mutations and polymorphisms. "Laboratory-on-a-chip" technology is being used to automate assay procedures, increase analysis rates, eliminate the use of radioactive isotopes, and reduce the consumption of limited DNA samples. Current efforts are focused on development of multiplexed microdevices for PCR-based analysis of simple tandem repeat markers used in the genetic and physical mapping of the mouse genome.

Mouse Physiological Monitoring Microchip (Nance Ericson) - Unique integrated circuitry is being created to provide a capability for large scale monitoring of mouse physiological parameters related to physiological or behavioral abnormalities due to mouse mutations. Application of this new technology to genome studies will accelerate mass specimen screening by providing automated detailed observation and reporting of multiple key physiological parameters, such as body temperature, heart rate, physical activity level, location and motion patterns using a microscale implanted device.

High-Throughput Tomographic Imaging of Mouse Phenotypes (Michael Paulus) - A new high-resolution high-throughput automated 3D mouse imaging and screening methodology is being created which will rapidly identify, quantify and record subtle phenotypes in mutagenized mice, based on a novel x-ray imaging technology with 50 micron spatial resolution and a 3D dataset acquisition time of <1 minute per mouse.

Development of DNA Sequencing and Automated Genotyping Infrastructure (Richard Mural) Basic capabilities for DNA sequencing and automated genotyping are being developed to support mouse genetics and molecular biology. Not only are these capabilities necessary for addressing basic problems in genetic and molecular biology but they will also and flexibility to the program for initiating new projects. Having the DNA sequence of regions which are being mutagenized will also be important to mutation detection.

Event-Based Flexible Automated DNA Sequencing (Reid Kress) - By moving automated DNA sequencing from a sequential, batch-mode process to a continuous mode hands off process we will be able to maximize the efficiency and cost-effectiveness of our limited DNA sequencing resources at ORNL.

Bioinformatics Support for Gene Function (Ed Uberbacher) - A comprehensive system to collect and assemble information relevant to the function of newly discovered genes is being constructed for application to sequenced human and mouse regions in the Functional Genomics Initiative. This includes support for initial gene finding and informatics-based characterization, and compilation of experimental information generated related to mouse mutations, gene and protein expression studies, and computational and experimental information derived from numerous external databases and other model organisms. The resulting catalog of gene function information will be linked to genome-wide browsing being produced by the Genome Annotation Consortium.

Direct Visualization of Regulatory Protein-DNA complexes and Mutations using AFM (David Allison) - High-throughput atomic force microscopy will be used to precisely locate and visualize proteins bound to individual DNA molecules or genes, including complexes related to regulation of gene expression and repair enzymes site-specifically bound to lesion sites. Automated image analysis and interpretation systems are being applied to locate and characterize the protein-DNA complexes.

Mass Spectrometry for DNA Analysis (Michelle Buchanan) - To complement the genome-wide mutation screening efforts, we are examining parameters in mass spectroscopy-based DNA analysis. Our initial efforts have focused on optical imaging and diagnostics of the matrix assisted laser desorption ionization (MALDI) plume, and investigations of the crystallization of DNA for the MALDI matrix. In parallel, we are examining methods for higher-throughput DNA analysis necessary for large-scale genome-wide polymorphism analysis.

Defining the Function of the Human DNA Repair Protein XPF by Comparative Genomics and Protein Chemistry

Sandra L. McCutchen-Maloney, Mark Shannon,
Jane Lamerdin and Michael P. Thelen
Molecular and Structural Biology Division,
Lawrence Livermore National Laboratory,
Livermore, CA 94550
smaloney@llnl.gov

The human XPF gene is required for the cellular response to DNA damage caused by ultraviolet radiation and mutagenic chemicals. Mutation in this gene leads to one form of the disease syndrome xeroderma pigmentosum. Homologs of the XPF protein are present in fly, worm, plant, yeast and archea, indicating conservation of an essential function throughout all of evolution. The expression of mouse XPF in pachytene spermatocytes corroborates other indications that XPF has an additional role in genetic recombination during meiosis. Information from all of the homologs has been valuable in directing experiments to determine the specific characteristics of the XPF protein. We have demonstrated the predicted endonuclease activity, protein-protein interactions, and DNA binding in the recombinant XPF protein that was overexpressed and purified from *E. coli*. Fragments of XPF obtained during purification retain endonuclease activity. This result, together with the sequence analysis of homologous proteins, has lead us to produce a series of deletion fragments of XPF in order to identify functional domains. The use of comparative genomics coupled with protein chemistry has thus enabled us to begin defining the function of this protein that is crucial to human health.

Work was funded by NIH grants to SLM-M and MPT; research was performed under the auspices of the U.S. DOE by LLNL under contract No. W-7405-ENG-48.

Direct Isolation of Functional Copies of the Human HPRT Gene by TAR Cloning Using One Specific Targeting Sequence

Natalay Kouprina*, Lois Annab**, Michael A. Resnick*, J. Carl Barrett** and **Vladimir Larionov***
*Laboratory of Molecular Genetics and
**Laboratory of Molecular Carcinogenesis,
National Institute of Environmental Health Sciences, Research Triangle Park NC 27709, USA

Recently we demonstrated that transformation-associated recombination (TAR) in the yeast *Saccharomyces cerevisiae* can be used to selectively isolate single copy genes, BRCA2 and BRCA1, from total human DNA as large circular YACs [1]. The TAR cloning method is based on co-penetration into yeast spheroplasts of gently isolated genomic DNA along with the vector DNA that contains 5' and 3' sequences specific for gene of interest, followed by recombination between the vector and the human DNA to establish a YAC [2]. We investigated whether a single copy gene could be isolated directly from total human DNA by TAR cloning using only one piece of its sequence information. A TAR cloning vector was constructed that contained a small amount of 3' HPRT sequence and an Alu repeat. Transformation with the vector along with human DNA led to the selective isolation of large circular YACs containing the entire HPRT gene. YACs up to 400 kb were generated that extended from the unique position of 3' HPRT to various Alu's similar to "genome walking". Based on transfection of the NeoR-retrofitted YAC clones into mouse cells, the YACs contained the functional HPRT. Use of the common Alu repeat as a second targeting sequence greatly expends utility of TAR cloning. We propose that a new TAR cloning approach is readily applicable to direct isolation of single copy genes from mammalian genomes.

[1] Larionov, V., Kouprina, N., Solomon, G, Barrett, J. C. and Resnick, M. A. Direct isolation of human BRCA2 gene by transformation-associated recombination in yeast. Proc. Nat. Acad. Sci. USA, 94, p.7384-7387, 1997.
[2] Larionov, V., Kouprina, N., Graves, J., X.-N., Chen, Julie R. Korenberg and Resnick, M. A. Specific cloning of human DNA as YACs by transformation-associated recombination. Proc. Nat. Acad. Sci. USA, 93, p. 491-496, 1996.

Mobile Elements and DNA Integration

Jerzy Jurka
Genetic Information Research Institute, Palo Alto, USA
e-mail: jurka@gnomic.stanford.edu

During the last two years or so, our DOE-sponsored research led to the discovery and characterization of DNA targets for retroposon integration (1,2). These targets are most likely recognized and cut by L1-ORF2 enzymes already existing in mammalian cells and have been postulated to be hot spots for homologous recombination (3). Recently, our team in collaboration with Dr. Sun-Yu Ng, has demonstrated that the preferred integration targets are also hot spots for homologous recombination which is increased by up to two orders of magnitude in test cancer cell lines (4). In another collaborative effort, we have discovered that mobile elements integrate primarily at kinkable DNA sites (5).

We continued our repeat annotation service to the community (6), as well as, discovery and analysis of new repetitive families (7,8).

1. Jurka, J., Klonowski, P. J. Mol. Evol. 43, 685-689 (1996)
2. Jurka, J. Proc. Natl. Acad. Sci. 94, 1872-1877 (1997)
3. Jurka, J. Site-Directed Recombination. U.S. Patent No. 08/643,886
4. Ng, S.-Y., Ma, J., Jurka, J. (in preparation)
5. Jurka, J., Klonowski, P., Trifonov, E.N. Mammalian Retroposons Integrate at Kinkable DNA sites (submitted)

6. CENSOR server: http://www.girinst.org
7. Jurka, J.,Kapitonov, V.V., Klonowski, P., Walichiewicz, J. and Smit, A.F.A. Genetica 98, 235-247 (1996)
8. Kapitonov, V.V. and Jurka, J. DNA Sequence (in press)

Study of Electric Field-Induced Transformation of *Escherichia coli* with Large DNA Using DH10B Cells

Alexander Boitsov[1,2], Boris Oskin[1], Pieter J. de Jong[2]
[1] Saint Petersburg State Technical University, Department of Biophysics, St. Petersburg 195251, Russia, [2] Roswell Park Cancer Institute, Department of Human Genetics, Elm and Carlton St., Buffalo, NY 14263
boitsov@dejong.med.buffalo.edu

Construction of large-insert libraries in bacterial hosts, such as those using PAC/BAC vectors and DH10B cells, has been limited by the inefficient transformation of *Escherichia coli* with large DNA molecules by electroporation. The goal of this project was to elucidate the mechanism and kinetics of the entrance of large DNA molecules into DH10B cells on exposure to electric fields and to exploit this information to increase the efficiency of PAC/BAC library construction. The essence of the approach used is the independent optimization of the three distinct stages of electrotransformation (ET): i) electrophoretic movement of DNA towards cells in cell-DNA suspension, ii) permeabilization of cell wall (production of electric pores) and iii) electrophoretic permeation of DNA molecules into the cells through electric pores. This was feasible due to the apparatus constructed in Saint Petersburg State Technical University that can generate multiple independently-regulated electric pulses. The results obtained previously with different *Escherichia coli* strains, other than DH10B, have revealed that the optimal time of electrophoretic permeation of DNA molecules into the cells through electric pores is proportional to the square root of the molecular length of supercoiled plasmids in the range from 7 up to 250 kb. This explains the failure of previous large DNA ET work that had been based on results with smaller DNA. To more fully characterize the ET process of DH10B cells with large DNA, we have determined the effects of each of the electric field parameters on transformation with respect to plasmid size (and topology). Plasmids ranging in size up to 250 kb have been used, and the resulting transformed molecules have been examined to verify their fidelity. The data which will be presented shall show that kinetics of DNA penetration into DH10B cells on exposure of electric field differ drastically from those obtained for other used *E. coli* cells (C600, HB101, K802, DH5a). First of all no clear dependence on molecule size of supercoiled plasmids was observed. The preliminary interpretation of the results is: the size of appearing electropores in DH10B is much bigger than that in other *E. coli* cells. It explains high ET efficiency of DH10B cells with large supercoiled plasmids which can be achieved (up to 108 transformants per mg of 250 kb plasmid) and which is, on a molar basis, only several times lower than the highest ET efficiency with small pUC18 plasmid.

Acknowledgment: This work has been sponsored in part by a DOE humanitarian grant OR0033-93CIS015 awarded to A.B. (boitsov@dejong.med.buffalo.edu) and by DE-FC03-96ER62294 to PdJ (pieter@dejong.med.buffalo.edu)

Progress Towards the Construction of BAC Libraries from Flow Sorted Human Chromosomes

Jonathan L. Longmire, Nancy C. Brown, Evelyn W. Campbell, Mary L. Campbell, John J. Fawcett, Phil Jewett, Mary Maltbie, and Larry L. Deaven
Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545.

Because of the advantages of large insert size and stability associated with BAC cloning systems, we have attempted to adapt the pBelloBAC vector for use with flow sorted human chromosomes. Compared to making genomic libraries (where DNA mass is not limited) the cloning of chromosomes into the BAC vector is very challenging due to the fact that only microgram quantities of DNA can be obtained even after extensive periods of sorting (months). The technical challenges involved in making chromosome-specific BAC libraries include developing methodologies to 1) allow efficient recovery of flow sorted chromosomes into agarose plugs; 2) predictable partial digestion of small masses of chromosomal DNA embedded in agarose; and 3) improving BAC cloning efficiencies to allow construction of multiple representation libraries from microgram quantities of chromosomal DNA.

We have found that partial digestions using HindIII can be performed in a predictable and reliable manner on microgram quantities of genomic DNA by carefully controlling incubation time and enzyme concentration-to-DNA mass ratios. In addition, these small amounts of partially digested genomic DNA can be size selected using PFG electrophoresis and cloned into the BAC vector with efficiencies that are sufficient for producing multiple representation chromosome libraries (103-104 cfu per mg starting DNA; average insert size approximately 90 kb).

Improvements have also been made in the collection of flow sorted chromosomes prior to DNA isolation. The previously used method for collecting chromosomes involved sorting into an agarose-coated tube until the tube was full of sheath fluid and then harvesting the relatively small number of chromosomes by centrifugation. We have developed a new method that allows larger numbers of chromosomes to be sorted into a single agarose coated tube. This is accomplished by using a series of centrifugation, decanting, and re-sorting steps to 'stack' chromosome pellets within the tube. A final spin followed by brief melting and regelling of the agarose is used to embed the chromosomes within the agarose plug. Higher DNA yields have resulted from using the stacked pellet method. However, chromosomes were lost during this process because the yield of chromosomal DNA was not directly proportional to the number of chromosomes that were originally sorted. This loss of chromosomes during the embedding step represents the single major problem that remains to be solved in order to allow the production of chromosome-specific BAC libraries. This work was supported by the US DOE under contract W-7405-ENG-36.

Construction of Human Chromosome 5 and 16 Specific Libraries by TAR Cloning

N. Kouprina, M. Campbell, J. Graves, E. Campbell, L. Meincke, J. Tesmer, N. Brown, J. Fawcett, P. Jewett, R. K. Moyzis, N. Doggett, L. Deaven, and V. Larionov
Los Alamos National Laboratory, Life Sciences Division and Center for Human Genome Studies, National Institute of Environmental Health Services, Laboratory of Molecular Genetics, Research Triangle Park, NC 27709
larionov@niehs.nih.gov

Transformation-associated recombination (TAR) was exploited in yeast to the selective isolation of human DNAs as circular YACs from monochromosomal mouse/human hybrid cell lines. Chromosome 5 and 16 specific YAC libraries were produced from the hybrid cell lines Q826-20 and CY18, respectively, using the F-factor based TAR vectors containing human Alu repeats. The presence of an F-factor origin in the TAR vectors provides the opportunity for transfer of generated YACs to *E. coli* to produce BACs. Although <2% of the DNA in the hybrid cells was human as many as 80% of the transformants had human DNA YACs based on colony hybridization of a representative number of clones with both human and mouse probes. Thus, the level of enrichment of human DNA to mouse was nearly 3,000-fold. The YAC libraries of chromosome 5 and 16 consist of 299 and 1320 clones, respectively, with an average size of 150 kb. Approximately 266 chromosome 5 and 748 chromosome 16 specific BACs were obtained after electroporation of YACs into *E. coli*. Based on the size of the clones generated from chromosome 5, the YAC and BAC libraries are ~0.24X and ~0.21X, respectively, in chromosomal coverage. Based on the size of the clones generated from chromosome 16, the YAC and BAC libraries are ~2.4X and ~1.2X, respectively, in chromosomal coverage. The chromosomal distribution of 41 TAR BACs each from chromosomes 5 and 16 was evaluated by fluorescence in situ hybridization. The distribution of FISH signals was random along the length of each chromosome. We concluded that TAR cloning may provide an efficient means for generating YAC/BACs from specific chromosomes.

Construction of Genome-Wide Physical BAC Contigs Using Mapped cDNA as Probes: Toward an Integrated Bac Library Resource for Genome Sequencing and Analysis

Steve C. Mitchell, Diana Bocskai, Yicheng Cao[2], Robert Xuequn Xu, Mei Wang, Troy Moore[3], So Hee Dho[4], Enrique Colayco[2], Christie Gomez, Gabriella Rodriguez, Annabel Echeverria, Melvin I. Simon[2], and **Ung-Jin Kim**[2]
[2] Biology, California Institute of Technology
[3] Research Genetics, Huntsville, AL
[4] Seoul National University, Seoul, Korea

The goal of the human genome project is to characterize and sequence entire genomes of human and several model organisms, thus providing complete sets of information on the entire structure of transcribed, regulatory and other functional regions for these organisms. In the past years, a number of useful genetic and physical markers on human and mouse genomes have been made available along with the advent of BAC library resources for these organisms. The advances in technology and resource development made it feasible to efficiently construct genome-wide physical BAC contigs for human and other genomes. Currently, over 30,000 mapped STSs and 27,000 mapped Unigenes are available for human genome mapping. ESTs and cDNAs are excellent resources for building contig maps for two reasons. Firstly, they exist in two alternative forms – as both sequence information for PCR primer pairs, and cDNA clones – thus making library screening by colony hybridization as well as pooled library PCR possible. We are now able to screen genomic libraries efficiently for large number of DNA probes by combining over 100 cDNA probes in each hybridization. Second, the linkage and order of genes are rather conserved among human, mouse and other model organisms. Therefore, gene markers have advantages over random anonymous STSs in building maps for comparative genomic studies.

As a preliminary work for the ongoing "BAC-EST" project, we are currently screening our human BAC libraries with thousands of cDNA probes. We have thus far used over 3,000 Unigene probes and the number will increase to 7,000 in this year. Our goal is to screen the library with at least 27,000 markers, most of which are in the form of cDNA probes. This represents at least 1 marker per every 100 kb of euchromatin regions. We plan to deconvolute the positive BACs to each marker by sorting the library into groups of BACs that are positive to specific pools, arraying each group on small hybridization filters, then hybridizing the filters with individual probes. We are also determining the end sequences for the positive BACs. BAC end sequence (BES) information can be extremely useful to precisely align these mapped BAC clones against any known sequence contigs by means of sequence match. Putative contigs or clone overlaps identified by markers or sequence match are verified via restriction fingerprint analysis. The BAC clone resources integrated with physical mapping information will be useful for building sequence-ready contigs on any chromosomal region.

Current Status of the Integrated Chromosome 16 Map

N. A. Doggett, L. A. Goodwin, J. G. Tesmer, L. J. Meincke, D. C. Bruce, M. R. Altherr, R. D. Sutherland, U.-J. Kim, and L. L. Deaven
Los Alamos National Laboratory, Los Alamos, New Mexico 87545 and California Institute of Technology, Pasadena, California 91125
doggett@gnome.lanl.gov

We have previously reported on the construction of an integrated physical map of human chromosome 16 (Doggett et al., Nature 377:Suppl:335-365, 1995). This map was constructed against a framework somatic cell hybrid breakpoint map which divides the chromosome into 90 intervals. The physical map consists of both a low resolution YAC contig map and a high resolution cosmid/P1/BAC contig map. The low resolution YAC contig map is now comprised of 900 CEPH megaYACs, and 300 flow-sorted 16-specific miniYACs that are localized to and ordered within the breakpoint intervals with 1150 STSs. (These include 200 megaYACs and 300 STSs which were incorporated from the Whitehead Institute's total genome mapping effort.) The YAC/STS map provides practically 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 and 400 genetic markers--as part of an ongoing effort to incorporate all available loci into a single map of this chromosome. A high resolution 'sequence ready' cosmid contig map consisting of 4000 fingerprinted cosmids assembled into contigs covering 60% of the chromosome is anchored to the YAC and cytogenetic breakpoint maps via STSs developed from cosmid contigs and by hybridizations between YACs and cosmids. Current work is focused on completing the 'sequence ready' map using a combination of cosmids and BACs. IRS-bubble PCR products from a minimally tiling set of YACs are being hybridized to the chromosome 16 cosmid library to localize cosmids to gaps in the existing map; and over 200 BACs, identified by library screening are now linked to the cosmid map. Several large contigs have been completed across disease gene regions in collaboration with several other investigators by supplying the available map resources (YACs, STSs, and cosmid contigs) and high density cosmid filter arrays for cosmid walking and YAC to cosmid hybridization experiments. The largest of these is a 4 Mb restriction mapped 'sequence ready' cosmid contig extending proximally from the p telomere. A 3.0 Mb region of this contig, extending proximally from the PKD1 locus is the substrate for a finished sequencing effort underway at Los Alamos. Supported by the US DOE, OBER under contract W-7405-ENG-36.

Construction of Sequence-Ready Physical Map of Human Chromosome 16p13.1-11.2 Region

Yicheng Cao, Diana Bocskai, Steve C. Mitchell, Robert Xuequn Xu, So Hee Dho#, Jun-Ryul Huh#, Byeong-Jae Lee#, Anna Glodek*, Mei Wang, Enrique Colayco, Gony H. Kim, Christie Gomez, Gabriella Rodriguez, Judith G. Tesmer**, Annabel Echeverria, Robin Hua Li, Melvin I. Simon, Norman A. Doggett**, Mark D. Adams*, and **Ung-Jin Kim**
* The Institute for Genomic Research, Rockville, MD
** Human Genome Center, LANL, Los Alamos, NM
# Seoul National University, Seoul, Korea

Extensive physical mapping efforts and advances in automated sequencing technology have resulted in the initiation of genomic sequencing of large human chromosomal regions. Currently, both NIH and DOE are supporting several centers in the U.S. to begin sequencing the genomes of human and model organisms systematically and in massive scale. In the past 1.5 years, Caltech has been building physical contig maps on the 20 Mbp region of the human chromosome 16p arm (16p13.1-11.2) jointly with TIGR and LANL as a pilot experiment to generate sequence-ready physical map using large insert human BAC libraries (A, B and C) that have been constructed at Caltech.

First, the pooled library A (with 3.5 X genomic coverage) was screened with 98 ordered STS primer pairs taken from the integrated chromosome 16 YAC-STS map constructed by LANL. In this initial screening, 77 STS markers successfully identified 184 positive BACs. Positive BACs were characterized by checking multiple single colonies per clone, and restriction fingerprint analysis. For the clones to be sequenced, FISH mapping and genomic Southern hybridization steps were added for the verification of the chromosomal localization and genomic colinearity. Inserts from these positive BACs were used for screening library B and C (approximately 10X genomic coverage) by colony hybridization. The libraries were also screened with approximately 90 Unigene cDNA probes that have been localized to the 20 Mb region, and with approximately 250 Unigene cDNA probes mapped to the other regions on chromosome 16. Shotgun clones derived from the ends of completely sequenced BACs were used as probes to efficiently identify BACs that overlap minimally with the sequenced BACs. We have thus far identified nearly 1,000 putative BACs belonging to the 20 Mb region, and more than 2,000 BACs over the entire chromosome 16. End sequences were determined from all of these BACs. The BES (BAC end sequence) have been used to align these clones against the sequenced BACs by sequence match, thus allowing rapid and precise determination of the extent of overlaps between clones. The putative overlaps from the sequence match are verified by restriction fingerprint analysis.

Currently we are building BacDB database, a modified version of ACeDB.4_1, by entering and organizing all information related to human BAC clones and physical mapping data that are available from Caltech, LANL, TIGR, as well as public resources. The database contains available BAC related data and mapping information, STSs, ESTs, BES, and completed BAC sequences. BacDB will not only serve as an integrated database for mapping and sequencing, but will be a tool for the efficient identification of sequence-ready clones.

WEB sites
Caltech: http://www.tree.caltech.edu
TIGR: http://www.tigr.org/tdb/humgen/humgen.html
LANL: http://www-ls.lanl.gov

Large-Scale BAC End Sequencing to Aid Sequence-Ready Map Construction

Mark D. Adams, Steve Rounsley, Casey Field, Jenny Kelley, Steve Bass, Brook Craven, and J. Craig Venter
The Institute for Genomic Research, Rockville, MD 20850
mdadams@tigr.org

Libraries constructed in BAC vectors have become the choice for clone sets in high throughput genomic sequencing projects because of their higher stability as compared to their YAC or cosmid counterparts. We have proposed the use of BAC end sequences as a primary means of selecting minimally overlapping clones for sequencing large genomic regions. A necessary prerequisite of this is the collection of end sequences from all the clones in deep coverage BAC libraries. This is now being pursued for both the human and *Arabidopsis* genomes. BAC vectors are based on the *E. coli* F-factor replicon and offer strict copy number control limiting the number of BACs to 1-2 copies per cell. However, in addition to minimizing the chances of chromosomal rearrangements, the low copy number also poses a challenge for high throughput direct sequencing of the BAC clone ends because of the difficulty in obtaining sufficient quantities of high quality template from standard minipreps.

We have developed reliable approaches for both a multiprep for BAC DNA purification and a protocol for the direct sequencing of BAC DNA using Dye Terminator chemistry. The combination of these two methods allows us to produce daily about 400 high quality BAC end sequences with an average edited length of 400 bases using 4 ABD 377 sequencers and a small team of personnel. To aid in sample tracking and high throughput processing, the prep is processed completely in a 96 well format, from clone storage and growth through DNA purification, isopropanol precipitation, and final resuspension. Sequencing reactions are also processed in a 96 well format, including the removal of excess dyes before loading onto the sequencers. Both high throughput methods are amenable to future automation. Our methods will be presented along with discussion regarding the advantages and disadvantages of various methods we tried.

The Sequence Tagged Connector (STC) Approach to Genomic Sequencing: Accelerating the Complete Sequencing of the Human Genome

Gregory G. Mahairas, Keith D. Zackrone, Stephanie Tipton, Sarah Schmidt, Alan Blanchard, Anne West, Joe Slagel and Leroy Hood
Department of Molecular Biotechnology,
University of Washington, Seattle, WA 98195

The STC approach has been proposed as an attractive strategy to provide a sequence ready scaffolding for the efficient and directed sequencing of the complete human genome [(1)]. This effort has been undertaken through a collaborative effort between the California Institute of technology, TIGR and the University of Washington, and funded through the U. S. Department of Energy. The approach entails the sequencing of the ends of 300,000 Bacterial Artificial Chromosomes (BACs) that constitute a 20X deep Human DNA library to construct a sequence ready scaffold of the human genome.

At the Univ. of Washington we have assembled a high throughput automated end sequencing and fingerprinting process with its associated informatics. BAC clones are robotically inoculated from 384 well plates into 4 ml 96 well culture format, grown and the BAC DNA robotically extracted using AutoGen 740 robots. BAC template DNA from the AutoGen is then robotically transferred into 96 well microtiter plates from which DNA sequencing and fingerprinting reactions are setup. DNA fingerprinting is performed using conventional agarose electrophoresis, digestion with a single restriction enzyme (EcoRV) followed by automated imaging and analysis. DNA sequencing is performed using PE-ABD High Sensitivity dye primers and ABI 377 DNA sequencers. Laboratory protocols, automated data production, data processing, quality control measures and LIMS will be described in detail. During a 50 day period the STC laboratory sequenced 23317 BAC ends (STCs). 19224 (82.4%) where greater than 100 bp non trimmed and the average nontrimmed read length was 388 bp for a total of 7.46 Mb (.25%) of the genome. 29 % of the STCs contained repetitive DNA but less than 11% where entirely repeat. 12% of the repetitive DNA were LINE sequence, 4.6% LTR, 6.7% SINE sequence and 1.3% of the STCs contain a microsatellite or simple sequence repeat. The total G + C content was 40% and the average CpG content was .28, both expected numbers for human genomic DNA. 224 STCs had CpG scores of 1 representing CpG islands. 3103 STCs (16.8%) hit the EST, non-redundant nucleotide or Sixframe database. 1103 STCs hit the EST database (DB), 517 of which hit only the EST; 1087 STCs hit the nr nucleotide DB, 471 of which hit the nr nucleotide DB only, and 913 STCs hit the nr protein DB, 500 of which hit only the nr protein DB. 181 STCs (1%) hit all three databases, 131 hit nr nuc. and nr protein DBs, 101 hit the EST and nr prot. DBs, and 304 hit EST and nr nucleotide DBs, i.e., 4% hit more than one of these DBs and probably represent genes.

[1] Venter, J. C., Smith, H. O., and Hood, L. (1996) Nature 381: 364-366

Amplification and Sequencing of End Fragments from Bacterial Artificial Chromosome (BAC) Clones by Single-Primer PCR

Joomyeong Kim, Ethan A. Carver, and Lisa Stubbs
Human Genome Center, Lawrence Livermore National Lab., PO Box 808, L-452, Livermore, CA 94551
Joo_Kim@b361.llnl.gov

PCR-based methods have provided invaluable tools for the analysis of large genomic clones, such as YACs and BACs, that comprise the bulk of most existing and emerging genomic physical maps. However, applications of the available methods are often limited, or made more difficult to apply, because of the need for significant identity between primers and sequences located on both sides of a targeted site. We have developed a method that permits target sequences to be exponentially amplified, with a high degree of purity and specificity, from low-complexity templates when only a single sequence-specific anchor primer is present in the mixture. Amplification is efficiently driven by specific binding of the primer to one end of the target locus, with reverse priming initiated at nearby regions containing only a 5-6 bp sequence match with the anchor oligonucleotide. Using standard vector-derived primers, we have applied the single-primer protocol to amplify end-fragments directly from a number of different BAC-containing colonies, and have generated high quality DNA sequence information from the PCR mixtures without the need for further purification. This work introduces single-primer PCR as a simple, efficient and convenient alternative to existing methods for isolation of sequence-ready end fragments and other sequences from within BACs and other large-insert genomic clones.

Construction of a High Resolution P1/PAC/BAC Map in the Distal Long Arm of Chromosome 5 for Direct Genomic Sequencing

Ze Peng, Steve Lowry, Duncan Scott, Yiwen Zhu, Eddy Rubin and **Jan-Fang Cheng**
Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
jcheng@mhgc.lbl.gov

One of the JGI genomic sequencing targets is the distal 45 Mb of the long arm of human chromosome 5. This region was chosen because it contains a cluster of cytokine growth factor (IL3, IL4, IL5, IL9, IL12, IL13, GM-CSF, FGFA, M-CSF) and receptor genes (GRL, ADRB2, M-CSFR, PDGFR) and is likely to yield new and functionally related genes through long range sequence analysis. This region is relatively rich in disease-associated genes including susceptibility to asthma, several autosomal dominant corneal dystrophies, low-frequency hearing loss, dominant limb-girdle muscular dystrophy, Treacher Collins syndrome, and myeloid disorders associated with the 5q- syndrome.

The mapping strategy is based on a combination of both hybridization and PCR approaches. Inter-Alu fragments generated from non-chimeric YACs covering 3-5 Mb of DNA were used to isolate regionally specific P1/PAC/BAC clones. These clones were sized by pulsed-field gel electrophoresis, and their map locations were confirmed using fluorescent in situ hybridization. We estimated that approximately 60% of the clones representing a targeted region in the P1, PAC or BAC libraries were identified by this approach. Overlaps between P1s, PACs and BACs were mainly established by PCR using STSs generated from ends of clones. Contigs were further oriented using STSs developed from known genes, ordered markers, and ends of P1s, PACs, BACs and YACs. The STS content mapping also allowed us to identify new clones to fill gaps.

We have used this strategy to map 198 P1s, 60 PACs and 1,407 BACs so far in the region of 5q23-q35. These clones were linked by 1,321 STSs to form 74 contigs. The contigs are approximately 0.2-4 Mb in size. The average density of STSs is about 3.7 per 100 Kb in the proximal 20 Mb and about 2.3 per 100 Kb in the distal 25 Mb. Most STSs were derived from the ends of BACs which allowed detection of 3.4% chimeras by PCR analysis. The clone coverage of the proximal 20 Mb is over 95% and the coverage of the distal 25 Mb is between 70-80% at different locations. To date, 121 P1/PAC/BAC clones spanning the proximal 10 Mb were in the pipeline for production sequencing. This clone map with STS information is distributed through our Web site (http://www-hgc.lbl.gov/sequence-archive.html) and is being updated periodically.

An Integrated Genetic and Physical Map of Human Chromosome 19: A Resource for Large Scale Genomic Sequencing and Gene Identification

L.A. Gordon, A. Georgescu, M. Christensen, S. Ross, L. Woo, L.K. Ashworth, H. W. Mohrenweiser, A.V. Carrano and A. S. Olsen.
Human Genome Center, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550
gordon2@llnl.gov

We have developed an integrated genetic and physical map for human chromosome 19 consisting of metrically ordered, well-annotated cosmid/BAC contigs that provide a critical resource for sequencing.

At approximately 65 Mb, chromosome 19 represents 2% of the haploid genome; it is the most GC-rich human chromosome, suggesting an especially high gene density. The current map consists of 185 ordered cosmid/BAC contigs of average size 225 kb (range 40 kb to 3.3 Mb) spanning a total of 42 Mb, i.e. over 75% of the non-centromeric portion of chromosome 19. An additional 8 Mb of small EcoRI mapped cosmid contigs, average size 80 kb, have not yet been incorporated into the ordered map. The order of the constituent contigs have been determined by standard FISH techniques applied to a series of chromatin targets with increasing resolution. High-resolution FISH to decondensed human sperm pronuclei establishes genomic distances between 286 ordered FISH markers in 19p and q, thereby providing a "metric" framework that links the ordered contigs to the cytogenetic map. Complete digest EcoRI maps have been constructed for all contigs, which provide validation of the contig assembly and constituent clones, as well as an indication of contig size and extent of clone overlap.

The map currently includes 15 Mb of restriction mapped contigs greater than 500 kb (average size 1 Mb), which provide ideal substrates for large-scale genomic sequencing of this chromosome. About 7.5 Mb have been sequenced or are currently in the sequencing pipeline. The high depth of coverage (average 5X) and mix of cosmid and BAC clones enables selection of an optimum set of spanning clones with minimum overlap for sequencing.

The map is extensively annotated, with over 300 genes/cDNAs and 180 polymorphic markers that have been localized at the clone, and occasionally restriction fragment, level. Placement of the genetic markers in the physical map demonstrates excellent correspondence with existing genetic maps and provides relative order of many markers that cannot be distinguished by recombination. This map provides a unique resource for identification of disease genes mapped to this chromosome.

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

Mapping by Two-Dimensional Hybridization

Cliff S. Han, Mark O. Mundt, Linda J. Meincke, Judy G. Tesmer, Robert K. Moyzis, Larry L. Deaven and **Norman A. Doggett**
Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545

The framework genome-wide physical maps have largely been constructed with YAC clones. YACs however, are often unstable and chimeric, and because of the difficulty to isolate cloned YAC DNA in a pure form, are unsuitable for DNA sequencing. Therefore, BAC libraries, which retain the advantages of large insert size, are stable, and easy to manipulate, were constructed. Physical mapping with these libraries is now underway and is a critical component to large scale genomic sequencing.

Clone based physical maps have been constructed by many methods, including clone fingerprinting and STS content mapping. We are developing an alternative method which is applicable to small to moderate complexity clone libraries such as chromosome specific cosmid and BAC libraries and plasmid subclone libraries of BACs. This two-dimensional hybridization method is based on clone hybridization with pooled clone DNA as probes. The principal experimental design is as follows: First, several grids of the library are made. Second, DNA of the clones are pooled by a two-dimensional strategy (rows and columns) and purified by subtractive hybridization to remove both low abundance and high abundance repetitive elements. Third, the grids are hybridized separately with the row and column pooled DNAs. Positive hybridizing clones that are in common between the row and column pool probes will overlap with the clone that is in the intersection point with these pools. We are using the program MAP (written by Mark Mundt) to construct contigs from the two-dimensional hybridization results.

This method was tested with 350 cosmids, chosen from the chromosome 16 specific cosmid library by grid hybridization with 3 YACs. Overlaps identified by two-dimensional hybridization were confirmed by restriction fingerprinting. We are currently implementing this approach for rapid construction of contigs from a chromosome 16 specific BAC library and for the ordering of plasmid subclones of cosmids and BACs into minimal tiling sets prior to their sequencing. Supported by the US DOE, OBER under contract W-7405-ENG-36.

Estimates of Gene Densities in Human Chromosome Bands

Norman A. Doggett, Robert D. Sutherland and David C. Torney
Theoretical Division, Life Sciences Division, and OHER Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM.

ISCN 1995 established a new set of Human chromosome ideograms that comprise 850 metaphase-chromosome bands, differentiated by five shades of staining intensity [ISCN (1995), Mitelman, F. (ed), S. Karger, Basel; and Francke, U. Cytogenet. Cell Genet. 65 206-219 (1994)]. The five band shades are referred to as white, light gray, medium gray, dark gray, and black. These shades, presumably, reflect different underlying states of chromatin. The publication of the mapping of 16,000 partially sequenced genes, or cDNAs, to a framework radiation hybrid map of the Human genome has established the chromosome assignment for as much as 20% of all Human genes [Schuler et al., Science 274 540-546 (1996)]. In this work, the genes were observed to be distributed non-uniformly along the chromosomes. Because none of these cDNAs were localized to a specific chromosome band, no conclusions were drawn about gene densities in the different types of bands.

To establish a relationship between gene density and band type we have used the proportion of each type of band on each chromosome to estimate the respective gene densities, by optimizing the consistency of the model with the numbers of cDNAs on the 22 autosomes. In detail, the theoretical prediction for the number of genes on chromosome number j equals â5=1 aipij, in which the ai are gene density parameters for band type I and the pij are the proportions for the five types of bands on this chromosome. The statistical analysis used involves integrating over the posterior joint distribution for the parameters, given the linear model and the cDNA data. In one data-analysis we grouped the white and light gray bands together, assigning them a shared gene-density, and also grouped the two darkest bands together, thereby reducing the number of gene-density parameters to three. The inferred gene density for the white &

light gray bands equaled 7.3/Mb, with a standard deviation of 0.3, the inferred gene density in gray bands equaled 1.45/Mb, with a standard deviation of 1.1, and the inferred gene density in gene density in dark gray and black bands equaled 0.45/Mb, with a standard deviation of 0.4. If we knew the total number, T, of genes in the Human genome, the expected gene densities are, of course, our estimates multiplied by T/16,000. Reasonable choices for T range from 60,000 to 100,000. We will present the gene density estimates for all types of bands, plus confidence intervals.

It is reasonable to conclude that the expected gene density decreases with band darkness. These results have practical implications for large scale sequencing of the Human genome and have ramifications for the study of the evolution of genomes.

Mapping and Functional Analysis of the Mouse Genome

Dabney K. Johnson, Edward J. Michaud, and Monica J. Justice
Mammalian Genetics Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-8080
johnsondk@ornl.gov

As sequencing of the human genome progresses, the role of the mouse as proxy mammal for functional studies will become crucial. We at ORNL have designed a comprehensive program that will combine the power of our unparalleled mouse 'mutation machine' with our massive computational capability in bioinformatics and our integrated technology development effort in detection and analysis of new mutant mouse phenotypes. Our goal is to assign function to all the genes that reside in defined segments of the mouse genome, via efficient identification of phenotypic changes in behavior, biochemistry, and gene expression that accompany induced genetic changes.

The creation of functional maps of the mammalian genome must keep pace with the rapidly expanding physical and transcriptional maps. High-efficiency mutagenesis in mouse spermatogonial stem cells, using the supermutagen N-ethyl-N-nitrosourea (ENU) to induce point mutations in single genes, will provide the functional information for given segments of DNA sequence within the physical map. Other mutagens, for made-to-order mutagenesis, are also being tested, as are DNA-repair deficient mouse strains as mutagenesis targets. For an immediate target, we propose to 'saturate' a long and well-characterized chromosomal deletion with ENU-induced mutations to develop a gene-by-gene phenotype map of the region uncovered by the deletion. This efficient experimental protocol will generate phenotypes evident in deletion hemizygotes after only the second generation post-treatment, and all progeny in that generation will be 'color-coded' for instant genotyping. Submapping of phenotypes will be quite efficient because we have existing mutant stocks that carry many additional deletions whose endpoints nest within the large deletion. Physical and transcriptional mapping within this set of nested deletions are well under way.

We will also create similar deletion reagents in the mouse cognate for a human chromosomal region already being sequenced; good choices would be mouse chr 16/human chr 21, or mouse chr 8/ human chrs 16 and 19. A complex of nested deletions will be generated molecularly using the Cre/loxP approach, or via radiation-induced mutations in ES cells.

Pilot experiments conducted here at ORNL demonstrated that high-efficiency ENU mutagenesis generated numerous visible and lethal mutations that fall within a large deletion encompassing the p locus in mouse chr 7. These experiments resulted in the isolation and fine localization of at least seven new loci in 1244 gametes tested; since only lethal and visible phenotypes were under scrutiny, more subtle mutations/alterations were missed. We propose to extend this analysis to a second p-region deletion, p30PUb, that has been shown to contain multiple interesting phenotypes. These phenotypes have been submapped to specific intervals between deletion breakpoints, but we now must create the necessary intragenic mutant alleles for any genes in the deletion. We expect also to generate new phenotypes as we disable additional genes within the large deletion. DNA sequence of the deleted region will be determined in conjunction with the JGI to facilitate gene identification and mutational analysis.

Additional crucial components of our strategy are automation of phenotype screening by microfabrication of subdermal sensors, computer-aided imaging of both skeletal and soft tissues, and improvements in through-put for behavior-testing paradigms and GC/MS analysis of blood, urine, and breath for biochemical alterations. Changes in gene expression will be detected by the use of chip-based DNA/RNA analysis, and mass-spectrometry-based protein analysis from mutant vs normal cell populations.

We are also developing sperm-freezing protocols for mutation preservation and distribution, and efficient artificial insemination for recovery.

This work is supported by the U.S. Department of Energy FWP ERKP260 under contract no. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corporation.

The Albino (c) Region of Mouse Chr 7: A Model for Mammalian Functional Genomics

M.J. Justice[1], E.M. Rinchik[2], D.A. Carpenter[1], S.E. Thomas[1], and D.K. Johnson[1]
[1] Mammalian Genetics Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831
[2] Sarah Lawrence College, Bronxville, New York 10708

The mouse, with its powerful genetic tools and its extensive comparative molecular linkage map with the human, is a useful model organism to study mammalian gene function. Multiple mutant alleles of genes can be derived by mutagenesis that may reflect loss of function, partial loss of function, or gain of function. The entire series of alleles must be studied together to dissect the function of the gene.

Overlapping deletions obtained at albino (c) are useful genetic reagents for functional genomics. Saturation mutagenesis with ethylnitrosourea (ENU) utilizing the deletions at the c region on mouse Chromosome 7 revealed many new functional units that reflect single gene changes (Rinchik et al. 1990; Rinchik et al. 1995). The region is homologous to human Chromosome 11q13-q21, and is linked to the mouse homologue of human oculocutaneous albinism, type 1A. Because of the nature of the phenotypic screen, many of the new mutants die as embryos. The region is particularly valuable for functional studies because of the variety of genetic reagents, including overlapping deletions and point mutations, that are available. Our focus is a group of alleles isolated at a locus (axis) that affects the development of the body axis. The homozygous phenotypes of six alleles at axis that are likely to be point mutations range from early prenatal lethality to adult viability. The baseline function of axis is demonstrated by two alleles that arrest during formation of the primitive streak. However, two other alleles have a severely disorganized body axis later in development. One allele exhibits a variety of neural tube defects, including exencephaly. Together, these observations suggest that the axis locus functions in the formation of the rostral-caudal body axis and neurulation. Intriguingly, homozygotes of one allele of axis survive to adulthood, and have axial skeletal abnormalities. Complete complementation studies of these six alleles reveal complex genetic characteristics such as intragenic complementation among two of the lethal alleles and a maternal effect of the viable allele. The maternal effect is a phenotype of severe neural kinking, accompanied by somite abnormalities, suggesting again, a role in the determination or maintenance of the integrity of the body axis. The varied features of axis will provide important insights into the mechanism of gene function at this complex locus.

The chromosomal region that includes axis contains other genes that affect body axis and skeletal development. Predictions of the potential role of a gene or genes at axis will be presented, as well as a view of the functional organization and possible interactions of other loci in the region. These analyses will give us essential data for subsequent large-scale expansions of the functional map of the mouse genome in parallel with human gene maps using additional induced and targeted deletions combined with chemical mutagenesis.

E.M. Rinchik, D.A. Carpenter & P.B. Selby. Proc. Natl. Acad. Sci. USA 87, 896-900 (1990). E.M. Rinchik, D.A. Carpenter & M.A. Handel. Genetics 92, 6394-6398 (1995).

This work is supported by the U.S. Department of Energy FWP ERKP260 under contract with Lockheed Martin Energy Research Corporation.

Comparative Mapping and Expression Study of Two H19q13.4 Zinc-Finger Gene Clusters in Man and Mouse

Joomyeong Kim, Mark Shannon, Linda Ashworth, Elbert Branscomb, and Lisa Stubbs
Human Genome Center, Lawrence Livermore National Laboratory, Livermore, CA 94551.
kimj@bioax1.bio.ornl.gov

One of the larger syntenically homologous blocks of human and mouse genomes includes the long arm of human chromosome 19 and the proximal portion of mouse chromosome 7. As part of an extensive comparative genome mapping of human chromosome 19, we have targeted a region spanning approximately 2.5 Mb near the telomere of H19q. As a first step for characterization of this region, we have assigned several known genes and cDNA sequences to the 19q13.4 physical map. Most genes assigned to this region are C2H2-type, zinc finger (ZNF)-containing genes, which include ZNF134, ZNF154, ZIK1, C2H2-25, and one EST (T26651). These genes are all localized to a centrally-located contig (577/1514) and KRAB (Kruppel-associated box)-type ZNFs. To provide clues for the potential role of these genes, the expression-patterns of these ZNFs have been determined; most genes are expressed ubiquitously in all tissues examined but the highest expression level for each gene has been observed in different tissues. Using an interspecific backcross system, we have mapped ZNF134-related sequences in the mouse and confirmed that the mouse genome has similar zinc-finger genes clusters in proximal Mmu 7. We are currently constructing mouse BAC contigs and, at the same time, isolating the mouse homologs or related genes for the human ZNFs from these contigs.

Recently, we have also assigned the human homolog of a mouse imprinted gene, Peg3 (paternally expressed gene 3), to a contig (174) located approximately 1Mb proximal of a ZNF134 contig. Studies of several imprinted domains, including Prader-Willi and Angelman syndrome-region (H15q11-13)/central Mmu 7 and Beckwith-Wiedemann syndrome-region (11p15.5)/distal Mmu7, indicated that genomic imprinting is generally conserved among mammalian species and also that imprinted domains are large--spanning distances ranging from several hundred kilobases to two megabases. Due to these observations, we have decided to characterize PEG3/Peg3-containing regions in both human and mouse. In human, the adjacent contig to a PEG3-contig appears to have numerous ZNFs. In mouse, we have isolated two zinc finger genes located very close to Peg3 and are currently investigating the imprinting status and genomic organization of these genes relative to Peg3.

One Origin of Man: Primate Evolution Through Genome Duplication

Julie R. Korenberg, Xiao-Ning Chen, Steve Mitchell, Rajesh Puri, Zheng-Yang Shi and Dean Yimlamai
Medical Genetics Birth Defects Center, The CSMC Burns & Allen Research Institute, UCLA School of Medicine, Los Angeles, CA
jkorenberg@mailgate.csmc.edu

Chromosome duplication is a force that drives evolution. We now suggest that this may also be true of the primates and that the resulting duplications in part determine the spectrum of human chromosomal rearrangements. To investigate the existence and origin of duplications in the human genome, and their consequences, 5,000 bacterial artificial chromosomes (BACs) were mapped at 2-5 Mb resolution on human high resolution chromosomes by using fluorescence in situ hybridization. A subset of 469 of these was defined that generated two or more signals, excluding those located in regions of known repeated sequences, viz., the regions of centromeres, telomeres and ribosomal genes. Although a subset of these multiple site BACs represent the chimeric artifacts of cloning, derived from two different chromosomal regions, others reflect regions of true homology in the human genome.

Two questions were considered; first, the extent to which the multiple sites of hybridization of single BACs within single chromosomes reflected the breakpoints of naturally occurring human inversions, and second, the extent to which these same multiple hybridization points reflected the chromosomal inversion points in primate evolution.

For human inversions, the results of the analyses revealed a total of 124 BACs (2.5%) mapping to two or more sites on the same chromosome, of which 81 (65%) mapped to one of 27 distinct human inversion sites, the largest share of which recognized the well-established pericentromeric inversions of chromosomes 1, 2, 9, and 18, as well as the paracentric inverted region of chromosome 7q11/q22. From this, we infer that meiotic mispairing involving the homologous regions may be responsible for the inversions.

With respect to primate evolution, a significant proportion of inversion breakpoints that characterize the chromosomal changes seen in the evolution of the great apes through man, are also reflected in the distribution of BAC multiple intrachromosomal sites. Further analyses of the 29 independent BACs recognizing the pericentromeric region of human chromosome 9 suggest at least three classes, two of which recognize only single sites in Pan troglodytes. These data suggest that inversions occurring through primate evolution may generate small duplications that, although they can cause chromosomal imbalance in single individuals, they also provide the additional genetic material for speciation.

Third-Strand In Situ Hybridization (TISH) to Non-Denatured Metaphase Spreads and Interphase Nuclei

Marion D. Johnson and Jacques R. Fresco
Princeton University/Washington Road/Princeton, NJ 08544
mjohnson@crick.princeton.edu

An in situ methodology employing solution conditions has been developed for binding oligodeoxyribonucleotide 'third-strands' to chromosomal DNA targets in non-denatured protein-depleted metaphase spreads and interphase nuclei. Third-strand in situ hybridization (TISH) was performed on slides at pH 6.0 using a dual psoralen-and biotin-modified 17-nt pyrimidine-rich third strand to target a unique multicopy sequence in human chromosome 17 alpha-satellite (D17Z1 locus). UVA photofixed third strands, rendered fluorescent by FITC-labeled avidin, are reproducibly centromere-specific for chromosome 17, and visible without amplification in human lymphocyte and somatic cell hybrid spreads and interphase nuclei. Two D17Z1 haplotypes, one positive and the other negative for third-strand binding, were identified in three combinations (+/+, +/-, and -/-). Third-strand probes specific for unique multicopy alpha-satellite targets in human chromosome X and 16 have also been developed. Similar alpha-satellite target sequences have been identified in 22 of the 24 human chromosomes, making centromere-specific chromosome identification by TISH applicable to virtually to all human chromosomes. The technology is presently being applied to chromosomes of other eukaryotes. TISH has potential diagnostic, biochemical, and flow cytometric applicability to native metaphase and interphase chromatin.

New Male-Specific, Polymorphic Tetranucleotide Microsatellites from the Human Y Chromosome

P. Scott White[1], Owatha L. Tatum[1,2], Larry Deaven[1], Jonathan L. Longmire[1]
[1]Genomics Group and Center for Human Genome Studies, Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545 and [2]Dept. of Biological Sciences, Texas Tech University, Lubbock, TX 79409.

Human genetic polymorphisms are valuable for extracting information about population structure and evolutionary histories. Clonally inherited DNA such as mitochondrial or Y chromosome-specific DNA is of particular use due to lack of recombination. Microsatellites have been used to reconstruct human evolutionary histories. Until recently, only six male-specific tetranucleotide repeats have been publicly available as PCR markers. We have developed six additional microsatellite markers using a cosmid library of flow-sorted human Y-chromosomes. These microsatellites are tetranucleotide (GATA)n repeats of nine to twelve repeat units each, and are polymorphic among unrelated individuals. All markers were analyzed using Applied Biosystems Genescan fluorescent fragment sizing. At least three alleles were identified for each marker when diverse genomic DNAs were used as PCR template. The allele sizes range from 162 to 372 nucleotides. An additional marker was identified, having polymorphic alleles in both males and females of sizes 217-237. These markers fall into sets of non-overlapping alleles that allow for efficient gel multiplexing with fluorescent dyes. Because six of these markers are male-specific or have male-specific alleles, they are valuable for evolutionary and population studies where non-recombining DNA is desired.

In Vivo Libraries of the Human Genome in Mice as a Reagent to Sift Genomic Regions for Function

Rubin, E.M., Zhu, Y., Fraser, K., Ueda, Y., Smith, D.J., Symula, D., Cheng, J.F.
Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
emrubin@lbl.gov

Libraries of all or part of the mammalian genome have been propagated in single cells and have been used as tools in gene discovery through in vitro analyses. We have expanded upon this concept by the creation of panels of YAC and P1 transgenic mice containing defined contiguous regions of the mouse or human genome. Since each library member contains a large (80-700 kb) transgene, together several megabases of contiguous DNA from a defined region of the genome can be propagated using the mouse as a host. We have successfully used such libraries to sift through large genomic regions and to localize and clone genes based on phenotyping members of the library. Examples of our use of these libraries to link sequence with function include:

(1) Biological annotation of human 5q31 genomic sequence data. Computational analysis of 1.2 Mb of sequence from human 5q31 generated as part of the JGI sequencing program has revealed multiple putative genes and exons. As a tool to validate the computationally predicted novel genes in this region, and to determine their site and timing of expression, we have created and analyzed a 1.5 Mb in vivo library of human 5q31 and documented the expression patterns of the newly identified genes in the YAC transgenics.
(2) Fine mapping a 5q31 QTL for asthma. Several human studies have mapped a major QTL determining IgE levels in asthmatics to 5q31. Through analysis of IgE levels in members of the 5q31 in vivo library following an airway irritant we have identified a single YAC noted in two separate founder lines to be associated with a marked decrease in IgE levels.

We are in the process of looking at mice containing fragments of this YAC to move from this disease associated 5q31 QTL to the causative gene.

(1) In vivo complementation for cloning mouse mutations. We developed an in vivo library of the 550 kb region to which the mouse recessive neurological mutation vibrator had been localized by meiotic mapping. Through the in vivo complementation of the phenotype with a member of the library, we have been able to fine map and than clone the gene (PITP-N) responsible for the disorder.

(2) Identifying genes contributing to defects in cognition on chromosome 21. We have created a 1.8 megabase in vivo library of human chromosome 21q22.2 in a panel of YAC transgenic mice. Analysis of these animals with regard to learning and behavior identified a 550 kb YAC responsible for specific deficits. Through fragmentation of the YAC we have been able to identifying a gene (GIRK) whose altered level of expression is responsible for behavioral abnormalities in mice, a gene whose altered expression has also been linked to learning defects in Drosophila.

These studies on panels of transgenic mice containing large inserts have effectively enabled phenotypic assays at the organismal level to be performed on many genes at once. This, in essence, constitutes a multiplex analysis that permits increased throughput of data collection relating sequence to phenotype.

Automatic BAC Contig Assembly by Three-Color Fluorescent Restriction Fragments

Y. Ding, M. Johnson, Y. Chen, J. Colayco, J. Melnyk, S. Khan, D. Gilbert, and **H. Shizuya**
Division of Biology, Caltech, Pasadena, CA 91125, PE Applied Biosystems, Foster City, CA 94404
shizuya@caltech.edu

Bacterial Artificial Chromosomes (BACs) are extensively used for large-scale mapping and sequencing. In order to identify and characterize BAC clones for contig assembly, we have developed a rapid fingerprinting method using fluorescently labeled dideoxyadenosine triphosphates ([F]ddATP). Taq FS incorporates [F]ddATP at the first nucleotide of a 5' overhang generated by Hind III digestion. The labeled fragments are further digested by four-base cutters to generate even smaller fragments (less than 500 bp) for visualization on sequencing gel. The reaction uses [F]ddATP labeled with one of three mobility-matched fluorescent dyes for fragments of each four-base cutter. They allow to multiplex the analysis in a single gel lane. The fourth color in the same lane is used to identify the location of the fragments of known molecular weights and then to calculate the size of each fragments based on that result. Fragment size data assigned by Genescan (ABI) are converted into FPC (Fingerprinted Contigs, Sanger Center, UK) format and then electronically transferred to FPC for automatic contig assembly. We have tested the system on one of the most well characterized BAC contigs of human chromosome 22; 96 BACs from the chromosome 22 regions where large-scale sequencing is underway. Without manual intervention, the system did not produce false overlap in all 96 BAC clones, and assembled 16 contigs and 31 singletons. The accuracy of overlaps constructed by the method is comparable to that established using BAC end sequence, and compared with completely sequenced BACs.

Fragments patterns determined by this method provide each BAC with digital fingerprint, and may be stored for searching clones with minimal overlap, and for identification of clones.

Construction and Characterization of New BAC Libraries

Y. Sheng, Y-J. Chen, C. Neal, and **H. Shizuya**
Division of Biology, Caltech, Pasadena, CA 91125
shizuya@caltech.edu

Human BAC clones have been extensively used in a variety of research areas of Human Genome Project because of their stability in *E. coli*, easy handling of BAC DNA, and relatively large insert size. Over the past four years, we have generated over 400,000 BAC clones from two individuals, and arrayed them in 384-well microtiter plates to organize these clones. Recently we initiated a new round of construction of BAC libraries for the community involved in the high throughput sequencing efforts. In order to build high quality libraries, we have implemented many checks and extensive experiments to test for the degree of representation and the degree to which BACs accurately reflect the human genome. We extend our improved quality control procedures to the new library making endeavor. For these libraries, we made a new generation BAC vector, pIndigoBAC45 which gives much darker blue colored colonies on the X-gal plates. This feature enables us to identify clones with inserts more accurately, resulting low percentage of empty clones, and to shorten the time required for library construction. To estimate the number of empty wells and the number of clones that lack inserts, we stamp all the clones from each 384-well microtiter plates onto LB agar-plates. This detects mistakes made during colony picking procedure. To check for chimerism, co-habitation of wells by multiple clones, and representation of the human genome, we end-sequence BACs and carry out RH mapping based on PCR using primers designed by the sequences. In collaboration with Mark Adams at TIGR, we have sequenced thus far both ends about 1,000 BAC clones in the newly constructed BAC library. Furthermore, we plan to compare new libraries with previously constructed BAC libraries by testing with the same probes used for these libraries. We will report the progress of these library construction and discuss quality of these BAC clones.

Completing the BAC-Based Physical Map of Human Chromosome 22

Holger Schmitt, Mitzi Shpak, Yan Ding, Melvin Simon, and **Hiroaki Shizuya**
Division of Biology, Caltech, Pasadena, CA 91125
shizuya@caltech.edu

The major goals of the Human Genome Project are the identification and the localization of 50,000-100,000 genes expected in the human genome, the generation of physical maps for each individual chromosome, and finally, the determination of the nucleotide sequence for the entire 30,000 megabase pairs of DNA. There is a clear need to develop reliable clone resources which are accurately mapped and at the same time can provide templates for direct use in large scale sequencing. A new generation of cloning system, the Bacterial Artificial Chromosomes (BACs), has been developed in our lab and is now extensively used throughout the community in a variety of research areas.

To initially demonstrate the usefulness of the BAC system for long range physical mapping, a scaffold integrated contig map of the entire long arm of chromosome 22 was constructed. Individual clones were ordered into contigs by fingerprint analysis and placed along the genomic stretch according to their content of genetic anchorpoints. The map consists of more than 700 BAC clones and spans the length of approximately 45 megabase pairs. Each BAC clone is further characterized by fluorescence fingerprinting and end-sequencing the inserts. Sequence information of the clone ends is presently used to select new clones from the 15x coverage BAC library, which extend contigs and close the gaps.

Our ultimate goal is to generate a well characterized collection of BACs, which provides complete physical coverage of the entire long arm of chromosome 22 with minimal overlaps. This map will serve as an excellent resource to discover all of the transcripts mapped on the chromosome, and can readily be used for cost-efficient genomic DNA sequencing with minimal redundancy.

Collaborators in our current mapping project are Dr. N. Blin, Univ. of Tuebingen, and Dr. E. Meese, Univ. of Saarland.

A Subdermal Physiological Monitoring System For Mass Screening Mice In Gene Expression Studies

M. N. Ericson, D. K. Johnson, R. S. Burlage, T. L. Ferrell, D. E. McMillan, K. G. Falter, A. D. McMillan, S. F. Smith, G. E. Jellison, and C. L. Britton, Jr.

Oak Ridge National Laboratory, Oak Ridge, TN
37831-6006
ericsonmn@ornl.gov

Researchers at the Oak Ridge National Laboratory are developing a highly automated integrated-circuit based research tool for subdermal monitoring of physiological parameters in mice used for gene expression studies. Application of this new instrumentation capability to genome studies will accelerate mass specimen screening by providing automated detailed observation and reporting of multiple key physiological parameters of interest. Body temperature, heart rate, physical activity level, movement trajectory, and possibly blood pressure will be measured by an implanted integrated-circuit based instrument containing multiple integrated sensors. Measured data will be transmitted periodically via wireless techniques for subsequent data processing, visualization, fusion, and storage. The integrated sensor/telemetry package will be low-cost, reusable, and sufficiently miniaturized to be directly injectable. The system will provide detailed parameter measurement and analysis capabilities not presently available to genomics researchers. Advanced multi-parametric data presentation will permit improved detail and accuracy in high-volume phenotype screening and increased detectability of subtle genetic defects. This paper will present preliminary information on parameter measurement methods, sensor selection, instrument and system architectures, instrument miniaturization techniques, and data processing methods.

HLA Genotype Analysis of Chronic Beryllium Disease Sensitivity

P. Scott White, Michelle Petrovic, Owatha L. Tatum, Nancy Lehnert, Usha S-Nair, Zaolin Wang, Larry L. Deaven, and Babetta Marrone
Los Alamos National Laboratory, Life Sciences Division and Center for Human Genome Studies, Los Alamos, New Mexico 87545

Beryllium alloys are used in several industrial processes and products, including the synthesis of components of nuclear weapons. Chronic beryllium disease (CBD) affects a percentage of human individuals exposed to airborne particles of beryllium. CBD is an autoimmune disease affecting the lungs of susceptible individuals. Antigen-presenting cell surface proteins have been the focus of investigation into possible genetic susceptibility to this disease. It was discovered previously that certain HLA-DPß1 alleles correlated with the development of CBD. Because the suspected allele is also found in normal populations at frequencies of over 40% it was necessary to collect sequence information from more individuals. In order to determine if the implicated genotype, with a Glu-69 mutation, was the sole contributing mutation we sequenced over 30 individuals for exon II of this locus. More than 90% of disease individuals possessed the Glu-69 mutation, although most of these were in the heterozygous state. This larger data set supports previous results, but because of the high frequency of this mutation in normal individuals it will require more extensive investigation to determine if other contributing genetic factors are present. In addition to this locus, we are sequencing two other MHC Class II HLA loci, HLA-DQß1 and HLA-DRß1 for CBD correlating genotypes. Screening for susceptibility to CBD will require robust assays, and the development of a genetic screen is one goal of this research.

A New Program to Develop a High-Resolution, High-Throughput Tomographic Imaging System for Mutagenized Mouse Phenotype Screening

M.J. Paulus, H. Sari-Sarraf, D.K. Johnson, D.H. Lowndes, M. L. Simpson, C.L. Britton, Jr., F.F. Knapp, Jr., J.S. Hicks
Oak Ridge National Laboratory, Oak Ridge, TN
37831-6006
paulusmj@ornl.gov

The Oak Ridge National Laboratory has recently begun the development of a new high-resolution, high-throughput 3-D mouse imaging and computer-aided screening methodology to rapidly identify, quantify and record subtle phenotypes in mutagenized mice. The research goals for this program are to develop a novel x-ray imaging technology with <50 mm spatial resolution, a 3-D data acquisition time of <1 minute per mouse and an estimated cost of a few dollars per image. A key element of this new system will be a novel cadmium zinc telluride detector operating in pulse counting mode. This new detector technology provides spatial resolutions and x-ray energy discrimination capabilities unattainable with traditional x-ray computed tomography detectors. Due to its high atomic number, the new detector is also suitable for traditional nuclear medicine studies. Additionally, new image processing and pattern recognition algorithms will be developed for the system to assist researchers in identifying phenotypes. In this paper we present the objectives and approach for this research program and some preliminary data.

Quantitative Detection of Nucleic Acids by Invasive Cleavage of Oligonucleotide Probes

Jeff G. Hall, Andrea L. Mast, Victor Lyamichev, James R. Prudent, Michael W. Kaiser, Tsetska Takova, Bob Kwiatkowski, Bruce Neri, and **Mary Ann D. Brow**
Third Wave Technologies, Inc., 2800 S. Fish Hatchery Rd., Madison, WI 53711
madbrow@twt.com

We have developed an enzymatic assay for direct, sensitive and quantitative nucleic acid detection. This assay is based on cleavage of a unique secondary structure that can be formed between two DNA probe oligonucleotides and a target nucleic acid of interest. The assay depends on the coordinate action between the two synthetic oligonucleotides. By the extent of their complementarity to the target strand, each of these oligonucleotides defines a specific region of the target strand. These regions are oriented such that when the two oligonucleotides are hybridized to the target strand, the 3' end of the upstream oligonucleotide overlaps with the 5' end of the labeled downstream "signal" oligonucleotide. The resulting structure is recognized by a structure-specific nuclease, which cuts the signal oligonucleotide to release a labeled fragment. When the reaction contains an excess of the signal molecules and is performed at elevated temperature to promote rapid dissociation and association of these molecules, the cleaved signal oligonucleotide is readily replaced by an intact copy so that the process can be repeated. In this way many signal molecules are cleaved for each copy of the target nucleic acid. The amount of target present may then be calculated from the yield of cleavage product, the rate of product accumulation and the time of incubation.

We have found that the accumulation of cleaved product is linear over both time and target concentration, with the rate of accumulation dependent on the turnover rate of the enzyme. With a turnover rate of about 35 cleavage events per minute, our current system permits signal to be amplified by more than 1000-fold, compared to single round hybridization, in 30 minutes, thus allowing the quantitative detection of sub-attomole levels of target nucleic acid.

Quantitative detection of nucleic acids in this fashion has several advantages over other methods of oligonucleotide-based detection. Foremost, because the cleavage requires the precise coordination and hybridization of two oligonucleotides, this reaction has a high level of specificity for the intended target sequence. The specificity of the detection is further enhanced because the investigator can select the site of cleavage in the signal molecule by designing the appropriate amount of overlap with the upstream oligonucleotide. The production of a discrete cleavage product of expected size allows that product to be more easily distinguished from the products of oligonucleotide destruction that may arise from thermal degradation or from nuclease contaminants in diagnostic test samples. Further, the use of oligonucleotides that are mostly or completely composed of DNA, rather than RNA, eliminates background that could arise from target-independent degradation by ribonucleases. Finally, because each cleavage event is dependent on the presence of the actual target material, and not on the products of the cleavage reaction, contamination by material carried over from completed detection reactions cannot induce additional signal in subsequent reactions. We have applied this method to direct detection of DNA targets such as DNA viral genomes, and to the detection of mRNA for monitoring of gene expression. Judicious placement of the oligonucleotide pair around splice junctions allows mRNA detection without prior destruction of genomic DNA. The products of the cleavage reaction can be analyzed by gel electrophoresis, or by non-gel methods such as capture to solid support. We will show how additional post-cleavage manipulation of the products can lower the limit of detection by 2 to 3 orders of magnitude when compared to gel-based readout.

Quantitative DNA Fiber Mapping (QDFM) Techniques for Physical Map Construction and Quality Control*

Heinz-Ulrich G. Weier, Stanislav Volik, Jenny Wu, Thomas Duell, Mei Wang, Ung-Jin Kim[1], Jan-Fang Cheng and Joe W. Gray
Resource for Molecular Cytogenetics and Human Genome Center, Life Sciences Division, University of California, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, and [1]California Institute of Technology, Pasadena, CA
weier@rmc.lbl.gov

The construction of high resolution physical maps and definition of a minimal tiling path are indispensable for directed approaches to large scale DNA sequencing. Similarly, closure of gaps and completion of shot-gun sequencing projects depends on knowledge of the physical location and size of gaps. We applied 'Quantitative DNA Fiber Mapping (QDFM)', an optical procedure for mapping based on hybridization of fluorescently labeled probes on to individual stretched DNA molecules, for construction of high resolution physical maps and definition of minimal tiling paths as well as for quality control of sequencing templates derived from human chromosome 20. Digital image analysis allowed localization of probes with near kilobase(kb)-resolution in intervals of several hundred kb[1,2]. When the technique was applied to construct physical maps for regions on the proximal long arms of human chromosomes 11 and 22 [3], respectively, we encountered numerous unstable yeast artificial chromosome (YAC) clones. Deletions in these YACs prohibited their use for map construction. We therefore modified our mapping scheme and prepared DNA fibers comprised of genomic DNA. This proved to be a rapid method for determination of clone overlap and genomic distances between genes or markers. Furthermore, the genomic fibers provide a 'gold standard' for validation of clones and their contigs as well as the delineation of deletions in large clones. Compared to mapping on to fibers that were prepared from cloned fragments and purified by pulsed field gel electrophoresis, genomic DNA fibers are expected to significantly shorten the mapping cycle time, because slides carrying these fibers can be prepared in large batches and stored. The use of genomic fiber slides for mapping will also facilitate the implementation of standardized protocols that are amenable to automation and increase the mapping throughput. Using clones derived from selected regions on human chromosomes 5 and 16, respectively, we are presently evaluating the utility of genomic fibers for clone/contig validation.

*Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research, Department of Energy, under contract DE-AC-03-76SF00098.

[1] H.-U.G. Weier et al. Human Molecular Genetics 4, 1903-1910 (1995).
[2] M. Wang et al. Bioimaging 4, 73-83 (1996).
[3] T. Duell et al. Genomics (in press).

Analysis of DNA Sequence Copy Number Variation in Breast Tumors and HPV16-Immortalized Cell Lines Using Comparative Genomic Hybridization to DNA Microarrays

D. G. Albertson[1], R. Segraves[2], D. Sudar[1], S. Clark[2], C. Collins[1], C. Chen[2], W.-L. Kuo[2], D. Kowbel[1], S. H. Dairkee[3], I. Poole[4], M. Dürst[5], J. W. Gray[1,2] , D. Pinkel[1,2]
[1] E. O. Lawrence Berkeley National Laboratory, [2]University of California San Francisco, [3]Geraldine Brush Cancer Research Institute, [4]Vysis, Inc., [5]Deutsches Krebsforschungszentrum
albertson@cc.ucsf.edu

Gene dosage alterations underlie many diseases. For example, variations in DNA sequence copy number are associated with a significant proportion of the genetic aberrations involved in cancers, and also with certain developmental abnormalities. Comparative genomic hybridization (CGH) has proven to be an effective method for detecting and mapping these genetic alterations. In CGH total genomic DNA from a test specimen and a normal genomic reference DNA are labeled with different fluorochromes and hybridized to normal metaphase chromosomes. The ratio of the fluorescence intensities at a location on the chromosomes is approximately proportional to the ratio of sequences in the test and reference genomes that bind there. The use of metaphase chromosomes as the hybridization target has previously limited the resolution of CGH to 10-20 Mb. However, we have now implemented a new form of high resolution CGH by replacing the normal metaphase spread with an array of genomic cosmid, P1, and BAC clones (Sudar et al., these Abstracts). This approach provides a resolution at least a factor of 100 better than standard CGH, as it is determined only by the size and spacing of the target genomic clones. Thus, measurements of copy number can be made at low resolution on clones spaced at several Mb, or at high resolution using closely spaced or overlapping clones. We have performed both low and high resolution analyses of the DNA sequence copy number variation occurring on chromosome 20 in breast cancer and in pre-cancerous models. A low resolution, 'scanning' array of the entire chromosome was constructed from genomic clones spaced at ~1-3 Mb intervals on human chromosome 20. The analysis of breast tumors and breast tumor cell lines revealed at least five independent regions of copy number increase and a region of decrease on 20q that were present in various combinations. Thus, multiple, interacting genes involved in breast and perhaps other cancers may be located on 20q. Two of these regions were also recurrently present at elevated copy number in HPV-immortalized keratinocytes, suggesting that the processes of tumorigenesis in vivo and immortalization of cells in culture may proceed through common pathways of amplification and overexpression of certain genes mapping to these two loci. High resolution analysis was also performed on breast tumors with elevated copy number occurring within a ~1 Mb region at 20q13.2. The array was composed of contiguous and overlapping clones from a contig that has been constructed spanning this region (Collins et al., these Abstracts). For some tumors, a constant level of elevated copy number was observed across the region. However, in others, an abrupt variation in copy number was recorded, which mapped the boundaries of different levels of amplification to within a fraction of a BAC or P1 clone. These copy number alterations measured with array CGH are concordant with data obtained by using the array target clones as probes for interphase fluorescent in situ hybridization. However array CGH appeared to be both more quantitative and substantially faster, since only one hybridization was required to obtain the data at all loci. The increasing availability of genomic resources and the technology to print arrays with more than 104 elements/cm2 (Sudar et al., these Abstracts) make it reasonable to consider performing genome-wide analyses of copy number variation using arrays that would provide 1 Mb or better resolution for the entire human genome.

Molecular Anatomy of a 20q13.2 Breast Cancer Amplicon

C. Collins[1], J. Rommens[2], D. Kowbel[1], G. Nonet[1], L. Stubbs[3], M. Shannon[3], M. Wernick[1], J. Froula[1], G. Hutchinson[4], T. Godfrey[5], D. Polikoff[1], T. Cloutier[1], K. Myambo[1], C. Martin[1], M. Palazzolo[1], Dan Pinkel[1,5], D. Albertson[1,5], and J.W. Gray[1,5]

[1] Lawrence Berkeley National Laboratory, Berkeley, CA

[2] Hospital for Sick Children, Toronto, Canada

[3] Lawrence Livermore National Laboratory, Livermore CA

[4] RabbitHutch Biotechnology, B.C. Canada

[5] University of California Cancer Center, San Francisco

collins@rmc.lbl.gov

High level amplification of chromosome 20 band 13.2 occurs in 10% of primary breast tumors and correlates with poor prognosis in node negative patients. This amplification is also detected in numerous other solid tumors including bladder, brain, colon, head and neck and melanoma. We hypothesize that selection for overexpression of one or more genes encoded within this amplicon drives the 20q13.2 amplification event. To fine map the amplicon and identify the hypothesized oncogene(s) a 1.0 Mb interval spanning the amplicon was cloned in contiguous BAC, PAC, and P1 clones.

To fully explore the genomics of the 20q13.2 amplicon we have sequenced ~ 80% of the 1 Mb contig and analyzed it extensively for genes using a suite of bioinformatics tools. This combined with exon trapping and cDNA direct selection has led to the discovery of four genes ZABC1, ZABC2, NABC1, PIC1-related and a new cyclophilin pseudogene. ZABC2 may be the ortholog of the Drosophila melanogaster homeotic gene tea shirt. PIC1-related is ~ 93% identical to PIC1 or sentrin, a gene that encodes a ubiquitin-homology domain protein. NABC1 encodes a novel cytoplasmic protein of unknown function. ZABC1 is especially interesting. The cDNA sequence of ZABC1 is predicted to encode a putative transcription factor containing eight C2H2 zinc finger domains. To manage this data and make it available to collaborators we have developed a tailored ACEDB database accessible via the World-Wide Web (Davy et al, these abstracts).

A ~260 kb "minimum common amplicon" was defined by performing interphase FISH using P1 and BAC probes from the sequence-ready contig on ~300 primary tumors. ZABC1 is centrally located in the 260 kb critical region. Quantitative PCR and Northern analysis were employed to analyze the expression of this gene in breast cancer cell lines and primary tumors. Expression of ZABC1 is elevated in all cell lines and tumors in which it is amplified. High level expression of ZABC1 was found in the breast cancer cell line 600MPE, a notable finding because 600MPE is not amplified at the ZABC1 locus. This suggests alternative mechanisms may elevate ZABC1 transcripts in some tumors. ZABC1 is the only transcribed sequence identified that maps in the 260 kb critical region with a pattern of expression so strongly correlated with copy number. The murine ortholog of ZABC1 has been cloned to study its expression in murine mammary tumors and its normal spatial and temporal pattern of expression. We have now constructed a BAC contig across the mouse ZABC1 locus. It is our goal to sequence this contig and identify conserved noncoding regulatory elements through comparative sequence analysis. BACs from the contig will also be used to determine if ZABC1 is amplified in murine mammary tumors. It is anticipated that the combination of comparative and functional genomics will elucidate key regulatory elements of the ZABC1 gene and lead to insights regarding the mechanism of amplification at 20q13.2. In addition, comparative genomic hybridization to DNA microarrays is being used to map this amplicon at unprecedented resolution revealing new regions of consistent copy number abnormalities and suggesting additional co-selected and counter-selected genes involved in the evolution breast tumors (Albertson et al., these abstracts).

Supported by grants from US DOE contract DEAC0376SF00098, USPHS grants CA44768, CA45919, CA52807 and Vysis.

3D Genetic Analysis of Thick Tissue Specimens

SJ Lockett[+], C Ortiz de Solorzano[+], A Jones[+], E Rodriguez[+], K Chin[°], C Fernandez[°], D Sudar[+], D Pinkel[+°], JW Gray[+°]
[+]Resource for Molecular Cytogenetics, Life Sciences Division, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA 94720, [°]Cancer Center, University of California, San Francisco, CA 94143
sjlockett@lbl.gov, CODeSolorzano@lbl.gov, ALJones@lbl.gov, enrique@rmc.lbl.gov, chin@cc.ucsf.edu, carlosf@cc.ucsf.edu, d_sudar@lbl.gov, pinkel@cc.ucsf.edu, gray@cc.ucsf.edu

The Resource for Molecular Cytogenetics is developing computer assisted microscopy and image analysis techniques to allow combined genotypic and phenotypic analysis of intact cells in tissue. The long range goal is to obtain quantitative information about the copy number of DNA sequences, levels of expression of RNA and proteins and their distributions in cells; to morphologically characterize cells, nuclei and other organelles, and to analyze the cellular organization of tissue. This information will contribute to our understanding of functional processes in normal and diseased tissue involving candidate genes. Some of these genes will have been discovered using genome-wide surveillance techniques, such as array-based comparative genomic hybridization(CGH).

The technical procedure we have adopted for this project, which is at an early stage, is as follows: Adjacent thin (4 micron) and thick (30 micron) sections are cut from cancer specimens. The thin sections are used for standard histological staging, while the thick sections are used for quantifying the specific molecular species of interest in the individual cells. Thick sections are employed, because they contain intact cells, thus permitting accurate quantification at the individual cell level, and the cellular organization of the tissue is preserved. However, such sections require careful optimization of fluorescent in situ hybridization (FISH) and immunocytochemical procedures for labeling the particular molecular species, followed by 3D (confocal) microscopic image acquisition and 3D image analysis. We have developed several image analysis programs for this project. The first interactively enumerates punctate FISH signals in the individual intact nuclei of thick sections, and the second registers adjacent thick and thin sections. It is thus now possible to use this procedure to study the relationship between the copy number of specific genomic sequences in individual cells and histological stage of the tissue. In a preliminary application to a breast biopsy specimen, we demonstrated that histologically benign regions contained cells with two copies of a chromosome 1 alpha-satellite probe, whereas invasive cancer regions had variable copy numbers per cell of the probe. We are continuing these experiments in order to determine the degree of genetic heterogeneity in tumor cells and surrounding histologically normal tissue.

The image analysis programs mentioned above limit the procedure to the enumeration of punctate FISH signals in each cell. In order to expand the range of applications, we have developed algorithms for segmenting individual cell nuclei within thick sections. This enables quantification of diffuse molecular markers inside nuclei, the size and shape of nuclei, the spatial positions of FISH signals inside the nuclei, and the spatial relationships of cells to each other in the tissue. The input to the algorithms is a 3D image of nuclei labeled with a DNA counterstain. The first of the algorithms automatically thresholds the image into regions of background and nuclei. The next algorithm is a custom-designed, interactive 3D rendering program, which the analyst uses to inspect each nuclear region and indicate if each region is a single nucleus or a cluster of nuclei. Clusters are then divided by an automatic algorithm, which employs a variant of the Hough transform to shrink nuclei and consequently separate them. The resulting divided regions are presented to the analyst, who indicates if they are nuclei, are still clusters, or have been incorrectly divided and should be rejoined. The alternating steps of automatic cluster division and human inspection are repeated as many times as necessary to correctly segment all nuclei in the image.

This work was supported by the US DOE contract DEAC0376SF00098, NIH grant CA67412 and a contract with Carl Zeiss Inc.

Informatics

Informatics at the Center for Applied Genomics

Donn Davy, Tom Cloutier, Colin Collins, Manfred Zorn, Joe Gray
University of California at San Francisco/
Lawrence Berkeley National Laboratory
DFDavy@lbl.gov

The Resource for Molecular Cytogenetics/Center for Applied Genomics, in collaboration with the University of California at San Francisco-Cancer Center, has molecularly characterized and sequenced an interval of chromosome 20 band 13.2 found amplified in 10% of primary breast tumors and correlated with poor prognosis in node-negative patients. The goal is to determine the complete genomic organization of the one-mega base interval spanning this amplicon. In a follow-on project, the group will select DNA clones from all known oncogenes for use in a DNA chip, to be used as a diagnostic tool. Collaborators conduct their work in San Francisco, Berkeley, Vancouver, Toronto and Finland: sites so geographically distant as to make travel between them inconvenient and inefficient. Data concerning physical map assembly, genomic sequencing, Northern hybridization, tumor mapping, public database searches and extensive sequence annotation efforts including graphical maps and detailed data on clones, loci, exons, sequences and cDNA must be made equally available to all collaborators independent of location. The computing technology brought to bear on this problem combines a standard World-Wide Web sever engine with a tailored ACEDB database and a number of supporting CGI scripts connecting Unix file system documents with retrieved data from the database and with related data both local and at remote sites on the WWW. The ACEDB-style database (CAGdb) is accessible three ways; as a normal X-window application, as an aceclient/aceserver application, and via a set of interface scripts, from the CAG Web page. Additional WWW services include the display of Genotater (LBNL-developed sequence annotation visualization tool) images, XGrail (ORNL gene-prediction tool) images, a set of HTML pages of annotated sequence, a second set of HTML pages providing first-pass results of dbest and non-redundant nucleotide public database searches, the Northern hybridization images, two maps graphically representing the region of interest and connections to related databases. For tracking progress on the oncogene project, a second web server and database were deployed using off-the-shelf Java-language Rapid Application Development (RAD) tools. It is used to report progress in selecting targets, developing primers, and extracting clones, and in keeping track of collaborators' suggestions for further oncogene chip targets.

The Genome Channel and Genome Annotation Consortium

Ed Uberbacher, Richard Mural, Manesh Shah, Ying Xu, Sheryl Martin, Sergey Petrov, Jay Snoddy, Morey Parang - Oak Ridge National Laboratory
Manfred Zorn, Sylvia Spengler, Donn Davy - Lawrence Berkeley National Laboratory
Terry Gaasterland - Argonne National Laboratory
Peter Schad - The National Center for Genome Resources
Stan Letovsky, Bob Cottingham - The Genome Database
David Haussler - University of California Santa Cruz
Pavel Pevzner - University of Southern California
Chris Overton - University of Pennsylvania
ube@ornl.gov

Human and model organism sequencing projects will soon be producing data at a rate which will require new methods and infrastructure for users to be able to effectively view and understand the data. A multi-institutional project was recently funded to provide large-scale analytical processing capabilities and we will present the results of several pilot efforts related to this project. The goals of the project are as follows:

- Provide an environment where annotation can be constructed based on multiple interoperable analysis tools and significant available computing power.
- Provide an environment where characterization of long genomic sequence

regions can be facilitated and analysis can maintained and updated over time.
- Provide an interactive graphical environment where predictions, features and evidence from many tools can be combined by users into high-quality annotation and visualized by the community.
- Provide high-throughput automated analysis methods which can be configured by genome centers for their use in constructing annotation and facilitating data submission.
- Provide high-quality annotation to large genomic sequence regions which would otherwise go unannotated.
- Provide the community with the best sequence level view of genomes possible.

The components of this system are a number of services, a broker that oversees the distribution and management of tasks and a data warehouse, with services implemented using distributed object technology. Multiple gene prediction is accomplished using several gene finding tools including the GRAIL-EST system and gene annotation from databases such as Genbank is also captured. The data warehouse supporting the Genome Channel view is updated daily by automated Internet agents and event triggers which facilitate analysis procedures. Real time operation of the Genome Channel browser will be demonstrated. A more detailed description of the basic components follows:

The Genome Channel

The Genome Channel provides a graphical user interface to comprehensively browse and query assembled sequence placed in the public domain by the Human Genome Project and sequencing of model organisms. It is a JAVA interface tool which relies on a number of underlying data resources, analysis tools and data retrieval agents to provide up-to-date view of genomic sequences, as well as computational and experimental annotation. Navigation from a whole chromosome view to contigs provided by sequencing centers allows one to zoom in on regions of interest to see information about clones, markers, ESTs, computationally and experimentally determined genes, the sequence and sequence source information, related homology and functional information, and hyperlinks to numerous underlying primary data resources.

Analysis Methods

Current analysis methods which combine pattern recognition with EST information and protein similarities are capable of accurate and automated analysis of large genomic regions containing complex multiple gene structures. Analysis methods will include the GRAIL EST/Protein homolog system, Procrustes (Pavel Pevsner et al.), and Genie (David Haussler et al.) as well as other tools. The results of multiple tools can be viewed in a common environment, combined mathematically in user specified ways, and used as the basis for the automated or interactive construction of annotation.

Data Mining Agents

Maintenance of an up to date description of genomic regions will be based on the use of data mining agents formulated for particular information resources. These agents will make use of several different technologies, such as OPM (Victor Markowitz), Kleisli (Chris Overton), and database indices to facilitate meaningful information retrieval from remote sources. We expect new information related to each gene or genome region to continue to be discovered and actively linked in a long-term ongoing process. The current state of these links will be maintained in the data warehouse.

The Data Warehouse

The data warehouse provides and maintains a snap shot of the current view or status of the genomes or genomic regions analyzed in the project. The warehouse is designed to facilitate rapid access by users, visualization tools, and analysis systems. The views contained in the warehouse will be constructed and maintained by processes within this project (such as sequence analysis and information retrieval agents), with additional help from central databases like GDB. It will contain and make available a synthesized best view of genomes from multiple underlying sources.

Internet Object Request Broker

Services for data input, analysis, visualization and submission will be facilitated with a distributed underlying Internet architecture using CORBA with an object request broker to manage processes. Compute platforms, analysis servers, databases, etc. will be at a variety of locations and in some cases duplicated depending on need. Specialized computing hardware will be used to facilitate some tasks.

This project is jointly sponsored by the Computational Grand Challenge program of the Office of Computational and Technology Research and the Human Genome Program of the Office of Biological and Environmental Research of the Department of Energy.

GRAIL and GenQuest Sequence Annotation Tools

Ying Xu, Manesh B. Shah, J. Ralph Einstein, Morey Parang, Jay Snoddy, Sergey Petrov, Victor Olman, Ge Zhang, Richard J. Mural and **Edward C. Uberbacher**
Computational Biosciences Section, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-1060
GRAILMAIL@ornl.gov

Our goal is to develop and implement an integrated intelligent system which can recognize biologically significant features in DNA sequence and provide insight into the organization and function of regions of genomic DNA. GRAIL is a modular expert system which facilitates the recognition of gene features and provides an environment for the construction of sequence annotation. The last several years have seen a rapid evolution of the technology for analyzing genomic DNA sequences. The current GRAIL systems (including the e-mail, XGRAIL, JAVA-GRAIL and genQuest systems) are perhaps the most widely used, comprehensive, and user friendly systems available for computational characterization of genomic DNA sequence. In the past 2 years of the project we have:

- Developed improved systems for recognition of exons, splice junctions, promoter elements and other features of biological importance, including greater sensitivity for exon prediction (especially in AT rich regions) and robust indel error detection capability.
- Developed improved and more efficient algorithms for constructing models of the spliced mRNA products of human genes.
- Developed and implemented methods for the analysis and visualization of sequence features including poly-A addition sites, potential Pol II promoters, CpG islands and repetitive DNA elements.
- Designed and implemented new methods for detecting potential sequence errors which can be used to "correct" frameshifts, add quality assurance to sequencing operations, and better detect coding regions in low pass sequences such as ESTs.
- Developed systems for a number of model organisms including mouse, *Escherichia coli*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae* and a number of microbial genomes.
- Implemented methods for the incorporation of protein, EST, and mRNA sequence evidence in the multiple gene modeling process.
- Constructed a powerful and intuitive graphical user interface and client-server architecture which supports Unix workstations and JAVA Web-based access from many platforms.
- Improved algorithms and infrastructure in the genQuest server, allowing characterization of newly obtained sequences by homology-based methods using a number of protein, DNA, and motif databases and comparison methods such as FASTA, BLAST, parallel Smith-Waterman, and algorithms which consider potential frameshifts during sequence comparison.
- An improved "batch" GRAIL client allows users to analyze groups of short (300-400 bp) sequences for coding character (with frameshift compensation

options) and automates database searches of translations of putative coding regions.
- Provided support for GRAIL use in more than a thousand laboratories and at a rate of over 4000 analysis requests per month.

The imminent wealth of genomic sequence data will present significant new challenges for sequence analysis systems. Our vision for the future entails incorporation of a more sophisticated view of biology into the GRAIL system. Computational systems for genome analysis have thus far focused on generic or textbook-like examples of single isolated genes which can be described fairly simply using the most usual assumptions, and fall far short of the intelligence necessary to interpret complex multiple gene domains. In its next phase the GRAIL project will involve the development of new pattern recognition methods and modeling algorithms for DNA sequence, expert systems for interpretation using experimental evidence and comparative genomics, and interoperation with other tools and databases. More specifically we will focus on several development areas:

(1) Improved accuracy of feature recognition and greatly increased tolerance to sequencing errors,
(2) development of technology to describe the structure and regulation of large, complex genomic regions containing multiple genes,
(3) automated and interactive methods for the incorporation of experimental evidence such as ESTs, mRNAs, and protein sequence homologs in multi-gene domains (GRAIL-EXP),
(4) more comprehensive feature recognition and increased biological sophistication in the areas of expression and regulation,
(5) capabilities for direct comparison of genomes,
(6) a comprehensive suite of microbial genome analysis systems,
(7) infrastructure for use of high-performance computing systems and specialized hardware to facilitate analysis and annotation of large volumes of sequence data,
(8) improved interoperation with other tools, databases and methods for integrating information from multiple sources, particularly within the Genome Annotation Consortium framework and Genome Channel, and
(9) continued community and user support, technology transfer, and educational outreach.

These developments will enable GRAIL to become more comprehensive and biologically sophisticated, and yet remain a user-friendly analysis environment which can be used interactively or in fully automated modes.

GRAIL and genQuest related tools are available as a Motif Graphical client (anonymous ftp from grail.lsd.ornl.gov (134.167.140.9)), through WWW interfaces (URL http://www.compbio.lsd.ornl.gov/), or by email server (grail@ornl.gov) and genQuest at (Q@ornl.gov). Communications with the GRAIL staff should be addressed to GRAILMAIL@ornl.gov. (Supported by the Office of Health and Environmental Research, United States Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.)

The Stationary Statistical Properties of Human Coding Sequences

David C. Torney, Clive C. Whittaker, and Gouchun Xie
Los Alamos National Laboratory, Theoretical Division, Los Alamos, New Mexico 87545, and Merck and Company, Inc., West Point, Pennsylvania
dct@lanl.gov

We analyzed approximately 0.5 Mb of in-frame, nonredundant coding sequences from the NFRES database. These were reversibly encoded as binary sequences, using 1 1 for A, 1 -1 for C, -1 1 for G and -1 -1 for T. This binary encoding is appropriate for characterizing all stationary statistical features of the data, either in terms of moments or of cumulants.

Moments are expectations of products of digits. Cumulants are constructed from the moments, aiming to remove features that could have been predicted from subsets of the digits. The simplest cumulants are covariances of two digits, vanishing if these are independent. In fact, the cumulants offer a more concise description of the statistical properties of coding sequences than do the moments.

Turning to the covariances of digits from coding sequences, as each codon is represented by six binary digits, there are 36 types of pairs of digits for separate consideration. Some of these pairs exhibit a nonzero asymptote as the separation between the two digits increases. Most pairs exhibit interesting transients, out to about 50 bases. We will show plots of these 36 types of covariances.

We will also report results for cumulants of larger numbers of digits. For cumulants of three digits, the asymptotes are all effectively zero, as the spacings increase indefinitely, except for those cases in which two digits are derived from one base position. Similarly, the cumulants of four digits that have the largest asymptotic values are the ones in which these four are derived from two base positions. We will show plots for the cumulants corresponding to the nine pairs of codon bases. The cumulants of six digits, corresponding to three bases, appear to have a zero asymptotic value, as the two spacings between the bases increase.

A complete list of the nonzero cumulants—a tabulation of the stationary statistical properties of human coding sequences—is in hand. This tabulation could greatly facilitate the classification of anonymous DNA sequences and provide a natural starting point for the non-stationary analysis of DNA sequences.

WIT/WIT2: A System for Supporting Metabolic Reconstruction and Comparative Analysis of Sequenced Genomes

Ross Overbeek,* **Natalia Maltsev**,* Gordon Pusch,* and Evgeni Selkov * **

* Mathematics and Computer Science Division, Argonne National Laboratory, and ** Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, 142292 Pushchino, Moscow region, Russia
maltsev@mcs.anl.gov

The WIT/WIT2 system has been developed to support metabolic reconstruction from sequenced genomes. Specifically, it supports

1. derivation of initial assignments of function to ORFs for an organism,
2. use of these assignments to construct an initial estimate of the metabolic pathways present in the organism,
3. use of consistency analysis to refine the functional assignments, and
4. provide a framework for presenting and refining the emerging metabolic models for a set of organisms.

WIT2 is a UNIX-based system that is made available to anyone wishing to support Web-based access to genomic sequence data. It includes in the standard distribution a set of integrated genomes from the public archives. For each of the distributed organisms, the user has access to the ORFs, RNAs, contigs, function assignments, and asserted pathways that characterize the current state of the analysis of the genome. The user has the option of adding new public or proprietary genomes, and then analyzing the new genomes based on clustering of ORFs with the distributed genomes. The system supports both shared and non-shared annotation of features and the maintenance of multiple models of the metabolism for each organism. WIT2 comes in two parts: a Web-based system offering access to the data and a set of batch tools that offer extensible query access against the data. The Web-based tools integrate WIT2 with other sources of data on the Web, while the batch tools allow one to do pattern matching, extract regions of sequence for analysis using other tools, look for operons, and so forth.

Currently, the released system offers access to data for the following organisms:

Archaeoglobus fulgidus, Caenorhabditis elegans, Deinococcus radiodurans, Escherichia coli, Haemophilus influenzae, Helicobacter pylori, Methanobacterium thermoautotrophicum, Methanococcus jannaschii, Mycobacterium tuberculosis, Mycoplasma genitalium, Mycoplasma pneumoniae, Rhodobacter capsulatus SB1003, *Saccharomyces cerevisiae, Synechocystis sp.,* and *Treponema pallidum.*

The release at Argonne National Laboratory is accessible via [1].

Acknowledgment. This work was supported by U.S. Department of Energy under Contract W-31-109-Eng-38.

[1] http://www.mcs.anl.gov/home /overbeek/WIT2/CGI/user.cgi

Internet Release of the Metabolic Pathways Database, MPW

Evgeni Selkov, Jr.,* Yuri Grechkin,* and Evgeni Selkov * **
* Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, 142292 Pushchino, Moscow region, Russia, and ** Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, 9700 S. Cass Ave., MCS-221, IL 60439-4844, USA
selkovjr@turtle.stack.net

The Metabolic Pathway Database, MPW [1, 2], a subset of EMP [3. 4], plays a fundamental role in the technology of metabolic reconstructions from sequenced genomes under the PUMA, WIT, and WIT2 systems [5-7]. It is the largest and the most comprehensive metabolic database, including some 2,800 pathway diagrams covering primary and secondary metabolism, membrane transport, signal transduction pathways, intracellular traffic, translation, and transcription.

The MPW diagrams were originally encoded and distributed as a formatted ASCII text [2]. In the current public release of MPW [8], the encoding is based on the logical structure of the pathways and is represented by the objects commonly used in electronic circuit design. This facilitates drawing and editing the diagrams and makes possible automation of the basic simulation operations such as deriving stoichiometric matrices, rate laws, and, ultimately, dynamic models of metabolic pathways. Individual pathway diagrams, automatically derived from the original ASCII records [1, 9], are stored as SGML instances supplemented by relational indices. An auxiliary database of compound names, encoded as SMILES strings [10], is maintained to unambiguously connect the pathways to the chemical structures of their intermediates. In accordance with the IUPAC nomenclature for chemical compounds, the current release supports super- and subscripts, Greek letters, Italic fonts, etc.

Acknowledgment. This work was supported by U.S. Department of Energy under Contract W-31-109-Eng-38, and award OR00033-97CIS001.

References

1. Selkov E., Galimova M., Goryanin I., Gretchkin Y., Ivanova N., Komarov Y., Maltsev N., Mikhailova N., Nenashev V., Overbeek R., Panyushkina E., Pronevitch L., Selkov E., Jr. Nucleic Acids Res., 1997, 25 (1), 37-38
2. http://www.biobase.com/emphome.html /homepage.html/pags/pathways.html
3. Selkov E., Basmanova S., Gaasterland T., Goryanin I., Gretchkin Y., Maltsev N., Nenashev V., Overbeek R., Panyushkina E., Pronevitch L, Selkov E., Jr., Yunus I. Nucleic Acids Res., 1996, 24 (1), 26-29
4. http://www.biobase.com/EMP
5. http://www.mcs.anl.gov/home/compbio/PUMA/ Production/ReconstructedMetabolism/ reconstruction.html
6. http://www.mcs.anl.gov/home/compbio/ WIT/wit.html
7. http://www.mcs.anl.gov/home/overbeek/ WIT2/CGI/user.cgi
8. http://beauty.isdn.mcs.anl.gov/MPW
9. http://www.cme.msu.edu/MPW/
10. http://www.daylight.com/dayhtml/smiles/ index.html

Metabolic Reconstruction from Sequenced Genomes

Evgeni Selkov,*+ Natalia Maltsev,* and Ross Overbeek*
* Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, 9700 S. Cass Ave., MCS-221, IL 60439-4844, USA,
+ Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, 142292 Pushchino, Moscow region, Russia
evgeni@mcs.anl.gov

With the availability of increasing numbers of complete genomes, the possibility of developing accurate models of metabolism for these organisms becomes of central interest. We have initiated a project to "reconstruct the metabolism" of organisms from the sequence data supplemented by available biochemical and phenotypic data, and we have developed initial reconstructions for a number of organisms. These reconstructions for over twenty organisms (some based on incomplete sequence data) have been made available via the WIT and WIT2 systems [1-3].

The reconstructions are based on the collection of metabolic pathways, MPW. This collection now includes over 2,800 diagrams and is continually being enhanced. Each diagram represents a grouping of functional roles, and these groupings provide a resource in analyzing the functional assignments made to ORFs in newly-sequenced genomes; when several functions in a pathway have been clearly identified, more focused analysis on the "missing functions" provides a powerful means of improving function assignments that were originally made without access to an overall understanding of the metabolism of the organism.

The actual process of metabolic reconstruction involves a number of steps which we describe in detail. We have made the WIT/WIT2 system available to support such efforts, but the actual process is independent of this software and will undoubtedly be adopted by other efforts based on the same overall goal of using sequence data as a foundation to develop an accurate model of the metabolism of an organism.

We are entering a new phase of the project in which substantial benefit can be achieved via our growing understanding of which parts of metabolism are universal, of gene families (allowing more rapid development of initial function assignments made to ORFs), and of insights achieved from one genome that have obvious application in others. What is emerging is not just a large number of isolated metabolic portraits for a set of diverse microbial organisms, but rather an integrated understanding of the evolution of metabolism and the technology for developing higher-level functional models.
Acknowledgment. This work was supported by U.S. Department of Energy under Contract W-31-109-Eng-38.

References

1. http://www.mcs.anl.gov/home/compbio/WIT/wit.html
2. http://www.mcs.anl.gov/home/overbeek/WIT2/CGI/user.cgi
3. http://beauty.isdn.mcs.anl.gov/WIT2.pub/CGI/org.cgi

From Genomic Sequence to Protein Expression: A Model for Functional Genomics

Tracy J. Wright, Simone Abmayr, Min S. Park, Darrell O. Ricke, Cleo Naranjo, Becky Welsh-Breitinger, Karen Denison and **Michael R. Altherr**
Genomics Group, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

The goal of Functional Genomics is to provide useful and effective annotation of genomic sequence data. While effective annotation is likely to be defined by the needs of the end user, there are a variety of routine procedures that geneticists employ in the characterization of any gene or hypothetical gene. Included in these collections of procedures are techniques to identify: genetic variation (or polymorphism), transcript complexity and distribution, as well as protein coding capacity. However, the first step in the process is the identification of putative coding segments within a genomic sequence. This can be accomplished by interrogating the expressed sequence tag data base (dbEST) with the genomic sequence of interest; or by submitting the genomic sequence to a gene identification program like GRAIL in a effort to identify potential transcriptional domains. Sequences and clones identified in this way should be considered putative genes until their functional significance is substantiated by additional biological data.

We have initiated a process to evaluate the biological function of putative genes identified in a large segment of genomic sequence. The genomic sequence we have focused on represents a portion (~200 kbp) of a 2.2 Mbp sequence contig derived from a gene rich region on human chromosome 4 (4p16.3). This region is significant because it has been demonstrated to represent the smallest region of overlap for deletions in Wolf Hirschhorn syndrome (WHS) patients. WHS is a multiple anomaly dysmorphic malady characterized by mental and developmental defects. Due to the complex and variable expression of this disorder, it is thought that WHS is a contiguous gene syndrome with an undefined number of genes contributing to the phenotype. The analysis of genomic sequence data by BLAST, FASTA and GRAIL identified a number of putative transcription units. Potential coding segments were subjected to full insert sequencing of cDNA (when available) and Northern blot analysis. These initial analyses identified two putative coding segments with a high probability (intron::exon structure and positive Northern results) of representing bona fide genes. In addition, highly similar clones have been isolated from mouse providing additional support that these represent true coding segments. The human coding segments have been cloned into expression vectors and will ultimately be used to generate antibodies to confirm the existence of the hypothetical proteins.

This work has allowed us to evaluate coding potential and possible function of genes identified by genomic sequencing efforts. It has identified a number of potential bottlenecks and allowed us to conceptually develop a scheme to provide functional annotation for genomic sequence.

Support: DOE Contract W7405-ENG-36 and LANL LDRD funds.
E-mail Contact: ALTHERR@LANL.GOV
Day Phone: (505)665-6144

FAKtory: A Customizable Fragment Assembly System

Eugene W. Myers, Susan J. Larson, Brad W. Traweek, Kedarnath A. Dubhashi
University of Arizona, Tucson, AZ
slarson@cs.arizona.edu

FAKtory supports a large range of sequencing protocols and specialized fragment database information. Customizations include a processing pipeline for clipping, tagging, and vector trimming stages (prescreeners), and a configurable fragments database. Each pipeline stage can run in Automatic, Supervised, or Manual mode depending on the degree of user control desired. The FAKII Fragment Assembler provides high-sensitivity overlap detection, near-perfect multialignments, alternate assemblies, and support of user specifiable assembly constraints. While initial development on FAKtory emphasized its customizability, recent work provides sophisticated prescreeners, an editor for contig layout manipulation, a Finishing editor, and I/O filter capabilities for easy translation of data formats and linking to post-analysis programs.

A pipeline can include any number of prescreener stages for vector removal, data trimming, and tagging. Each prescreener includes a number of recognizers, which locate within a fragment selected intervals, frequencies of specified bases, trace signal characteristics, overlap with reference sequences, or matches to regular expressions. Because alternate assemblies may be generated for a given data set, FAKtory provides a Layout Edit panel for comparing and interacting with the potential solutions. Portions of contigs can be locked together, split apart, or checked for possible joins to other contigs. Constraints may be added or deleted, and reassembly with additional data generates additional assemblies to consider.

A Finishing editor displays a multialignment with a scrollable canvas of all trace data for the current location. FAKtory allows automatic tabbing to the previous or next unedited problem in the contig, minimizes the number of keystrokes needed in an editing sweep, and allows unlimited undo. All edited regions are shown to indicate finishing progress.

Divide-and-Conquer Multiple Sequence Alignment

Dan Gusfield, **Jens Stoye**
Department of Computer Science, University of California, Davis, CA 95616, USA
stoye@cs.ucdavis.edu

We present a fast heuristic algorithm for the simultaneous alignment of multiple sequences which provides near-to-optimal results for sufficiently homologous sequences.

The algorithm makes use of the optimal alignments of all pairs of sequences which give rise to secondary matrices containing additional charges imposed by forcing the alignment path to run through a particular vertex of the distance matrix. From these "additional-cost" matrices, we compute suitable positions for cutting all of the given sequences simultaneously, thus reducing the problem of aligning a family of sequences in a divide-and-conquer fashion to aligning two families of sequences, each of approximately half the original length. When, after re-iterating this division procedure sufficiently often in a recursive manner, the subsequences are sufficiently short, these are aligned optimally.

This procedure allows us to align simultaneously up to twelve amino acid sequences of the usual length (< 500) within a few minutes. The poster will also present several results concerning running time and memory usage as well as the quality of the obtained alignments.

Acknowledgment:
Part of this work was supported by Dept. of Energy grant DE-FG03-90ER60999 and by the German Academic Exchange Service (DAAD).

Sequence Assembly Validation by Restriction Digest Analysis

Eric C. Rouchka and **David J. States**
Washington University
ecr@ibc.wustl.edu states@ibc.wustl.edu

DNA sequence analysis depends on the accurate assembly of fragment reads for the determination of a consensus sequence. Genomic sequences frequently contain repeat elements that may confound the fragment assembly process, and errors in fragment assembly may seriously impact the biological interpretation of the sequence data. Validating the fidelity of sequence assembly by experimental means is desirable. This report examines the use of restriction digest analysis as a method for testing the fidelity of sequence assembly. A dynamic programming algorithm to determine the maximum likelihood alignment of error prone electrophoretic mobility data is derived and used to assess the likelihood of detecting rearrangements in genomic sequencing projects.

Restriction digest fingerprint matching is an established technology for high resolution physical map construction, but the requirements for assembly validation differ from those of fingerprint mapping. Fingerprint matching is a statistical process that is robust to the presence of errors in the data and independent of absolute fragment mass determination. Assembly validation depends on the recognition of a small number of discrepant fragments and is very sensitive to both false positive and false negative errors in the data. Assembly validation relies on the comparison of absolute masses derived from sequence with masses that are experimentally determined, making absolute accuracy as well as experimental precision important. As the size of a sequencing project increases, the difficulties in assembly validation by restriction fingerprinting become more severe. Simulation studies are used to demonstrate that large-scale errors in sequence assembly can escape detection in fingerprint pattern comparison. Alternative technologies for sequence assembly validation are discussed.

Segmentation Based Analysis of Genomic Sequence

Eric C. Rouchka and David J. States
Washington University
ecr@ibc.wustl.edu states@ibc.wustl.edu

The human genome is patchy and non-uniform in composition. We have developed a method for analyzing genomic sequence based on segmental variation in sequence composition using a heuristic algorithm employing classic changepoint methods and log-likelihood statistics. Our approach models the genome as composed of linear segments each characterized by its own compositional characteristics. The most informative description of a sequence is that description that maximizes the likelihood of the observed sequence given the model while simultaneously minimizing the cost of specifying the model. A Java interface has been developed to aid in visual inspection of the segmentation results (http://www.ibc.wustl.edu/~ecr/CPG/segment.html). The software has been tested using CpG dinucleotide distribution as a well-established example of biologically significant non-uniform composition.

Regions of DNA rich in CpG dinucleotides, also known as CpG islands, are often found upstream of the transcription start site in both tissue specific and housekeeping genes. Overall, CpG dinucleotides are observed at a density of 25% the expected level from base composition alone, partially due to 5-methylcytosine decay . About 56% of human genes have associated CpG rich islands. Since CpG dinucleotides typically occur with low frequency, CpG islands can be distinguished statistically in the genome. Our model is tested using several sequences obtainable from GenBank, including a 220 Kb fragment of human X chromosome from the filanin (FLN) gene to the glucose-6-phosphate dehydrogenase (G6PD) gene which has been experimentally studied. Results demonstrate a breakpoint segmentation that is consistent with observable manual analysis.

In addition to segmenting DNA according to the location of CpG islands, other compositions are explored as well. Among these are C + G content, mononucleotide content, and dinucleotide content. Segmentation according to higher order oligomers is also under consideration.

Swedish and Finnish Quality Based Finishing Tools for a Production Sequencing Facility

Matt P. Nolan, Jane E. Lamerdin, Stephanie A. Stilwagen, Glenda G. Quan, Ami L. Kyle, Anthony V. Carrano
Joint Genome Institute, Lawrence Livermore National Laboratory
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 ~40KB cosmid 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.

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. Within each problem region, we quantify the problem with respect to our QbDsFc, the phrap consensus quality values and whether or not the consensus sequence overlaps a database of repeats. We can generate a list of sequences (candidates for resequencing) intersecting the problem region. For each sequence, we identify where the read intersects the region. Within the intersection we compute the mean and sigma of the phred basecall quality values, quantify how well it matches the consensus sequence, and compute a value proportional to resequencing suitability.

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. The SaF applications consist of ~8000 lines of C++ and ~1000 lines of perl. We run the script Swedish-setup (~4000 lines of perl) to coordinate the conversion of assembly data through multiple formats to construct a CAF file representation of the phrap assemblage which the SaF applications input. One application produces a file used to direct a robotic workstation to rearray clones to prepare chemistries for finishing reads. We expect other new SaF functionality to migrate into our production environment to meet the ten fold increase projected for JGI sequencing in the upcoming year.

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

Informatics to Support Increased Throughput and Quality Assurance in a Production Sequencing Facility

Arthur Kobayashi, David J. Ow, Tom Slezak, Mark C. Wagner, Matt P. Nolan, T. Mimi Yeh, Stephan Trong, Anthony V. Carrano
LLNL HGC; Joint Genome Institute
kobayashi1@llnl.gov

We are developing an integrated informatics infrastructure to support the increasing throughput and quality assurance demands of our production sequencing facility. The LLNL Human Genome Center utilizes a modified shotgun-based approach to sequence human chromosome 19 and targeted gene regions of interest. To date we have finished over 1.5 Mb of high-quality genomic sequence, including a contiguous 1-Mb region. Our laboratory strategies and protocols are described in more detail in other abstracts in these proceedings (e.g., Lamerdin, McCready, et al).

To support this increase in sequence data volume, we have designed and implemented our informatics system to support our current sample prep and sequencing strategy: bulk generation of dye primer and dye terminator reads early in the random phase, followed by increasing automation in the pre-finishing and directed closure phases. We have developed a number of software programs (predominantly Sybperl or PerlTk) which are integrated through our Sybase relational database.

To support increased throughput, we have developed Sybperl programs to generate HTML WWW forms to create and manage sequencing projects, track samples, and create sample sheets. We have also implemented an automated sample file sorting system that analyzes and distributes sample files from our 14 ABI sequencers, which currently are loaded and run twice each day. Analysis results for *E .coli* contamination, sequence read length, and percent vector are archived in our relational database for reports, trend analysis, and troubleshooting. We have recently completed a rearraying procedure for the pre-finishing and gap closure phases using laboratory robots, which are also integrated into our database and sample-tracking system.

To support quality assurance procedures, we have developed an automated reporting system which uses archived analysis results as well as project information to generate and distribute a series of reports which include an estimate of sequencing efficiencies and project consensus quality. We also have provided WWW access to project directories, which display current quality plots and assembly information for each project. In addition, we have developed a suite of tools which can be used for analyzing specific projects for sequencing efficiencies, read length, etc. We plan to continue to develop and expand our informatics capabilities as we continue to significantly scale up our throughput over the next several years.

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.

Software Tools for Data Analysis in Automated DNA Sequencing

Michael C. Giddings, Jessica Severin, Michael Westphall, and Lloyd M. Smith
University of Wisconsin-Madison Chemistry Dept., 1101 University Ave., Madison, WI 53703
giddings@whitewater.chem.wisc.edu

A crucial component of the automated DNA sequencing process is the analysis software. The software analysis can be roughly divided into four primary steps: gel analysis, base calling, assembly, and finishing. In each of these steps, the software is responsible for the difficult task of accurately analyzing large quantities of complex data and deriving information that is useful to end users of the data.

We have developed software for gel analysis and base calling utilizing a cross-platform, modular, object oriented architecture. The core of this software is BaseFinder, a framework for trace processing, analysis, and base-calling. BaseFinder is highly extensible, allowing the addition of trace analysis and processing modules without recompilation. Powerful scripting capabilities combined with modularity allow the user to customize BaseFinder to virtually any type of trace processing. It currently runs on Windows/NT (Microsoft) and OpenStep/Mach (NeXT), with ongoing work to port it to additional operating systems, including: Solaris (Sun), Rhapsody (Apple), and Linux (GNU/freeware). The Solaris is in progress and is expected to be completed shortly.

Base calling is currently performed by an adaptive, iterative algorithm which can be quickly tuned to work on data from a large variety of sequencing instruments. The base calling module, when applied in combination with appropriate pre-processing of the data, provides fast, accurate base labeling along with confidence measures on the bases called. It has been extensively tested with data from a number of machines, including the ABI 373A with and without the "stretch" (long gel) upgrade, our in-house Horizontal Ultrathin Gel Electrophoresis (HUGE) system, our vertical slab-gel scanning system, and a capillary sequencing system.

We are also working on a project to incorporate these programs and software developed elsewhere, along with a relational database engine, utilizing distributed object technology. This system has the promise of providing a robust means of dealing with the large quantities of complex data that must be handled in a genome sequencing center, while providing a simple and consistent view of the entire process to a human operator.

Treasures and Artifacts from Human Genome Sequence Analysis

Darrell O. Ricke and Larry L. Deaven
Los Alamos National Laboratory, Center for Human Genome Studies, Los Alamos, New Mexico 87545
ricke@telomere.lanl.gov

A very interesting aspect of the Human Genome Project (HGP) is the analysis of the sequences being generated. Sequence analysis of genomic sequence data is yielding treasures and interesting artifacts. Human genomic sequence analysis finds significant similarities to (1) known genes, (2) known repetitive sequences (Alu, L1, etc.), (3) human and mammalian ESTs, (4) trapped exons, (5) gene orthologs and paralogs (both at the DNA and protein level), (6) tRNAs, (7) snRNAs, (8) rRNAs, (9) CpG DNA sequences, etc. Multiple significant protein sequence similarities have been observed between translated human genomic regions (i.e., exons) and protein sequences from drosophila, yeast, and *C.elegans*. Similarities with less than 90% identity between a human genomic sequence and a human EST sequence usually represent two members of a multigene family or a possible pseudogene. Surprisingly, human genomic DNA includes multiple regions with significant similarities to mitochondrial DNA. Analysis of over 6 megabases of human genomic sequence data has detected 244 EST clusters, 21 known genes, 55 novel genes, and 35 trapped exons. These and many other treasures await us as the HGP progresses.

Identification of the treasures in the genomic sequence data is confounded by the multiple artifacts that are encountered. Artifacts that need to be ignored include: low complexity similarities, similarities to unannotated repeats, similarities to unannotated contaminating sequences (vector, yeast, and *E.coli*), exon predictions within repetitive sequences, etc. Many EST, intron, and genomic sequences contain novel repeats that have not been previously described or annotated. One hallmark of a novel repeat in an EST is the lack of sequence similarity between the rest of the EST sequence and genomic sequence surrounding the region of similarity. One particularly interesting artifact is the presence of a human Alu repeat in a bacterial database sequence (pva). Most likely this represents a bacterial sequence contaminated with human sequence. However, horizontal DNA transfer can not be ruled out. Other interesting examples include similarities between human DNA and plant DNA database sequences.

http://www.jgi.doe.gov and
http://www.chgs.lanl.gov

Annotating and Masking Repetitive Sequences with RepeatMasker

Arian F.A. Smit and Phil Green
Human Genome Center, Department of Medicine,
University of Washington, Seattle, WA 98195

Probably over half of mammalian genomic DNA is comprised of sequences that still can be recognized to be derived from transposable elements. These interspersed elements as well as simple repeats and low complexity DNA need to be masked before performing database searches, since interesting sequence similarities may be overlooked amidst the many spurious matches to these repeats. Furthermore, recognition of interspersed repeats dramatically can improve gene prediction and interspecies sequence comparison and is useful as a finishing tool in genomic sequencing. RepeatMasker is a widely used program to annotate and mask repetitive DNA sequences. Here, we report on recent improvements in the program and databases as well as development currently in progress. We also present examples of the usefulness of RepeatMasker in database searches, gene prediction and evolutionary studies. The program can be run locally on any computer with sufficient memory, or can be accessed via the web at http://ftp.genome.washington.edu/cgi-bin/RepeatMasker or by e-mail at repeatmasker@ftp.genome.washington.edu.

Why Is Basecalling Hard to Do Well? Sources of Variability in DNA Sequencing Traces and Their Consequences

David O. Nelson
Human Genome Center
Lawrence Livermore National Laboratory
daven@llnl.gov

Terence P. Speed
Statistics Department
University of California, Berkeley

"Basecalling" is the process of inferring the sequence of a segment of DNA from data arising from electrophoresis traces. Assigning a realistic, probabilistic quality measure to the inferred bases in a DNA sequence requires knowledge of the variability of the data, given the underlying sequence. Treating the sequencing process as an attempt to decode a message in a digital communications system provides us with a framework for analyzing that variability, quantifying it, assigning it to various features of the system, and assessing the performance consequences.

We have examined traces of known sequences from locally produced ABI 377 data to assess the extent of variability in DNA traces. We present data showing the variability of interpeak distance and peak width as a function of base number. We also present data showing the relationship between base number and the amount of information in the signal available for basecalling. We show that, as the number of bases called increases, the primary stumbling block to accurate basecalling is similar to the problem of "synchronization" in a digital communications system. We examine the relationship between the probabilistic quality scores produced by Phil Green's basecaller "phred" and the local resolution of the signal, defined for a consecutive pair of peaks as the ratio of the interarrival time between the two peaks and the sum of the scale parameters for the peaks. We derive a simple linear model for the probability that the resolution is less than one. Our data shows that beyond a few hundred bases phred quality scores appear to be driven mainly by the probability that the resolution of the signal is less than one. These statistical characteristics of the underlying signal help to explain the abruptness of the well-known "phase transition" from high quality to low quality decisions seen in DNA sequence data.

A Statistical Model for Basecalling

Lei Li, David O. Nelson, and Terence P. Speed
Department of Statistics, University of California, Berkeley
lilei@stat.berkeley.edu, daven@llnl.gov, terry@stat.berkeley.edu

Base-calling is that part of automated DNA sequencing which takes the time-varying signal of four fluorescence intensities and produces an estimate of the underlying DNA sequence which gave rise to that signal.

We approach the problem of base-calling from a statistical perspective, hoping to make use of a statistical model to call bases, and also to attach suitable measures of uncertainty to the bases we call. We model the automated Sanger sequencing process by the following steps. First, the underlying sequence of bases is encoded by a hidden Markov model into a virtual signal, consisting of four spike trains. Second, each component of the virtual signal is distorted by a slowly-changing point spread function, to represent the diffusion effect of electrophoresis. Third, mobility shifts are applied to the components separately, with the result being the (model) time varying concentrations of the four bases. Finally, these dye concentrations are converted into fluorescence intensities by the application of an instrument-dependent cross-talk matrix.

Signals simulated according to this model exhibit many, but certainly not all of the features found in real sequencing traces. We hope to have captured the important ones. Our base-calling strategy is to invert the process described in the model. Specifically, we combine algorithms for color separation, mobility adjustment and deconvolution with a hidden Markov model decoder. The output of this analysis will be a "best" estimate of the sequence of bases that gave rise to the data. In addition, for each base called, the analysis will provide the marginal probability of any alternative base call at that position.

We will report our progress in the design of the hidden Markov model, and the deconvolution and color-separation steps.

An Expert System for Base Calling in Four-Color DNA Sequencing by Capillary and Slab Gel Electrophoresis

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

An important consideration in fluorescent DNA sequencing strategies is to extract as much information as possible from each run. In previous work in which we described the sequencing of more than 1000 bases per run by capillary electrophoresis (Carrilho et al., Anal. Chem. 1996, 68, 3305-3313), one characteristic contributing to the extended read length was the use of sophisticated base calling algorithms. We have recently developed a new base calling method that shows promise to increase read lengths for both capillary and slab gel (e.g., ABI) electrophoresis by decreasing the number of errors at long migration times. For example, errors between 800 and 1100 bases in the above cited paper are reduced by 40% relative to the graph-theoretic method employed at that time. Accuracy in this late region is better by the new procedure than by any other one we have tested to date. Data processing with the new system begins by determining the dye spectra from the raw data file in order to perform color separation. This step is followed by baseline subtraction, and bases are subsequently assigned in a moving time window, roughly ten bases wide. Using a set of empirical rules, each channel in the window is divided into the smallest sections likely to contain at least one call, and then a second set of rules is applied to determine the final calls. Confidences on the base assignments are supplied by statistical correlation of the variables used in the rules, such as peak height and width, with errors made on known sequence. The system has been applied to four-dye separations using four-channel or full-spectral detection. Results on a large body of data will be shown, along with comparisons to several widely used base callers.

This work is being supported by D. O. E. grant #DE-FG02-90ER60895.

Web-Based Tools for the Analysis and Display of DNA Trace Data

Judith D. Cohn, Mark O. Mundt, A. Christine Munk, Larry L. Deaven and Darrell O. Ricke
Los Alamos National Laboratory, Los Alamos, New Mexico 87545
cohn@lanl.gov

The Human Genome Project worldwide is currently in transition towards a new era of vastly increased production of DNA sequences. Anticipating the requirements of increased sequence production, the Center for Human Genome Studies (CHGS) at Los Alamos began work in 1996 on a new suite of integrated, web-based tools for processing and evaluating sequence data. Major development goals were: 1) to limit the need for human intervention in the analysis process; 2) to design highly modular software, which could run on multiple platforms while accessing a single, centralized (though possibly distributed) database; and 3) to build user-friendly GUI applications. In order to meet these goals, we chose the Java programming language as our primary development environment. Currently, a number of pieces are in every day use and will be described here as well as in other posters/presentations by our Informatics Team.

Working applications for the analysis and display of DNA trace data from automated DNA sequencing instruments include Trace Viewer, Base Caller, Sequence Trimmer and Production Statistics. While these applications have been designed to work with data from a variety of sources, they have been implemented thus far for use with data from ABI sequencing instruments stored in a central flat-file database. We are in the process of moving the data to an Informix object-relational database.

The Trace Viewer displays trace data from a single sample file at multiple resolutions and includes such features as semantic zooming and coordinated scrolling. At present, it is possible to view ABI, phred and our own base calls along with associated quality numbers and output from the Sequence Trimmer.

The CHGS Base Caller was designed to replace our use of the ABI base caller. At the time, we did not have access to the phred base caller. Designing our own base caller, however, opened up the possibility of achieving additional goals: 1) more accurate base calling under a variety of conditions (e.g. dye-primer versus dye-terminator, new dye sets, new polymerases, new sequencing instrumentation, etc); 2) a robust set of quality numbers at each base position, which can be used to fine tune analysis further down the pipeline, e.g. assembly or primer picking. To date these quality numbers include an overall quality assessment (similar to phred), a quality number for each base and a gap statistic.

The Sequence Trimmer performs two automated functions, which together define a "clear" region for each trace sequence. The first function is to designate a "high quality" region. This is determined using the overall quality scores produced by our base caller and can be modified by manipulating one or parameters. The second function is to trim vector sequence. Vector trimming has been optimized to recognize even very poor quality sequence if it appears at the appropriate position in the sequence. The vector trimming function is also used to confirm the status of gel control sequences.

Joint Genome Institute's (JGI) Informatics Plans and Needs

Darrell O. Ricke, Tom Slezak, Sam Pitluck, and Elbert Branscomb
Joint Genome Institute
ricke@telomere.lanl.gov

The DOE funded human genome centers at LANL, LBNL, and LLNL are joining together to form the DOE Joint Genome Institute (JGI). Under the JGI, the informatics teams at LANL, LBNL, and LLNL are working together to focus on meeting JGI needs. The JGI informatics needs are considerable. JGI projects include scaling up both shotgun and transposon-based sequencing, scaling up sequence ready physical map production, new functional genomics projects, systems and data integration, and building software systems for the new JGI Production Sequencing Facility (PSF). To illustrate this complexity, functional genomics projects will generate annotation information for JGI sequenced regions of the human genome in the form of (1) human cDNA sequences, (2) mouse cDNA sequences, (3) sequences for targeted syntenic mouse genomic regions, and (4) gene expression data. To manage data and software distributive across four sites, integrated databases and WWW software projects are being designed and planned. While these integrated systems are being put into place, existing software and systems will continue to be used and in some instances enhanced to meet production scale-up needs. A summary of plans, status of major projects, and informatics needs will be presented.

See URL: http:\\www.jgi.doe.gov

Restriction Map Display on the World Wide Web

Mark C. Wagner, Thomas R. Slezak, Arthur Kobayashi, David J. Ow, Linda K. Ashworth, Laurie A. Gordon, Anne S. Olsen, Anthony V. Carrano
Joint Genome Institute, Lawrence Livermore National Laboratory, Livermore, CA 94550
wagner5@llnl.gov

The amount of data generated by a Genome Center precludes anything but a graphical interface to the database. The graphical user interface used by the Human Genome Project at Lawrence Livermore National Laboratory predated the World Wide Web, being written for an environment consisting solely of Unix workstations, and was sufficient to meet the needs of our local researchers and more sophisticated collaborators.

However, current collaborative agreements require immediate public release of data, and this has necessitated a shift from a specific type of hardware platform to a much wider audience. The advent of Web based technologies (particularly Java) have made it possible to write interfaces to our database for public use, without the necessity of dealing with software upgrades and hardware dependency problems. We have written a WWW version of the Restriction Map display of our Genome Browser software which will enable access to our database from anywhere on the Internet, and from any type of system supporting Web browsers.

This software will serve the clone resources task of the Joint Genome Institute (JGI) in addition to the Genome Center at LLNL. It is our intention that this software will become a prototype for a larger package that will include other types of displays for data generated by the JGI.

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.

The BCM Search Launcher — Providing Enhanced Sequence Analysis Search Services

Kim C. Worley, Pamela A. Culpepper, Daniel B. Davison
Department of Molecular and Human Genetics and Department of Cell Biology, Baylor College of Medicine, Houston, TX
kworley@bcm.tmc.edu

We provide Genome Program investigators 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 additional hypertext links to provide easy access to Medline abstracts, links to related sequences, and additional information which can be extremely helpful when analyzing database search results.

For frequent users of sequence analysis services, the BCM Search Launcher Batch Client provides access to all of the searches available from the BCM Search Launcher web pages in a convenient drag and drop (on the Macintosh) or command line (Unix) interface. The BCM Search Launcher Batch Client is a Unix and Macintosh application that automatically 1) reads-in 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.

One of the tools unique to the Search Launcher is BEAUTY, our Blast Enhanced Alignment Utility. BEAUTY makes it much easier to identify weak, but functionally significant matches in BLAST protein database searches. BEAUTY generates an alignment display showing the relative locations of annotated domains and the local BLAST hits in each matched sequence, greatly facilitating the analysis of BLAST search results. Recent improvements make BEAUTY searches available for DNA queries (BEAUTY-X) and for gapped alignment searches (using WU-BLAST2). In addition, new releases of the Annotated Domains database used with BEAUTY are produced for each full GenBank release. These up-to-date versions of the database present annotation information for many more sequences than previous editions. From the Search Launcher, users can submit sequences to the NCBI's BLAST network server to search the non-redundant, daily-updated database, and have their search results returned with BEAUTY displays added.

Our future development will focus on the analysis of large scale genomic sequences to support the efforts of the Genome Annotation Collaboratory.

This research is supported by a grant from the U.S. Department of Energy Office of Health and Environmental Research (DE-FG03-95ER62097/A000), and grants from the National Human Genome Research Institute, National Institutes of Health (1F32-HG00133-03, 1R01-HG00973-03).

SubmitData Data Submission Framework

David Demirjian, **Sushil Nachnani**, Manfred Zorn
Lawrence Berkeley National Laboratory
sknachnani@lbl.gov

SubmitData is a data translation and submission framework. It is being built to provide a common user representation to public genome databases having different internal formats. SubmitData parses in the schema of a particular database and provides the user with a forms-like display to enter his or her data. The user can define a number of different types of variables to incorporate data files generated by another program e.g. Excel. SubmitData will replace the variables with actual fields read in from these data files, when submitting a transaction. Thus the user can define a template consisting of fields having a constant value along with fields with variables defined as their value. This template can then be used to process a number of data files over a period of time. SubmitData also takes care of translating the transaction into a format expected by the specific database.

The framework implemented initially in Smalltalk and subsequently in the Java programming language contains five main categories of classes: data representation, user interface, parser/builder, printer and batch processing. The data representation objects are the unchanging internal common representation of objects and fields in the various databases. The user interface objects serve as the views of the data objects in a forms-like display. The parser/builder classes are responsible for parsing the schemas of public databases and building the definitions for the common data objects. The printer set of classes translate the internal data object representation into a format required by a particular public database for submitting a transaction. The batch process objects allow the user to define different types of variables and incorporate data files for submitting batch transactions.

Graphical Ad hoc Query Interface for Federated Genome Databases

Dong-Guk Shin,[1] Lung-Yung Chu,[1] Wally Grajewski,[1] Joseph Leone,[2] Thomas Barnes,[2] and Rich Landers,[2]
[1] Computer Science & Eng., University of Connecticut, Storrs, CT 06269-3155
[2] CyberConnect EZ, LLC, Storrs, CT 06268
shin@cse.uconn.edu

We have been developing the Graphical SQL Query Editor capable of aiding genome scientists in learning and/or examining third-party database schemas in a relatively short time and assisting them in rapidly producing correct SQL queries. Specifically, our goal has been to allow a user to form an SQL query within a 5 - 10 minute time frame despite a lack of familiarity with the schemas of the public federated genome databases. Using the SQL Editor, genome scientists can construct queries targeting not only a single database but also distributed queries targeting multiple databases. Currently the SQL Editor interlinks GDB, GSDB, EGAD and SST at TIGR, MGD at Jackson Laboratory, and CHR 12, the chromosome 12 database at Yale University.

The SQL Editor is a client program written in Java which can be downloaded as an applet via the Internet or an Intranet. Database schema information is imported by clicking on the button representing a given database. For example, clicking on the EGAD button causes the EGAD table list to be imported, on-the-fly, from the remote server at TIGR. Similarly, the list of attributes for each table can also be imported, on-the-fly, with a single mouse click. Users express SQL queries by browsing imported database schema information, choosing the items of interest, and graphically specifying SQL selection clauses, restriction conditions, and join links. These join links are capable of linking tables in different databases, and are the means of constructing distributed queries.

One unique feature of this interface is the SchemaViewer. Through the SchemaViewer, the system can suggest to the user semantically correct ways of joining tables. This feature, which we call join path discovery, can even discover join paths that cross database boundaries (i.e. distributed join paths). Join path discovery is supported by a meta-data repository, which is constructed for each participating genome database in the federation. SchemaViewer's Selection, Zooming and Abstraction features allow the user to browse, and choose among multiple system-discovered join paths. Once a join path is chosen, it can be exported to the main SQL Editor window for refinement into a complete query. The SQL Editor also includes other features designed to enhance the user's query formulation.

Acknowledgements
(1) The author's work was supported in part by the NIH/NHGRI Grant No. HG00772-03.
(2) The author's work was supported in part by the DOE SBIR Phase II Grant No. DE-FG02-95ER81906.

Towards a Comprehensive Conceptual Consensus of the Expressed Human Genome: A Novel Error Analytical Approach to EST Consensus Databases

Robert Miller, John Burke, Alan Christoffels and **Winston Hide**
South African National Bioinformatics Institute. The University of the Western Cape, Private Bag X17, Cape Town, South Africa
winhide@sanbi.ac.za

Human gene data is currently mostly available in fragments in the form of expressed sequence tags (ESTs) and only a relatively minor fraction of all human genes are completely sequenced. The fragmentary but redundant nature of ESTs provides a large but complex and error prone resource for gene discovery.We have developed a novel set of highly portable tools to manufacture and process a database of publicly available assembled consensi of Human ESTs and alignments. The database represents an easily distributable, core information resource upon which a comprehensive knowledgebase can be built. We maximise the coverage of accurate high quality consensus sequences of human genes, perform error compensation and analysis, and provide a measure of error so that we can derive the best possible estimate of the makeup of the human gene set from the data as it becomes available. The system has been designed to derive extended consensus representation of the expressed human genome.

The database differs markedly from indices such as TIGR Gene Index (1), and also databases of clusters of ESTs such as UniGene (2) because it does not discard noisy information. Instead, all possible information is used for clustering using d2-cluster (Burke, Davison, Hide, in prep), a global word based clustering methodology. The "dirty" information is carefully checked for useful constituent subsequences. As a result, extended gene consensi are manufactured containing both high quality and poor quality regions; these are duly annotated. The database has relational access to annotated alignments via the Genome Sequence Database (3). Alignment of the clusters can be on a system with a specific consensus builder using a combination of two error analysis systems: DRAW and CONTIGPROC. Each entry contains all fragments and isoforms of the gene in serial association, separated by spacers. Thus the maximum possible consensus for each gene exists, providing a useful reagent for functional analysis. Comparison of gene sequence, aligned clusters and "alternative splice" frequencies from STACK now allows a more comprehensive understanding of the nature of expressed genes to be performed. We are in the process of discovering what is artifact, and what is genome biology.

References
(1) http://www.tigr.org/tdb/hgi/
(2) http://www.ncbi.nlm.nih.gov/UniGene/index.html
(3) http://www.ncgr.org/gsdb/

Query and Display of Comprehensive Maps in GDB

Stanley Letovsky, Robert Cottingham and GDB Staff
Genome Database, Johns Hopkins University, Baltimore MD
letovsky@gdb.org

GDB, the human Genome Database (http://www.gdb.org), stores whole chromosome and regional maps of the human genome from a variety of sources. Mapping methologies representing in the database include linkage, radiation hybrid, content contig, and various forms of cytogenetic mapping. An important class of queries against GDB look for markers in some region of interest, possibly combined with additional restrictions on the types of markers or their properties. It is useful for such positional queries to be able to search all maps, regardless of type, source, or whether or not they happen to contain the markers used to specify the region of interest. We do this by combining the various maps of each chromosome into a single comprehensive map; positional queries are then expressed against the coordinates of the comprehensive map.

The comprehensive maps are generated by a novel integration algorithm that constructs nonlinear warpings of maps in order to bring common markers into correspondence. The comprehensive maps produced by this algorithm are a significant improvement over the linear deformation method used previously for this purpose. The improvement translates into an increased accuracy of positional querying.

The comprehensive maps can also be viewed using our Mapview display program, which is accessible from the Web. A recent software enhancement allows the results of queries on any type of mapped object (gene, clone, amplimer, etc.) to be displayed as dynamically generated maps, as well as in tabular form. These map displays are based on the comprehensive maps. Queries which can be expressed in this fashion include "find genes in region", "find sequenced clones in region", "find polymorphic amplimers in region", and so on. As the content of GDB is extended in various areas, it becomes possible to ask other queries using this same basic capability. For example, we are currently adding the capability to store gene function and expression data; when this is in place it will be possible to ask for a map of genes that are highly expressed in a given tissue, or genes that have a certain function.

GSDB - New Capabilities and Unique Data Sets

A. Farmer, C. Harger, S. Hoisie, P. Hraber, D. Kiphart, L. Krakowski, M. McLeod, J. Schwertfeger, A. Siepel, G. Singh, M. Skupski, D. Stamper, P. Steadman, N. Thayer, R. Thompson, P. Wargo, M. Waugh, J.J. Zhuang, and **P.A. Schad**

The Genome Sequence DataBase (GSDB), at the National Center for Genome Resources (NCGR), is moving forward to provide the research community, with new data access capabilities and unique data sets. GSDB has improved sequence data access by creating a web-based program (Excerpt) that allows researchers to extract desired portions of sequences from the database; designing and implementing a GSDB flatfile that displays all of the data which can be represented in text format; and improving a web-based query tool (Maestro). In addition, major improvements have been made to the software suite that imports data from the IC databases into GSDB.

GSDB has created unique data sets by constructing alignments to high-profile sequences (such as the complete *E. coli* genome with over 5000 alignments to other *E. coli* sequences in the database); and by constructing and maintaining discontiguous sequences that represent sequence-based chromosome maps (such as human chromosome X which is comprised of over 1400 sequences markers and several larger genomic sequences). In addition, there are ongoing projects to improve the quality of data within the database, such as the identification and subsequent removal of vector contamination from the 5' and 3' ends of sequences.

These tools and data are available to the public from the GSDB web site (www.ncgr.org/gsdb).

GSDB has been gaining momentum over the past six months and will continue to develop new and innovative ways to access data in the database. GSDB will also continue to create unique data sets to provide to the public. Planned enhancements include the development of a web-based sequence viewer to replace Annotator, improvements to Excerpt and Maestro, collaborations with JGI and ORNL, and the incorporation of unique data sets such as SANBI's STACK data.

Database Transformations for Biological Applications*

G. Christian Overton, **Susan B. Davidson**, Peter Buneman
Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104
Tel: (215) 898-3490; Fax (215) 898-0587
{susan,peter,coverton}@central.cis.upenn.edu

The overall goal of the Kleisli project is to develop a suite of tools to perform data source transformation and integration. The central components under development are the high-level query language and system, CPL/Kleisli, a general schema description language and a transformation constraint language, TSL. (Development of TSL is funded through an NSF grant.) The two main tools--Morphase and CPL--complement one another; Morphase is a heavier-weight system designed to transform an entire database or dataset according to user specifications and constraints written in TSL, a language which allows one to naturally express a broad class of such operations. Conversely, CPL is of most use precisely when there are multiple distributed source databases and transforming each in its entirety to a uniform representation would be infeasible. In such situations, CPL permits complex queries over the distributed heterogeneous data sources, performing integration "on the fly", while simultaneously allowing powerful structural transformations to be done. CPL's extensible query optimizer ensures that this all takes place in a timely fashion.

Together, the tools can be used either to perform data integration by providing dynamic user-defined views, or to create specialized data warehouses. This layer can then be used underneath data mining, OLAP or other decision making tools.

Recent developments include:

- Coupling CPL to OPM-Based Databases. CPL/Kleisli can now query against GDB 6.0's Object Broker Server (documented at http://gdbwww.gdb.org/ob/top.html) using the OPM query language. Having reimplemented our existing Web-based queries to take advantage of GDB 6.0's richer OPM-based schema, we envision as a next step embedding some or all of OPM's query translator in the CPL optimizer in a modular fashion. Since OPM queries are ultimately translated to one or more relational SQL queries by the OPM tools anyway, we can perform this translation at an earlier stage of the query process, producing SQL to which CPL can apply its maximal subquery migration capabilities, pushing to the servers operations which would otherwise have to be performed locally. Furthermore, opening up the query in this manner allows CPL to apply other RDBMS-specific techniques, such as semijoins, essentially leveraging the work already done on relational multidatabase optimization to OPM queries.
- Complex Object Libraries. As a first step towards migrating CPL onto more ubiquitous and widely-accepted language platforms we have implemented a prototype set of complex object manipulation libraries in each of the C++, Perl, and Java(TM) languages. These libraries allow application programs written in any of the target languages to dispatch queries to a central CPL server (in the first stage of migration the server will still run under ML) and receive query results in an abstract representation appropriate to the specific host language, with an API which is uniform across all of them. The Perl and C++ complex object libraries are currently being used in an automated annotation testbed, GAIA (http://agave.humgen.upenn.edu/gaia/). This approach allows CPL queries to be seamlessly integrated as part of the annotation process.
- Local Storage Management. The customizable optimizer of CPL employs a multi-pass source-to-source rewriting strategy to attempt to minimize both the response time and intermediate storage consumption of queries. However, to be able to efficiently store and retrieve ever-larger intermediate query results, preferably along with one or more indices, we are currently exploring the use of a more sophisticated local backing store using either the freely-available SHORE system, the relational Sybase database system, or perhaps OPM.
- User Interfaces. As a first step towards developing a simple and powerful user interface to CPL, we have created a number of Web-based stereotypic queries. To date, we have focused on the implementation and optimization of queries which were chosen to test system performance in formulating and executing distributed queries while providing functionality to a growing user community. Eight major classes of queries are available with new ones being added by request from the user community.

These queries, shown below with the databases accessed indicated in parenthesis, demonstrate how Kleisli can be used to integrate data stored in disparate formats and physical locations. More details on the queries can be found at the Kleisli homepage, http://agave. humgen.upenn.edu/cpl/cplhome.html.

- "Complete Genome" query, e.g., "Return all complete mitochondrial genomes larger than 20kb." (GSDB.)
- EST Location query, e.g., "Find the location of a mapped EST." (dbEST or GenBank or GSDB and GDB.)
- Gene/Location query; i.e., "Find protein kinase genes on human chromosome 4." (GDB.)
- Sequence/Size query; i.e., "Find mapped sequences longer than 100,000 base pairs on human chromosome 17." (GDB and GSDB.)
- Mapped EST query; i.e., "Find ESTs mapped to chromosome 4 between q21.1 and q21.2." (GDB and GSDB.)
- Primate alu query; i.e., "Find primate genomic sequences with alu elements located inside a gene domain." (BLAST and GSDB.)
- BLAST Sequence/Feature query; i.e., "Find sequence entries with homologs of my sequence inside an mRNA region." (BLAST and GSDB.)
- Human genome map search; i.e., "Find human sequence entries on human chromosome 22 overlapping q12." (GDB, GSDB and ASN.1 GenBank.)

*Supported by a grant from the Department of Energy, 94ER61923.

Recent Kleisli References

"Querying an Object-Oriented Database Using CPL," S.B. Davidson, C. Hara and L. Popa. Proceedings of the Brazilian Symposium on Databases (October 1997).

"WOL: A Language for Database Transformations and Constraints," S.B. Davidson and A. Kosky. Proceedings of the International Conference of Data Engineering, April 1997 (Glasgow, Scotland).

"BioKleisli: A Digital Library for Biomedical Researchers," S.B. Davidson, C. Overton, V. Tannen and L. Wong. Journal of Digital Libraries 1:1 (November 1996).

"A Data Transformation System for Biological Data Sources," P. Buneman, S.B. Davidson, K. Hart, C. Overton and L. Wong. Proceedings of VLDB, Sept. 1995 (Zurich, Switzerland).

"Challenges in Integrating Biological Data Sources," S.B. Davidson, C. Overton and P. Buneman. J. Computational Biology 2 (1995), pp 557-572.

"Semantics of Database Transformations," A. Kosky, S.B. Davidson and P. Buneman. Semantics of Databases, edited by L. Libkin and B. Thalheim.

"Transforming Databases with Recursive Data Structures," A. Kosky. PhD Thesis, December 1995.

Exploring Heterogeneous Biological Databases with the OPM Multidatabase Tools

Victor M. Markowitz, I-Min A. Chen, Anthony Kosky, Ernest Szeto
Lawrence Berkeley National Laboratory, Berkeley, CA 94720
VMMarkowitz@lbl.gov

The Object-Protocol Model (OPM) data management tools provide support for rapid development, documentation, and flexible exploration of scientific databases. Several archival molecular biology databases have been designed and implemented using the OPM tools, including the Genome Database (GDB) and the Resource Center Primary Database (RZPD) of the German Human Genome Project, while other databases, such as the Genome Sequence Database (GSDB) and NCBI's Genbank have been retrofitted with semantically enhanced views using the OPM tools.

The multidatabase OPM tools provide powerful facilities for: (1) assembling (federating) heterogeneous databases into a multidatabase system, while documenting their structure and inter-database links; (2) processing ad-hoc multidatabase queries via uniform OPM interfaces; and (3) assisting scientists in specifying and interpreting multidatabase queries. Incorporating a database into a multidatabase system involves constructing one or more OPM views of the database and entering information about the database and its views into an Multidatabase Directory (see http://gizmo.lbl.gov/DM_TOOLS/OPM/MBD/MBD.html). This Directory records information necessary for accessing and formulating queries over the component databases, including: general information required for accessing the database; structural information on the schemas of each database; and information on known links between databases, including semantic descriptions of the links, and data manipulations necessary in order to traverse such links.

Queries over an OPM based multidatabase system are expressed in an extension of the OPM query language, that includes additional constructs necessary for accessing multiple databases. Multidatabase queries are processed by generating queries over individual databases, and combining the results using a local query processor. A Java graphical multidatabase query construction tool provides support for dynamic construction and automatic generation of Web query forms that can be then either used for further specifying conditions, or can be saved and customized. The OPM multidatabase tools have been applied to several projects, including the construction of a federation of molecular biology databases involving GDB, GSDB, and Genbank (http://gizmo.lbl.gov/jopmDemo/gdbs_mqs.html), and are currently used for setting up a federation of biocollections databases and a federation of several databases involved in the German Human Genome Project.

The main problem we are currently addressing consists of identifying, documenting, and formally defining a comprehensive set of relevant inter-database links that drive multidatabase queries and applications.

This work is 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 Contract DE-AC03-76SF00098.

The OPM Data Management Toolkits Anno 1997

Victor M. Markowitz, I-Min A. Chen, Anthony Kosky, Ernest Szeto, William Barber, Thodoros Topaloglou
Lawrence Berkeley National Laboratory, Berkeley, CA 94720
VMMarkowitz@lbl.gov

The Object-Protocol Model (OPM) data management tools provide support for rapid development, documentation, and exploration of scientific databases. These tools are based on OPM, an object data model that is similar to the ODMG standard, but also has additional constructs for modeling scientific data. Databases designed in OPM can be implemented with commercial relational DBMSs, using the OPM Database Development Toolkit that includes OPM schema translators for generating complete DBMS database definitions from OPM schemas, a Java based OPM schema editor and browser, tools for specifying and maintaining multiple OPM views, and OPM schema publishing tools for documenting OPM databases in a variety of formats and notations. OPM schemas can be also retrofitted on top of existing relational databases or structured files defined using notations such as the ASN.1 data exchange format. Native or retrofitted OPM databases can be queried using the OPM Database Query Toolkit that includes OPM query language translators that interpret queries expressed in the ODMG compliant OPM query language (OPM-QL) and translate them into the languages supported by the underlying DBMS. A Web-based OPM query interface allows graphical construction of ad-hoc OPM queries and can be used for generating Web query forms.

An OPM Multidatabase Toolkit contains tools that support: (1) assembling heterogeneous databases into an OPM based multidatabase system, while documenting their schemas and inter-database links; (2) processing ad-hoc multidatabase queries via uniform OPM interfaces; and (3) assisting scientists in specifying and interpreting multidatabase queries.

Several archival molecular biology databases have been designed and implemented using the OPM tools, including the Genome Database (GDB) and the Resource Center Primary Database (RZPD) of the German Human Genome Project, while other biological databases, such as the Genome Sequence Database (GSDB) and NCBI's Genbank, have been retrofitted with semantically enhanced OPM views. The OPM multidatabase tools have been applied for constructing a molecular biology database federation and for providing a common interface on top of a variety of different biological databases.

Current OPM work includes further development of the OPM multidatabase tools and extending the OPM toolkit to support complex data types, such as DNA sequences and 3-dimensional crystallographic data, irrespective of the underlying DBMS facilities.

Examples, documentation, and papers regarding the OPM toolkits are available at http://gizmo.lbl.gov/opm.html.

This work is 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 Contract DE-AC03-76SF00098.

Quality Control in Sequence Assembly Analysis

Mark O. Mundt, Judith D. Cohn, Tracy L. Ricke, P. Scott White, Larry L. Deaven and Darrell O. Ricke
Los Alamos National Laboratory, Life Sciences Division and Center for Human Genome Studies, Los Alamos, New Mexico 87545
mom@telomere.lanl.gov

As participants in the Human Genome Project produce more sequence data at higher rates, there is less chance that any given base pair will be manually edited by interactive means. Moreover, assembly decisions made by automatic programs must be scrutinized in a more efficient manner. To further complicate matters, natural variation among individual human sequences is at least five to ten fold higher than the desired quality standard for finished sequence. On average, we expect to see approximately one polymorphism in every 1000 base pairs, while the community's sequencing error standard is now projected at one in every 10,000 base pairs. Assessment of quality and variation must therefore be efficiently automated.

At the Center for Human Genome Studies in Los Alamos, we are using our own base caller and quality values to derive input for the TIGR assembly program. Qualitative comparisons are then made between individual cosmid clone assemblies and ones using overlapping clones. Phred and phrap results are also contrasted by means of algorithms operating on a novel, standard assembly format. Differences in assemblies are now automatically detected, and our future objectives would include using quantitative measures and adding repeat profiles to make decisions favoring the results which best meet the goals of the project. These tools along with graphical analysis displays are being integrated into a web-based Java system.

In addition to quality assessment, we are concurrently searching for single nucleotide polymorphisms (SNPs). Because of the higher mutation rate relative to other sequences (17 to 25 fold higher) we are presently targeting CpG islands in our 7q telomeric sequence for resequencing. For this automated process, we have developed software that targets regions of high densities of CpG dinucleotides, and we use Whitehead Institute's Primer3 software to design PCR and sequencing primers. These are used to amplify and sequence genomic DNA templates from several diverse individuals in search of SNPs. Low quality regions of sequence are targeted in a similar manner, and the multiple assembly information is used to target weak joins.

In conclusion, an automated system for designing PCR and resequencing primers, targeting specific regions to assess sequence quality and natural variation, is being implemented and optimized.

Automating the Detection of Human DNA Variations

Scott L. Taylor, Mark J. Rieder, and **Deborah A. Nickerson**
Department of Molecular Biotechnology,
University of Washington, Seattle, WA 98195
debnick@u.washington.edu

Fluorescence-based sequencing is playing an increasingly important role in efforts to identify DNA polymorphisms/mutations in genes of biological and medical interest. We have developed a computer program known as PolyPhred that automatically detects the presence of single nucleotide substitutions in their heterozygous state using fluorescence-based sequencing of PCR products. The operation of PolyPhred will be described as well as its integration with the Phred base-calling program, the Phrap assembly program, and the Consed viewing program. Additionally, we will illustrate how Consed can be leveraged to display a set of highly annotated reference sequences that greatly simplifies the analysis of DNA variations with respect to existing information on gene structure, PCR primers, and previously known DNA polymorphisms or mutations. Lastly, we will document the ease and speed of performing high quality and accurate fluorescence-based resequencing on long-tracks of mitochondrial and nuclear DNA as well as the application these new tools to automatically find and view DNA variations within these sequences.

Fast and More Accurate Distance-Based Phylogenetic Construction

William J. Bruno, Aaron L. Halpern, Nicholas D. Socci
Los Alamos National Laboratory
billb@lanl.gov

Phylogenetic reconstruction is relevant to genomic analysis whenever profile methods are used to assess gene function, because accurate profile construction depends on historical relationships [Bruno, 1996]. The Neighbor-Joining algorithm of Saitou and Nei [1987] has the advantage of being fast enough for constructing a profile from hundreds of sequences, but it does not do the best job of constructing the correct tree, and in fact we show it to be biased.

Gascuel recently [1997] introduced the BIONJ method, which improves the Neighbor-Joining method by using a weighted average to compute the new distances. BIONJ is still biased, and performs exactly the same as Neighbor-Joining in the case of 4 taxa.

We introduce a new, weighted neighbor-joining method called Weighbor. This method uses weights that accurately reflect the exponential dependence of variances and covariances on distance. The weights are used both in determining which pair is joined and in computing new distances. As a result, the method is not biased, and gives better results than Neighbor-Joining, even in the case of 4 taxa. The current implementation of Weighbor has a computational complexity of N^4, but an N^3 version, comparable in speed to Neighbor-Joining, is underway.

Bruno, W. J., "Modeling Residue Usage in Aligned Protein Sequences via Maximum Likelihood," Mol. Biol. Evol. 13:1368-75 (1996).

Gascuel, O., "BIONJ: an Improved Version of the NJ Algorithm Based on a Simple Model of Sequence Data," Mol. Biol. Evol. 14:685-95 (1997).

Saitou, N. and M. Nei, "The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees," Mol. Biol. Evol. 4:406-25 (1987).

Determining the Important Physical-Chemical Parameters Within Various Local Environments of Proteins

Jeffrey M. Koshi and William J. Bruno
Los Alamos National Laboratory, Theoretical Biology and Biophysics
jkoshi@lanl.gov

In this work the importance of various physical-chemical parameters of the amino acids are analyzed in a position specific manner for various protein secondary structure and surface accessibility classes. This is done on the basis of a previously published maximum likelihood method that finds the optimal site-specific residue frequencies for a set of aligned homologs. These vectors representing the site-specific residue frequencies can be broken up into subsets based on position within various secondary structures and surface accessibility. An analysis of the eigenvectors (PCA analysis) of the resulting covariation matrix shows the most important physical-chemical parameters for a given subset of data. Preliminary results indicate the importance of hydrophobicity, agreeing with the conclusions of many other researchers, and we are investigating the importance of size, charge, and aromatic rings at specific positions within various secondary structures.

Work is also in progress to expand the model of evolution used in the above work. Rather than looking only at amino acid frequencies, we are implementing a method that uses a limited, amino acid mutation matrix optimized for each position in a set of aligned homologs. This more realistic model of evolution is still computationally feasible and should improve the performance of the model in phylogenetic tree reconstruction, homolog detection, analysis of the importance of physical-chemical parameters, or any other application.

Improving Software Usability and Accessibility

Ryan Carroll and Ruth Ann Manning
ApoCom Inc., 1020 Commerce Park Drive, Suite F, Oak Ridge TN 37830-8026
carroll@apocom.com

Over recent years the number of software tools being produced to assist in research related to the Human Genome Project has increased rapidly. However, most of these tools are currently being under-utilized because either they are not available on a wide variety of computer platforms or they have an interface that cannot be learned without a significant time investment. An additional shortcoming that is inherent in most of the available sequence analysis tools is the use of opaque pattern analysis systems such as statistical, syntactic, Markovian, or artificial neural network methods. Although such systems can be very accurate, the reasoning methods they use cannot be described intuitively. This means that the user has very little information (typically only a single numerical score) to assist in appraising the value of system predictions. ApoCom is addressing these problems by implementing comprehensive Java and CORBA interfaces through which it will make accessible a large assortment of computational and database related tools. ApoCom is also developing a fuzzy logic-based gene hunting algorithm.

bioWidgets: Visualization Componentry for Genomics

Steve Fischer, Jonathan Crabtree, Mark Gibson and **G. Christian Overton** (PI)
Department of Genetics, University of Pennsylvania; Philadelphia, PA 19104
{sfischer,crabtree,gibson,coverton}@cbil.humgen.upenn.edu

bioWidgets is a software package for the rapid development and deployment of graphical user interfaces (GUIs) designed for the scientific visualization of molecular, cellular and genomics information. The overarching philosophy behind bioWidgets is componentry: that is, the creation of adaptable, reusable software, deployed in modules that are easily incorporated in a variety of applications, and in such a way as to promote interaction between those applications. This is in sharp distinction to the common practice of developing dedicated applications. The bioWidgets project additionally focuses on the development of specific applications based on bioWidget componentry, including chromosomes, maps, and nucleic acid and peptide sequences.

The current set of bioWidgets has been implemented in Java with the goal in mind of delivering local applications and distributed applets via Intranet/Internet environments as required. The immediate focus is on developing interfaces for information stored in distributed heterogeneous, databases such as GDB, GSDB, Entry, and ACeDB. The issues we are addressing are database access, reflecting database schemas in bioWidgets, and performance. We are also directing our efforts into creating a consortium of bioWidget developers and end-users. This organization will create standards for and encourage the development of bioWidget components. Primary participants in the consortium include Gerry Rubin (UC Berkeley), Nat Goodman (Jackson Labs), Stan Letovsky (GDB) and Tom Flores(EBI).

Current progress includes the development of an Inter-widget Communication package. This package consists of a set of Java(tm) classes and interfaces, and is based on the Java(tm) JDK 1.1 Delegation Event Model. It defines a set of inter-widget events that control: (1) mutual selection of elements in multiple widgets and (2) coordinated scrolling and zooming of multiple widgets. We have used this package in our implementation which integrates our genome viewer and our sequence viewer.

We have also developed an object-oriented data specification for the Sequence and Map widgets. This specification will more generally apply to widgets that display sequence and sequence annotation. With the increasing use of object-oriented databases, object brokers and standardized remote object formats (CORBA and JAVA's RMI), application developers using our bioWidgets will have the capability to serve data objects directly to the widgets. Our object-oriented specification is written using Java(tm) interfaces, and will likely migrate to CORBA IDL. The interfaces we define are as generic as possible, with the goal of allowing diverse applications to use them.

The Inter-widget communication model, component-based design and object oriented data format are intended in the near future to be used with the Java Beans(tm) technology. This will allow users of bioWidgets to incorporate the widgets into an application simply by interacting with an application builder tool rather than writing actual integration code.

The current implementation includes widgets which display Sequence, Map, Blast results, Chromosomes and Sequence alignments.

For more details, please see http://agave.humgen.upenn.edu/bioWidgetsJava/.
*Supported by a grant from the Department of Energy, 92ER61371.

Research on Data and Workflow Management for Laboratory Informatics

Nathan Goodman
The Jackson Laboratory
nat@jax.org

We have been pursuing a strategy for laboratory informatics based on three main ideas: (1) component-based systems;(2) workflow management; and (3) domain-specific data We have been pursuing a strategy for laboratory informatics based on three main ideas: (1) component-based systems; (2) workflow management; and (3) domain-specific data management. The workflow and data management software we have developed pursuant to this strategy are called LabFlow and LabBase respectively. LabFlow provides an object-oriented framework for describing workflows, an engine for executing these, and a variety of tools for monitoring and controlling the executions. LabBase is implemented as middleware running on top of commercial relational database management systems (presently Sybase and ORACLE). It provides a data definition language for succinctly defining laboratory databases, and operations for conveniently storing and retrieving data in such databases.

Both LabFlow and LabBase are implemented in Perl5 and are designed to be used conveniently by Perl programs. The total quantity of code is modest, comprising about 10,000 lines of Perl5. The software is freely available and redistributable (see http://goodman.jax.org for details), but be forewarned that this is research software and is incomplete in many ways.

Method of Differentiation Between Similar Protein Folds

I. Dubchak, I. Muchnik, S. Spengler, M. Zorn
Lawrence Berkeley National Laboratory, Berkeley, CA 94720
ildubchak@lbl.gov

Predicting the protein fold and implied function for a target sequence whose structure is unknown is the problem of significant interest. The information derived from such a prediction is substantial, guaranteed by the similarity between the three-dimensional (3D) structures and the functions of class members. Prediction is especially complex when a distinction of one particular fold from other highly similar three-dimensional folds is needed.

For the development of the prediction technique we used a particular example of a separation of 4a-helical cytokines (fold of interest, FI) from similar folds. We applied the method based on global descriptors of a protein in terms of the biochemical and structural properties of the constituent amino acids [1]. Neural networks were used to combine these descriptors in a specific way to discriminate members of the FI. The following steps were necessary:

1. Defining the neighborhood of the FI, i.e. spatially close protein folds which are hard to distinguish from the FI.
2. Collecting the non-redundant set of proteins to represent a wide range of available members of protein folds in the context of the comprehensive Structural Classification of Proteins (SCOP).
3. Selecting several attributes which represent various groups of physico-chemical and structural properties.
4. Analyzing the prediction accuracy achieved by each parameter set in order to chose the most efficient sets for separation of the FI from a particular neighbor.
5. Voting among predictions made by different sets of parameters to achieve higher reliability of prediction.

The developed procedure is simple and efficient. Further improvement of the prediction method greatly depends on the possibility of automation of the steps 1-3 and growth of existing databases.

1. I. Dubchak, I. Muchnik, S. R. Holbrook (1995). PNAS 92, 8700-8704.
2. Murzin, A. G., S. E. Brenner, T. Hubbard and C. Chothia. (1995). J. Molec. Biol. , 247: 536-540.

Ethical,
Legal,
and
Social Issues

Geneletter: An Internet Newsletter on Ethical, Legal, and Social Issues in Genetics

Dorothy C. Wertz, Philip P. Reilly, Robin J.R. Blatt
The Eunice Kennedy Shriver Center for Mental Retardation, Inc.
dwertz@shriver.org

Geneletter has reached a wide audience, with 348,412 hits *Geneletter* has reached a wide audience, with 348,412 hits and 54,083 user sessions between September 18, 1996 and October 6, 1997. Current readership is about 9000 user sessions per month (323 per day), with an average length of 7 minutes. Readers range from elementary school students to graduate and professional school students, and include people with genetic conditions, lawyers, and medical professionals. 17% are international, including Canada, Australia, United Kingdom, Sweden, Japan, france, Germany, Singapore, Malaysia, New Zealand, Israel, Norway, Brazil, and Italy, as the leading nations. Our content, in the seven issues published to date, includes ethics, science, medicine, law, international views, education, society, book reviews, and a digest of the news about genetics. Reader interest has focused on cloning, the genetics of homosexuality, "The Calico Cat, the Munchkin, and Achondroplasia," "Adam, Eve, and Mitochondria," and basic information about genetic testing. Reader queries have focused on popular science issues (cloning, Jurassic Park), accuracy of paternity testing, causes of miscarriage, insurance coverage, and material for student papers, rather than on ethical concerns. This project demonstrates the feasibility of using the Internet to educate people about genetics.

Communicating Science in Plain Language: The Science + Literacy for Health: Human Genome Project

Maria Sosa, Judy Kass, and Tracy Gath
American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005

Recent literacy surveys have found that a large number of adults lack the skills to bring meaning to much of what is written about science. This, in effect, denies these adults access to vital information about their health and well-being. To address this need, the American Association for the Advancement of Science (AAAS) has developed a 2-year project to provide low-literate adults with the background knowledge necessary to address the social, ethical, and legal implications of the Human Genome Project.

With its Science + Literacy for Health: Human Genome Project, AAAS used its existing network of adult education providers and volunteer science and health professionals to pursue the following overall objectives: (1) to develop new materials for adult literacy classes, including a high-interest reading book, a short video providing background information on genetics, a database of resources, and fact sheets to assist other organizations and researchers in preparing easy-to-read materials about the HGP, and (2) to develop and conduct a campaign to disseminate the materials to libraries and community organizations carrying out literacy programs throughout the United States. To introduce the materials to low-literate adults, workshops using the materials were conducted in Washington, DC, Baltimore, MD, Chicago, IL, Miami, FL, and Cleveland, OH. In 1997, thousands each of the book and video, both titled *Your Genes, Your Choices*, have been sent to literacy educators, community colleges, church groups, libraries, and other organizations around the country. In each workshop, a genetic counselor or other genetic professional was present to answer questions and provide insight into genetic research and issues. The entire book is also available on the Web at http://ehr.aaas.org/ehr/books/index.html. Our model for helping scientists communicate in simple language will has impact beyond classrooms and learning centers. Since not every low-literate adult is enrolled in a literacy class, we developed a model that reaches out to community groups providing health services. These groups have indicated that easy-to-read materials on genetics are not only desirable but necessary; indeed, the groups we worked with often received requests for information on heredity and genetics. *Your Genes, Your Choices* enables medical and scientific organizations to communicate more effectively with economically disadvantaged populations, which often include a large number of low-literate individuals.

* 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 contract DE-FG02-95ER61988.

The Arc's Human Genome Education Project

Sharon Davis, Ph.D.; Leigh Ann Reynolds, Project Associate
The Arc of the U.S. 500 E. Border Street, Suite 300 Arlington, TX 76010
lreynold@metronet.com

New genetic findings from the Human Genome Project (HGP) pose unique ethical questions and legal/social concerns to those with disabilities and their family members. Many disorders associated with the disability of mental retardation have genetic causes, with Down syndrome and fragile X syndrome being the most common.

In an effort to begin addressing these complex issues, The Arc of the U.S. (the largest voluntary organization on mental retardation in the country with 140,000 members) developed a series of reports, fact sheets and a training package for use by the organization's leadership to educate members about these important and timely issues. The materials were distributed to all 1,100 chapters of The Arc and made available through the internet. Some topics covered include:

- An Introduction to Genetics and Mental Retardation
- Genetic Discrimination: Why Should We Care?
- Genetic Testing, Screening and Counseling – An Overview
- Protecting Genetic Privacy

The aim of this educational effort is, first, to provide a basic understanding of genetic inheritance and mental retardation. Next, unique concerns people with mental retardation and their families may face in light of new genetic research are addressed. Finally, inherited syndromes associated with mental retardation are highlighted to provide a practical example of how someone with this disability is affected by the HGP and to report on the latest research in genetic therapy.

The Arc's Human Genome Education Project: Examining Genetic Ethical, Legal and Social Issues is an interactive and comprehensive training package which includes a detailed script for the workshop leader, a pre and post-questionnaire to measure change in opinion and understanding of the issues, background information for the presenter, print copies of overheads, a project description to use in promoting the workshop, a 15 minute video and other handouts. The most important objective of the training (discussing highly complicated and sometimes controversial issues) is achieved through the use of case scenarios. Through this challenging exercise, members actively learn how difficult it can be making decisions regarding genetic testing, therapy and discrimination on behalf of themselves or their children who have disorders either caused by or associated with mental retardation. By simplifying medical terminology and concepts, identifying core issues most threatening to those with disabilities and utilizing practical case scenarios, members of The Arc are better equipped to make informed decisions and educated opinions on ethical, legal and social issues impacting the lives of people with mental retardation resulting from the HGP.

Measuring the Effects of a Unique Law Limiting Employee Medical Records to Job-Related Matters

Mark A. Rothstein, J.D., Steven G. Craig, Ph.D., Betsy D. Gelb, Ph.D.
University of Houston
mrothstein@uh.edu

Under the Americans with Disabilities Act, after a conditional offer of employment an employer is permitted to conduct medical examinations of unlimited scope and may require the release of all medical records in the possession of the individual's health care providers. Allowing unlimited access to personal medical records not only facilitates surreptitious discrimination, but it invades the privacy of individuals and discourages at-risk individuals from undergoing genetic testing in the clinical setting.

Pursuant to a law enacted in 1983, Minnesota is the only state to restrict the scope of all medical inquiries by employers to matters that are strictly job-related and consistent with business necessity. If this law has had no adverse effects on employers or other parties, then it may serve as a model for protecting genetic privacy in the workplace.

We are using the following three main methods to measure the effects of the law: (1) reviewing and analyzing all of the cases filed under this law in the last five years; (2) surveying human resource managers in Minnesota (using Ohio as a control state) to learn their opinions about limitations on the scope of preplacement medical examinations; and (3) doing an economic analysis in Minnesota (with Ohio as the control state) of workers' compensation claims rates, employee turnover rates, productivity rates, health insurance claims rates, and other data.

The data collection phase is well underway and should be completed by the end of 1997. Analysis of the data will help determine whether this legislative approach is a viable alternative to current proposals for prohibiting genetic discrimination in employment.

The Genetics Adjudication Resource Project

Franklin M. Zweig, Ph.D., J.D.

The Genetics Adjudication Resource Project (GARP) has been funded by the Ethical, Social and Legal Issues (ELSI) Program of the DOE Human Genome Project to provide foundation education for 1,000 judges of federal and state courts in genetics, molecular biology and biotechnology. The Project's over-arching objective is to familiarize judges, who have little scientific background, with the concepts, research findings, and vocabularly they are likely to encounter in civil lawsuits and criminal prosecutions. Expected evidence flows from parties' pleadings and from expert witness testimony. ELSI matters flow from the conferences' active, case-based, problem-solving curriculum.

The project's operational objective is to provide judges with tools by means of which to exercise their gatekeeping duties for scientific evidence. Those duties include, but are not limited to, determination of the fitness of the evidence for presentation to juries.

The Conference series was preceded by four preparatory "working conversations" ("WC's") intended to perfect a durable, effective, and adaptable educational technology template. One of the "WC's" was hosted by the Lawrence Livermore National Laboratory in November, 1996. The first large scale conference of 106 judges and 40 faculty was held in May, 1997 at Airlie House, Airlie, Virginia for judges of courts located in the Greater Washington, D. C. region.

The Project has been approved for conduct of six conferences in 1997 and 1998. Included are the National Capital Area, New England, Chicago, and Mountain West regions, and two conferences scheduled for June 27-July 3 and August 1-6 at Orleans, Cape Cod. Three additional conference are on the drawing boards for the Mid-Atlantic region in 1998 and in 1999 for the Southeast regions, and for the California State Courts.

The GARP has also produced a Handbook for Judges for Cases Involving Genetics, Molecular Biology and Biotechnology Evidence. Under production is a nine cassette videofilm primer. Another primer has been produced by EINSHAC within the courts' own publication system. The Judges' Journal, a quarterly published by the American Bar Association, produced a dedicated theme issue late Summer, 1997 entitled "Genetics in the Courtroom." This primer is divided into scientific and adjudication perspectives.

Key to the GARP's conferencing success is the recruitment, training and deployment of science advisors. Scientists reside with the judges and the informal interaction is a key supportive device in making this unfamiliar subject material patent to lay judges. DOE scientists and ELSI personnel have been instrumental in this effort. GARP deploys one scientist for every four judicial participants. GARP has developed a training program for scientific faculty so that differences in communication content and style rooted in the judicial culture can be anticipated and managed. The training strategy is based upon the different professional paradigms (mental worlds unique to adjudication and to science) highlighted by the literature and experienced in the 1995 and 1996 WC's.

Science advisors are now being sought for the 1998 and 1999 judicial conferences. Six hour training seminars will be scheduled conveniently. GARP promises an interesting experience with an avid group of learners whose day-to-day responsibilities can make important contributions to our society's adaptation to advances in human genetics.

(DE-FG02-96ER62170, conducted by EINSHAC, the Einstein Institute for Science, Health and the Courts, Washington, D.C.)

Upstream Patents and Downstream Products: A Tragedy of the Anticommons?

Michael A. Heller & Rebecca S. Eisenberg[1]
University of Michigan Law School

The past twenty years have witnessed significant privatization of pre-market or "upstream" stages of biomedical research. In an earlier era, such research was typically performed in the public sector and made freely available in the public domain to "downstream" users. Today, public and private institutions are increasingly likely to patent biomedical research discoveries that are several steps removed from end product development. Patent rights have been credited with a conspicuous increase in private funding for biotechnology research. Yet paradoxically, a proliferation of patent rights in upstream discoveries could create barriers to the development of new pharmaceutical products further downstream in the R&D process.

In this article we propose a theory of *anticommons property* to explain how too many upstream intellectual property rights may lead to too few downstream products. Anticommons property may be understood as the mirror image of commons property. In the familiar *tragedy of the commons*, too many owners have the privilege to use scarce resources, and the property is prone to overuse. By contrast, in a *tragedy of the anticommons*, too many owners have the right to exclude others from using a scarce resource, and the resource is prone to underuse. Anticommons property may arise whenever governments define new property rights that are too fragmented. Empty storefronts in Moscow provide one stark example of this phenomenon. Transition regimes charged with privatization have endowed multiple owners with overlapping rights in each storefront, so no owner holds a useable bundle of rights to convey to an entrepreneur who wishes to set up shop. As a result, scarce property is wasted. Once an anticommons emerges, collecting rights into useable bundles can be brutal and slow.

Like post-socialist transition regimes, intellectual property systems are constantly creating new property rights. An anticommons could arise if multiple owners obtain fragmented patent rights that are difficult to assemble into usable bundles. The anticommons model provides one way of understanding a widespread intuition that issuing patents on gene fragments makes little sense. Depending on their scope, patent rights in gene fragments could lead to the emergence of a genomic anticommons in which a product developer requiring use of a full-length gene or a set of polymorphic markers useful in diagnosing disease might be stymied by the costs of bundling licenses from many patent owners. Up to a point, privatization may enhance efficiency by spurring investment in upstream research. But privatization can go astray. When the legal system creates too many fragmented intellectual property rights, a tragedy of the anticommons could block the development of new pharmaceutical products.

[1] Supported by DOE Grant No. DE-FG02-94ER61792.

The Science and Issues of Human DNA Polymorphisms

David Micklos, Mark Bloom, Scott Bronson, and John Kruper
DNA Learning Center, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724
micklos@cshl.org

This ELSI training 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. By targeting motivated biology faculty who currently perform student laboratories with viral and bacterial DNA, this program offers a cost-effective means to bring high school biology education up-to-the-minute with genomic biology. In October-November 1997, we are instructing the first of 12 workshops nationwide at Mt. Sinai School of Medicine (New York), Boston University School of Medicine (Massachusetts), and Canada College (California).

The program is based on lab and computer technology, developed at the DNA Learning Center and the University of Chicago, that makes human DNA "fingerprinting" by polymerase chain reaction (PCR) accessible and affordable for high school use. Program participants learn simplified lab techniques for amplifying two types of chromosomal polymorphisms – an Alu insertion (TPA-25) and a VNTR (D1S80). 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.

Workshop participants amplify their own polymorphisms from DNA obtained from rapid preparations of buccal cells and hair sheaths. The loci are amplified using rapid cycling profiles and analyzed on agarose gels. To further reduce the cost of experiments, we have developed the Biogenerator, the first inexpensive ($900) thermal cycler licensed for precollege use. This "Rube Goldberg" thermal cycler articulates with a Macintosh computer and gives results comparable to commercial machines.

Using a facility at our WWW site (darwin.cshl.org), the Alu insertion data are further used as an entree into human population genetics and genome diversity. The "Student Allele Database" has forms for entering student-generated data, as well as archival data from populations around the world. Several statistical functions are available: testing Hardy-Weinberg equilibrium within a single population, measuring genetic distance between two populations, and comparing two populations using contingency chi-square. A "Monte Carlo" generator shows the effect of genetic drift in small populations.

During the summer, we developed reliable methods for generating mitochondrial DNA sequence from buccal and hair samples. We hope to introduce this technology at as many DOE workshop sites as possible. Workshop participants can then use their own mitochondrial DNA sequence as an entree to modern bioinformatics. Our WWW site has a step-by-step template for analyzing mitochndrial DNA sequence – including similarity searches, multiple sequence alignments, a recreation of the Neandertal DNA analysis, and the identification of the Romanov family remains. Ultimately, we envision students preamplifying their mitochondrial DNAs and performing dye terminator reactions at their own schools. The ready-to-sequence DNAs would then be sent to regional centers for sequencing and the results would be posted by Internet.

Thus, we are striving to develop a robust and accurate analog of human genome research that allows students to use their own chromosomal and mitochondrial DNA polymorphisms as the basis of explorations into contemporary genomic biology.

Implications of the Geneticization of Health Care for Primary Care Practitioners*

Mary B. Mahowald, John Lantos, Mira Lessick, Robert Moss, Lainie Friedman Ross, Greg Sachs, Marion Verp
Department of Obstetrics and Gynecology and MacLean Center for Clinical Medical Ethics, University of Chicago, Chicago, IL 60637
mm46@midway.uchicago.edu

Phase I (fall 1995): Generic topics in genetics in primary care presented to broad audience
Ten sessions: Goals, methods, & achievements of HGP; Typology of genetic conditions; Scientific, clinical, ethical, and legal aspects of gene therapy; Concepts of disease, Genetic Disabilities; Gender and socio-economic differences; Cultural and ethnic differences; Directive or nondirective genetic counseling.
Transcripts of presentations prepared for revision by authors

Phase II (Jan.-Mar. 1996): Grand rounds in specific areas of primary care:
Topic: What every general practitioner should know about the new genetics

- Pediatrics — Stephen Friend
- Obstetrics/gynecology — Joe Leigh Simpson
- Medicine — Tom Caskey
- Family medicine — Loralane Lindor
- Nursing — Colleen Scanlon

Transcripts of presentations prepared for revision by authors. Bibliography developed on generic issues in genetics, and issues specific to areas of primary care

Phase III (Apr., 1996)
Policy issues presented by Sherman Elias and George Annas
Syllabi and chapters developed by each primary care team + policy team

Phase IV (Oct.-Dec. 1996)
Series on genetics in primary care for Clinical Ethics fellows and Robert Wood Johnson Clinical Scholars, presented by each team of primary caregivers (Co-PI + fellow)
Teaching sessions for faculty, house staff, students led by Co-PIs + fellows

Phase V (April 1997)
National Conference on The New Genetics in Primary Care, keynoted by Victor McKusick, with CME/CNE workshops for primary caregivers
Outreach (Summer 1997 and beyond)
CME conference, AMA planning for conference on genetics in primary care,
Teaching sessions on genetics for new clinical ethics fellows,
Preparation of materials for publication

* Supported by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research, U.S. Department of Energy, DE-FG02-95ER61990

Confidentiality Concerns Raised by DNA-Based Tests in the Market-Driven Managed Care Setting

J.S. Kotval, D. Dewar, and S. Brynildsen
School of Public Health, University at Albany, One University Place, Rensselaer, NY 12144-3456
E Mail: jsk03@health.state.ny.us

Previous policies to protect the confidentiality of DNA-based tests have centered on protecting the dispersion of genetic information. In general, the goal has been to keep information from passing outside a health care institution to third parties that might discriminate against patients in employment, access to health care and other areas of civic life. We are focusing our attention on the threats to patient welfare created within the setting of the market-driven managed care organization (MCO). This setting presents unique ethical dilemmas, since physicians (and, often, personnel at testing laboratories) are employees of the MCO, and since the payor and provider functions are contained within the same entity. In this context, the institutional imperative of cost savings in order to capture market share could lead to discrimination in health care access (either through outright denial of enrollment or prohibitive premiums) if DNA-based tests reveal the likelihood of future high-cost illnesses. Our group -- which spans the disciplines of genetics, medical ethics, health economics and the law -- seeks to (i) formulate an ethical construct for medical confidentiality with a view to defining its core ethical functions and its limitations in the changing health care system; (ii) examine the practices and institutional imperatives of market-driven MCOs to understand the context in which the DNA-based tests would be used and to assess the cost of confidentiality measures; (iii) identify gaps in the policy framework that could allow misuse of confidential medical information within the MCO; and (iv) make recommendations to remediate these gaps in policy in a manner that is applicable and practical to MCOs. Progress to date includes a preliminary formulation of confidentiality concerns raised in the market-driven managed care setting and the design of a survey instrument to assess the institutional and marketing practices of MCOs. We are paying specific attention to the current practices of MCOs that encroach on the core ethical concerns of medical confidentiality.

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

Sara L. Tobin, Ph.D., M.S.W.* and Ann Boughton#
*Stanford University and #Thumbnail Graphics
tobinsl@leland.stanford.edu and
annie@telepath.com

The knowledge gained from the Human Genome Project has the potential to correlate molecular diagnosis with effective treatments and to lead the way to novel medical interventions. The molecular tools that have emerged from genetic studies are changing the face of medical practice and inaugurating a transitional period that will be uncomfortable for both physicians and the public. There will be marketing pressures, health care industry changes, uneven supporting resources, variable training of physicians, and limited public understanding. We have designed an interactive, multimedia CD-ROM course to ease this transition for the majority of physicians, who have received little or no training in clinical applications of molecular genetics because the field has developed so recently. The courseware, entitled "The New Genetics: Courseware for Physicians. Molecular Concepts, Applications, and Implications," will provide accredited continuing education for practicing physicians through the Stanford University Office of Postgraduate Medical Education.

It is important for physicians to understand the modern clinical applications of molecular genetics for several reasons. First, physicians will be explaining genetic tests and their implications to their patients, selecting specific tests, and interpreting the results. Second, physicians must be able to work productively with other health professionals, including genetic counselors and psychologists. Third, standards of practice that will govern the application of molecular genetic diagnosis to patient care are currently under development, and physicians need to contribute to this evolution. Fourth, a lack of training in modern genetics prevents many physicians from understanding much current medical research. Finally, physicians with training in molecular medical genetics can serve as informed resource persons and enhance the level of public understanding of and appreciation for the Human Genome Project in their communities.

The CD format confers multiple advantages for continuing medical education. The delivery of course materials via CD-ROM frees the physician from a presentation schedule and does not require travel and time away from a busy practice. CD's function as an improved teaching resource because of their interactivity and multimedia capability. We have designed a streamlined, accessible navigational system tailored to the needs of the busy (and possibly computer-naive) physician. Engaging interactive features and animations have been created to convey complex concepts. The course content is supervised by a Board of Advisors. While the Internet is currently too slow to serve as the primary delivery system, we are programming the CD to accept updates and supplements from the Internet or from a subscription floppy.

The development of a prototype version of the courseware is funded by the Department of

Energy, and our current draft version of the courseware will be demonstrated at the Workshop.

Electronic Scholarly Publishing: Foundations of Classical Genetics

Robert J. Robbins
Fred Hutchinson Cancer Research Center, 1124 Columbia St., LV-101, Seattle, WA 98104
rrobbins@fhcrc.org
http://www.esp.org

With the continuing success of the Human Genome Project (HGP), more and more people are interested in the project and wish to understand 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 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.

Interest, however, is not enough. Real understanding of the results of HGP research requires some familiarity with basic genetics notions. In particular, most of the methods and findings of molecular genetics are essentially inaccessible to those without appropriate training. On the other hand, both the methods and results of early work in classical genetics can be appreciated by virtually anyone: cross a black mouse with a white mouse, count the progeny of different colors, then try to figure out what might be going on.

A familiarity with the basic notions of classical genetics is essential for a simple appreciation of the significance of the HGP, and a more detailed knowledge of classical genetics (including an appreciation of the question, What is the chemical nature of the gene?), can provide a basis for the genuine understanding of HGP findings. Gaining this basic understanding of classical genetics is becoming more difficult, even for science majors. The runaway success of modern molecular genetics is driving the detailed presentation of classical genetics out of most text books, and the original literature is increasingly difficult to obtain.

To address these problems, we have established an electronic educational resource at which classic literature (both papers and monographs), that established the foundations of modern genetics, is being republished on-line, freely accessible to all. Works are being made available in a variety of formats including simple HTML, Adobe PDF files containing high-quality typeset republications, and Adobe PDF files containing image facsimiles of the original publication. The fundamental work of Gregor Mendel, for example, is available as both a typeset republication and as an image facsimile of the original 1865 publication.

Funding for the project was received earlier this year, and work thus far has involved (1) moving the original prototype site from Johns Hopkins to Seattle, (2) redesigning the site to make it easier to understand and use, and especially to prepare it to handle more data and more traffic, (3) establishing high-efficiency systems for converting paper publications into electronic form, and (4) acquiring material for republication. When full-scale publication begins in January, we anticipate publishing the equivalent of a 25-page paper every day.

The Community College Initative

Sylvia J. Spengler* and Laurel Egenberger**
*Life Sciences Division, **Center for Science and Engineering Education. E. O. Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720
sjspengler@lbl.gov

The Community College Initiative prepares community college students for work in biotechnology. A combined effort of Lawrence Berkeley National Laboratory (LBNL) and the California Community Colleges, we aim to develop mechanisms to encourage students to pursue science studies, to participate in forefront laboratory research, and to gain work experience. The initiative is structured to upgrade the skills of students and their instructors through four components.

Summer Student Workshops: Four weeks summer residential programs for students who have completed the first year of the biotechnology academic program. Ethical, legal and social concerns are integrated into the laboratory exercises and students learn to identify commonly shared values of the scientific community as well as increase their understanding of issues of personal and public concern. In the two year period of the grant, we have involved over twenty students. Students in the second summer have had laboratory interships.

Teacher Workshop Training: Seminars for biotechnology instructors to improve, upgrade, and update their understanding of current technology and laboratory practices, with emphasis on curriculum development in current topics in ethical, legal, and social issues in science. These workshops have involved the students as well.

Sabbatical Fellowships: For community college instructors to provide investigative and field experience in research laboratories. During the fellowship, teachers also assist in development of student summer research activities.

This work was supported by the Director, Office of Energy Research, Office of Biological and Environmental Research, Human Genome Program, ELSI program, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

The Hispanic Educational Genome Project

Margaret C. Jefferson, Mary Ann Sesma, and Patricia Ordonez
Department of Biology & Microbiology, California State University, Los Angeles CA 90032
mjeffer@flytrap.calstatela.edu

The primary objectives of this grant are to develop, implement, and distribute culturally competent, linguistically appropriate, and relevant curricula that leads to Hispanic student and family interactions regarding the science, ethical, legal, and social issues of the Human Genome Project. By opening up channels of familial dialogue between parents and their high school students, entire families can be exposed to genetic health and educational information and opportunities. In addition, greater interaction is anticipated between students and teachers, and parents and teachers.

Each participating high school has taken different approaches to exposing students and parents to the science and ELSI of HGP. Some schools have divided students into research teams from various levels of biology curricula with each team analyzing small segments of their own DNA. Other schools have science classes following the BSCS HGP-ELSI curricula or components of the University of Washington High School Human Genome Program. Still other schools have utilized the materials that were developed by various Los Angeles Unified School District science teachers. Several other classes (e.g., Spanish classes, English classes, Journalism classes, and Social Science classes) at each participating school have developed ELSI newsletters for distribution to the parents. In addition, we have math and computer science classes from one of the schools helping in the construction of our web page. Each participating school is also expected to have parent focus groups which usually meet once per month in the evening to discuss various genetic health issues and implications of HGP.

Genes, Environment, and Human Behavior: Materials for High School Biology

Joseph D. McInerney and Michael J. Dougherty
5415 Mark Dabling Blvd., Colorado Springs, CO 80918-3842
jmcinerney@BSCS.org

The Biological Sciences Curriculum Study (BSCS) has begun the development of an instructional module on genetics and human behavior for use in introductory high school biology. 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 project employs the process of curriculum development that BSCS has refined continually since the inception of the organization in 1958. In addition, development of the module is drawing upon the experience BSCS acquired during the development, distribution, and implementation of three genome-related instructional modules between 1991 and 1996. This experience includes 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.

In late July 1997, BSCS completed the first of two writing conferences during which experts in behavioral and medical genetics, ethics, and high school biology teaching produced a draft module containing instructional activities for students and extensive background materials for teachers. The draft instructional activities are designed to help students move through the following ideas:

1. variation in behavior exists in populations;
2. there are 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.

To test the effectiveness of the draft module in helping students understand how genes and environment influence human behavior, BSCS pilot tested several activities in October. Following analysis of these data, BSCS will refine the materials, conduct a complete field test, convene a second writing conference, produce the final module, and distribute the module free of charge to all interested high school biology teachers.

Microbial Literacy Collaborative

Cynthia Needham
Lahey Hitchcock Medical Center, Burlington, MA
cynthia.a.needham@hitchcock.org

The Microbial Literacy Collaborative is a partnership of organizations dedicated to enhancing public understanding of microorganisms and the roles they play in sustaining the planet. Partners include the American Society for Microbiology, Baker & Simon, Oregon Public Broadcasting, the Association of Science and Technology Centers, and the American Association for the Advancement of Science. The MLC's initiatives include

- *Intimate Strangers: Unseen Life on Earth*, a television series airing in 1999
- Formal college and precollege educational programs, designed specifically to support and enhance teaching in settings where educational resources are limited
- Informal learning programs specifically targeted to youths at risk developed and disseminated in conjunction with the Association of Science Technology Centers (ASTC) Youth Alive! project, and the American Association for the Advancement of Science (AAAS) Black Churches Project.
- An interactive Website for linking scientists with students, their parents and teachers and fostering wide access to resources developed through the preceding initiatives.

The scientific messages delivered through these projects will promote a balanced view of our interactions with our microbial partners on the planet. One of the MLC's primary goals is to dispel public anxiety about microorganisms which has been created through intensive media focus on microbes as disease agents. Viewers and participants will be introduced to broad concepts such as

- The important microbial role as bioremediators and how we can use their natural processes to help clean up our environment,
- The biotechnological potential for microbial products and processes and how they can better our daily lives
- The microbial basis of all ecosystems and how we can sustain our environment while still developing the economic potential of our natural resources
- The linkages between global climate changes and microbial processes and how increasing our understanding of these linkages can influence public policy
- New strategies for understanding and controlling infectious diseases and how we can take advantage of these new developments to improve the human condition throughout the world.

Project Status:

The three initiatives of the MLC received a guarantee for complete funding in May, 1997. The intellectual framework for the science documentary was established at a seminal meeting of the Science Advisory Group held in Woods Hole, MA, three years prior to funding. The production staff officially began work Oct. 6 and are presently participating in "Microbiology school" and beginning to formulate story lines for the four part series to reflect the scientific content.

Planning for the Annenberg sponsored telecourse began mid-summer, with a meeting of the Science Education Advisory group at OPB. The group is nearing completion of curricular content and learning objectives for an accredited college level telecourse, both of which will be finalized at an upcoming meeting in October.

The Advisory Group to the informal set of initiatives met in September for a 3 day planning meeting to discuss the hands on activities that will be developed to accompany each of their dissemination plans.

The official kickoff meeting for the MLC will take place at the end of the month at Mt. Hood, Oregon, where all parties will come together to validate timelines and recognize who the responsible parties are for various aspects of each of the three initiatives.

High School Students as Partners in Sequencing the Human Genome

Maureen Munn, Leroy Hood
Department of MBT, University of Washington,
Box 352145, Seattle, WA 98195
mmunn@u.washington.edu

The High School Human Genome Program (HSHGP) encourages high school students to think constructively about the scientific and ethical issues of genomic research by enabling them to participate in both. This program supports many of the teaching objectives presented in the National Science Education Standards (1996), including meeting the content standards for genetics education, teaching science through inquiry and developing a learning community of teachers and scientists to promote better science education. Participating students learn about many career options in science through discussions with scientist mentors who assist during classroom experiments, and the lab experiences help to prepare them for future employment.

A. Program modules. The DNA Synthesis experiment is an introductory experiment that helps students learn about DNA structure and replication and develop their laboratory skills. During the DNA Sequencing experiment, students sequence a region of chromosome 5 that is involved in a hereditary form of deafness. This project is made possible through a collaboration with Eric Lynch and Mary-Claire King from the Departments of Genetics and Medicine at the University of Washington. The Ethics unit, which focuses on presymptomatic testing for Huntington's disease, helps students develop the skills to define ethical issues, ask and research relevant questions about a particular topic and make justifiable ethical decisions. The module was developed by Sharon Durfy and Robert Hansen from the UW Department of Medical History and Ethics.

B. Teacher Preparation and Classroom Implementation. Local, regional and national teachers attend a week-long summer workshop, which provides training in program modules, informal seminars and discussions of relevant topics.

C. Equipment Kit Loans. During the academic year, local teachers are provided with the necessary equipment, supplies and technical assistance to carry out the classroom experiments. Teachers from distant sites receive DNA templates and primers and ongoing technical advice. This program is currently serving 32 high school teachers at fifteen schools in Washington state and 13 other teachers nationally.

D. The HSHGP web-site (http://hshgp.genome.washington.edu). This site is intended as a resource for teachers and students everywhere and contains the following:

Program modules. These are available on-line or in a downloadable version. On-line DNA assembly and data analysis. This tutorial enables students to carry out the assembly of the sequencing data from the web-site, using the demonstration version of the DNA assembly program, *Sequencher* and a folder of student data files.

Virtual DNA sequencing. This tutorial enables classrooms that are unable to do the sequencing experiment to do many aspects of the sequencing process by providing scans of student sequencing ladders and highlighting the portions of our teaching modules that can be used to simulate DNA sequencing.

Future tutorials on the Web-site:
On-line research projects. In the future, we will develop tutorials that help students access DNA and protein databases and related analysis software and understand what the analysis software is doing. Once students understand these tools, they will be able to design and explore their own plan research problems.
Exploring ethical issues related to genomic research. We plan to develop additional modules that emphasize how the decision making process can be used to examine any ethical issue.

Communication among program participants. Discussion boards will be set up so that participants can discuss technical problems and solutions, ask research questions and exchange classroom tips.

E. Program Evaluation: Preliminary results of program evaluation will be discussed.

Your World/Our World — Exploring the Human Genome Creating Resource Materials for Teachers

Jeff Alan Davidson
Pennsylvania Biotechnology Association, Alliance for Science Education
PA_Biotech@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 studentsto 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 now creating 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 assembling and reviewing for usefulness materials from other publishers and by continuing the development of new materials by PBA to create a comprehensive supplemental materials package that provides resources in several different media. Second, by running 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.

This project is targeted at middle and high school teachers and students for several reasons - most of the biological information studied and learned in the United States occurs at this level, students are generally very interested in biology and science at these levels, and teachers can greatly assist students in considering this material, but need more support in teaching about the HGP and ELSI.

Phase I Materials are expected to include:

- A Further Exploration of the Science and Ethics, Legal and Societal Issues of Genomics
- Reference Materials for Genomics
- Teaching Plans, Colorful Overheads & Assessment Materials
- Biological Animations (Computer Animations available on Video & CD that illustrate important ideas)
- Experiments, Activities and Demonstrations
- Science Fair or Research Project Ideas
- Theatrical Vignettes that explore important questions in ELSI of HGP

Phase II Materials are expected to include:

- Teaching Plans
- Graphics, Art, Photographs
- 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

Human Genome Teacher Networking Project

Debra L. Collins, M.S., and Rebecca Knetter, Ph.D.
Genetics Education Center, University of Kansas Medical Center, Kansas City, KS 66160-7318
dcollins@kumc.edu

Families, health care providers, and the general public are all increasingly aware of the human genome project discoveries. However, many do not have a background on basic genetic information, and therefore are not aware of, or prepared for, the ethical, legal, and social implications of this new technology and the applications. Our program helps prepare high school students for their future through updated information and resources from their biology teachers. Each teacher in this project spent time in educational activities over a two-year period, learning about updated information through two one-week workshops, preparing updated lesson plans, presenting peer teacher programs, and networking with other teachers, genetic professionals, and ELSI experts.

During 1993-1997, 177 teachers attended a series of Human Genome Teacher Networking education workshops which addressed the applications of Human Genome Project technology, with a focus on ethical, legal, and social implications (ELSI). After the teachers attended summer workshops, they used new materials with students, then conducted peer and community education programs, and contacted genetic and ELSI experts to enhance their classroom teaching. The networks which developed between teachers, liaisons with genetic professionals, peer teacher programs, and on-line computer communications helped educators and their students obtain current human genetics information.

We analyzed the improvement in teacher confidence and preparedness and measured student achievement as a result of the summer workshops. Following the workshops, teachers were prepared and confident teaching complex genome technology and applications ($p > .05$). They acquired new information to expand their knowledge of human genetics and integrate the complex information into existing science curricula. Some teachers developed new school courses. Student achievement was significantly increased (p>.05) due to teacher attendance at the workshops. The teachers gained a new awareness of the scientific as well as the personal aspects of the genome information.

Participants were 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 than of comparable students whose teachers did not attend the workshop.

Teachers, as part of their biology curricula, can integrate genome project concepts, and help students understand the ELSI issues which will be important in their future. Teachers are enthusiastic about education workshops, they increase the amount of curricula time devoted to genome/ ELSI projects, and can present information at an appropriate pace for students. The studentis science literacy is increased on timely topics, and they have knowledge of internet and other resources to answer new questions, not available in current published textbooks.

Workshop resource materials, lesson plans, the mentor network, and teachers who agreed to have their names listed on the internet are available though the web site for this project: http://www.kumc.edu/gec (Genetics Education Center). Links are provided to other ELSI sites, career information, and other genetic resources.

Funded by DOE #DE-FG02-92ER61392

A Question of Genes

Noel Schwerin
NoelEye Documentaries, 1327 Church Street, San Francisco, CA 94114
415.282.5620
schwerin@slip.net

A two-hour national PBS special (aired September 16, 1997), *A Question of Genes* looks for the first time at the ethical, social and legal implications of genetic testing. *A Question of Genes* enters the lives of a few individuals and families as they confront genetic risks for conditions like heart disease, Alzheimer's, breast cancer and genetic birth defects. *A Question of Genes* explores the profound challenges genetic information makes to a person's sense of self, family and future. Closely observing regular people over several years, *A Question of Genes* takes us inside the decisions and dilemmas of a range of personalities and perspectives: a woman who had a preventive mastectomy after losing three sisters to breast cancer tries to make sense of her just-discovered results; a poor African-American mother struggles to know more about the genetic legacy to her daughter; a physician who administers genetic tests wrestles with ethical dilemmas about his patients' privacy and rights; a pregnant woman makes startling decisions about the fate of her unborn children.

In seven stories told by the participants themselves, *A Question of Genes* captures both the profound emotional drama as well as the enormous social, legal and ethical implications of the powerful new technology of genetic testing. *A Question of Genes* was produced and directed by Noel Schwerin for the Chedd/Angier Production Company and Oregon Public Broadcasting.

The DNA Files: 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
strp@aol.com

The DNA Files is a series of nationally distributed public radio programs furthering public education on developments in genetic science. Program content is guided by a distinguished body of advisors and will include the voices of prominent genetic researchers, people affected by advances in the clinical application of genetic medicine, members of the biotech industry, and others from related fields. They will 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 will target educators, scientists, and involved professionals. Ancillary educational materials will be distributed in paper and digital form through over collaborative organizations and in fulfillment of listener requests.

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 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. A live, two-hour call-in show will cover the broad scope of human genetic research and its applications. It will include a discussion with experts, and a chance for listeners at home to make comments and ask questions.

Nine one-hour documentaries will provide the basic science of DNA, genes and heredity, while illustrating the accompanying social and ethical issues. "DNA and the Law," for example, reviews the scientific basis for genetic fingerprinting and looks at cases of alleged genetic discrimination by insurance companies, employers and others. "Gene Therapy: Medicine for Our Genes" takes on popular descriptions of genetic therapy, derived from stories like Jurassic Park, and attempts to help us realistically understand the science, as well as its promise, limits, and social implications. Other shows include "The Commercialization of Genetics," "Prenatal Genetic Testing: Better Babies Through Science," and "Genetic Evolution, Diversity and Kinship."

*Supported by ELSI grant DE-FG03-95ER62003 from the Office of Health and Environmental Research of the U.S. Department of Energy.

Human Genome Project Education & Outreach Component

Molly Multedo
Self Reliance Foundation, 121 Sandoval Street,
3rd Floor, Santa Fe, NM 87501 (505) 984-0080

Self Reliance Foundation (SRF), in collaboration with Hispanic Radio Network (HRN) and the National Center for Genome Resources (NCGR) has developed Spanish radio programming and outreach services which will help inform Hispanics on the ethical, legal, and social issues related to the Human Genome Project and motivate them to access the resources available for further education on these issues. Funding by DOE-ELSI supports the production of 50 new episodes of "Buscando la Belleza" (BB) for broadcast over the period of two years; extensive, national outreach and referral services and a linked Web site.

Infrastructure

Human Genome Management Information System

Betty K. Mansfield, Anne E. Adamson, Denise K. Casey, Sheryl A. Martin, **John S. Wassom**, Judy M. Wyrick, Laura N. Yust, Murray Browne, and Marissa D. Mills
Life Sciences Division; Oak Ridge National Laboratory; 1060 Commerce Park; Oak Ridge, TN 37830
bkq@ornl.gov

The Human Genome Management Information System (HGMIS), established in 1989 by the Human Genome Program Task Group of the DOE Office of Biological and Environmental Research, helps DOE fulfill its commitment to informing scientists, policymakers, and the public about the program's funded research. Through its communication role, HGMIS increases the use of resources generated in the Human Genome Project, reduces duplicative research efforts, and fosters collaborations and contributions to biology from other research disciplines.

HGMIS products, including the Web sites and newsletter, have won technical and electronic communication awards. In May 1997, DOE acknowledged the newsletter's value by presenting an exceptional service award to HGN's managing editor at a symposium celebrating 50 years of biological and environmental research.

Communicating scientific and societal issues to nonscientist audiences contributes to increased science literacy, thus laying a foundation for more informed decision making and public-policy development. For example, since 1995 HGMIS has been participating in a project to educate judges on the basics of genetics and gene testing to prepare for the coming flood of cases involving genetic evidence.

Information Resources

In keeping with its goals, HGMIS produces the following information resources in hard copy and on the Web:

Publications. A quarterly forum for interdisciplinary information exchange, *Human Genome News* (HGN) uniquely presents a broad spectrum of genome-related topics to a widely diverse body of 14,000 domestic and foreign HGN subscribers. HGMIS also produces the DOE *Primer on Molecular Genetics*, progress reports on the DOE Human Genome Program, Santa Fe contractor-grantee workshop proceedings, 1-page topical handouts, and other related resource documents.

Document Distribution. In addition to HGN, HGMIS has distributed more than 65,000 copies of items requested by subscribers, meeting attendees, and managers of genetics meetings and educational events.

Electronic Communication. Since November 1994, HGMIS has produced a comprehensive, text-based Web server called Human Genome Project Information, which is devoted to topics relating to the science and societal issues surrounding the genome project. The HGMIS Web sites contain more than 1700 text files that are accessed over 1.2 million times a year. Each month, about 10,000 host computers connect directly to the HGMIS site and through more than 1000 other Web sites.

All HGMIS documents are published on the Web site, along with other DOE-sponsored documents pertaining to the Human Genome Project. Collaborating with the Einstein Institute for Science, Health, and the Courts, HGMIS helps to produce CASOLM, the online magazine for judicial education in genetics and biomedical issues. HGMIS also maintains the Genetics section of the Virtual Library from CERN (Switzerland) and the DOE Human Genome Program pages and moderates the BioSci Human Genome Newsgroup.

Information Source

HGMIS staff members answer individual questions and supply general information about the Human Genome Project by telephone, fax, and e-mail and refer other questioners to experts in the Human Genome Project. HGMIS displays the DOE Human Genome Project traveling exhibit at occasional scientific conferences and genome-related meetings, makes presentations to educational, judicial, and other groups, and strives to strengthen the content relevancy of its services.

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. 9/97

DOE Joint Genome Institute Public WWW site

Robert Sutherland*, Linda Ashworth+, Mark O. Mundt*, Sam Pitluck$, Tom Slezak+, and Darrell O. Ricke*
Joint Genome Institute: *Los Alamos National Laboratory, +Lawrence Livermore National Laboratory, $Lawrence Berkeley National Laboratory

The Joint Genome Institute (JGI) is integrating its information on its public WWW site. This information will combine data from LANL, LBNL, LLNL, and the new JGI PSF (Production Sequencing Facility). While this WWW site is under development, its planned content includes: (1) information about the JGI and its member institutions, (2) links to member institution's WWW pages and other relevant WWW sites, (3) physical maps for regions being sequenced by the JGI, (4) links to entries submitted to public databases, (5) links to the JGI FTP server, (6) finished sequence data with associated quality information and feature annotations, (7) JGI informatics, (8) functional genomics information, and (9) information and WWW-links to promote public education and understanding of Genetics and the Human Genome Program. This work is funded by the United States Department of Energy.

URL: http://www.jgi.doe.gov

Human Genome Program Coordination Activities

Sylvia J. Spengler, Kelcey J. Poe, and Janice L. Mann
Human Genome Program Office, 459 Donner Laboratory, E. O. Lawrence Berkeley National Laboratory, 1 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 the Associate Director, OBER, Dr. A. Patrinos. Members of the Human Genome Program Management Task group are ex officio members, as are members of the Biological and Environmental Research Advisory Committee's subcommittee on the Human Genome. 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 a informatics, technologies, 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 OHER 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 U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

Appendices

Appendix A: Author Index

First authors are in **bold**.

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.