CONF-9411116

DOE Human Genome Program

Contractor-Grantee Workshop IV

Santa Fe, New Mexico

November 13-17, 1994

BIOLOGY · PHYSICS · MATHEMATICS · ENGINEERING · CHEMISTRY

Please address queries on this publication to:

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

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: 615/576-8401.

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

CONF-9411116

Contractor-Grantee Workshop IV
November 13-17, 1994
Santa Fe, New Mexico

Date Published: October 1994

Prepared for the
U.S. Department of Energy
Office of Energy Research
Office of Health 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 37831-6050

Managed by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
for the
U.S. DEPARTMENT OF ENERGY
UNDER CONTRACT DE-AC05-84OR21400

Contents

Workshop Agenda v

Poster Presentation Times viii

Introduction To The Santa Fe Workshop ix

Abstracts

Mapping 1–75
Informatics 76–122
Sequencing 123–167
Instrumentation 168–183
Ethical, Legal, and Social Issues 184–201
Infrastructure 202–204

Appendices

A. Author Index A1
B. National Laboratories Index B1
C. Anticipated Workshop Attendees C1

DOE Human Genome Program Contractor–Grantee Workshop IV

Santa Fe, New Mexico
November 13–17, 1994

Agenda

Plenary sessions are in the Eldorado; poster sessions are in the Hilton. Each speaker and demonstration in the plenary sessions will have an abstract number and a poster associated with the talk. Abstract numbers indicated in parentheses. Schedule correct as of October 14, 1994. Agenda subject to change.

Sunday, November 13, 1994

2:00–8:00p	On-Site Registration	Eldorado
6:00–8:30p	Reception	Eldorado Ballroom

Monday, November 14, 1994

8:30a–3:30p	On-Site Registration Continues	Eldorado
8:30a	Introductions	

Gene and Genome Mapping I — **R.K. Moyzis, chair**

9:00	Progress on Chromosomes 16 and 5 N.A. Doggett, M. Altherr, D.L. Grady (1, 36–41) and (52–54)
10:25	Coffee Break
10:55	Progress on Chromosome 19 L.K. Ashworth, H.W. Mohrenweiser, B. Brandriff, T. Slezak (5, 29–31, 42–50)
12:20p	National Laboratory Gene Library Project L.L. Deaven (25), J.C. Gingrich (26)
12:40	Lunch
2:00–3:00	Poster Session Setup

Gene and Genome Mapping II — **A.V. Carrano, chair**

3:00	D.L. Nelson (56)
3:20	K.H. Fasman (76)
3:40	A. Gnirke (61)
4:00	M.D. Zorn (94)
4:20	L.J. Stubbs (31)
4:40	M.I. Simon (20)
5:15–7:30	Poster Session I & Computer Demonstrations — Hilton

Tuesday, November 15, 1994

Gene and Genome Mapping III — **C.T. Caskey, chair**

8:30a	J. Bashkin, D. Barker (130)
8:50	J.T. Petty (179)
9:10	R. Lipshutz (75)
9:30	F.M. Zweig (184)
9:50	M.C. Jefferson (186)
10:10	A.F. Westin (197)
10:30	Coffee Break
11:00	A.J. Cuticchia (77)
11:25	J.-M.H. Vos (18)
11:45	R.P. Woychik (33)
12:05p	J. Gray (66)
12:25	Lunch

Sequencing I — **M. Narla, chair**

3:00	R.S. Eisenberg (201)
3:30	L.E. Hood (147, 192)
4:10	M.J. Palazzolo (145)
4:35	R.K. Wilson (149)
5:00–7:30	Poster Session II & Computer Demonstrations — Hilton

Wednesday, November 16, 1994

Sequencing II — **D. Kingsbury, chair**

8:30a	C.A. Fields (83)
8:55	D.J. States (105)
9:15	E.W. Myers (119)
9:35	M.J. Cinkosky (79)
10:00	E.C. Uberbacher (103)
10:30	Coffee Break
11:00	A.D. Mirzabekov (124)
11:20	G.A. Evans (125)
11:40	C.R. Cantor (34)
12:00	J.M. Jaklevic (174)
12:20p	Lunch

(Wednesday, cont.)

Sequencing III — **C. Cantor, chair**

3:00	S. Tabor (157)
3:25	C.H.W. Chen (139)
3:45	C.M. Nelson (141)
4:05	P. Williams (140)
4:25	R.D. Smith (143)

5:00–7:30	Poster Session III & Computer Demonstrations	Hilton

Thursday, November 17, 1994

Sequencing IV — **L.M. Smith, chair**

8:30a	N.J. Dovichi (131)
8:55	R.A. Mathies, A.N. Glazer (133)
9:20	B.L. Karger (132)
9:45	R.B. Weiss (148)
10:05	J. Balch (134)
10:25	Coffee Break
11:00	M.S. Westphall (135)
11:25	R.A. Guilfoyle (156)
11:45	L. Ulanovsky (164)
12:05	J. Kieleczawa (161)
12:25	Summary
12:45	Lunch

Poster Presentation Times*
Hilton Ballroom

The number assigned to the investigator's abstract determines when the poster will be presented.

Divide abstract number by 3.

Results with a remainder of 1 (Poster session I, Monday 5:15–7:30 p.m.)
Results with a remainder of 2 (Poster session II, Tuesday 5:00–7:30 p.m.)
Results with a remainder of 3 (Poster session III, Wednesday 5:00–7:30 p.m.)

*Posters should be mounted and ready for display before Session I begins. All posters will remain up throughout the meeting.

Introduction to the Santa Fe Workshop

Welcome to the fourth Contractor-Grantee Workshop sponsored by the Department of Energy (DOE) Human Genome Program. As all funded projects are represented here, this gathering offers valuable opportunities for scientists, program managers, and invited guests to review the program's content and progress and to assess its direction. We also encourage investigators to take this opportunity to discuss their successes and challenges and initiate new collaborations. DOE program management strongly encourages collaborations among investigators and genome centers.

Much progress has been achieved since the last workshop held in February 1993. High-resolution physical mapping of chromosomes 16 and 19 is now virtually complete, with results soon to be published. Cosmid, PAC, and BAC libraries produced by the program are having a major impact on genome activities around the world. The National Laboratory Gene Library Project has achieved its Phase Two goals of producing cosmid libraries for each human chromosome. As technology improves and more groups become involved, DNA sequencing is starting to pick up; a stretch of about 685 kb from the human T-cell receptor is now the longest contiguous sequence entered in the databases. I believe we can expect rapid progress in sequencing technology in the near future. We have also made considerable progress in the rapid analysis of newly developed sequence data. Of particular note is the online implementation of Smith-Waterman analysis at the GRAIL-genQUEST server at Oak Ridge National Laboratory.

Managing and increasing access to the rising tide of genome data and independently maintained databases that contain crucial mapping, structural, and other biological data are critical challenges. Late last year, after an extensive review of the program's informatics activities, DOE expanded its mission by establishing the Genome Sequence Data Base (GSDB), which now functions both as a service facility and research resource. GSDB is housed at the newly established National Center for Genome Resources in Santa Fe. A major current goal is to facilitate close cooperation between GSDB and the Genome Data Base located at John Hopkins University. Thus, the program is moving closer to its vision of a system of interlocking community databases that will allow easy access to independently maintained databases containing relevant biological data.

Of the 204 abstracts in this book, some 200 describe the genome research of DOE-funded grantees and contractors located at the multidisciplinary centers at Lawrence Berkeley Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory; other DOE-supported laboratories; and more than 54 universities, research organizations, and companies in the United States and abroad. Included are 16 abstracts from ongoing projects in the Ethical, Legal, and Social Issues (ELSI) component, an area that continues to attract considerable attention from a wide variety of interested parties. Three abstracts summarize work in the new Microbial Genome Initiative launched this year by the Office of Health and Environmental Research (OHER) to provide genome sequence and mapping data on industrially important microorganisms and those that live under extreme conditions. Many of the projects will be discussed at plenary sessions held throughout the workshop, and all are represented in the poster sessions. Following up the successful debut at the last workshop, two informatics resource rooms will again be set up and maintained to allow researchers to exhibit new resources and software capabilities.

OHER would like to extend thanks to all contributors for their efforts in moving the Human Genome Program toward its goals. We anticipate that this will again be a very interesting and productive meeting, and special thanks go to all who have contributed to its organization.

David A. Smith, Director
Health Effects and Life Sciences Research Division
Office of Health and Environmental Research

Mapping

This page intentionally left blank.

Efforts Toward the Development of a Transcription Map of Human Chromosome 16

The Los Alamos Genomics Group. Presented by Amanda Ford and Cleo Naranjo.

Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545.

Individual chromosomes provide the skeletal framework on which genomic data is organized. As the physical map and an ordered assemblage of molecular clones for chromosome 16 nears completion, we have embarked on the construction of an 'expressed sequence map' of the chromosome. To this end we have chosen the strategy of exon amplification to identify expressed sequences on chromosome 16. Our strategy employs 96-well plate pools of DNA from the flow sorted and arrayed chromosome 16 cosmid library as the substrate for exon trapping. At present we have generated 1800 exon clones from 75 of the 150 plates in the chromosome 16 library. We have sequenced over 500 of these clones and determined that approximately 60% appear unique. These sequences are being mapped to specific locations on chromosome 16 using a panel of somatic cell hybrids and molecular clones. Furthermore, these sequences are being subjected to database analysis to determine whether they represent previously characterized genes or contained conserved motifs that might provide some insight as to their biological function. Once completed, our map will serve two functions. First, it will allow us to pursue a detailed study of chromosome structure and gene organization. Lastly, it will provide investigators who have previously identified linkage to disease genes using chromosome 16 genetic reference markers, additional markers and candidate genes to analyze in their families.

Prepared by M.R. Altherr

Gene Expression in Hydatidiform Mole

Lobb, Rebecca, Cheryl Lemanski, Karen Denison, and Joe Gatewood

Genomics and Structural Biology Group; LS-2, MS 880;Los Alamos National Laboratory; Los Alamos, New Mexico 87545

Complete hydatidiform mole (CHM) is an unusual pregnancy of paternal genetic origin. The majority (90-95%) of CHM pregnancies result from fertilization by a single haploid sperm. Disomy is restored by duplication resulting in a 46, XX karyotype homozygous at all loci. CHM is characterized by complete absence of a fetus, excess trophoblastic growth, and the propensity to become invasive or malignant. Our historical interest in this tissue is based on the unusual genetic origin. We have hypothesized that an unusual chromatin organization in human sperm i s responsible for programming paternal gene expression in early development. This hypothesis is being testing by determining the chromatin compartmentalization properties of genes expressed in this tissue. The current focus of this project is cDNA sequence sampling, and mapping and full length sequencing of selected cDNAs.

Homology comparisons indicate the majority of cDNAs with homology to known genes can be categorized as structural proteins, growth factors and receptors, maternal mitochondrial transcripts, and immune system proteins. The structural genes include laminin, keratin, actin, collagenase, and fibronectin. The immune system genes include beta-2-microglobulin, chaperonin, docking protein, tumor necrosis factor receptor, and HLA Class I light and heavy chains. The majority (>80%) of the cDNAs represent new genes. Over twenty cDNAs have been sequenced.

Research funded by U.S. Department of Energy under Contract W-7405-ENG-36.

Fluorescent *in situ* Hybridization Mapping of cDNA Clones from an Unusual Human Library

Karen Denison, Joe M. Gatewood
Genomics and Structural Biology; LS-2, MS M880; Los Alamos National Laboratory; Los Alamos, New Mexico 87545.

Hydatidiform mole results from an exclusively paternal fertilization event and is characterized by unusual trophoblastic growth without an apparent fetus. Incidence is approximately 1/1000 pregnancies. Tissue from a 14-week pregnancy was used for the production of size selected, directionally cloned cDNA libraries. Over thirty selected clones from one library were mapped using fluorescent *in situ* hybridization (FISH) to human R-banded metaphase chromosomes after the method of Korenberg, *et al.*[1]. Six of these clones represent known genes as described in GenBank. Five of these previously unmapped genes (M-type pyruvate kinase[2], ribosomal protein S3[3], human growth hormone hGH[4], rho GDP-dissociation inhibitor 2[5], and GATA-binding protein 2[6]) have been assigned and the assignment of the sixth gene (type IV collagenase[7]) has been refined.

[1] Korenberg, J.K., Chen, X.N., Doege, K., Grover, J. and Roughley, P.J. (1993). Assignment of the human aggrecan gene (AGC1) to 15q26 using fluorescent *in situ* hybridization analysis. *Genomics* **16**:546-548.
[2] GenBank accession X56494.
[3] GenBank accession X55715.
[4] GenBank accession J00148, K00612.
[5] GenBank accession X69549.
[6] GenBank accession M68891.
[7] GenBank accession J03210, J03070, J05471.

Human cDNA Mapping by Hybridization to High-density megaYAC Dot Blots.

Lucy Ling, Betty Borsody, JoAnn Dubois, Jonathan Norcross, Kathy Falls, Romina Bashirzadeh, Liam Haveran, Teresa Kanjuparamban, Ron Lundstrom, and Donald T. Moir,

Collaborative Research, Inc., Waltham, MA 02154

The emerging YAC-based physical map of the human genome provides a unique substrate for assignment of genes to chromosomal locations. The resulting gene map will be useful for positional cloning strategies and for improving the resolution of comparative mapping between organisms such as mouse and human. We are developing technology for high throughput, large scale mapping of human cDNAs by hybridization to high density filter grids of DNA from megaYAC clones. Our approach involves redundant pooling of megaYAC DNAs and of cDNA probes to reduce errors and save labor. For example, two 96-well source plates of megaYAC DNA are pooled into one 96-well target plate in a manner which represents each clone in three different (**r**=3) pools of six clones each (**p**=6). This reduces the number of required dots by a factor of two (**p/r**) but at the same time provides multiple signals for each authentic hybridization to a megaYAC. Probes are also pooled in a similar manner to reduce the number of hybridizations required. We have completed isolation of about 100 ug of total yeast DNA from each of 9,000 megaYAC clones (CEPH plates 887-980) to provide 3-fold coverage of the human genome. Despite the chimeric nature of many megaYACs, calculations indicate that 62% of cDNAs will be mapped unambiguously to the correct chromosomal location by hybridization to megaYACs representing 3 human genome equivalents. Multiple possible chromosomal locations, including the correct location, will be obtained for most of the remaining probes. Results from probing of filters containing DNA from 3,000 megaYAC clones (one genome equivalent) in a 3x3 array will be described. Random cDNA probes (provided by G. Lennon) were from a normalized human infant brain cDNA library prepared by Bento Soares (Columbia U.), gridded by Greg Lennon (Livermore), and sequenced by Charles Auffray (Généthon). Probes for known members of multi-gene families were used to examine the extent of cross hybridization. Positive dot blot signals are translated into unique megaYAC addresses by computer, and genetic/physical map positions of resulting megaYACs are determined from the Généthon QUICKMAP database and the Whitehead Institute/MIT Center for Genome Research data releases. Experiments were designed to address the following questions: (1) What is the rate and accuracy of cDNA mapping achievable with this approach? (2) How interpretable are results obtained with probes representing members of homologous multi-gene families? For example, can the multitude of hybridization signals obtained from such probes be disambiguated into unique megaYAC addresses? And, are signals resulting from hybridization to the authentic gene distinguishable from signals resulting from hybridization to related gene family members? (3) Is it possible to map cDNA probes containing repetitive elements by pre-reassociation of the probe with Cot-1 DNA prior to hybridization? (4) Can low-stringency inter-*Alu* element PCR products from megaYAC DNA be substituted for total yeast DNA from megaYAC clones as a substrate for cDNA mapping by hybridization? What fraction of cDNA probes find hybridization targets among the inter-*Alu* PCR products of megaYACs? Answers to these questions will determine the feasibility of this approach for rapid construction of a total human gene map. This work was supported by DOE grant DE-FG02-92ER61399.

Gene Isolation From Human Chromosome 19

Greg Lennon, Dominique Giorgi, Kimberly Lieuallen, Len Pennacchio, Christa Prange, and Sylvie Rouquier

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore California 94550.

The goal of this effort is to isolate, sequence, and map coding regions located on human chromosome 19. First, 34 cDNAs selected by direct screening of cDNA libraries with ch 19 cosmids have been verified to correspond to ch 19 genes, and their corresponding genomic loci have been mapped. Second, hybrid selection experiments using flow-sorted ch 19 and arrayed cDNA libraries have resulted in cDNA sublibraries highly enriched for sequences from chromosome 19. A combination of both sequence and hybridization analysis of over 100 sublibrary clones indicates that over 50% have cognate sequences on chromosome 19. Some of the more than 60 new genes are being characterized in detail, including a candidate gene for two disorders linked to 19p, pseudoachondroplasia and multiple epiphyseal dysplasia.

The need for high-quality, arrayed cDNA libraries has also led us to form (in conjunction with C. Auffray) the Integrated Molecular Analysis of Gene Expression (IMAGE) Consortium. IMAGE is a collaborative group of over 27 laboratories working on the characterization of clones from shared arrayed cDNA libraries. Information derived from the study of specific clones is shared through the use of a collaborative database established and maintained (by M. Boguski and C. Tolstochev) at the NCBI. This database periodically transfers data directly to the public domain database dbEST. Most work to date has focused on the normalized infant brain cDNA library (from Dr. M. Soares, Columbia Univ.) as arrayed at LLNL, and then replicated and distributed worldwide. This array consists of 40,000 clones; over 10,000 single pass sequences have been generated, and over 1,000 cDNAs have been chromosomally mapped by Consortium members. We are currently arraying other high-quality normalized cDNA libraries, and invite the participation of laboratories willing to abide by the Consortium guidelines.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

Mouse RNAs for the Determination of Stage- and Tissue-Specific Expression Profiles for Human Genes Throughout Development

Dabney K. Johnson[1] and Lisa J. Stubbs

Mammalian Genetics and Development Section, Biology Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37830

Easy access to tissues for the isolation of mRNA at all developmental stages makes the mouse an ideal model organism for determining the expression profile of a gene throughout the life of a mammal. For cell populations, like those of preimplantation-stage embryos, that are prohibitively difficult to obtain in usable quantities, cRNAs generated from stage-specific cDNA libraries (1) can fill the same role.

We propose, first, to make a series of polyA+ Northern blots that represent the spectrum of mouse development from the fertilized egg to a complete panel of adult tissues. Egg, 2-cell, 8-cell, and blastocyst cRNAs will be generated via *in vitro* transcription from existing cDNA libraries, while postimplantation mRNAs will be isolated from whole or dissected fetuses, neonates, juveniles, and adults. Given the extensive mouse-human sequence homologies found in most coding sequences, these Northern blots should allow the determination of the expression profile for most human genes in a comparable mammalian system. In addition, aliquots of these same stage-specific RNAs will be preserved for use in RT-PCR reactions in cases where gene-specific primers can be used.

(1) Rothstein, J.L., Johnson, D.K., DeLoia, J.A., Skowronski, J., Solter, D., and Knowles, B.B. (1992) Gene expression during preimplantation mouse development. *Genes and Devel.* **6**, 1190-1201.

Integrated physical mapping of human cDNAs.

Mihael H. Polymeropoulos

National Center for Human Genome Research, NIH Bldg49 4A66, Bethesda Md20892.

In the effort to characterize human cDNA derived partial sequence tags, we have developed STSs for 1,000 human cDNAs. With the use of human-rodent somatic cell hybrids, these cDNAs have been assigned to a chromosome. Although the information of the chromosome of origin is useful, it is the fine resolution mapping of the expressed sequences that maximizes their use human disease and related mapping efforts. In order to develop high resolution expression maps of the human genome, we have completed a pilot project for the high resolution mapping of chromosome 8 derived ESTs. In this project we screened the CEPH megabase YAC library for YAC addresses with each chromosome 8 STS. Two addresses were obtained for each cDNA and the database was searched for overlapping YACs at levels two and three. The original YAC addresses and the overlapping YACs were database searched for the presence of microsatellites with chromosome 8 origin. These searches were performed with the computer program ***yacsr.*** Of the chromosome 8 cDNAs mapped in this manner 60% were linked to a microsatellite on levels one or two. The relative locations were also verified by the use of chromosome 8 deletion hybrids. With the increasing density on YACs typed for genetic markers the success in fine mapping of the cDNAs will continue to increase. This mapping approach will provide powerful means of gene mapping by providing simultaneously a physical clone address and a location on the genetic map.

Utilization of normalized libraries in cDNA selection experiments to identify transcribed sequences in genomic DNA

[1]Marcelo Bento Soares, [1]Maria de Fatima Bonaldo, [1]Pierre Jelenc, [1]Lee Lawton, [1]Long Su & [2]Argiris Efstratiadis
[1]Department of Psychiatry and [2]Genetics & Development, Columbia University and [1]The New York State Psychiatric Institute

We have further optimized the method that we developed for construction of normalized directionally cloned cDNA libraries [1] and we have used it to generate a human fetal liver and spleen library of 5 million recombinants. In order to assess the quality of this normalized library we performed a number of screenings (colony hybridization) of both starting and normalized libraries with cDNA probes representing mRNAs that occur at a wide range of frequencies in the starting library. The results that were obtained clearly indicated that normalization was successful.

We have also used this library in cDNA selection experiments to isolate transcribed sequences from within a 600 kb YAC contig spanning the Spinal Muscular Atrophy critical region [2]. Approximately 30 different cDNAs were identified and confirmed to map to this region, which corresponds to an average spacing of one gene every 20 kb. cDNA selection was carried out according to a method that we developed [3] and successfully utilized to isolate a number of chromosome 13-specific cDNAs.

As part of the quality control that we routinely perform on our libraries, we have generated a total of about 1,100 ESTs, all of which have been deposited in Genbank. These sequences were derived from a number of libraries that were constructed while we were developing this method for normalization, as well as from an infant brain and the fetal liver and spleen normalized libraries.

[1] Soares, M.B., Bonaldo, M.F., Su, L., Lawton, L. & Efstratiadis, A. Construction and characterization of a normalized cDNA library (1994). Proc. Natl. Acad. Sci. USA **91**(20), 9228-9232.

[2] Kindly provided by Dr. Conrad Gilliam, Department of Genetics & Development, Columbia University.

[3].Bonaldo, M.F., Yu, M-T., Jelenc, P., Brown, S., Su, L., Lawton, L., Deaven, L., Efstratiadis, A., Warburton, D., & Soares, M.B (1994). Selection of cDNAs using chromosome-specific genomic clones: application to Human Chromosome 13. Hum. Mol. Genet. **3** (9).

Application of Fluorescence in situ Hybridization in Genome Analysis

Barbara J. Trask[1], Hiroki Yokota[1], Cynthia Friedman[1], Hillary Massa[1], Eric Green[2], Jim Evans[3], Antonia Martin-Gallardo[4], Janey Youngblom[5], and Ger van den Engh[1]

[1]University of Washington, Seattle, WA; [2]National Center for Human Genome Research-NIH, Bethesda, MD; [3]University of North Carolina, Chapel Hill, NC; [4]Centro Nacional de Biotecnologia, Madrid, Spain; California State University, Turlock CA.

We report on the application of fluorescence in situ hybridization (FISH) to several areas of genome research:

A) Correlating and confirming maps. The extensive cross-correlation of the cytogenetic map with the genetic and physical maps of chromosome 7 will be shown as an example.

B) Mapping relative to chromosome rearrangements. This application will be illustrated with results on patients with split-hand/split-foot syndrome. YACs from 7q were ordered by FISH relative to deletion and translocation breakpoints in ~10 patients. The results order the breakpoints along the chromosome and demonstrate that the critical region for this disease spans 500 kbp.

C) Characterizing large-scale polymorphisms. We have characterized a 40-kbp region in different individuals by a combination of FISH, PCR, and hybridization to flow-sorted chromosomes. This sequence is found near the ends of several human chromosomes, but on only one chimp or gorilla chromosome. The region is polymorphically present on additional chromosomes in some individuals. The frequency of polymorphic variants differs among reproductively isolated populations. Limited sequencing indicates that the region may contain functional genes.

D) Studying the organization of the interphase nucleus. We have made >25,000 pair-wise distance measurements in G1 interphase nuclei between DNA sequences separated by 0.15 to 190 Mbp on three different human chromosomes. Our results are consistent with a model in which interphase chromatin is organized in ~5-Mbp-sized flexible loops held together along a supple backbone-like structure The results have important consequences for the use of FISH as a mapping tool.

This work was funded by the U.S. DOE grant FG06-93ER61553, B. Trask, P.I.

Production of probes at the Resource for Molecular Cytogenetics.

W.-L. Kuo[1,3], C. Collins[2], L. Daneshvar[1], K. Greulich[2], D. Kowbel[2], J. Marstaller[2], D. Pinkel[1,2], L. Riedell[1], F. Shadravan[2], M. Wang[2], U. Weier[2], P. Yue[1], M. Zorn[2], and J. Gray[1,2].

LBL/UCSF Resource for Molecular Cytogenetics, [1]Dept. Laboratory Medicine, University of California, San Francisco, CA 94143-0808 and [2]Lawrence Berkeley National Laboratory, Berkeley, CA. [3]Corresponding author.

The LBL/UCSF Resource for Molecular Cytogenetics has been created to develop probes and associated technologies to facilitate molecular cytogenetic analyses. One goal is to develop probes optimized for use in fluorescence in situ hybridization (FISH). The majority of probes are being selected to contain specific genes or genetically mapped polymorphic loci distributed at ~5 Mb intervals over the human genome.

Human probes are selected from chromosome-specific cosmid libraries, YAC libraries and the Du Pont P1 library. The P1 genomic library is our primary source for PCR-based screening using gene or locus-specific primer pairs. Selected clones are mapped to human metaphase chromosomes by FISH. A QUantitative Image Processing System (QUIPS) developed in the Resource maps probes according to fractional location relative to the p-terminus.

To date, 411 clones have been mapped. 233 clones (218 P1's, 11 YACs and 4 cosmids) were selected for 133 loci defined by STS. These include approximately 40 genes. The rest are anonymous clones (75 P1's and 103 cosmids). Extensive probe sets have been mapped to chromosomes 3, 10, 17, 20 and 21. Our current emphasis is on isolating P1 probes at chromosomal loci known or suspected to display copy number alteration or rearrangement in human disease. So far we have isolated P1's for tumor suppressor genes and oncogenes (p53, c-MYC, GLI, SIS, E-cadherin), translocation breakpoints of clinical significance (PML, RARA, TCRA, ETO, ALL), and regions involved in contiguous gene syndromes (Angelman, Prader-Willi, Cri du chat, Wolf-Hirschhorn and DiGeorge syndrome). Information on probes developed by the Resource will be made available through GDB and through the Resource Internet Web server. A mechanism for prompt, low cost distribution of the probes is being established.

This work was funded by the Office of Health and Environmental Research, Department of Energy, under contract DE-AC-03-76SF00098 and Imagenetics.

Reagents For Understanding And Sequencing The Human Genome

[1]J. R. Korenberg, [1]X.N. Chen, , [1]S. Gerwehr and
BAC: [1]S. Mitchell, [1]R. Hubert, [2]U-J. Kim, [2]H. Shizuya, [2]M. Simon, [1]K. Yamakawa.
cDNA: [3]M. Adams, [4]K. Becker, [5]K. Denison, [4]P. Drew, [5]J. Gatewood, [6]G. Guellaën, [7]L. Hood, [8]D. Hwang, [7]G. M. Huang, [8]C.-C. Liew, [3]J. C. Venter
[1]Cedars-Sinai Medical Center & UCLA, Los Angeles, CA; [2]California Institute of Technology, Pasadena, Ca.; [3]Institute for Genomic Research, Gaithersburg, MD; [4]Neuroimmunology Branch, NINDS; [5]Los Alamos National Laboratory, Los Alamos, NM; [6]INSERM U 99 Hospital Henri Mondor 94010. Creteil, France; [7]University of Washington - Seattle, Wa; [8], University of Toronto, Canada.

An ultimate goal of the human genome project is to rapidly identify a subset of the 50-100,000 human genes that is responsible for genetic and acquired disease. Toward this goal, we have focused work in two areas: defining and mapping the genes and providing arrays of genomic reagents for gene finding, sequencing and molecular cytogenetics.

I. Human cDNA Mapping using Fluorescence in situ Hybridization (FISH)

Technology has been developed and applied for the rapid assignment of cDNAs to single sub-bands (2-6 Mb regions) of high resolution human chromosomes. An international consortium of 6 groups has been established and has resulted in the mapping of 240 new genes from 5 tissue-specific cDNA libraries to single human chromosome regions. These include testis, fetal brain, fetal and adult heart, bone marrow, and prostate, in addition to members of the family of zinc finger-motif-containing genes. The techniques developed eliminate the confusion of pseudogenes, confirm and provide anchor points for the large fragment maps, and rapidly and unambiguously provide candidate genes for diseases mapped in the region. Further, the throughput would allow low-cost (<$100) assignment of 8-10,000 genes in 3 years by a single production group.

II. Towards a BAC Map of the Human Genome

Although the construction of YAC libraries has clearly changed the paradigm for generating complete genome arrays, there is a basic, widespread need for large fragment stable reagents covering the genome. To address this need, the laboratory has begun the elaboration of a genome-wide array of BACs using a random, cost-efficient strategy to obviate the issues of YAC chimerism and rearrangement. The ultimate goal is to provide stable reagents for 70% of the genome. These arrays would be produced by using current techniques of multiplex FISH to assign 24-30,000 BACs to genomic regions at a resolution of 2-6 Mb. This resource will provide unambiguous reagents for gene identification and sequencing, will fix the cytogenetic to the molecular maps, are readily integrated by STS and hybridization analyses within a 2-5 Mb region and provide anchor points for generating complete contigs, easily combined with other large-fragment libraries including YACs, PACs, P1s, cosmids, and fosmids Toward this goal, we have assigned more than 1,000 BACs, that cover more than 80% of all chromosome bands and have identified a subset of 148 BACs containing chromosome-specific human alpha-satellites and have identified BACs for 7 human telomeres. Chimerism in the library has been evaluated and estimated at less than 5%. These reagents now provide a broadly accessible resource for rapidly generating contigs of any genomic region, allowing the study of human centromeres and telomeres, and facilitating the pace of human gene discovery. This requires knowledge of the cDNAs and their map positions, an array of stable genomic reagents that cover most of the genome and are suitable for sequencing, and a high resolution genetic map. The goal of the research in this laboratory has been focused on characterizing the cDNAs and and providing their potential relationship to human genetic disease.

Strand-Specific In Situ Hybridization: Application of the Method

Julianne Meyne, Edwin H. Goodwin, Susan M. Bailey,
Loanne R. Smith, Denise I. Quigley and Robert K. Moyzis

Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545

Strand-specific fluorescence in situ hybridization (FISH) is a modification of the standard FISH method that allows hybridization of a single-stranded deoxyoligomer probe to only one chromatid of a metaphase chromosome. This is accomplished by removing one strand of the DNA helix from each chromatid. Because the method allows determination of the relative chromosomal orientation of highly repetitive, tandemly organized DNA sequences it is also referred to as CO-FISH (Chromosome Orientation FISH) [1,2]. We have hybridized a variety of repetitive DNA sequences localized within the pericentromeric regions of mammalian chromosomes using the CO-FISH method. All of the tandemly repeated satellite sequences studied to date have been oriented in the head-to-tail fashion except the heterochromatic C-band positive region of the long arm of human chromosome 9. The satellite 2 and 3 sequences in this region are arranged in mixed orientation.

CO-FISH has been a very reliable method to detect compound lateral asymmetry in the heterochromatic C-band positive region of the long arm of human chromosome 1. This polymorphic region exhibits a variety of simple and compound lateral asymmetry combinations, as revealed by studies of chromosomes from both cultured fibroblasts and lymphocytes. Similar patterns are observed using repetitive probes present in the pericentromeric regions of other mammalian species.

When slides prepared using the CO-FISH method are hybridized with both a repetitive DNA probe and a probe for the telomere, one can also determine the direction of the DNA sequence and its complement. It is known that the G-rich strand of the telomere overhangs the 3' end of the DNA duplex within each chromatid. Therefore, by selecting the appropriate single-stranded telomere probe one can label either the 3' or 5' end of each chromatid and determine the direction of that strand. Centromeric twists and sister chromatid exchanges can be discerned by analyzing the relative frequencies of the observed hybridization patterns.

[1] Goodwin, E.H. and Meyne, J. (1993) Strand-specific FISH reveals orientation of chromosome 18 alphoid DNA. Cytogenet Cell Genet 63: 126-127.

[2] Meyne, J., Goodwin, E.H. and Moyzis, R.K. Chromosome localization and orientation of the simple sequence repeat of human satellite I DNA. Chromosoma 103: 99-103

This work was supported by grants from the U.S. Department of Energy.

TIME-GATED FLOW ANALYSIS OF HUMAN CHROMOSOMES. V.V. Zenin[#], N.D. Aksenov[#], A.N. Shatrova[#], Y. V. Kravatsky[*], A. Kuznetsova[*], L. S. Cram[^], A. I. Poletaev[*]; [#]Institute of Cytology RAS, St. Petersburg; [*]Engelhardt Institute of Molecular Biology, RAS, Moscow; [^]Los Alamos National Laboratory, Los Alamos, NM 87545.

For chromosome analysis and sorting we have made several modifications to our flow cytometer and chromosome isolation procedures. An electronics modification of our serial flow-sorter (ATC-3000, ODAM-Bruker) was made to provide the option of measuring the inter-event time interval in the microsecond to millisecond time range. The inter-event time interval can be used as an independent parameter in flow cytometry data acquisition. A stepper motor driven sample delivery system (for achieving calibrated flow rates) is used in conjunction with a capillary high-gradient mitotic cell breaking device to disrupt mitotic cells and release intact metaphase chromosomes. We have found that one can get stable staining of human chromosomes within mitotic cells with both Hoechst 33258 and Chromomycin A3 without significant cell membrane rupture using saponin as the cell membrane affecting substance. Two versions of a mitotic cell breaking device were tested: electromagnet sheering and a high-gradient curved capillary. Both devices proved to be compatible with most of serial flow cytometers able to perform flow analysis of chromosomes. Time-gated data acquisition provides a possibility to filter out signals of chromosomes arising from single cells from contaminating signals: cell debris and randomly appearing single chromosomes. This can be done both in real time and on list mode data files. The specific software provides the possibility of doing different filtering procedures with list mode data to get the differential information about intracellular chromosome sets. Future work will target optimization of the new components and the procedure of measurement. The realization of this approach might provide an efficient analysis of different chromosome aberrations. This work was supported by grants from DOE and Russian State Program "Human Genome".

Hybridization Technology Development at the Resource for Molecular Cytogenetics.

H.-U. Weier[1,2], C. Thompson[2], M. Wang[1], K. Greulich[1], C. Collins[1], J. Gray[1,2] and D. Pinkel[1,2].

Resource for Molecular Cytogenetics, [1]Life Sciences Division, Lawrence Berkeley Laboratory, Berkeley, CA and [2]Department of Laboratory Medicine, University of California, San Francisco, CA.

The advancement of fluorescence in situ hybridization (FISH) techniques is a major goal of the Resource for Molecular Cytogenetics. We have focused on techniques to facilitate high resolution DNA probe mapping and ordering and to improve genetic analysis of human tissue samples.

Probe mapping and ordering: High resolution mapping and ordering of DNA molecules is required for tasks such as contig assembly. With current techniques it is often difficult or extremely laborious to establish the relative locations of subfragments of a large molecule, or to determine the amount of overlap of two elements in a contig. FISH has the ability to visualize hybridization to single extended DNA target molecules. Current techniques involve the use of genomic target DNA released from cell nuclei and spread on slides. We have employed techniques developed by collaborators (A. Bensimon, Institut Pasteur, Paris) that permit linear extension of individual cloned molecules on activated glass substrates by covalent attachment of one end. Hybridization of probes as small as 2kb can detected and the positions of hybridization sites determined to within 1kb.

FISH in thick tissue sections: FISH has the potential to provide genetic information from individual cells. Analysis of cells in context with surrounding cells in tissue samples provides information on the spatial distribution of genetic status, such as the evolution of genetic changes in tumors. Such studies require hybridization to tissue specimen that are thick enough to contain a large number of intact cells. We have developed protocols employing direct fluorochrome labeled probes that permit hybridization with centromeric probes throughout 20μm thick formalin-fixed/paraffin-embedded clinical specimens. Image analysis algorithms to obtain genetic information from such three dimensional specimens are under development (see Lockett et. al., this meeting).

This work was funded by the Office of Health and Environmental Research, Department of Energy, under contract DE-AC-03-76SF00098 and Imagenetics.

CGH Imaging for Routine Use

Damir Sudar[1,5], Jim Mullikin[1], Steve Lockett[1], Jim Piper[3,1], Gus van der Feltz[4,1], David Kaszuba[2], Mark de Kanter[4,1], Dan Pinkel[2,1], Joe Gray[2,1]

[1]Resource for Molecular Cytogenetics; MS 74-157; Lawrence Berkeley Laboratory; 1 Cyclotron Road; Berkeley, CA 94720. [2]University of California, San Francisco, CA. [3]Medical Research Council, Edinburgh, UK. [4]Delft University of Technology, Delft, The Netherlands. [5]Corresponding author

The Resource for Molecular Cytogenetics at LBL/UCSF uses CGH analysis for research into many different tumors and celltypes potentially containing genetic abnormalities. In a typical CGH application test and reference genomic DNA, each labeled with a different fluorochrome, are hybridized to normal metaphase chromosomes. The ratio of intensity of the 2 fluorochromes along the length of the chromosome is proportional to the local DNA copy number. Around 25 researchers routinely perform CGH experiments which puts special requirements on the systems used for the analysis. The typical rate of image acquisition is 100 Mbytes/day of compressed data on 2 digital imaging microscopes; 5 workstations are used for analysis and image data is maintained on multiple servers and storage systems. We describe the CGH imaging procedures in use at the Resource and their implementation for routine use.

Data is obtained from CGH experiments using a digital imaging microscope (QUIPS) which we developed. QUIPS emphasizes straightforward acquisition of multi-color fluorescence images of high quality. User-friendly programs have been developed for convenient and rapid acquisition of the required images which are automatically stored on the fileserver in a single dataset. Fully automated routines take care of migration of older datasets to a tape robot based storage system to make space for new data. Image data stays on the local fileserver for an average of 4 months which gives the researcher time to analyze it. Data archived on the tape robot system can be recalled overnight.

Images stored on the fileserver are recalled on one of the analysis workstations using programs under SCILimage. Individual fluorochrome images are read in from the dataset for visual analysis and for semi-automated processing according to Piper et al. [1]. Visual analysis consists of inspection of the individual fluorochrome images and multi-color compositions. Semi-automated processing begins with chromosome segmentation based on the counterstain and reference DNA images after background correction and is followed by a cluster detection and decomposition step [2]. Normalizers for each DNA image are calculated from the derived segmentation mask. Chromosome identification is done visually after image enhancement of the counterstain image. We are working on the integration of computer-assisted karyotyping techniques from standard cytogenetics to facilitate the identification of chromosomes. For each chromosome a local background is calculated and a profile is extracted for both hybridization images by integrating the background corrected and normalized pixel values along slices orthogonal to the chromosome axis. A division of the 2 profiles yields the CGH ratio profile for this chromosome

The ratio profiles for multiple chromosomes are averaged to reduce noise and displayed in a "copy number karyotype" with a choice of variability measures. A typical number of chromosomes used is 4. Significant deviations from the ratio value of 1.0 are noted as amplifications or deletions.

This work was funded by the US DOE contract DEAC0376SF00098.

[1] Piper J, Rutovitz D, Sudar D, Kallioniemi A, Kallioniemi O-P, Waldman F, Gray JW, Pinkel D. Computer Image Analysis of Comparative Genomic Hybridization. Cytometry 16 (in press)

[2] Ji L. Intelligent Splitting in the Chromosome Domain. Pattern Recognition 22: 519-532, 1989

High Resolution Mapping by Scanning Probe Microscopy

D. P. Allison, T. Thundat, and R. J. Warmack
Health Sciences Research Division
Oak Ridge National Laboratory
and
Colin Collins* and Joe Gray*#
**Lawrence Berkeley National Laboratory*
Resource for Molecular Cytogenetics
#University of California San Francisco
Division of Molecular Cytometry

The aim of this project is to evaluate the feasibility of using atomic force microscopy (AFM) to image protein specifically bound to large DNA molecules and to apply this technology to the rapid construction of "restriction mapped" long range P1, BAC and PAC contigs. Substantial progress toward this goal has been accomplished.

Recent improvements in our techniques for mounting DNA to pristine mica surfaces and the incorporation of critical point drying into our methodology have allowed us to image DNA relatively free from background noise. We have imaged large molecules, such as 50 kb lambda DNA and 80 kb P1 clones, and anticipate no problems imaging molecules up to 200 kb. Restriction mapping will be accomplished by identifying site specific binding of restriction enzymes that have been modified to bind *but not cleave* DNA. Model experiments using such a mutant EcoRI restriction enzyme have identified the single EcoRI site on pBS^+ and the two EcoRI sites on pMP32 plasmid DNA. Visualization of the restriction sites is greatly facilitated by biotinylating the enzyme-DNA complex and labeling with 10nm streptavidin-gold prior to image acquisition. In an extension of this technology we are attempting to map oligonucleotides to cloned DNA by RecA protein mediated hybridization and imaging RecA attachment sites or D-loop sites.

This technology offers the possibility to radically accelerate the construction of long range P1, PAC and BAC contigs for use in large scale directed sequencing and positional cloning projects. A single technician using a pair of AFM microscopes ($80,000) can be expected to image and assemble 50 Mb per year. Since AFM can potentially image both protein (transcription factors) and small oligonucleotides (corresponding to exons, STSs, and promoter elements) specifically bound to DNA, it may become possible to rapidly construct fully integrated physical and genetic maps for any chromosomal region.

Construction And Characterization of Region-Specific Microdissection Libraries for Human Chromosome 2

Fa-Ten Kao and Jingwei Yu
Eleanor Roosevelt Institute for Cancer Research, 1899 Gaylord St., Denver, CO 80206, and Department of Biochemistry, Biophysics and Genetics, University of Colorado Health Sciences Center, Denver, CO.

Using microdissection and microcloning techniques developed in this laboratory [1], we have constructed 4 region-specific libraries for the entire short arm of human chromosome 2: 2p23-p25 (designated 2P1 library), 2p21-p23 (2P2), 2p14-p16 (2P3), and 2p11-p13 (2P4). These libraries are large, comprising 300,000-1,000,000 recombinant microclones with a mean insert size of 250 bp (ranging between 50-800 bp). 40-60% of the microclones contain unique sequences. Southern blot hybridization showed that 54-86% of the single-copy microclones were derived from the respective dissected region. A subset of single-copy microclones from each library was characterized and the hybridizing HindIII fragments in the human genome determined, including 26 microclones for the 2P1 library, 60 clones for 2P2, 66 clones for 2P3, and 30 clones for 2P4. Details of these libraries have been described [2-5]. In addition, 14 of the 26 microclones in the 2P1 library were further mapped to the 2p23.3-p25.1 region by dosage analysis using a patient with an interstitial deletion of this region [2], and 10 of the 60 microclones from the 2P2 library were mapped to 2p21 using an interstitial deletion in 2p21 [4]. The 2P3 library and the single-copy microclones have been used in a team effort which led to the cloning of the hereditary nonpolyposis colorectal cancer (HNPCC) gene [6].

Six region-specific libraries for the entire long arm of chromosome 2, plus a library for the centromere region, are under various stages of construction and characterization. One library for the distal long arm, region 2q35-q37 (2Q1), has already been constructed and described [7]. 26 single-copy microclones from the library were characterized and 4 of these clones were further mapped to the 2q37 region using a cell hybrid containing only this region. This library has been deposited in the American Type Culture collection (ATCC No. 77419). Other libraries will be similarly deposited for general distribution to the genome community.

Region-specific libraries and the single-copy microclones from the libraries are useful resources for genome studies [8]. For example, the libraries can be used to screen for highly polymorphic microsatellite markers for high resolution linkage analysis and to narrow down the distance between a probe and a disease locus under study [6]. In addition, single-copy microclones with short inserts can be conveniently sequenced to prepare STSs, and it is relatively simple to isolate many single-copy microclones to provide sufficient STSs (e.g. 1 STS per 50-100 kb or less) for constructing high density physical maps and to prepare sequence-ready reagents for each dissected region of 10-20 mb.

[1] PNAS 88, 1844-1848, 1991.
[2] Hum. Genet. 93, 557-562, 1994.
[3] Somat. Cell Mol. Genet. 20, 133-136, 1994.
[4] Cytogenet. Cell Genet. in press.
[5] Somat. Cell Mol. Genet. in press.
[6] Cell 75, 1215-1225, 1993.
[7] Genomics 14, 769-774, 1992.
[8] BioEssays 15, 141-146, 1993.

Human Artificial Episomal Chromosomes (HAECs) for Cloning, Mapping and Functional Testing of Large Genetic Units in Human Cells

Tian-Qiang Sun[1], Michael Grosz[3], Zachary Kelleher[2] & Jean-Michel H. Vos[1-3]

[1]Department of Biochemistry and Biophysics; [2]Curriculum in Genetics and Molecular Biology; [3]UNC Lineberger Comprehensive Cancer Center, CB#7295, 349 LCCC, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA

Of approximately 100,000 human genes, only a few thousands have been cloned or mapped so far. For other chromosomal regions, such as those involved in DNA replication, chromatin packaging and chromosome segregation, much less is known. The construction of detailed physical maps is only the first step for the localization, identification and functional determination of genetic units in human cells. To study the function and regulation of human genes and other critical genomic regions which span hundreds of kilobase pairs of DNA, one has to be able to clone an entire functional unit as a single DNA fragment and transfer it into human cells. Current large-insert heterologous cloning systems, such as YACs, BACs and PACs, do not appear suitable for transferring and functional analysis of genetic units in human cells.

We have developed a human artificial episomal chromosome (HAECs) system, based on the latent replication origin of the large herpes Epstein-Barr virus (EBV) for the propagation and stable maintenance of DNA as circular minichromosomes in human cells [1]. Individual HAECs carried human genomic inserts ranging from 60 to 330 kb and appeared genetically stable. A HAEC library of 1,500 independent clones carrying random human genomic fragments with average sizes of 150-200 kb was established. The multicopy circular HAEC DNA was selectively recovered and used for restriction mapping, PCR amplification and fluorescent *in situ* hybridization. This autologous HAEC system, with human DNA segments directly cloned in human cells, provides an important tool for the mapping, sequencing and, most importantly, the functional study of large mammalian DNA regions [2]. Current efforts are focused on a) the packaging of the HAEC-based library as infectious EBV for shuttling large human genomic inserts in human cells [4], and b) the construction of human chromosome-specific HAEC libraries to complement the bacterial- and yeast-based mapping efforts.

[1] Sun T-Q and Vos J-MH (1994) "Human Artifical Episomal Chromosomes for Cloning Large DNA in Human Cells" Nature Genetics, in press.
[2] Vos J-MH (1994) "Herpesviruses as Genetic Vectors" in "Viruses for Human Gene Therapy" ed. J-MH Vos, Carolina Academic Press, Durham, NC, USA, pp. 109-140.
[3] Sun T-Q and Vos J-MH (1994) "Cloning Large DNA in Human Cells with the HAEC system" Methods in Molecular Genetics, ed. K. Adolph, Academic Press, San Diego, CA, Vol. 3, in press.
[4]Sun T-Q and Vos J-MH (1992) Packaging of 150-200 kb DNA as infectious Epstein-Barr virus, Int. J. Genome Res. 1: 45-57.

A Human DNA Library in Bacterial Artificial Chromosomes (BACs): Application to Physical Mapping of Chromsome 22

Hiroaki Shizuya, Ung-Jin Kim, Bruce Birren, Tania Slepak, Valeria Mancino, April Mengos and Melvin Simon

Division of Biology, 147-75, Caltech, Pasadena, California 91125

Two new BAC vectors have been constructed. They are derived from the original pBAC108L vector. pBeloBAC11 has a cloning site in the *lacZ*, therefore BACs are identified as white colonies. The positive selection pTrimBAC111 is based on the ability of growing in the presence of trimethoprim in the *thyA* mutant bacteria when DNA insertion occurs. Using these vectors, we have constructed total human BAC library containing 96,000 clones with average insert size of ~125kb (*i.e.*, 4X coverage).

The library can be accessed by hybridization of clones gridded at high density onto nylon filters in 5 x 5 configuration or by PCR with STS primers from BAC DNA of pooled clones. The library has been successfully probed with many known markers, YACs, Fosmids, and cosmids. Moreover, regions of the genome that are difficult to clone and maintain in YAC or cosmid vectors have been identified on BACs. Thus it appears to cover extensive regions of human DNA. Using the library we are constructing a high resolution physical map of human chromosome 22. One approach toward this goal is to identify as many chromosome 22 specific BACs as possible by probing with a variety of chromosome 22 markers and to walk from these anchor BACs. BAC inserts can be used to walk by hybridization in order to generate a long contig nucleated by the inserts (BAC to BAC walking). Fingerprinting these clones by restriction enzymes then confirms the contig, and identifies BACs located at the end of contigs for the next round of walking. After two rounds of BAC-to-BAC walking we collected more than 1,000 chromosome 22 specific BACs and constructed 69 separate contigs. Each contig spans about 300 kb on average, and therefore 20 mb of chromosome 22q has been covered, leaving, on average, a 350 kb gap between contigs. We are currently filling the gaps with a variety of methods.

Bacterial Artificial Chromosomes (BACs) for Mapping and Sequencing

Melvin I. Simon[1,4], Ung-Jin Kim[1], Hiroaki Shizuya[1], Bruce Birren[2], and Jerry Solomon[3]

[1]Biology Division, California Institute of Technology, Pasadena, California. [2]Whitehead Institute, Center for Genome Research, Massachusetts Institute of Technology, Cambridge Massachusetts. [3]Beckman Institute, California Institute of Technology, Pasadena, California. [4]Corresponding author.

Bacterial artificial chromosomes and Fosmids are stable, non-chimeric, highly representative cloning systems. The BACs maintain large fragment genomic inserts (100-300 kilobase pairs) that are easily prepared as DNA templates for sequencing. We have improved the methods for generating BACs and developed extensive BAC libraries. A BAC library in the mouse was prepared in less than two months; it consists of approximately 105,000 clones providing approximately four-fold coverage of the mouse genome. BAC libraries corresponding the human genome now provide more than 6X coverage. Libraries have been made with DNA from many other organisms.

In order to demonstrate the utility of the BAC system for mapping, we have applied and extended a variety of existing techniques to generate a physical map of chromosome 22. We have chromosome-specific-BACs that correspond to approximately three-fold coverage of the chromosome. These BACs have been arranged into contigs and we are currently engaged in closing the remaining gaps to finish a complete map of the 45 megabase Q-arm of chromosome 22 which will incorporate over 100 STS markers, cosmid and cDNA markers and will be correlated with yeast artificial chromosome maps that have been developed in other laboratories for chromosome 22. BACs provide a rapid, easy source for hybridization or PCR-STS mapping that can allow the complete coverage of all of the chromosomes. Furthermore, the BAC libraries represent general resources that can be interrogated with almost any form of DNA probe to generate 1-2 megabase contigs in any region of the genome. This will be illustrated by a number of examples. Furthermore, since the BACs are present as closed covalent circular DNA which is easily isolated from *E. coli*, the BAC contigs provide convenient stable, ordered sets of material for long range sequencing. The circular form of the plasmid also allows BACs to be easily manipulated by the techniques of bacterial genetics and a variety of transposons can be used to retrofit the BAC vectors with properties that allow specific inserts to be expressed in eukaryotic cells or to generate nested sets of deleted fragments of any genomic insert that provide for the rapid characterization of the insert.

We are currently designing a variety of transposon based methods for sequencing using BACs and Fosmids as well as long range PCR to map transposon inserts. These approaches can be taken together to form an integrated long range sequencing system that can eventually be automated. Progress with these approaches will be described.

Generation and Characterization of a Large Insert Chromosome 2 PAC Library

Denise Boehrer and Jeffrey C. Gingrich

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California, 94550.

As one part of the National Laboratory Gene Library Project we are generating large insert, ~50 kb to ~250 kb, bacterial clone libraries using the pCYPAC cloning vector developed at Livermore [1]. For a chromosome 2-specific PAC library, PAC clones are being generated from the hybrid cell line GM10826. DNA partially digested with *Mbo*I is size fractionated using PFGE and the DNA cloned into the *Bam*HI cloning site of the pCYPAC-2 vector. As with the bacteriophage P1 cloning system, sucrose is being used to select for recombinants. Ligation mixes containing clones over 80 kb average size are being plated and the chromosome 2 clones identified by colony hybridization using both human and hamster probes. To date, approximately 2,500 chromosome 2 clones, ~ 1X chromosome 2 coverage, has been arrayed into microtiter plates and we are continuing to array additional clones. The chromosome 2 representation of the library is being tested by PCR screening using STS primers distributed across the chromosome. Combining the chromosome 2 specific PAC library with the cosmid and fosmid library being generated here by the Gene Library Project (see abstract by Garnes et al.) in conjunction with STSs being developed here (see abstract by Gingrich et al.), and elsewhere, will allow for the construction of a high quality, high resolution physical map of chromosome 2.

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

[1] Ioannou et al. (1994) *Nature Genetics*, **6**, 84-89.

Construction of improved vectors for the preparation of PAC libraries

Eirik Frengen[1], Panayotis A. Ioannou[1,2], Chira Chen[1], Eugenia Pietrzak[1], Xiaoping Guan[1], Joe Capehart[1], Joel Jessee[3], Hans Lehrach[4] and Pieter J. de Jong[1]

[1]Department of Human Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263, [2]The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus, [3]BRL/Life technologies, Gaithersburg, MD, [4]Imperial Cancer Research Fund, London, UK.

Recently, we have developed new procedures for the cloning of large DNA fragments using a bacteriophage P1 derived vector, pCYPAC1 [1]. In view of the large sizes (up to 400 kb clones have been observed) and the single copy mode of maintenance, we have designated our clones as "PACs" for "P1-derived artificial chromosomes". A 3-fold redundant PAC library with average insert size of 120-130 kb has been prepared. The library is currently being characterized and copies of the library have been prepared and distributed to many genome centers to ensure wide-spread access to the library. Our experience with the current library is that chimerism is not a problem associated with this cloning approach, that the clones are very stable as compared to conventional cosmid clones and that the cloning procedures are not particularly recombinogenic.

Some of the anticipated uses of PAC clones include expression studies in mammalian cells and the preparation of transgenic animals for the analysis of genes carried on the PAC clone. To facilitate these studies, we are currently improving the PAC-vector as follows. The EBV oriP replication sequences are inserted to ensure stable episomal propagation when a PAC clone is transfected into a human cell for analysis of genes carried on the clone. An eukaryotic dominant-selectable marker gene (Blasticidin) under control of a "universal" promoter (β-actin promoter) is also inserted in order to enable selection for the desired mammalian clones. These improvements of the vector will be highly useful for expression cloning and selection. The PAC vector will also include sequences important to select for cre/loxP-based site-specific integration in mammalian chromosomes, genome targeting.

A conversion of YACs to PACs would be one strategy to obtain PAC-clones carrying large human genes. We have therefore constructed shuttle vectors that permits the conversion of linear YACs into circular PACs through *in vivo* homologous recombination procedures in yeast, for subsequent transfer to *E.coli*. The resulting PACs could then be purified from *E. coli* and transfected into a mammalian cell for further analysis of cloned genes.

Work was supported in part by a grant from the U.S. Department of Energy (ER61883, P.J. de Jong, P.I.).

[1] Ioannou, P.A., Amemiya, C.T., Garnes, J., Kroisel, P.M., Shizuya, H., Chen, C., Batzer, M.A. and de Jong, P.J. (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Gen.* **6**: 84-89.

A Panel Of Mouse/ Human Monochromosomal Hybrid Cell Lines Each Containing A Single Different Tagged Human Chromosome

Arbansjit K. Sandhu, G. Pal Kaur and Raghbir S. Athwal

Fels Institute for Cancer Research and Molecular Biology, Temple University School of Medicine
3420 N. Broad Street, Philadelphia, PA 19140

We are producing a panel of mouse/human monochromosomal hybrid cell lines each containing a single different gpt tagged human chromosome. Presence of the selectable marker on the human chromosome assures stable retention while cells are cultured in the medium containing mycophenolic acid (25ug/ml) and xanthine (70ug/ml, MX medium). This panel would provide a unique resource for gene mapping and gene isolation. In addition tagged human chromosomes can be selectively introduced into any cell type of interest for genetic analysis of complex phenotypes by complementation.

The experimental approach to produce these cell lines involves tagging of the chromosomes in normal diploid human cells with a retroviral vector. Since normal human cells have a limited life span, they are fused with mouse A9 cells to perpetuate the marked chromosomes. Tagged human chromosomes which are now present in mouse/human hybrid cells are transferred further to mouse and/or Chinese hamster cells by microcell fusion method. Monochromosomal hybrid cell lines thus recovered are analyzed by a battery of methods including G-11, chromosome painting and Alu-PCR Southern hybridization, to ascertain the identity and integrity of the transferred human chromosome. We have already produced and characterized monochromosomal hybrid cell lines for 19 different chromosomes. This partial panel is comprised of hybrid cell lines for chromosomes 1,2,3,4,5,6,7,8,9,10,11,12,13,14,16, 17,19,21 and X. In addition to the hybrids containing intact chromosomes, cell lines bearing chromosomal arms have also been recovered and characterized.

Construction of a 2 Mb YAC library in transgenic mice of human chromosome 21q22.2

Desmond J.Smith, Yiwen Zhu, Jan-Fang Cheng, Edward M. Rubin
Human Genome Center, Lawrence Berkeley Laboratory
1 Cyclotron Road, Berkeley , CA 94720

Libraries of the human genome have been made in bacteria, yeast and somatic cells. These libraries have been of use for the identification of genes based upon screens for sequence, expression and complementation of function *in vitro*. We are presently expanding this strategy by creating *in vivo* libraries of transpolygenic mice (mice with large transgene inserts likely to contain several genes). Genes within the introduced human DNA may be screened for through the assessment of altered phenotype in members of the library. We have constructed an initial library which focuses on the Down syndrome region of chromosome 21, a region syntenic to the segment of murine chromosome 16 where the mutation,*Weaver*, has been mapped. This murine mutation is characterized by cerebellar ataxia and hypotonia.

Using microinjection into fertilized mouse eggs, transpolygenic mice have been created containing each of one of the following overlapping YACs from 21q22.2: 230E8 (700kb), 141G6 (500kb), 152F7 (550kb) and 285E6 (400kb). The YAC DNA, now as part of the mouse genome, was mapped and in approximately 35% of the transpolygenic lines the full length YAC was present. Between two to five independent lines of mice containing an intact YAC was created for each of the 4 YACs from this region. The remainder possessed partial fragments of these YACs and these lines will be a useful as a fine structure genetic mapping resource. There is a gap in the contig between YACs 230E8 and 141G6. We have utilized the availability of overlapping P1s for this area to create multiple lines of transpolygenic mice containing 6 P1s which cover the gap. This was achieved by pooling the DNA of the P1 phage together before microinjection.

Together, the transpolygenic animals harbor about 2 Mb of human DNA from the Down syndrome region. Individual members of this *in vivo* library are being assessed for the phenotypic features associated with trisomy 21q22 (heart disease, hematopoietic abnormalities, dysmorphic facies, developmental delays etc.). In addition, since the DNA in the transpolygenic mice encompasses the entire genetic interval to which the *Weaver* gene has been mapped we are in the processes of mapping / cloning this gene by mating members of the *in vivo* library with Weaver homozygous animals and assessing the offspring for *in vivo* complementation of the Weaver phenotype.

Construction of DNA Libraries From Flow Sorted Human Chromosomes

Larry L. Deaven, Mary K. McCormick, Deborah L. Grady, Donna L. Robinson, Judy M. Buckingham, Nancy C. Brown, Evelyn W. Campbell, Mary L. Campbell, John J. Fawcett, Phil Jewett, Jonathan L. Longmire, Adelmo Martinez, Linda J. Meincke, Pat L. Schor, and Robert K. Moyzis, Center for Human Genome Studies and Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545

We have constructed a series of DNA libraries from flow sorted chromosomes. Small insert, complete digest libraries cloned into the EcoRI insertion site of Charon 21A are available from the American Type Culture Collection, Rockville, MD. Partial digest libraries cloned into cosmid (sCos1) or phage (Charon 40) vectors have been constructed for chromosomes 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 20, X and Y. Purity estimates by *in situ* analysis of sorted chromosomes, flow karyotype analysis, and plaque or colony hybridization indicate that most of these libraries are 90-95% pure. Additional cosmid library constructions, 5-10X arrays of libraries into microtiter plates, and high density membrane arrays of libraries are in progress.

Recently we have completed YAC libraries for chromosomes 5, 9, 16, and 21. These libraries are made from complete DNA digests using the rare cutters ClaI, SacII, EagI, or NotI/NheI. The average insert size is ~200 kb, and chimera frequencies are low (1-10%).

Libraries have also been constructed using M13 or bluescript vectors (chromosomes 5, 7, 17) to generate STS markers for the selection of chromosome-specific inserts from total genomic YAC libraries.

Because of the advantages of insert size and stability associated with BAC and PAC cloning systems, we are currently attempting to adapt pBAC108L and pCYPAC1 vectors for use with flow-sorted chromosomal DNA. This work was supported by the U.S. Department of Energy under contract W-7405-ENG-36.

Chromosome Specific Cosmid and Lambda Libraries from Flow-Sorted Chromosome

Jeffrey A. Garnes, Benjamin S. Wong, Wanda Johnson, Jerry Eveleth, Anne Bergmann, Richard Langlois, Anthony V. Carrano, and Jeffrey C. Gingrich

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

The National Laboratory Gene Library Project is responsible for the construction of human chromosome-specific cloned DNA libraries. The focus of phase II has been to construct partial MboI digest lambda (~10 kb) and cosmid (~ 40 kb) libraries for the entire human genome. The source of DNA for all of the libraries has been human chromosomes sorted from human/rodent hybrid somatic cell lines. Forty-five thousand cosmids cloned in the Lawrist 16 vector have been arrayed for chromosome 2 in addition to 23,000 fosmid clones. Approximately 63% of the clones in both libraries are derived from chromosome 2 and represent 7.5 chromosome equivalents. Approximately fifty thousand cosmids are being arrayed for chromosome 1 with additional DNA available for augmenting the library. The sort purity of this library as assessed by colony hybridization of a representative number of cosmids with both human and hamster DNA probes indicates that 61% of the sorted DNA is human. We are in the final stages of constructing chromosome-specific lambda libraries for chromosomes 1, 2 and 3. All of the lambda libraries will be deposited in the ATCC repository for distribution to the scientific community. This cooperative effort between Lawrence Livermore and Los Alamos National Laboratories has resulted in the production of large insert libraries for all of the human chromosomes. These libraries will continue to be invaluable resources in human genome mapping and analysis.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

Efficient Pooling of YAC Libraries

David J. Balding[1], David C. Bruce, William J. Bruno, Norman A. Doggett, Emanuel Knill, David C. Torney, and Clive C. Whittaker, Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, USA; [1]Queen Mary and Westfield College, University of London, London, UK.

We have developed general methods for designing and implementing pooling experiments for YAC libraries of all sizes and coverages, primarily for repeated STS screenings. Conventional pooling uses variations of multi-tiered row-and-column pools. Although implementation and interpretation of row-and-column pools is straightforward, we propose more efficient designs.

We have implemented combinatorial algorithms which can optimize pooling experiments, given the costs of constructing pools, testing pools and confirming candidate-positive clones. Random k-sets designs are particularly effective in minimizing the total number of pools. In random k-sets designs, every clone occurs in k of the pools, and the pool assays are performed in one pass. We sho how to evaluate the expected number of *definite-positive* clones—clones which must be positive based upon the pool assays—given the number of positive clones is binomially distributed. The parameter k can be chosen to maximize this expectation. For example, if a 33,000 clone library with tenfold coverage were pooled using a random ten-sets design with 170 pools, then approximately half of the positive clones would be definite-positive. If one used a random ten-sets design on 253 pools, then approximately 95 per cent of the positive clones would be definite-positive. Random k-sets designs have no *a priori* constraints on the number of pools or clones; they are easily generated on a computer. Furthermore, they can be designed to correct for false-negative and other pool-assay errors.

In addition, we can often get improved performance if there is an upper bound on the number of pools in which any two clones coincide. When each clone occurs in k pools, such designs are referred to as k-sets packings. We compare the performance of our designs and other implemented designs.

The implementation of efficient pooling strategies requires the use of robots. We are using a commercially available system (Packard multiPROBE 204) to pool two $\sim$ 1300 clone YAC libraries using optimized random four-sets designs. Using one robot, one of these libraries can be pooled into 47 pools in about 11 hours, with every clone occurring in four pools. Our first STS assay yielded the correct PCR product in eight of ten pools known to contain a positive clone. We have developed effective techniques for ranking the candidate-positive clones when the rates of false-negative and false-positive pool assays are appreciable; this work is described in detail in a separate abstract. Here we note that random k-set designs can be efficient even with typical error rates.

Generation of Transgenic Mice Carrying Centromeric Sequences

Ana V. Perez-Castro, Julie Wilson, Michael Altherr and Robert K. Moyzis.

Life Sciences Division, LS-2, MS-M880, and Center for Human Genome Studies; Los Alamos National Laboratory, Los Alamos NM 87545.

The centromere of mammalian and other complex eukaryotic chromosomes is characterized by the presence of varying amounts of non-transcribed repetitive DNA sequence.

The human satellite repeat $(GGAAT)_n$ is similar to the central region of the yeast centromere sequence CDE III. We have done PCR experiments which show that this repeat is also present in mice, flies and sea urchin, suggesting evolutionary conservation. It has been proposed that this specific repeat $(GGAAT)_n$, could be a component of the functional centromere (1).

In order to test this hypothesis in vivo, and to see any possible effects that this sequence might have on development and gene expression, we decided to introduce a 158 bp DNA fragment containing the human $(GGAAT)_n$ repeat in the mouse genome, using pronuclear injection.

We have generated 7 independent transgenic mouse lines carrying different number of copies of the $(GGAAT)_n$ repeat. The mice are apparently normal and we are currently mating them to see if their breeding capability has been affected, and if the second generation shows any abnormality. We will present our results and discuss the significance of our findings. We are also in the process of injecting other interesting centromeric repeats and evaluating their effects in the development of fertilized mouse oocytes.

(1) Grady, L., D. et al. (1992) Highly conserved repetitive DNA sequences are present at human centromeres. Proc. Natl. Acad. Sci. USA, **89**, 1695-1699.

COMPARATIVE MAPPING OF A CONSERVED ZINC-FINGER GENE CLUSTER IN MAN AND MOUSE

Mark Shannon[1,3], Michael L. Mucenski[1], Linda Ashworth[2], and Lisa Stubbs[1]

[1]Biology Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8080; [2]Human Genome Center, Lawrence Livermore National Laboratory; [3]Corresponding author.

A comprehensive program aimed at man-mouse comparative mapping of human chromosome 19 is being conducted as a collaboration between the Biology Division at the Oak Ridge National Laboratory and the Human Genome Center at the Lawrence Livermore National Laboratory. The broad objectives of these studies are to establish the structural relationships between the genes in regions of man-mouse homology and to use the mouse as a model system for determination of the functions of those resident genes. In this study, we extend comparative mapping to the detailed analysis of a 200-300kb region located near the human *XRCC1* gene in 19q13.2, that has been shown to carry a cluster of zinc-finger (ZNF) genes. We report the isolation of a genomic fragment from the region containing DNA that is not related to the ZNF-consensus sequence, but that is present in multiple copies within the gene cluster and conserved in a variety of vertebrate species. Sequence analysis has indicated that the genomic fragment encodes the A domain of the Kruppel associated box (KRAB), which is estimated to be present in one-third of the ZNF genes of the Kruppel (C2H2) type. Interestingly, this conserved human fragment detects a small number of independent loci on Southern blots of mouse genomic DNA, and interspecific backcross (IB) analysis has demonstrated that most or all of these cross-hybridizing sequences are located in the related *Xrcc1* region of mouse chromosome 7.

These results suggest that the human KRAB fragment may preferentially recognize genes within the homologous murine KRAB-containing ZNF gene cluster. This suggestion is supported by the fact that the related murine sequences are restricted to single, independent 225 kb fragments in *Mlu*I- or *Eag*I-digested genomic DNA. These data serve to define a maximum size for the mouse ZNF cluster which is comparable with the estimated length of the human 19q13.2 ZNF gene region. Using the human KRAB sequence as a probe, we have identified mouse cosmids spanning the murine ZNF gene region, as well as cDNA clones that may represent some of the estimated 10-20 independent ZNF genes in this region. Functional characterization of the murine ZNF genes, in conjunction with restriction mapping and sequencing studies of human chromosome 19q13.2 that are currently underway at LLNL, will eventually allow complete structural and functional analysis of this region of man-mouse homology.

This work was supported by USDOE under contract DE-AC0584OR21400 with Martin Marietta Energy Systems,Inc., and contract W-7405-ENG-48 with the Lawrence Livermore National Laboratory. M.S. is supported by a DOE Human Genome Distinguished Postdoctoral Fellowship.

DETAILED MAN-MOUSE COMPARATIVE MAPS OF HUMAN CHROMOSOME 19

[1,6]Johannah Doyle, [1]Estela Generoso, [1]William Dunn, [1]Beverly Stanford, [1]Ethan Carver, [2]Eugene Rinchik, [3]Susan Watt, [4]Wolfgang Zimmermann, [5]Linda Ashworth, [5]Greg Lennon, [5]Anne Olsen, [5]Susan Tsujimoto, [5]Harvey Mohrenweiser, [5]Brigitte Brandriff and [1]Lisa Stubbs

[1]Mammalian Genetics and Development Section, Biology Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8077; [2]Sarah Lawrence College, 1 Mead Way, Bronxville, NY 10708; [3]MRC Molecular Haematology Unit, John Radcliffe Hospital, Oxford OX3 9DU, UK; Dept. Immunobiology, University of Freiburg, Freiburg, Germany; [5]Human Genome Center, Lawrence Livermore National Laboratory, P.O. Box 808, L-452, Livermore, CA 94551; [6]Corresponding Author.

In order to set the stage for the exploitation of the mouse in the functional analysis of human chromosome 19, we have constructed series of detailed maps of related regions of the mouse genome. We have initiated these efforts by systematically localizing more than fifty conserved human chromosome 19 markers on the genetic map of the mouse, most representing genes that serve as regional anchors on LLNL's highly detailed physical map. Because both maps are centered upon the same series of markers, and because all murine markers have been mapped on a single interspecies backcross system, these data provide a measure of accuracy and internal consistency that is lacking from most published consensus maps. The large number and high density of markers that have now been genetically assigned to conserved regions has provided the foundation for initiation of physical maps in regions of special interest, allowing synteny relationships to be examined on the detailed molecular level.

While homologs of 19p genes are split into relatively small homology segments on several different mouse chromosomes (8, 9, 10 and 17), the entire length of 19q is represented within a 25 cM region of proximal mouse chromosome 7. Our results show that related genes are organized in a remarkably similar fashion along the lengths of these large, homologous regions, although with a few notable exceptions. A 4-5 Mb gene-rich region of human chromosome 19q13.3-13.4, for example, appears to have been 'transposed' from its relatively telomeric position to a murine region that is most closely related to the centromeric portion of the human chromosome; mouse equivalents of the clustered human PSG (pregnancy-specific glycoprotein) genes of human 19q13.2 are also located in completely different genomic environments in mouse and man. Both the similarities and the differences provide important insights regarding the evolution of this gene-rich human chromosome, and provide a powerful basis for the application of ES cell-based and molecular methods for assessing the functions of resident genes.

This work was supported by USDOE under contract DE-AC0584OR21400 with Martin Marietta Energy Systems,Inc., and contract W-7405-ENG-48 with the Lawrence Livermore National Laboratory.

DEVELOPMENT AND TESTING OF A METHOD FOR THE PURIFICATION OF EVOLUTIONARILY CONSERVED SEQUENCES FROM CLONED HUMAN DNA

[1,3]Lisa Stubbs, [1]Karen Glantz, and [2]Elbert Branscomb

[1]Mammalian Genetics and Development Section, Biology Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-8077; [2]Human Genome Center, Lawrence Livermore National Laboratory, P.O. Box 808, L-452, Livermore, CA 94551; [3]Corresponding author.

As a corollary to our groups' efforts to define genomic homologies between human chromosome 19 and related murine regions, we are interested in devising means through which these genomic relationships can be exploited for the identification of new human genes. This project has been designed to capitalize upon LLNL's collection of contiguous cosmid, YAC and other clones which span the length of human chromosome 19. A large number of cloned chromosome 19 genes and DNA markers have already been localized to specific clone contigs, but a vastly larger number of new genes remains undiscovered. In order to address the pressing problem of large-scale gene identification, we have developed a novel method to permit rapid and efficient identification of genes and other functionally-significant DNA sequences from cloned human DNA.

Our approach relies upon the fact that DNA sequences with important biological functions are most likely to be conserved throughout evolution. Because of this fact, the genes of mouse and man are, on the whole, very similar in DNA sequence; non-gene regions, by contrast, will generally vary greatly between two such highly divergent species. Our goal is thus to selectively clone the sequences which are most similar between mouse and human DNA. The approach focuses upon using conserved sequences in murine YAC or P1-based clones to "trap" similar sequences from human cosmids spanning homologous genomic regions, and is loosely based upon cDNA selection methods (such as those developed by M. Lovett and colleagues). To test and troubleshoot our 'conserved element mapping' strategy, we have worked with a pair of homologous cosmids carrying the mouse and human XRCC1 gene sequences. Since both mouse and human cosmids have been completely sequenced (by J. Lamerdin and colleagues, LLNL), the exact size, position, and degree of similarity of conserved DNA sequences are known. Initial results with this cosmid model system have demonstrated that XRCC1 exons, which occupy less than 5% of cosmid sequences, can be efficiently purified using this approach; significant contamination is derived only from Alu repeats, which are most likely binding to related B1 repeats of mouse. Experiments now in progress are aimed at improving the purity and yield of our products in the cosmid model system, and at expanding the length of scanned regions to include large human contigs and mouse P1-based contigs and YAC clones. Ultimately, we intend to apply these methods to the identification of genes and conserved regulatory elements along the length of human chromosome 19.

This work was supported by USDOE under contract DE-AC0584OR21400 with Martin Marietta Energy Systems,Inc., and contract W-7405-ENG-48 with the Lawrence Livermore National Laboratory.

FUNCTIONAL ANALYSIS OF GENES USING THE MOUSE AS A MODEL SYSTEM

Peter R. Hoyt[1], Theresa A. Banks[2*], Christopher Bartholomew[3], Patrick J. Blair[1], Virginia L. Godfrey[4], Marilyn K. Kerley[4], William H. Lee[4], Brigid L. M. Hogan[5], Barry T. Rouse[2], James N. Ihle[6], S.Steven Potter[7], and Michael L. Mucenski[4,8]

[1]University of Tennessee-Oak Ridge Graduate School of Biomedical Sciences, PO Box 2009, Oak Ridge, Tennessee 37831-8080 USA, , [2]Department of Microbiology, University of Tennessee, Knoxville, TN 37996 USA, [3]Beatson Institute for Cancer Research, Glasgow G611BD Scotland, [4]Biology Division, Oak Ridge National Laboratory, PO Box 2009, Oak Ridge, TN 37831-8080 USA , [5]Department of Cell Biology, Vanderbilt University Medical School, Nashville, TN 37232-217 USA, [6]Department of Biochemistry, St. Jude Children's Research Hospital, Memphis, TN 38105 USA, [7]Institute of Developmental Research, Children's Hospital Research Foundation, Cincinnati, Ohio 45229 USA, [8]Corresponding Author.

It has been shown that murine and human genomes are considerably conserved at the nucleic acid and gene linkage levels. These similarities make the mouse an excellent model for the physical and functional mapping of the human genome. As the murine homologues of interesting human genes are identified, we will use targeted mutagenesis technology of embryonic stem (ES) cells to generate mice so that the biological function of these genes may be determined. In the course of these studies, murine models for human diseases will be emphasized.

Our targeted mutagenesis laboratory has already generated null mutants for two genes which are evolutionarily conserved between mice and humans, the proto-oncogene *Evi-1*, and the cytokine tumor necrosis factor-beta (TNF-β), which is also referred to as lymphotoxin-alpha (LT-α). Detailed characterization of the homozygous mutant animals will be presented. The *Evi-1* gene appears to play a critical role in the development of multiple organ systems while LT-α functions in the development of secondary lymphoid organs in the mouse. These results show the utility of targeted mutagenesis technology which will enable us to determine the biological function of the murine homologues of human genes that may be involved in specific disease states. Future work will involve the analysis of genes identified through the physical mapping of the mouse genome within well defined mouse-human homology regions, a collaborative effort between the Mammalian Genetics and Development Section of the Biology Division of the Oak Ridge National Laboratory and the Human Genome Center of the Lawrence Livermore National Laboratory.

This work is supported by the USDOE under contract DE-AC0584OR21400 with Martin Marietta Energy Systems, Inc.

*Present Address: Viagene, Inc., San Diego, CA 92121-1204

Mouse Genome Studies Aimed at Deciphering Gene Function

Woychik, R.P., Stubbs, L., Mucenski, M.L., Moyer, J., Kwon, H., Richards, W.G., Yoder., B. and Wilkinson, J.E. Biology Division, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, TN 37831-807,

We have initiated a comprehensive program aimed at man-mouse comparative genome studies in the Biology Division at the Oak Ridge National Laboratory. As one component of this effort, transgenic and targeted mutagenesis strategies are being developed in mice to begin to explore the function of individual genes derived from the human and mouse physical mapping initiatives. To illustrate the nature of this research, we have developed a mutation in the mouse which corresponds to a gene in humans mapping to Chromosome 13 at position 13q12.1. This gene is over 100-kb in length and is comprised of 25 individual exons, most of which are less than 100 bp in length. The gene gives rise to a 2.8 kb mRNA in humans and is expressed with a broad tissue distribution. In the mouse, the gene produces a 3.2 kb mRNA which is also expressed with a wide tissue distribution. The difference in size between the mouse and human mRNA's is due to a deletion within the 3' untranslated region of the human transcript. Computer analysis revealed that the putative coding region has the potential to produce an 825 amino acid protein in humans that contains a repeated motif called the tetratrico repeat (TPR). Inactivation of this gene in the mouse was most revealing in that the animals developed polycystic kidney disease and a specific liver lesion among other defects. The disease in the mouse closely resembles human autosomal recessive polycystic kidney disease (ARPKD). This approach is allowing us to correlate specific genes with functions and disease conditions in humans. The submitted manuscript has been authored by a contractor of the U.S. Government under contract No. DE-AC05-84OR21400. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

Sequence-Directed DNA Manipulation for Enhanced Mapping and Sequencing

Charles R. Cantor[1], Natalia Broude, Takeshi Sano, Kazuhiro Tsukamoto, Marek Przetakiewicz, A. Burak. Dalan, Abha Chandra, Rhonda Harrison, Nikolay Bukanov, Ronald Yaar, Demetri Moustakas, Przemekyslaw Szafranski, and Cassandra L. Smith.

Center for Advanced Biotechnology and Department of Pharmacology, Boston University, College of Engineering, Boston MA 02215. [1]Corresponding author.

Most traditional DNA analysis is done based on fractionation of DNA by length. We have, instead, begun to explore the use of DNA sequences as capture and detection methods to expedite a number of procedures in genome analysis. In enzyme-enhanced sequencing by hybridization (SBH), a partially duplex DNA probe is used to capture the five complementary bases at the end of a single-stranded DNA target. DNA ligase can be used to covalently seal the target to the probe and ensure faithful detection of the target sequence. Subsequent DNA polymerase I extension of the probe, now serving as a primer, along the target, now serving as a template, allows additional bases of DNA sequence to be read and also further ensures the fidelity of the original detection. Pilot studies show that this procedure can, indeed, reveal the sequence of a DNA target. By incorporating length information about the target, many of the branch point ambiguities that occur in ordinary SBH can be eliminated.

While conceived as a direct DNA sequencing procedure, the SBH format we use may be even more useful as a device for the rapid preparation of DNA samples for fast serial methods like capillary arrays or mass spectrometry. For example, an array of only 1024 probes could capture and then generate sequence ladders from any arbitrary DNA sequence. Some of the ways in which this sort of capture device might be used in DNA sequencing will be illustrated. We have also begun to explore the use of sequence-specific DNA capture as a method for looking at specific DNA sequence differences. This could eventually lead to faster methods for genetic mapping.For effective sequence-specific capture it is vital that non-specific binding of target DNA to the substrate used for the probe array be minimized.We have been exploring a number of different surfaces and modes of attachment to try to achieve the very low backgrounds needed to handle large numbers of samples simultaneously. Some of the probe designs we are using would allow direct production of probe arrays by replication of a master array and transfer to a new surface.

Human-specific PCR is widely used in DNA mapping. The most common version of this technique is inter-*Alu* PCR. Two major limitations in this approach are the relatively small fraction of a human target sample that is amplified and the poor results generally obtained from regions of the genome in which *Alu* sequences are relatively sparse. Many of these problems would be eliminated if a relatively efficient single-sided *Alu* PCR procedure were used, instead of inter-*Alu* PCR. We have developed such a procedure based on assymetric PCR between an *Alu* primer and a ligated splint. The new procedure appears to give much better coverage when human YAC DNA is amplified. We have devised and tested several other new amplification methods including an efficient procedure for relatively uniform whole genome amplification and a non-enzymatic spatially-resolved DNA amplification method that might assist the analysis of sample arrays or chromosome spreads.

Marker development for the EPM1 region of human chromosome 21, q22.3. J.A. Warrington[1], K. O'Connor[1], S. Hebert[1], M. Harris[1], R. Goold[1], A.E. Lehesjoki[2], A. de la Chapelle[2], N. Stone[1], R. Myers[1] and D.R. Cox.[1]
[1]Stanford University, Stanford, CA., [2]University of Helsinki, Finland.

New STSs have been developed for a 0.9 Mb region of chromosome 21 that is not represented in existing YAC libraries using an efficient method that is generally applicable to any region of the genome. The region, 21q22.3 is of particular interest because the gene for progressive myoclonic epilepsy of the Unverricht-Lundborg type (EPM1) maps to this region. Until recently there were only three probes for the 1.3 Mb surrounding the EPM1 gene (D21S141, LJ112, LB2T). This very limited number of probes is problematic for obtaining clone coverage and for confirming map position of newly developed markers in the EPM1 region. To develop new markers, a somatic cell hybrid containing chromosome 21 as its only human complement (GMO8854) was digested with *NOT*1 and hybridized with D21S141. The fragment hybridizing with D21S141 was excised, amplified by *Alu* PCR and the amplification products were cloned and sequenced. Of the fifteen clones sequenced, four were duplicates and one consisted entirely of repeat sequence. STSs were developed for the remaining ten unique clones. To determine the map position of the new STSs, quantitative PCR was used in conjunction with whole genome radiation hybrid mapping. Quantitative PCR confirmed that the STSs mapped to appropriate sized PFGE fragments and whole genome RH mapping showed that the markers were linked and gave order and distance information. Three of the new STSs are in the EPM1 region, providing additional starting points for obtaining clone coverage and gene isolation. This combination of techniques for developing markers and confirming map position is an effective approach for obtaining probes and has general applicability for regions of the genome not represented in YAC or cosmid libraries.

Not funded by DOE.

Current status of the integrated physical map of chromosome 16.

D.F. Callen,[1] S. Apostolou[1], S. Lane,[1] S.A. Whitmore,[1] N.A. Doggett,[2] R.I. Richards,[1] J.C. Mulley,[1] G.R. Sutherland[1]

[1]Centre for Medical Genetics Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, 72 King William Road, North Adelaide, South Australia 5006. [2]Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

The integrated map of chromosome 16 is based upon our high resolution somatic cell hybrid panel for chromosome 16. Naturally occurring rearranged chromosome 16s were isolated in mouse/human somatic cell hybrids. The panel now consists of 78 hybrid cell lines representing 90 independently ascertained breakpoints on the chromosome. With the four fragile sites these divide the 85 Mb of euchromatin into 91 bins of average size 940 kb. There are four breakpoints in the 15 Mb of centromeric heterochromatin. Cloned DNA markers or STSs have been identified in 73 bins. To correlate the physical and genetic maps of chromosome 16 all available polymorphic markers typed on the CEPH panel of families were placed on the physical map. Gaps were targetted in the linkage map and closed by isolation and genotyping of additional PCR-based markers on these families. This resulted in the construction of a PCR-based linkage map which incorporated our 49 markers and 22 markers characterised by other laboratories, including those of the Weissenbach group. We coordinated the production of the CEPH consensus linkage map of this chromosome incorporating all PCR and non-PCR markers. This linkage map has an average marker distance of 2.8 cM and the largest gap is 11.2 cM. This already exceeds the revised 5-year Research Goal of the US Human Genome Project to have a 2-5 cM genetic map completed by 1995. All available data on markers and cloned genes from published sources, databases and unpublished data obtained as a result of our numerous collaborations have been placed on the integrated map. The map now includes 73 genes and expressed sequences, 132 microsatellite markers and 111 other DNA markers. Expressed sequences on chromosome 16 have been isolated from an hn-cDNA library of a hybrid containing only chromosome 16. Development of an integrated mega-YAC, sorted chromosome 16 YAC and cosmid contig map of the chromosome by LANL with our collaboration is nearing completion.

This work was funded by the DOE Genome Program Grant DE-FG02-89ER60863.

Progress with a strategy for cosmid closure of band 16q24.3.

D.F. Callen,[1] S. Apostolou[1], S. Lane,[1] S.A. Whitmore,[1] N.A. Doggett,[2] R.I. Richards,[1] J.C. Mulley,[1] G.R. Sutherland[1]

[1]Centre for Medical Genetics Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, 72 King William Road, North Adelaide, South Australia 5006. [2]Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

A strategy for contig closure, currently being implemented for band q24.3 of chromosome, involves the following steps: 1. saturation of the region with cosmids identified by probing the high density chromosome 16 specific cosmid arrays with all cloned sequences known to map to this region. 2. Identification of contigs to which any of the cosmids identified belong. 3. Verification of contigs by STS development and mapping. 4. Contig extension by development of STSs from the ends of contigs and from the 3' and 5' ends of cDNAs and the re-screening of arrays with these. We have targetted 16q24.3 since it is a gene rich region of considerable biological interest since it possesses a major gene involved in breast cancer and other tumours. *Alu*-PCR products from two somatic cell hybrids in which the only human chromosome material was the distal portion of 16q, were used to probe high density grids of fingerprinted cosmids. Each cosmid isolated was mapped back to the region by Southern analysis of hybrid panel DNA. 49 cosmids, many of which are members of unmapped contigs, have now been located within 16q24. STSs from the ends of selected cosmids are being isolated by inverse PCR and being used to screen for additional cosmids, verifying contig structure and screening YACs. It is estimated that 1.8 Mb or 45% of this region is now in mapped cosmids and cosmid contigs. Direct selection has been used to isolate pools of cDNAs from cosmids. These cDNA pools are being used to screen the gridded cDNA filters of LLNL. We are aiming to produce a verified minimum tiling path of cosmids over the region as a set of reagents in preparation for large scale sequencing of this chromosome.

This work was funded by the DOE Genome Program Grant DE-FG02-89ER60863.

MAPPING SEQUENCE TAGGED SITES TO DISTINCT SEGMENTS OF CHROMOSOME 16.

The Los Alamos Genomics Group[1] and D.F. Callen[2]. Presented by M.R. Altherr[1].

[1]Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM, USA; [2]Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, Adelaide, SA, Australia.

We have used a panel of 70 human::murine somatic cell hybrids containing discrete segments of human chromosome 16 to physically localize 310 sequence tagged sites (STSs) derived from an extensive collection of fingerprinted chromosome 16 specific cosmids. The distribution of these sequences along the chromosome appears to be random. Fewer than 10% of the sequences failed to unambiguously identify chromosome 16 locations. Along with the 229 PCR based probes currently listed in GDB this now represents an average spacing of one STS for every 176 kbp of chromosome 16. Our localized STSs provide physical anchors for both cosmid and YAC contigs. Additional STSs will be generated as a result of our chromosome 16 expressed sequence studies using clones resulting from exon amplification. As a consequence, any additional STSs added to the map by us will also represent potential coding sequences. We anticipate that these sequences will provide the starting point for a chromosome 16 transcription map.

An Integrated Map of Human Chromosome 16.

N.A. Doggett[1], D.F. Callen[2], M.R. Altherr[1], L.A. Duesing[1], J.G.Tesmer[1], L.J. Meincke[1], D.C. Bruce[1], A.A. Ford[1], D.C. Torney[1], R.D. Sutherland[1], M.G. Lowenstein[1], M.O. Mundt[1], W.J. Bruno[1], E.H., Knill[1], R.I. Richards[2], G.R. Sutherland[2], L.L. Deaven[1], and R.K. Moyzis[1]

[1]Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM, [2]Cytogenetics and Molecular Genetics, Woman's & Children's Hospital, Adelaide, SA, Australia

We have constructed an integrated physical/genetic/cytogenetic map of human chromosome 16. The framework for constructing this map is a high resolution cytogenetic breakpoint map derived from 78 mouse/human somatic cell hybrids and 4 fragile sites which divide chromosome 16 into 90 intervals of average size 1 Mb. The physical map consists of both a low resolution YAC contig map and a high resolution cosmid contig map. The low resolution YAC contig map is comprised of 515 CEPH MegaYACs, and 220 flow sorted 16-specific YACs that are localized to and ordered within the breakpoint intervals with 320 STSs. This YAC map provides nearly complete coverage of the euchromatic arms of the 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. To date, 300 of these contigs containing greater than 1600 cosmids have been merged with the integrated physical/genetic/cytogenetic map and thus regionally localized along the chromosome.

A highly informative microsatellite-based genetic map (developed at the Adelaide Children's Hospital) consisting of 78 PCR typable markers and having a 2.6 cM median (3.2 average) distance between markers is tightly integrated with the physical map because most of these GT/CA markers were developed from localized cosmids or screened against the YAC map, and by regional localization of all markers to the cytogenetic breakpoint map.

To integrate a transcription map of chromosome 16 with the existing physical/genetic/cytogenetic map, we have subjected 14,000 chromosome 16 cosmids to exon amplification, isolated 5000 exon clones and sequenced 500 of these clones. Over 500 genes, anonymous DNA markers and microsatellite repeats have also been incorporated as part of an ongoing effort to integrate all available GDB loci with the 16 map. This integrated map is facilitating the cloning of various disease genes and fragile sites on chromosome 16. Supported by the US DOE (W-7405-ENG-36).

Analysis of External YAC Data for Incorporation into the Chromosome 16 Map

R.D. Sutherland, R.M. Pecherer, L.A. Duesing, and N.A. Doggett

Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545.

Using our own chromosome 16 derived STS's and YAC pooling strategies, we have generated an STS-content map for 500+ mostly CEPH MegaYACs. The STS-content data that makes up this map is stored in our chromosome 16 relational database. Now that other laboratories have made their YAC, overlap-pair, STS, and primer data publicly available, we have incorporated this data into our database. We have taken advantage of the relational database structure to design complex queries to interrelate LANL and external data to expand coverage and close gaps in the 16 map. The starting point for STS-content mapping was a cosmid contig map covering 60% of chromosome 16 with 550 cosmid contigs (islands). STS's were developed from the largest of these contigs and used for somatic cell hybrid localization and PCR-based screening of CEPH MegaYACs. Single YAC bridges between islands could be produced directly from the STS content of YACs. Thus, a single YAC bridge exists when STS's from two different islands hit the same YAC. Using external YAC overlap-pair data (generated primarily from the fingerprinting of YAC clones at Genethon), we could evaluate double and triple YAC bridges between islands. A double YAC bridge occurred when two STS's hit separate YACs but these YACs directly overlapped. A triple YAC bridge occurred when two STS's hit separate YACs, each of which overlapped the same third YAC. This process could continue on but we felt farther reaches were too tenuous. These new bridges increased chromosome coverage and contributed to gap closure. Another benefit of using YAC overlap data is that this provided additional information for the ordering of STS's and YACs within breakpoint intervals of the somatic cell hybrid map. External STS-YAC data was easily integrated with our map because a) many of the external STS markers were already localized to the somatic cell hybrid map, and b) our cosmid-derived STS's and external STS markers hit common YACs. The STS data also identified candidate chimeric YACs i.e., hit by both chromosome 16 STS's and STS markers from other chromosomes. Alu-PCR data was used in conjunction with the STS data (eliminating YACs with STS hits on other chromosomes) to create a list of candidate YACs for further evaluation for their placement on 16. Supported by the US DOE (W-7405-ENG-36).

Construction of an STS-Content YAC Map of Human Chromosome 16.

N.A. Doggett[1], L.A. Duesing[1], J.G.Tesmer[1], L.J. Meincke[1], M.R. Altherr[1], A.A. Ford[1], D.C. Bruce[1], D.C. Torney[1], E.H., Knill[1], W.J. Bruno[1], R.D. Sutherland[1], M.G. Lowenstein[1], M.O. Mundt[1], D.F. Callen[2], G.R. Sutherland[2], L.L. Deaven[1], and R.K. Moyzis[1]

[1]Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM, [2]Cytogenetics and Molecular Genetics, Woman's & Children's Hospital, Adelaide, SA, Australia

An STS-content YAC map has been constructed for human chromosome 16. The map consists of 515 CEPH MegaYACs, 220 flow sorted 16-specific YACs and 320 STSs, and provides nearly complete coverage of the euchromotic arms of this chromosome. STSs, which were derived largely from the end clones of cosmid contigs were first localized to a high resolution cytogenetic breakpoint map derived from 78 mouse/human somatic cell hybrids and 4 fragile sites which divide chromosome 16 into 90 intervals of average size 1 Mb. In the later stages of mapping STSs were derived from the ends of YAC clones which bordered gaps or from genes and markers that were known to lie near gap regions. The CEPH MegaYAC library was pooled into 95 top level pools based on a plate shuffling pooling scheme, and 672 secondary pools based on a plate shuffling and row column pooling scheme. An individual positive YAC is identified after screening the primary pool and a minimum of 14 secondary pools. For the flow-sorted 16 specific YAC library, a four-sets packing design was implemented which involved screening 47 pools in a single level to identify positive clones. The STS content map has been assempled in a spreadsheet program with YACs and STS oriented along the x and y axis respectively. In this manner STSs (rows) and YACs (columns) could easily be moved as newer data permitted refined ordering. This hand construction proved superior than automated methods such as **segmap** because available algorithms did not easily support the known ordering of STSs which had been localized to the breakpoint map. In the current map there is evidence for deletions occuring in 108 MegaYACs or ~ 21% of the MegaYACs. Due to the redundancy of coverage in the YAC map however, coverage exists for nearly all of these deleted regions in at least one YAC.

Supported by the US DOE (W-7405-ENG-36).

Development of a Multi Level Resolution Map of the p arm of Chromosome 19 by Integration of FISH, YAC, Cosmid Contig, and Restriction Map Data

Susan A. Allen, Anca M. Georgescu, Maria T. de Jesus, Susan M. G. Hoffman, Gregory G. Lennon, Anthony V. Carrano and Anne S. Olsen

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

A comprehensive map of the p arm of chromosome 19 is being constructed integrating data from various methods. Each form of data contributes to the integrated map at different levels of resolution. An initial framework map was developed by fluorescence *in situ* hybridization (FISH) of 80 cosmid markers for which order and relative distance was determined along the 20 Mb span of 19p (see poster by Brandriff et al). Most of the ordered cosmid markers are members of cosmid contigs assembled by the fingerprinting procedure that constitutes the foundation of the chromosome 19 map. Cosmid contigs in close proximity were often merged by directed walking in cosmid, BAC, PAC or P1 libraries. In addition, cosmid walking often incorporated previously unmapped contigs into the map.

Remaining gaps in the cosmid contig map are initially being spanned by YACs isolated by STS screening and Alu-PCR hybridization. At present, a total of 106 STSs, including 26 polymorphic markers, have been developed from ordered cosmids or derived from outside sources. These STSs have been subsequently integrated into the map (see poster by Tsujimoto et al). Over 90 YACs have been added to the ordered map, thereby reducing the 80 ordered cosmid contigs to less than 36 YAC-cosmid islands. These islands span an estimated 15 Mb or about 75% of the p arm. Alu-PCR products from YACs have been hybridized to cosmid colony filters to incorporate additional cosmid contigs into the map and to close gaps between YACs.

At the highest level of resolution, EcoRI restriction maps have been constructed for cosmid contigs spanning an accumulated 5.6 Mb (see poster by Burgin et al). This serves to verify the established contigs and determine the length of the DNA spanned by overlapping clones. This procedure will eventually localize individual genes and genetic markers with an average resolution of 10 Kb.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

A Metric Physical Map of Human Chromosome 19

Linda K. Ashworth, (ashworth1@llnl.gov), Joe Balch, Mark Batzer, Brigitte Brandriff, Elbert Branscomb, Emilio Garcia, Jeff Garnes, Jeff Gingrich, Jane Lamerdin, Richard Langlois, Greg Lennon, Ray Mariella, Harvey Mohrenweiser, Anne Olsen,Tom Slezak, and Anthony V. Carrano

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

The chromosome 19 physical map constructed at Lawrence Livermore National Laboratory is a high resolution cosmid-based map spanning the entire chromosome. The foundation of the map consists of a set of 802 cosmid contigs assembled by an automated restriction fragment fingerprint analysis of cosmids from a chromosome 19-specific library. These contigs span an estimated 54 Mb, or >95% of the chromosome excluding the centromere. Over 400 of the contigs have been mapped by fluorescence in situ hybridization (FISH) to metaphase bands. About 160 contigs have been further ordered along the chromosome by a high resolution FISH mapping technique in which the distance between cosmid markers is also determined (see poster by Brandriff, et. al). This ordered FISH map provides a `to-scale' framework, or metric map, upon which to anchor cosmid contigs. By identifying large insert clones (YACs, BACs, PACs and P1s) that span gaps between the ordered cosmid contigs, we have decreased the number of ordered islands to 72. These islands cover a cumulative length of 38 Mb, or about 75% of the non-centromeric part of the chromosome, with the largest island spanning over 10 Mb (see poster by Elliott, et. al.). YACs are isolated by either of two procedures: STS screening of total genomic YAC libraries or hybridization of Alu-PCR products from cosmid contigs to Alu-PCR products from a chromosome 19-enriched YAC sub-library provided by Genethon. Large insert clones spanning gaps are being hybridized back to the cosmid library to identify additional cosmids/contigs located in the gaps. Overlap between adjacent contigs in these regions can often be determined by EcoRI mapping. In some cases cosmid walking experiments are needed in order to achieve continuity at the cosmid level. To date, over 30 Mb of EcoRI restriction maps have been generated (see poster by Burgin, et. al.). In addition, over 400 genes and genetic markers have been localized on cosmids, of which nearly 300 have been incorporated into the metric map (see poster by Tsujimoto, et. al.). Genomic sequencing is being done in selected regions of interest (see poster by Lamerdin, et. al.), particularly on three DNA repair genes found on the chromosome. Software has been designed (see poster by Wagner, et. al.) that integrates and displays cosmid, YAC, FISH, and restriction maps, as well as sequence, hybridization, and screening data.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory, contract no. W-7405-ENG-48.

Human Chromosome 19: A Fluorescence *in situ* Hybridization Map with Genomic Distance Estimates for 196 Sequential Intervals Separating 198 Ordered Cosmid Reference Points and Spanning 50 Mb

Brigitte F. Brandriff, Laurie A. Gordon, Anne S. Olsen, Anthony V. Carrano, Anne Bergmann, Mari Christensen, Anne Fertitta, Linda Danganan, Denise Lee, Linda K. Ashworth, David O. Nelson, and Harvey W. Mohrenweiser

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore California 94550.

A physical map of human chromosome 19 has been constructed by fluorescence *in situ* hybridization of a set of reference cosmids to metaphase chromosomes, somatic interphases, and sperm pronuclear interphases. Genomic distance estimates were derived by reference to a standard curve relating known genomic distances and physical distances measured in pronuclei [1]. The map spans 50 Mb and was generated with multiple, partially overlapping estimates of genomic distances for 196 intervals separating 198 sequentially ordered cosmid reference points. The average distance separating pairs of cosmids in the p arm was 250 kb, with a range from 50 to 700 kb; in the q arm, 270 kb with a range of 50 to 840 kb. The mapped elements included cosmids positive for 33 genes or gene family members and 5 polymorphic markers in the p arm; and in the q arm, cosmids positive for 50 genes or gene family members, and 26 polymorphic markers. Overall, there were 22 cosmids from 19p13.3; 19 cosmids from p13.2; 21 cosmids from p13.1; 18 cosmids from p12; 20 cosmids from q12; 27 from q13.1; 23 from q13.2; 23 from q13.3; and 25 from q13.4. An analysis of replicate genomic distance estimates generated for a subset of 45 cosmid pairs showed that over 98% of the variation between measurements was due to properties inherent in the decondensed pronuclear chromatin with less than 2% due to incidental experimental factors. Results of a comparison between a subset of genomic distance estimates generated by FISH in pronuclei and genomic distance estimates for the same intervals obtained by restriction mapping indicated close agreement between the two sets of data in most cases. The FISH map is being utilized as a metric backbone for other chromosome 19 mapping efforts (see related LLNL posters).

This work was performed under the auspices of The U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no.W-7405-ENG-48.

[1] Brandriff et al. (1994), *Genomics,* in press.

Chromosome 19 closure: High Resolution Physical Map of the Entire q12 and q13.1 Bands

Jeffrey Elliott[1], Ann Gorvad[1], K. Soliman[2], Brian Cheney[1], Linda K. Ashworth[1], Matt Burgin[1], Jane S. Lamerdin[1], Greg Lennon[1], Anthony V. Carrano[1], and Emilio Garcia [1,3]

[1]Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550. [2]Department of Plant and Soil Science, A&M University, Huntsville, Alabama. [3]Corresponding author.

A 12.5 Mb, high-resolution map of a region that encompasses the entire q12 and q13.1 bands of human chromosome 19 has been constructed. The ordered clone map has been obtained starting from a foundation of cosmid contigs assembled by automated fingerprinting and localized to the cytogenetic map by fluorescence in situ hybridization (FISH) [1,2]. Clonal continuity of the map has been achieved by binning and linking the pre-mapped cosmid contigs by means of YACs. The map consists of a single YAC contig comprised of 145 YAC members (minimal spanning path of 18 YACs) linking more than 100 cosmid contigs. Nineteen STSs associated with genetic markers or derived from FISH-mapped cosmids have been placed on the map. Clonal continuity has been obtained by linking a series of pre-mapped cosmid contigs by means of YACs. In the first closure step, a series of YACs has been obtained by either STS-PCR screening of 75,000 YACs from the CEPH YAC library [3,4] or by hybridization to a 800 member chromosome 19-specific Mega YAC sub library (Genethon) [5] with Alu polymerase chain reaction (Alu-PCR) probes obtained from our cosmid contigs. In a second step, the YACs obtained by this combined procedure have been used as templates to generate Alu-PCR probes against chromosome 19-specific cosmid arrays. This two step approach enables unambiguous identification of the cosmid-contig-YAC overlaps. Seventy seven percent (9531 Kb) of the map obtained has been validated by EcoRI restriction mapping. This type of map provides a level of resolution and order of magnitude higher than those obtained from YACs alone.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

[1] Brandriff, B.F., Gordon, L.A., Fertitta, A., et al (1994) *Genomics*. In Press.
[2] Trask, et al (1993) *Genomics*, **15**, 133-145.
[3] Albertson, et al (1990) *PNAS*, **87**, 4526.
[4] Dausset, et al (1992) *Behring Inst. Mitt*, **91**, 13-20.
[5] Chumakov, et al (1992) *Nature Genetics*, **1**, 222-225.

Chromosome 19 Closure: High Resolution Physical Map of an Eight Megabase Region Extending from q12 to q13.1

Emilio Garcia, Jeff M. Elliott, Ann Gorvad, Linda K. Ashworth, Matt Burgin, Greg Lennon, Jane S. Lamerdin and Anthony V. Carrano

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

We report the construction of a high-resolution physical map of an 8 Mb region that encompasses the entire q12 and most of the q13.1 bands of human chromosome 19. The ordered clone map has been obtained starting from a foundation of cosmid contigs assembled by automated fingerprinting and localized to the cytogenetic map by fluorescence in situ hybridization (FISH). Clonal continuity of the map has been achieved by binning and linking the pre-mapped cosmid contigs by means of YACs. The map consists of a single YAC contig with 108 YAC members (minimal spanning path of 14 YACs) linking 66 cosmid contigs. Eighteen STSs associated with genetic markers or derived from FISH-mapped cosmids have been placed on the map. Clonal continuity has been obtained by linking a series of pre-mapped cosmid contigs by means of YACs. In the first closure step, a series of YACs has been obtained by either STS-PCR screening of 75,000 YACs from the CEPH YAC library or by hybridization to a 800 member chromosome 19-specific MegaYAC sub library (Genethon) with Alu polymerase chain reaction (Alu-PCR) probes obtained from our ordered cosmid contigs. In a second step, the YACs obtained by this combined procedure have been used as templates to generate Alu-PCR probes against chromosome 19-specific cosmid arrays. This two step approach enables unambiguous identification of the cosmid-contig-YAC overlaps. Eighty percent (6 Mb) of the map obtained has been validated by Eco R I restriction mapping. This type of map provides a level of resolution an order of magnitude higher than those obtained from YACs alone.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-705-ENG-480.

Incorporation of FISH Reference Clones, Genetic Markers and Genes into the Physical Map of Human Chromosome 19

Susan Tsujimoto, Annette Swartz, Laurie Gordon, Anthony V. Carrano, Brigitte Brandriff, and Harvey W. Mohrenweiser

Human Genome Center; L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory; Livermore, California 94550.

The usefulness of a physical map is dependent on the number of attributes assigned to and positioned along that map. A densely populated physical map of ordered genes and genetic markers allows the map to be combined with other types of maps and data. Our cosmid contig map of human chromosome 19 covers 54 MB, ~95% of the non-centromeric region of the chromosome, and contains >200 genes, >50 ESTs, and >100 polymorphic markers, all of which have been assigned to cosmids that have been cytogenetically localized. Over 80 of the >100 currently available, highly informative genetic markers have been assigned to cosmids. Many of these markers were mapped to cosmids using a strategy that exploits the unique sequence on either side of a microsatellite repeat.

Of the localized genes and genetic markers assigned to cosmids, 160 of the markers, representing 120 contigs, have been incorporated into a detailed flourescent *in situ* hybridization (FISH) order map. The order of and distance between cosmids was determined by FISH, using G1 interphase pronuclei of sperm as targets. The average distance between FISH mapped cosmids is ~250kb, and fewer than 15 gaps >500kb (maximum 800kb) remain to be resolved. The FISH ordered cosmids then served as anchor points amongst which an additional 60 markers from linkage and restriction fragment mapping data could be binned and ordered. A partial order tree generated by an automated computer algorithm incorporates these 220 markers, and serves as the backbone for our integration process of genetic and linkage information.

Discrepancies in the estimated physical and genetic distances are noted. The genetic distance between some markers is 15 times (15cM) the estimated physical distance (1MB); the total genetic distance is only ~2X the physical size of the chromosome. The currently available genetic markers are sufficient to provide a map with very few gaps greater than 1MB, although recombinational "hotspots" distort the linkage map.

Significant strides toward complete integration of the genetic and physical maps have been made. The commonality of reference points enhances the utility of our cosmid clone map of human chromosome 19 as a resource for positional cloning of disease genes and chromosome breakpoints or translocations.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

EcoRI Restriction Mapping of Chromosome 19

Matt Burgin, Linda K. Ashworth, Stephanie Johnson, Laurie Gordon, Susan Hoffman, Anthony V. Carrano

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

The foundation chromosome 19 physical map is comprised of a set of overlapping cosmids clones assembled into contigs by automated fluorescence-based fingerprinting, which spans about 95% of the chromosome. Restriction mapping is being used to validate contig construction, close gaps between selected cosmid contigs, validate distance measurements obtained by FISH, and guide clone selection for genomic sequencing. Cosmids are digested to completion with EcoRI, and simultaneously ligated to a fluorescence-labeled oligomer. Fragments are separated using the PE/ABI 362 Gene Scanner. Fragment sizes are calculated by comparison to known fragments labeled with a different fluorophore and run in the same lane as the sample. Data are transferred directly to the chromosome 19 database. Restriction maps are assembled manually from the fragment data. Similar EcoRI mapping has been successfully done with BAC, PAC, fosmid, and P1 phage clones.

Over 170 restriction maps ranging from 45-1,020 Kbp have been assembled, representing coverage of approximately 60% (30 Mb) of the non-centromeric part of the chromosome. The maps span 87 genes and 137 genetic markers. In some cases, genes, markers or probes are associated with specific EcoRI fragments by southern hybridization. All restriction maps and associated hybridization data are displayed graphically using Browser and incorporated into the partial order map of the chromosome. Extension and consolidation of restriction maps is being directed in part by chromosome walking, and is aided by mapping larger insert clones when available.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

Distribution and Organization of Zinc Finger Genes on Chromosome 19

Susan M. G. Hoffman, Chris Amemiya, and Harvey W. Mohrenweiser

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore , California 94550

Characterized genes containing zinc finger motifs are known to encode transcriptional regulators, and are of critical importance in developmental processes. It has been suggested that at least 100 of these genes, or 10-20% of the 300-700 ZNF loci the human genome is estimated to contain, are located on chromosome 19. Since typical ZNF loci contain large introns and cover 30-40 kb of genomic sequence, 5-10% of the 60 megabase chromosome 19 is expected to consist of these genes. A conserved probe from the H-C link between zinc fingers was used to screen an ~10X cosmid library of chromosome 19, and hybridized to more than 1000 cosmids. These cosmids were assembled into contigs and localized cytogenetically by FISH mapping. The ZNF genes are clustered physically, probably reflecting their proliferation via tandem duplications; currently, six major clusters of ZNF genes are known on chromosome 19, in p13.3, p13.2, p12, q13.1, q13.2, and q13.4.

At least 40-50 closely related ZNF loci of the KRAB family are clustered on 19p12. Most of this ZNF gene group falls within a four-megabase region of continuous clonal coverage we have assembled; these four megabases span two-thirds of the cytogenetic p12 band. About half of these loci are very similar to ZNF 91, which maps near the centromere in p12. This cluster of ZNF genes overlaps with several oncogenes at the p12-p13.1 boundary; otherwise, only ZNF genes have been localized to p12.

Another family of highly related ZNF genes, the ZNF 58 group, is located on p13.2, where at least six "58-like" loci are spaced along a contig of ~260 kb. A third group, the cluster on q13.4, contains MZF-1, an important regulator of hematopoietic development, plus at least one other ZNF gene of related function.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

Cosmid Contig Map of a 900-kb Candidate Region for Finnish Congenital Nephrosis in 19q13.1

Anne Olsen[1], Anca Georgescu[1], Laurie Gordon[1], Matt Burgin[1], Marjo Kestilä[2], Minna Männikkä[2], and Karl Tryggvason[2]

[1]Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550. [2]Biocenter and Department of Biochemistry, University of Oulu, Oulu, Finland.

Congenital nephrotic syndrome of the Finnish type (CNF) is an autosomal recessive disease with an incidence of about 1 in 8000 in Finland. It is characterized by massive proteinuria and signs of nephrosis at birth. The basic defect is unknown, and the disease represents a unique biological model for kidney dysfunction. Based on results of linkage analysis in 17 Finnish families, the CNF locus was mapped to chromosome 19 between D19S416 and D19S224 [1].

In order to define the physical location of the candidate disease gene region, the genetic markers used for linkage analysis were integrated into the physical map of chromosome 19. PCR primers for microsatellite markers were hybridized to colony filters of a chromosome 19-specific cosmid library, and positive clones were confirmed by PCR. Cosmids corresponding to genetic markers in this region were analyzed by high resolution fluorescence in situ hybridization (FISH) to pronuclei (see poster by Brandriff et al.) to determine the physical order of, and distance between, these markers. The results indicated that the markers were located in q13.1 in the order: cen - (S213, S248, S416) - S425 - (S208, S191) - S224 - S220 - tel. The distance from S416 to S224 was an estimated 2.2 Mb. When the physical order data was combined with additional linkage results, the disease gene location could be further restricted to an 860-kb region between S208 and S224.

Cosmid contigs previously established by fingerprinting were integrated into the ordered FISH map. The contigs were analyzed by EcoRI mapping to verify contig assembly and to determine the distance spanned. Probes from the ends of contigs were used for cosmid walking to extend and merge the previously established contigs. Most of the region from S208 to S224 is currently covered with four contigs of 85-, 318-, 171- and 272-kb. Genes mapped to these contigs include COX6B, COX7A1, MAG, ATP4A and APLP1, as well as two anonymous cDNAs. Further walking in cosmid, BAC and PAC libraries should close the three small gaps remaining between contigs.

New polymorphic markers are being developed from cosmids in the candidate region. Linkage analysis with two of these markers has further decreased the size of the candidate region to less than 750 kb. One of the new markers shows no recombination with the CNF gene in any of the families and may lie close to the disease locus.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

[1] Kestilä et al. (1994) Congenital nephrotic syndrome of the Finnish type maps to the long arm of chromosome 19. *Am. J. Hum. Genet.* **54,** 757-764.

Generation and High Resolution Mapping of New STSs Derived From Chromosome 2 Microdissection Clones

Jeffrey C. Gingrich, Jeffrey A. Garnes, Jane E. Lamerdin and Lisa K.Scheidecker

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

In order to generate new resources for the mapping of human chromosome 2 we are generating new STSs from chromosome 2 microdissection clones. A series of microdissection clone libraries spanning ~20 Mb regions of chromosome 2 are being generated by Dr. F.-T. Kao and colleagues at the University of Colorado. From each library a number of unique sequence clones were identified and shown to be derived from chromosome 2. We have taken unique sequence clones from the 2Q1 and 2P1 libraries, libraries derived from the two telomeric regions of the chromosome, and have determined their DNA sequence. STS primer pairs were then determined using the Primer program. The STS primers were then tested against DNA from human, hamster and a monochromosomal hybrid cell line containing chromosome 2, GM10826. To date, 35 new STSs have been generated, ~1 per Mb of the microdissected regions of the chromosome. Additional STSs will be generated from 8 other microdissection libraries from chromosome 2 as we obtain them. The STSs are being further localized on the YAC map of chromosome 2 by testing them against a set of regionally assigned YACs from the CEPH data set. The information obtained gives both an indication of the location of the STSs relative to other mapped markers on the chromosome and provides further validation of the CEPH data on chromosome 2.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

Analysis of Randomly Cloned Chromosome 5 DNA Sequences for Similarities to GenBank Sequences

Robert Wagner, Deborah Grady, Julie Houts, Donna Robinson, Joe Gatewood, Sue Thompson, Larry Deaven, and Robert Moyzis.

Life Sciences Division and The Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

In order to generate a framework STS map of human chromosome 5, approximately 761 random DNA segments from chromosome 5 (Chr5seqs) were cloned in M13mp18. They averaged 250bp in length, and were sequenced with a Dupont Genesis 2000 automated sequencer. A total of 39 (5.9%) Chr5seqs were found to have open reading frames (ORFs) using the GRAIL program; five of these were all L1 and one was all Alu. After eliminating from the total 761 all those sequences that were all Alu, L1, vector or not sequenceable, 445 of the remaining 72% (545) sequences, were subjected to similarity determinations against all sequences in GenBank using the INHERIT program. An INHERIT similarity score of 200 (corresponding to approximately 90% similarity) or better with various GenBank sequences was found for 42% (187) of these 445 sequences. About 100 of these were also subjected to analysis with the BLASTGA program searching for similarities of length greater than 60% that had identities greater than 70% of alignment length. The combined results of both searches may be summarized as follows: (1) Five sequences have been identified that have significant similarities to genes known to be on chromosome 5, and map to loci previously identified for them. This is as expected given that approximately one percent of the genes likely to be present on chromosome 5 have been sequenced. (2) At least ten Chr5seqs show significant similarity for specific regions of a number of different seemingly unrelated genes. (3) A number of known repeats such as pTR5 have been identified which have ORFS. (4) Several Chr5seqs showed close similarity with protein encoding sequences of genes in GenBank not previously known to have loci on Chromosome 5. As an example one sequence with an ORF has about 90% similarity for the gene HUMHSD that encodes delta-5-3-beta-hydroxy-delta-5-steroid dehydrogenase on 1p13.1. This sequence also shows similarity for cognate genes in four other mammals, indicating that an HSD family member or pseudogene resides on chromosome 5.

Low Resolution Physical Mapping of Human Chromosome 5 : Cloning the Cri-du-Chat Critical Regions

D.L. Grady[1], D.L. Robinson[1], J. Overhauser[2], S. Goodart[2], M. Feder[2], A. Simmons[3], M. Lovett[3], E. Nickerson[4], J.D. McPherson[4], J.J. Wasmuth[4], E.T. Peterson[1], L.A. Chasteen[1], L.L. Deaven[1], and R.K. Moyzis[1]

[1]Los Alamos National Laboratory, Los Alamos, New Mexico, [2]Thomas Jefferson University, Philadelphia, Pennsylvania, [3]University of Texas Southwestern Medical Medical Center, Dallas, Texas, and [4]University of California, Irvine, California.

Over 300 new STS's have been generated from flow sorted human chromosome 5 DNA. These STS's have been regionally assigned a p or q arm status. All p arm STS's have been regionally localized into 40 bins using a natural mapping panel of somatic cell hybrids. The majority of q arm STS's have been localized into 16 unique bins. The STS's appear to be uniformly distributed along the length of this 194 Mb chromosome. This current density of markers (1 STS 650 Kb) is sufficient to provide a framework map for YAC contig assembly in any region of chromosome 5. Our primary focus is the generation of a YAC contig of 5p, centered on the regions associated with the Cri-du-Chat Syndrome. Cri-du-Chat is the most common human terminal deletion syndrome (1/45,000 births). It involves deletion of regions of the short arm of human chromosome 5. Clinical features include growth and mental retardation, microcephaly, hypertelorism, epicanthal fold, and the characteristic high pitched cat like cry. A correlation between clinical features and chromosome deletion patterns has led to the identification of two regions of the short arm that appear to be crucial for the manifestation of the complete Cri-du-Chat phenotype. Loss of a small region in 5p15.2 (Cri-du-Chat critical region) correlates with all the clinical features of the syndrome with the exception of the cat-like cry, that maps to the 5p15.3 (cat-like cry critical region). A complete, non-chimeric, YAC contig of the Cri-du-Chat critical region (2 Mb) has been identified and characterized. A YAC contig of the cat-like cry critical region has been constructed and is currently being characterized. Portions of this work were funded by the U.S. Department of Energy under contract W-7405-ENG-36.

Construction of a P1 Map in the Region of Chromosome 5q31-q35

Jan-Fang Cheng, Steve Lowry, Yiwen Zhu, Duncan Scott and Eddy Rubin

Human Genome Center, Life Science Division, Lawrence Berkeley Laboratory, MS 74-157, 1 Cyclotron Road, Berkeley, CA 94720

The mapping project at the LBL Human Genome Center has focused on generating a set of P1 clones providing a complete coverage of the target region so that cloned genomic fragments with minimal overlaps can be determined and selected as templates for production sequencing. The q31-q35 region of chromosome 5 was chosen as the primary target for sequencing because this region contains a cluster of growth factor or receptor genes and it is likely to yield new and functionally related genes through long range sequence analysis. We have selected the 1.1 Mb interleukin gene cluster region in 5q31 as a starting point because: (1) that it contains a number of genes with related biological function, therefore, it exists the high likelihood that other genes resulting from gene duplications will be identified through sequence analysis; (2) that the genes localized to this region are of considerable health and medical importance. The known genes regulate hematopoietic cell proliferation; (3) the fact that multiple, previously sequenced genes (IL3, IL4, IL5, IL13, IRF1 and CSF2) are dispersed throughout this region would serve as a quality control for production sequencing; (4) the fact that sequence and functional information about the genes already characterized from this region may well assist in the difficult task of assigning importance and function to genes derived from the production sequencing program.

There are three key experimental steps in our mapping strategy, and they were designed to generate a clone map for specific regions of a chromosome in a time- and cost-effective way. The first step is to use inter-Alu fragments generated from YACs to isolate regionally specific P1s covering the target region. The second step is to establish the order and overlaps of the isolated P1s in a clone-limited approach. The third step is to close gaps and verify the integrity of the cloned fragments.

Three non-chimeric YACs were identified to spin the 1.1 Mb interleukin gene region, and 54 P1s have been isolated using probes derived fron these YACs. Informative STSs were developed from 6 known genes and 19 ends of key P1s to resolve two contigs (700 Kb and 300 Kb in size). DNA sequences flanking the gap are being used to identify additional clones for closure. All STSs are being used to generate restriction maps from both genomic DNA and cloned DNA in this region. The restriction map comparison should allow us to identify gross structure rearrangement, if any, in the cloned P1 DNA.

Construction of a High Resolution P1 and cDNA Map in the Down Syndrome Region of Chromosome 21

Jan-Fang Cheng and Yiwen Zhu

Human Genome Center, Life Science Division, Lawrence Berkeley Laboratory, MS 74-157, 1 Cyclotron Road, Berkeley, CA 94720

A high resolution clone map is constructed in the "major Down syndrome region" of chromosome 21 for a number of reasons: (1) transgenic studies of Down syndrome phenotypes (see abstract by Smith et al.), (2) identification of the mouse *Weaver* gene by complementation analysis. (3) DNA templates for production sequencing of the genomic fragments yielded interesting phenotypes.

We have developed a two-tier hybridization method to screen a human P1 library (Du Pont Merck and Genome System Inc.) for clones in the region of 21q22.2-22.3. The goal is to construct a high resolution clone map providing continuous coverage of human DNA with P1 clones. A rapid initial screening of the P1 library is carried out by filter hybridization using pooled DNA probes derived from a specific chromosomal region approximately 3 Mb in length. We have included inter-Alu fragments derived from YACs, STSs, and mapped cDNA sequences in the probe pool to increase representation of the target region. The P1 filter contains DNA prepared from the pooled library (series B library) which was constructed by pooling 12 clones of the same row from the original library. The first tier hybridization therefore indicates the plate and row yielding positive clones. The twelve clones from each positive hybridization are re-gridded on filters for the second tier of hybridization using the same probe pool. The two tier screening produces a small set of clones (approximately 130 clones in a 3 Mb region with this 3.5 hit library) for contig construction. This process bypasses the slow and costly process of screening of the entire genomic library with every available STS. We have used this hybridization method to isolate 130 P1 clones in the major Down's syndrome (DS) region of chromosome 21 which comprises approximately 3 Mb of genomic DNA extending from D21S17 to ETS2. Overlaps between these P1's are determined using end probes generated from each P1 to cross-hybridize with DNA isolated from all isolates. This hybridization identifies overlapping P1's as well as gaps between two P1 contigs. The end probe which identifies a gap is used to probe the pooled P1 library again or to probe a chromosome 21 specific cosmid library for clones that could bridge the gap.

These P1 were used to isolate novel cDNAs by hybrid selections, and some of the P1s were used in creating transgenic animals for phenotypic studies. These results should impact on future targets of the production sequencing program.

Long range mapping and sequencing of the human X chromosome

D.L. Nelson[1], E.E. Eichler[1], B.A. Firulli[1], Y. Gu[1], G.B. Ferraro[1], B. Franco[1], A. Grillo[1], G. Borsani[1], M.C. Wapenaar[1], A. Ballabio[1], A.C. Chinault[1], E.J. Roth[1], H.Y. Zoghbi[1], F. Lu[1], M.A. Wentland[1], D.M. Muzny[1], J. Lu[1], R.L. Clingan[1], S. Richards[1], R.A. Gibbs[1],

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas

The human X chromosome is significant from both medical and evolutionary perspectives. It is the location of several hundred genes involved in human genetic disease, and has maintained synteny among mammals; both of these aspects are due to its role in sex determination and the haploid nature of the chromosome in males. We have addressed the mapping of this chromosome through a number of efforts, ranging from long-range YAC-based mapping to genomic sequence determination.

YAC mapping. We have constructed a 40 Mb physical map of the Xp22.3-Xp21.3 region, spanning an interval from the pseudoautosomal boundary (PABX) to the Duchenne muscular dystrophy gene. The map is composed of nearly 500 YAC clones derived from five different YAC libraries, including the two CEPH libraries, which contributed the majority of the clones. YAC screening has been carried out through use of two methods: PCR sampling of YAC DNA pools and hybridization of pooled Alu PCR products. Contigs were assembled primarily by STS content analysis, with subsequent verification of YAC overlaps via extensive Southern blot studies of YAC DNAs. Verification of marker order was by use of numerous somatic cell hybrids retaining translocated and deleted chromosomes derived from patients whose breakpoints are well mapped. The map is highly annotated, with 85 breakpoints defining 53 deletion intervals, 175 STSs (20 of which are highly polymorphic), and 19 genes.

Cosmid binning. Two regions of the Xp physical map have been converted to cosmid based maps using the Lawrence Livermore National Laboratories flow sorted X chromosome-specific cosmid library. YAC probes have been used to identify cosmids from the library, and these have been sorted with the use of adjacent YACs and YAC fragments derived by rare-cutter digestion and pulsed-field gel isolation. One of these efforts was recently published (1). A systematic effort to identify cosmids for this region, and particularly for the genes contained in the region is ongoing. Additional cosmid contig construction in distal Xq is described in the accompanying abstract (*Parrish et al.*).

Sequencing. An independently funded project awarded to RAG seeks to develop long-range genomic sequence for ~2 Mb of the human X chromosome. In support of this project, cosmids have been constructed and isolated for the 1.2 Mb region between FMR1 (the gene involved in fragile X syndrome) and IDS (iduronate sulfatase, Hunter syndrome) in Xq27.3-Xq28. This region is at the boundary of Xq27.3 and Xq28, and its sequence may provide interesting data regarding the nature of chromosome bands in addition to discovery of novel coding sequences (such as that involved in FRAXE related mental retardation). To date, the complete sequence of both the FMR1 and IDS genes have been determined (62 and 40 kb, respectively), along with ~250 kb of the interval between. Shotgun sequencing using the sequence-mapped gap strategy in conjunction with standard ABI fluorescent chemistry is being employed. Additional sequence in Xq28 has been determined, including that of a cosmid containing the three genes, ALD, DXS1357E and a creatine transporter. Curiously, this cosmid exhibits strong homology with a region on 16p11.1-11.2 and clones derived from the Los Alamos chromosome 16 library have been isolated and are being characterized for extent of similarity. This duplication appears to have a very recent evolutionary origin.

(1) Wapenaar et al. *Hum Mol Genet* **3**:1155-1161.

Cosmid Contig Construction and Gene Identification in Human Xq28

J.E. Parrish[1], E.E. Eichler[1], Y. Gu[1], B.A. Oostra[2], A.J.M.H. Verkerk[2], J. Reynolds[3], C.S. Richards[1], A.S. Spikes[1], L.G. Shaffer[1] and D.L. Nelson[1]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas; [2]Department of Clinical Genetics, Erasmus University, Rotterdam, The Netherlands; [3]Shodair Hospital, Helena, Montana.

Efforts to identify and characterize genes in human Xq28 have led to the construction of cosmid- based physical maps of several regions using clones derived from the Lawrence Livermore National Laboratories flow sorted X chromosome-specific cosmid library. Contigs have been seeded using cDNA clones; walking with cosmid ends has provided physical linkage between several genes, particularly in the gene-dense region between the opsin genes and G6PD. This region is fully covered by a contig of about 300 kb; a novel muscle gene has been identified and is currently being tested as a candidate for Emery-Dreifuss Muscular Dystrophy. Two other contigs of 150-200 kb each have been constructed in the L1CAM-HCF region and in the ALD region. An EST, DXS1012E, has also been used to seed a contig between the opsin genes and HCF. We are currently attempting to close the gaps between these four contigs. Further proximal, scattered cosmids have been identified using polymorphic markers and repeat sequences. Over 30 cosmids have been isolated for most markers known in the interval between FRAXA and IDS in efforts to support the long-range sequencing efforts of Dr. R. A. Gibbs. Efforts to identify potential transcripts at the FRAXE locus are ongoing and aided by the long-range sequence effort, however cosmids spanning FRAXE and FRAXA CGG repeats have not been found in this library.

A fourth large contig has been established in the proximal portion of Xq28. The original cosmid was identified by hybridization to a novel cDNA with homology to FMR1, the gene involved with Fragile X syndrome. We have characterized a GCC repeat in this cosmid which represents FRAXF, the third folate-sensitive fragile site to be defined in Xq27.3-q28. A 5kb *Eco* RI fragment of the cosmid detects a fragment increased by 900 base pairs in a mentally retarded male exhibiting 28% fragile site expression. By DNA analysis, this patient demonstrates normal alleles at both FRAXA and FRAXE. The mother of the proband carries an allele expanded to a lesser extent, in addition to a normal allele. The proband's DNA at this locus is methylated at three sites (*Sac*II, *Eag*I, and *Bss*HII) within the CpG island. Additional normal and retarded family members were tested both for fragile site expression and for expansion and methylation at this locus; the expansion and methylation are found only in individuals expressing the fragile site, suggesting that this clone does represent the fragile site. Inheritance of the fragile site does not show a direct correlation with the mental impairment in this pedigree.

The site of variation was localized to within 300 base pairs. By sequence analysis, the plasmid subclone contains $(CGG)_8$. PCR primers were designed to amplify across the repeat. Expanded alleles amplify poorly or not at all, which is consistent with difficulties observed with FRAXA. Alleles in the normal population vary from 6 to 29 repeats. Fluorescence *in situ* hybridization studies have been carried out using cosmids which overlap in the region of the fragile site; the results demonstrate unequivocally that this clone spans the fragile site in the patient mentioned above. Efforts to define the gene content of the region are under way, in order to determine the potential involvement of this repeat in regulation of genes in its vicinity. A cDNA has been isolated which lies just distal to the fragile site. Sequence analysis is currently under way; preliminary results suggest that this clone represents a member of the MAGE gene family of tumor antigens. Screening of the cosmid library with this cDNA produced a total of 70 positive cosmids, which likely represent the entire cluster (three known MAGE genes, plus the novel gene cloned in this effort). Current efforts involve the characterization of this genomic region, including the structure of the MAGE gene family and analysis of expression of the MAGE genes in individuals with expansions at FRAXF.

Towards an STS content map of human chromosome 11: Localization of 910 YAC clones and contigs and assembly of a first generation physical map.

Glen A. Evans[1], Julie M. Bailis[2], Sylvia Thomas[1], Jason V. Khristich[2], Karin Diggle[2], Yelena Marchuck[2], Joshua Tobin[2], Stephen P. Clark[3], Annie Rodkins[2], Stewart Marcano[2], Allan C. Churukian[2], Jane S. Hutchinson[2], Yalin H. Wei[2], Ron Scott[1], Kim Jackson[1], Lori Romberg[1], Shane Probst[1], David Burbee[1], Michael W. Smith[4], Licia Selleri[5], John Quackenbush[6] and Harold R. Garner[1]. [1]McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center at Dallas, 6000 Harry Hines Blvd., Dallas, TX 75235-8591; [2]Molecular Genetics Laboratory and Human Genome Center, The Salk Institute, La Jolla, California 92037; [3]Amgen, Amgen Center, MS 239, Thousand Oaks, CA 92320; [4]Biological Carcinogensis and Development Program, PRI/Dyncorp, Frederick Cancer Research and Development Center, National Cancer Institute, Frederick, MD 21702; [5]Department of Pathology, Stanford University School of Medicine, Stanford CA 94305; [6]Department of Genetics, Stanford University School of Medicine, Stanford CA 94305.

Physical mapping of human chromosomes at a resolution of the 100 kb to 1 mb will provide reagents for gene identification and templates for ultimately determining the complete DNA sequence. Sequence-tagged site (STS) content mapping coupled with large fragment cloning in yeast artificial chromosomes provides an efficient mechanism for producing first generation, low resolution maps of human chromosomes. Previously, we produced a set of standardized STSs for human chromosome 11 regionally localized by fluorescence *in situ* hybridization or somatic cell hybrid analysis. We used these, and other STSs to map over 900 YAC clones to chromosome 11, organized into 109 contigs. This data set spans 218 mb of coverage on the 126 mb chromosome, includes a large number of $(CA)_n$ repeats and other genetically defined markers, and represents a first order approximation of a physical map of human chromosome 11. This set of clones, contigs, genetic markers and associated STSs will provide the material for the production of a continous overlapping set of YACs as well for high resolution physical mapping based upon sampled sequencing, and ultra-rapid genotyping using DNA chips.

This work was supported by the DOE Genome Program, the National Center for Human Genome Research, and the G. Harold and Leila Y. Mathers Charitable Foundation. MWS was a Human Genome Distinguished potdoctoral fellow of the DOE and JQ is a Special Research Career Emphasis Awardee of NCHGR.

BUILDING A PHYSICAL MAP BY STS CONTENT MAPPING TO SUPPORT LARGE SCALE GENOMIC SEQUENCING: THE *DROSOPHILA* GENOME PROJECT

William J. Kimmerly, Karen E. Stultz, Keith Lewis, Veronica M. Lustre, Dazhong Sun, Christopher H. Martin, and Michael J. Palazzolo. Human Genome Center, MS 74-157, Lawrence Berkeley Laboratory, 1 Cyclotron Rd., Berkeley, CA 94720

We are constructing a bacteriophage P1-based physical map of the euchromatic genome of *Drosophila melanogaster*. The ultimate goals of our physical mapping project are to construct a physical map annotated with genetic and biological information, and to identify a minimal number of P1 genomic clones from the library that represent the genome as sequencing templates for the directed genomic sequencing project at LBL. The mapping approach utilizes sequence-tagged sites, or STS markers which fall into three classes: sequences from the ends of P1 genomic inserts, genomic sequences flanking rescued P elements from 2nd and 3rd chromosomal lethals, and known *Drosophila* genes. The major class of STS markers derive the ends of genomic inserts of individual clones in the library and are obtained by direct sequencing of full-length P1 genomic clones. The mapping strategy we are pursuing, referred to as double-end clone-limited, emphasizes the importance and utility of STS markers derived from the ends of clones in the library. This mapping strategy allows the assignment of all clones in the library to contigs using the fewest number of markers and generates contigs larger than a random or single-end approach.

For these experiments we are using a five-hit P1 genomic library (99% statistical coverage). Using the unique sequences defined as STS markers, primers are designed for use in a PCR-based mapping strategy. STS markers are mapped in the library by screening pools of P1 DNA by PCR in a two-tiered scheme. These mapping data are used to construct contigs of overlapping clones and to determine linkage of adjacent STS markers. Thus far we have mapped over 1500 STS markers. So far nearly 400 contigs have been constructed which together cover over 100 megabases, or about 75% of the euchromatic genome. Our goal at LBL is to map a total of 3000 STS markers towards a preliminary physical map by July 1995, at which time the clone-limited phase of the mapping project should be complete, and all clones in the library should be assigned to contigs. The next stage of the physical mapping project will attempt to join adjacent contigs into larger ones.

We have successfully built two multi-clone contigs of 350-400 kb that served as our initial targets for large-scale genomic serquencing. These targets were the *Bithorax* complex (BX-C) and the *Antennapedia* complex (ANT-C). Both complexes have been nearly completely sequenced and submitted to Genbank by the LBL directed genomic sequencing group. In collaboration with Michael Ashburner and colleagues in Cambridge, England we have focused much of our mapping effort on a 2 megabase region of chromosome 2 encompassing the polytene bands 34D-36A. This region contains the well-studied *Adh* gene, and is rich in genetic information with numerous chromosomal breakpoints, P element insertions, and lethal complementation groups. This region is now covered by a small number of large contigs and many P1 clones from this region are in the process of being sequenced at LBL.

The *Drosophila* Genome Project is a collaboration between the biology, informatics, and engineering groups at the LBL Human Genome Center. The project also includes Gerald Rubin of the University of California, Berkeley, Allan Spradling of the Carnegie Institute in Baltimore, MD who is contributing the genetic verification of P element lethal lines used in the project, and Dan Hartl of Harvard University in Cambridge, MA who has provided *in situ* hybridization data for over 2500 P1 clones in the library being used for STS generation.

Generating and Typing CAPS Genetic Markers in Dog

Mark W. Neff,[1, 2] Mike Strathmann,[1] and Jasper Rine [2]

[1]Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720. [2]Department of Molecular and Cellular Biology, University of California, Berkeley,, CA 94720.
EM: Neff@Mendel.berkeley.edu.

The behavioral and morphological traits that distinguish the various dog breeds are genetically determined as they have been fixed by artificial selection. A genetic map of the dog genome will make breed-specific traits amenable to quantitative trait loci (QTL) analysis. Presently, a genetic map of dog consisting of microsatellite markers is being assembled [1]. Although polymorphic and informative, these molecular markers have several genotyping limitations. They require polyacrylamide gel electrophoresis, radioactive- or fluorescence-based detection, and manual scoring. Furthermore, only a few markers can be typed per reaction such that genotyping efforts increase linearly with map resolution. For a quantitative genetics approach involving controlled matings, a less polymorphic, more easily typed marker might also be valuable.

A cleaved amplified polymorphic sequence (CAPS) is a RFLP encompassed by primer sites [2]. Although less polymorphic than microsatellite markers, CAPS markers may be typed simply by agarose electrophoresis and ethidium bromide staining. Genomic subtraction methods have previously been used to generate libraries of RFLP loci [3]. We report on the construction of a CAPS library based on an experimental Border Collie X Newfoundland cross. We also present progress on a novel method for genotyping many CAPS markers simultaneously. These advances should reduce the resources and effort necessary to construct and use genetic maps.

M.N. was supported by a DOE Human Genome Postdoctoral Fellowship, and M.S. was supported by an Alexander Hollander Postdoctoral Fellowship.

[1] E. A. Ostrander, G. F. Sprague and J. Rine (1993) Identification and characterization of dinucleotide repeat (CA)n markers for genetic mapping in dog. Genomics 16: 207-213.
[2] A. Konieczny and F. M. Ausubel (1993) A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. Plant J. 4: 403-10.
[3] M. Rosenberg, M. Przybylska and D. Straus (1994) "RFLP subtraction": a method for making libraries of polymorphic markers. Proc. Natl. Acad. Sci. USA 91: 6113-7.

Physical Mapping in Preparation for DNA Sequencing

Andreas Gnirke, Regina Lim, Gane Wong, Jun Yu, and Maynard Olson

Department of Molecular Biotechnology, FJ-20, University of Washington, Seattle, Washington 98195

We are concentrating on a set of methods to allow the precise physical mapping of substantial segments of human DNA. Initial mapping objectives are on the order of 1 Mbp, typically available as YAC contigs supported by STS-content maps. Our goal is to develop subclone collections and precise maps from this starting point. These clones and maps are designed to provide direct support for large-scale DNA sequencing .

The core methodology involves a method of restriction-site mapping that we refer to as multiple-complete-digest (MCD) mapping, which we have developed in collaboration with Will Gillett (Washington University Department of Computer Science). MCD mapping starts from fragment-size lists acquired with multiple restriction enzymes for each member of a highly redundant set of randomly generated subclones that collectively span the mapping objective. In general, enzymes with 6-bp specificity are employed when mapping cosmids subcloned from YACs, while enzymes with 4-bp specificity are employed when mapping small-insert clones subcloned from cosmids. Most experience to date involves mapping cosmids subcloned from YACs.

Digests are analyzed on standard agarose gels, which are post-stained with ethidium bromide or thiazole orange and imaged with a Molecular Dynamics FluorImager 575. We have developed software with which to extract fragment sizes from these images automatically. Fragment-size lists are assembled into MCD maps with software developed by W. Gillett. The basic concepts that underlie this software are (1) absolute compatibility between the maps and the fragment-size lists and (2) compatibility between the partial orderings determined independently in each digest domain for the ends of the clone inserts.

The only requirements for applicability of this approach are the ability to create a highly redundant set of subclones of a mapping objective in a vector system that allows complete-digest fingerprinting and the existence of a higher-level physical map of sufficient resolution to allow the contigs produced by MCD mapping to be ordered and oriented. The present emphasis on YAC contigs is simply motivated by the need for a convenient source of deep cosmid coverage of targeted mapping objectives, particularly coverage that is easily generated in order to allow free experimentation with subcloning methods. The long-term goal of this project is to use MCD mapping as the basic engine of an automated system for the analysis of whole chromosomes or genomes. Hence, there is a major emphasis on minimizing the need for expert human intervention in the data collection and analysis.

We have also developed methods based on the RARE (RecA-Assisted-Restriction-Enzyme) cleavage technique to accommodate gaps in these maps that are due to systematic absence of valid subclones from specific subregions. These techniques also allow STS-content maps to be converted to true, distance-calibrated physical maps.

IDENTIFICATION AND MOLECULAR CLONING OF A 20q13.2 AMPLICON IN BREAST CARCINOMA.

Colin Collins[1], Jeff Froula[1], Minna Tanner[3], Jan-Fang Cheng[1], David Kowbel[1], Farideh Shadravan[1], Chris Martin[1], Michael Palazollo[1], Mary Hintz[2], Ulli Weier[1], Wen-Lin Kuo[2], Olli Kallioniemi[3], Jeff Gingrich[4], Johanna Rommens[5] Dan Pinkel[1,2] and Joe Gray[1,2].

[1]Lawrence Berkeley Laboratory, Berkeley, California; [2]University of California, San Francisco, California; [3]Tampere University Hospital, Tampere, Finland; [4]Lawrence Livermore National Laboratory, Livermore, California; [5]Hospital for Sick Children, Toronto, Ontario.

Comparative genomic hybridization (CGH) has identified a previously undescribed region of increased copy number involving chromosome band 20q13.1-13.2 in 15-20% of primary breast carcinomas. The application of FISH to the study of tumor interphase nuclei using 33 locus specific cosmid and P1 probes distributed along chromosome 20 revealed amplification of band 20q13.2 in 35% of breast cancer cell lines and 8% of primary tumors, and localized the amplification event to the approximately 2 Mb interval defined by (Flpter 0.80-0.84.). This excludes all known genes in the region as candidates for the putative oncogene(s). It is hypothesized that selection for the overexpression of a novel oncogene(s) is driving the observed increase in copy number.

A 3 Mb 12 member YAC contig has been assembled that spans the region of increased copy number. Conversion of the YAC contig to a P1 contig is proceeding by three approaches. In the first approach, interAlu PCR reactions are performed on individual YACs and sufficient products sequenced to create 5-10 STSs per YAC clone. The DuPont P1 library is then screened for these STSs by the PCR. In the second approach, the entire interAlu PCR product is used as a hybridization probe against a gridded array of the DuPont P1 library. Finally, P1 ends are sequenced and STSs created for P1 library screening. These complementary approaches have yielded 52 P1 clones forming 12 contigs. STS content mapping, P1 Southern hybridizations, and P1 fingerprinting are being employed to merge the P1 clones into a single contig. All 52 P1s are being hybridized to tumor derived cell lines and a panel of primary breast tumors using interphase FISH to precisely define the minimum common region of increased copy number.

The direct selection of cDNAs and exon trapping are being performed to identify genes encoded within the 20q13.2 amplicon. In collaboration with the Human Genome Center at the Lawrence Berkeley Laboratory we have initiated the directed sequencing of a P1 contig localized at the postulated core of the 20q13.2 amplicon. It is expected that the combination of large scale directed sequencing, exon trapping and direct selection will culminate in the isolation of the hypothesized oncogene(s) and contribute to the development of novel molecular diagnostic and therapeutic approaches for the management of breast cancer.

This work was supported by grants from US DOE contract DEAC0376SF00098, USPHS grants CA44768, CA45919, CA52807 and Imagenetics.

Cloning of a Human Damage-specific DNA Binding Protein

Anne F. Nichols[1], Tom Brody[1], Scott Keeney[2], and Stuart Linn[1]

[1]Department of Molecular and Cell Biology; University of California; Berkeley, California 94720. [2]Department of Biochemistry and Molecular Biology; Harvard University; Cambridge, Massachusetts 02138.

DDB is a damage-specific DNA binding protein and potential DNA repair factor [1]. This specific binding activity is absent from cells of some individuals with Xeroderma Pigmentosum Group E [2], a human hereditary disease characterized by defective nucleotide excision repair and a high incidence of skin cancer. When DDB was purified to near homogeneity from HeLa cells, the specific DNA damage binding activity copurified with polypeptides of 124 and 41 kDa [3]. DNAse I footprints of the heterodimer of DDB on synthetic DNA substrates containing a single UV-photoproduct was different from that of the isolated p124 subunit. While the size of the footprint was unchanged, the appearance of an additional hypersensitive site in the presence of the heterodimer indicated that the p41 subunit is a functional subunit of the DDB protein [4].

We have isolated full-length cDNAs which encode both polypeptides of DDB from a human fibroblast λZAP library. There is no significant homology of the 41 kDa subunit with any proteins in current databases. However, the translated region of human p124 DDB cDNA has 98% homology with the monkey p127 UV-DDB isolated by Takao *et al.* [NewGenBank, ACC L20216]. Their amino acid sequences have almost 100% identity, with a single conservative substitution. The p41 subunit has been expressed in an *in vitro* rabbit reticulocyte lysate system. Polyclonal antibodies will be produced in rabbits, using over-expressed DDB polypeptides as antigens.

This work was funded by DOE Grant FG03-92ER61458. A.N. was supported in part by an appointment to the Alexander Hollaender Distinguished Postdoctoral Fellowship Program sponsored by the DOE, Office of Health and Environmental Research, and administered by the Oak Ridge Institute for Science and Education.

[1] Keeney, S., Eker, A.P., Brody,T., Vermeulen, W., Bootsma, D., Hoeijmakers, J.H., and Linn,S. (1994) Correction of the DNA repair defect of xeroderma pigmentosum group E by injection of a DNA damage-binding protein. *Proc. Natl. Acad. Sci., USA*, **91**, 4053-6.

[2] Keeney, S., Wein, H., and Linn, S. (1992) Biochemical heterogeneity in xeroderma pigmentosum complementation group E. *Mutation Res. DNA Repair*, **273**, 49-56.

[3] Keeney, S., Chang, G.J., and Linn, S. (1993) Characterization of a human DNA damage binding protein implicated in xeroderma pigmentosum E. *J.Biol. Chem.*, **268**, 21293-21300.

[4] Reardon, J.T., Nichols, A.F., Keeney, S., Smith, C.A., Taylor, J-S. Linn S., and Sancar, A. (1993) *J. Biol. Chem.*, **268**, 21301-8.

The role of recombination and *RAD52* in mutation of chromosomal size DNA transformed into yeast.

Vladimir Larionov[1], Joan Graves, Natalya Kouprina, and Michael Resnick.

Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, Box 12233 Research Triangle Park , NC 27709. [1]Corresponding author.

While transformation is a prominent tool for genetic analysis and genome manipulation in many organisms, transforming DNA has often been found to be unstable relative to established molecules. We determined the potential for transformation-associated mutations in a 360 kb yeast chromosome III composed primarily of unique DNA. Wild type and *rad52* *S. cerevisiae* strains were transformed with either a homologous chromosome III or a diverged chromosome III from *S. carlsbergensis*. The host strain chromosome III had a conditional centromere allowing it to be lost on galactose medium so that recessive mutations in the transformed chromosome could be identified. Following transformation of a RAD$^+$ strain with the homologous chromosome, there were frequent changes in the incoming chromosome including small and large mutations. Based on results with the diverged chromosome, interchromosomal recombinational interactions were the source of many of the changes. Even though *rad52* exhibits elevated mitotic mutation rates, the percentage of transformed diverged chromosomes incapable of substituting for the resident chromosome was not increased in *rad52* compared to wild-type strain, indicating that the mutator phenotype does not extend to transforming chromosomal DNA. Based on these results and our previous observation that the incidence of large mutations is reduced during the cloning of mammalian DNA into a *rad52* as compared to a RAD$^+$ strain [1], a *rad52* host is well-suited for cloning DNA segments in which gene function must be maintained.

Support was provided in part by the Department of Energy through an interagency agreement with NIEHS and grant (1-YO2-HG-60021-01) from the NIH Human Genome Center to M.A.R.

1. Kouprina, N., Eldarov, M., Resnick, M., Moyzis B., Larionov, V. A model system to assess the integrity of mammalian YACs during transformation and propagation in yeast. (1994) *Genomics*, **21**, 7-17

Ligase chain reaction using colorimetric detection*

Layne Huiet, Annette Tumolo, Luis Ugozzoli, Tony Reyes, Jimmie Lowry, Bruce Wallace and Frank Witney, Bio-Rad Laboratories, Hercules, CA 94547

Ligase chain reaction is an amplification method which involves the ligation of two sets of adjacent oligonucleotides. These pairs of oligonucleotides are complementary to one another and therefore provide the templates for the exponential amplification of the products after repeated rounds of ligation. Use of a thermostable DNA ligase allows the reaction to be cycled in conventional thermocyclers. The specificity of the reaction is such that only oligonucleotides which hybridize without mismatch at their junction are ligated. This requirement for exact base pairing can be used to discriminate single base differences in a target sequence. Here we will present a method for carrying out LCR reactions and the non-radioactive detection of the products of the reaction using both a synthetic model and a human genome model. This is carried out by binding the LCR reaction products containing a biotin moiety to a streptavidin coated microtiter well. An alkaline phosphatase/oligonucleotide conjugate is hybridized to the ligation products bound to the wells and utilized as the reporter molecule with NADPH or pNPP as the substrate. This system has the advantage of a nonradioactive detection which is not gel based, lending itself to automation.

*This work is not supported by DOE

Resource for Molecular Cytogenetics

C. Collins[2], L. Daneshvar[1], Karin Greulich[1], D. Kowbel[2], W.-L. Kuo[1], L Riedell[1], F. Shadravan[2], D. Sudar[2], J. Mullikin[2], S. Lockett[2], P. Yue[1], U. Weier[2], Manfred Zorn[2], D. Pinkel[1,2], J. Gray[1,2].

[1]Dept. Laboratory Medicine, University of California, San Francisco, CA and [2]LBL National Laboratory, Berkeley, CA.

The LBL/UCSF Resource for Molecular Cytogenetics has been created to facilitate the application of molecular cytogenetics in clinical and biological studies. Work is being pursued in three areas: Development and application of improved hybridization technology, selection of probes optimized for use in fluorescence in situ hybridization (FISH), and development of digital imaging microscopy.

Work this year on hybridization technology has focused on development and application of comparative genomic hybridization (CGH), hybridization to thick tissue sections, and hybridization to extended single DNA molecules. CGH allows genome wide screening for DNA sequence copy number aberrations and has been applied in collaborative studies for characterization of whole chromosome and segmental aneusomies in prenatal specimens and in cancers of the breast, prostate, ovary, colon, bladder, lung, skin, brain, and blood. Progress also has been made on development of CGH using arrays of DNA sequences rather than metaphase chromosomes as the hybridization target. FISH to thick tissue specimens permits analysis of the genetic status of individual cells within the tissue architecture, for example permitting study of the genetic evolution of cancer. Hybridization to extended DNA molecules permits rapid determination of the positions of sub segments of a larger molecule, and measurement of the amount overlap of contiguous clones, among other applications.

Human probes for FISH have been selected from chromosome-specific cosmid libraries, YAC libraries and the Du Pont P1 library with primary emphasis on P1 clones. Our goal is to develop probes distributed at ~5 Mb intervals over the entire genome that contain STSs defining genes or genetically mapped polymorphic loci. To date, 233 such clones (218 P1's, 11 YACs and 4 cosmids) have been selected for 133 loci, including 40 genes. Selected clones are mapped according to FLpter using FISH. An additional 178 anonymous clones (75 P1's and 103 cosmids) have been selected at random and mapped. Included are probes for tumor suppressor genes and oncogenes (p53, c-MYC, GLI, SIS, E-Cadherin), translocation breakpoints of clinical significance (PML, RARA, TCRA, ETO, ALL), and regions involved in contiguous gene syndromes (Angelman, Prader-Willi, Cri du chat, Wolf-Hirschhorn and DiGeorge syndrome).

A QUantitative Image Processing System (QUIPS) has been developed in the Resource to facilitate molecular cytogenetic studies. Software currently allows multi-color image acquisition, chromosomal probe mapping (relative to the p-terminus), comparative genomic hybridization, metaphase finding three dimensional image display and computer assisted hybridization domain enumeration. Software for rare event scanning and three dimensional image segmentation and automated domain enumeration is under development.

Information about probes, technologies and software developed by the Resource will be available via an Internet Web Server and, when appropriate, through GDB.

(This work was supported by the US DOE contract DEAC0376SF00098 and Imagenetics)

NONCANONICAL OCTANUCLEOTIDE SEQUENCES RECOGNIZED WITH THE POU DOMAIN OF OCT-2B FACTOR

Alexander G.Stepchenko, N.N.Luchina and Oleg L.Polanovsky

Engelhardt Institute of Molecular Biology; Russian Academy of Sciences; Moscow 117984, Russia.

The Oct proteins belong to the important class of transcription factors involved in the tissue specific gene expression, cell differentiation and development. All these factors contain a POU domain recognizing a conservative oct sequence ATGCAAAT in the regulatory regions of target genes.

With a random modification method it was shown [1] that noncanonical sequences are targets for mouse POU domain, Oct-1 and Oct-2B transcription factors. These sequences were subdivided into two groups: one group containing octanucleotide related sequences, and the other group having tetranucleotide TAAT. The methylation interference assay was used to establish what nucleotides of target sequences are in contact with the POU domain. This analysis revealed that the core (6-7 bp) of octamer related sequences was essential for protein/DNA interaction. In the second group of target sequences the nucleotides recognized with the POU domain may be separated with a spacer (up to 4 bp) [1, 2]. From obtained data follows that POU domain interacts with some targets as an entire entity and with others "homeo" type interaction is prevalent.

Our results indicate that the Oct proteins interact with canonical octanucleotide and partially degenerated oct-sequences. These data have greatly increased the number of potential targets for Oct proteins on DNA and changed our view on the gene-expression regulation by these protein factors.

This work was funded by the DOE Genome Program, A.S. and N.L.were supported by the Human Genome Program (Russia).

[1]. Stepchenko A.G. (1994) Noncanonical oct-sequences are targets for mouse Oct-2b transcription factor. FEBS Letters, 337, 175-178.

[2]. Stepchenko A.G., and Polanovsky O.L (1994) Oct-Proteins/DNA interactions. Submitted manuscript.

Chemiluminescent Detection of Multiplex Labeled Microsatellite Markers and DNA Sequences

Chris S. Martin, Laurie Butler, Greg Schneider and Irena Bronstein

Tropix, Inc., Bedford, MA 01730

We have developed a technique for sequential nonisotopic detection of multiple sets of DNA reaction products which are labeled with different haptens. This multiplex labeling approach utilizes hapten-specific alkaline phosphatase conjugates and chemiluminescent 1,2-dioxetane substrates. The chemiluminescent detection method generates intense light signals which can be easily imaged with standard X-ray film. For DNA sequencing, multiple primers, each with a unique ligand label, are incorporated in sequencing reactions, the products are separated, transferred to nylon membrane and detected by binding hapten specific alkaline phosphatase conjugates. We have used primers labeled with biotin, digoxigenin, fluorescein and 2,4-dinitrophenyl (DNP), enabling the acquisition of four images of DNA sequence data from a single nylon membrane. The need for large scale screening of polymorphic microsatellite markers for genetic mapping led us to adapt this technique for the detection of PCR amplified microsatellite markers. Individual sets of PCR primers labeled with each hapten are utilized to amplify different microsatellite repeat markers. The amplified markers for each genomic DNA sample are loaded in a single gel lane, electrophoretically separated, transferred to a nylon membrane and detected sequentially with hapten-specific alkaline phosphatase conjugates. Each of the four different haptens have been used for three amplimer pairs, to generate three different size fragments with each label. Thus, 12 different markers can be typed from a single gel lane. In addition, we have evaluated a new 1,2-dioxetane substrate, CDP-Star™, which generates extremely intense light signals. This high light output enables efficient detection of the chemiluminescent signals with low light sensitive CCD cameras and rapid acquisition of digitized images of the data suitable for automated analysis is possible. Chemiluminescent detection of overlapping sets of DNA fragments coupled with multiplex labeling is an efficient, non-radioactive method for PCR product detection and DNA sequence analysis.

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

INTERACTION OF DIMERIC INTERCALATING FLUORESCENT DYES WITH SINGLE-STRANDED DNA*

Hays S. Rye and Alexander N. Glazer, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720.

We have shown that a wide variety of suitably designed dimeric dyes form highly fluorescent complexes with double-stranded DNA (dsDNA) stable to electrophoresis,[1] and that such dsDNA-dye complexes can be exploited for the high-sensitivity determination of dsDNA concentration in solution[2], for multiplex detection of restriction fragments in slab gel and capillary electrophoresis[1,3-5], and in the quantitative study of protein-DNA interactions.[6] Others have demonstrated that these dimeric dyes can be utilized for the ultra-high sensitivity analysis of dsDNA content in cells by flow cytometry,[7] in the analysis of PCR products,[8] and in the examination of the static and dynamic properties of isolated DNA molecules.[9,10]

To extend the range of applications of such dye-DNA complexes, we have initiated studies of the interaction with single-stranded DNA (ssDNA) of dimeric dyes capable of bisintercalation, such as ethidium homodimer (EthD) and thiazole orange homodimer (TOTO), that form stable complexes with dsDNA.[3,11] These studies compared in detail the binding of these dyes to linearized M13 ss and dsDNAs and characterized both types of complexes by absorption, fluorescence, and circular dichroism spectroscopy. Surprisingly, M13 ss and dsDNAs bind the dyes with similar affinity at comparable numbers of high affinity sites. Per bound dye, the fluorescence emission of the ssDNA-dye complexes is about 3-fold weaker than that of the corresponding dsDNA complexes. Both TOTO and EthD form complexes with M13 ssDNA stable to electrophoresis. The ssDNA-dye complexes are more stable at high Na^+ concentrations than the corresponding dsDNA complexes. Such ssDNA-dye complexes, preformed at 1 dye per 15 bases, retain about 50% of the bound dye when challenged with a 600-fold by weight excess of unlabeled dsDNA. These preliminary results indicate that the applications of the fluorescent DNA-dye complexes can be extended to ssDNA-complexes.

*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-FG-91-61125.

[1]A.N. Glazer and H.S. Rye, *Nature (London)* **359**, 859-861 (1992).

[2]H. S. Rye, J.M. Dabora, M.A. Quesada, R. A. Mathies, and A.N. Glazer, *Anal. Biochem.* **207**, 144-150 (1993).

[3]H.S. Rye, S. Yue, D.E. Wemmer, M.A. Quesada, R.P. Haugland, R.A. Mathies, and A.N. Glazer, *Nucleic Acids Res.* **20**, 3803-2812 (1992).

[4]S.C. Benson, R.A. Mathies, and A.N. Glazer, *Nucleic Acids Res.* **21**, 5720-5726 (1993).

[5]H. Zhu, S.M. Clark, S.C. Benson, H.S. Rye, A.N. Glazer, and R.A. Mathies, *Anal Chem.* **66**, 1941-1948 (1994).

[6]H.S. Rye, B.L. Drees, H.C.M. Nelson, and A.N. Glazer, *J. Biol. Chem.* **268**, 25229-25238 (1993).

[7]G.T. Hirons, J.J. Fawcett, and H.A. Crissman, *Cytometry* **15**, 129-140 (1994).

[8]K. Srinivasan, S.C. Morris, J.E. Girard, M.C. Kline, and D.J. Reeder, *Appl. Theoret. Electrophor.* **3**, 235-239 (1993).

[9]I. Auzanneau, C. Barreau, and L. Salome, *Compt. Rend. Acad. Sci.., Ser III* **316**, 459-462 (1993).

[10]T.T. Perkins, D.E. Smith, and S. Chu, *Science* **264**, 822-826 (1994).

[11]A.N. Glazer, K. Peck and R.A. Mathies, *Proc. Natl. Acad. Sci. USA* **87**, 3851-3855 (1990).

PHOTO-ANCHORING PROBES FOR IN-SITU LOCALIZATION. A. I. Poletaev[*], T. V. Nasedkina[#], T. S. Godovikova[*] and D. G. Knorre[*]; [#]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, 117984, Moscow; [*]Novosibirsky Institute of Bioorganic Chemistry RAS, 630090, Novosibirsk, Russia.

A new version of of DNA probe localization technology on chromosomes and chromatin is suggested. It is based on use of the nucleotide analogs able to cross-link with matrixed DNA in the course of photo-reaction stimulated by light of appropriate wavelength. We used phenilazide active group (4-azidobenzoic, 4-azido-2-nitroaniline and 4-azido-2,3,5,6-tetrafluorobenzoic) linked in 5-position to deoxyuridine (PAdU). After formation of covalent links between the probe and chromosomal DNA, excess probe can be removed from the sample by extensive washing at high temperature (up to 70°C), thus providing a high contrast and a possibility of conducting the second hybridization. Two versions of this technology were proven: with polymerase reaction (Klenow fragment) primed by hybridized DNA probes of different length and conducting with photo-active analogs of dNTP; and by direct incorporation of photo-active nucleotides in the probes by nick-translation using two types of modified dNTP (photoactive and fluorescent). In the first case only the perfect duplexes formed by DNA probes with matrice supply the polymerase extension of the probes and become anchored to matrice after irradiation, that is more suitable for localization of small probes. Extensive washing removes most of the non-specific signals from the samples. The second approach is more simple for bigger probes. Both technologies will be illustrated by examples. The further development of this approach might provide a possibility of localization of very short probes both on chromosomes and on interphase chromatin. This work was supported by a grant from Russian State Program "Human Genome".

Comparative Analysis of Human DNA Variations by Fluorescence-based Sequencing of PCR Products

Pui-Yan Kwok,[1,2] Christopher Carlson,[1] Thomas D. Yager[3], Wendy Ankener,[1] and Deborah A. Nickerson[1]

[1]Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195
[2]Division of Dermatology, Washington University School of Medicine, St. Louis, Missouri 63110
[3]Division of Nephrology, Hospital for Sick Children, Toronto, Canada M56 1X8

Automated, direct cycle sequencing of purified double-stranded PCR products using Taq polymerase and fluorescently-labeled dideoxynucleotide terminators provides a robust and highly reproducible method to identify DNA sequence variations in sequence-tagged sites (STSs). We have developed a simple and sensitive approach that is capable of identifying a single heterozygote among homozygotes even when the standard base-calling program (ABI 373 analysis program) is unable to distinguish between these individuals. Using this approach, we have scanned for the presence of common DNA sequence polymorphisms in more than 200 STSs from the human genome. We have found that three criteria are important in accurately calling heterozygous bases: 1) a greater than 40% decrease in normalized peak height of the heterozygous nucleotide compared to the homozygous nucleotide, 2) the appearance of a new and significant peak at the location of the heterozygous nucleotide, and 3) a significant change in the normalized peak height of the nucleotide immediately adjacent to the polymorphic site. With these criteria, we now routinely detect heterozygous bases with negligible false-positives. We have also extended the use of quantitative peak height analysis by comparing sequencing traces from a heterozygous individual to that obtained from a pooled DNA sample. This comparison provides a system for rapidly determining allele frequencies for single nucleotide polymorphisms in a population of interest, and also provides a method for rapidly comparing allele frequencies in regions of genetic or evolutionary interest among different human populations.

Alu Repeats: Source for the Genesis of Simple Sequence Repeats

Santosh S. Arcot [1], Hernan Bazan [2], Prescott L. Deininger [2] and Mark A. Batzer [1]

[1]Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550. [2]Department of Biochemistry, Louisiana State University Medical Center, 1901 Perdido St., New Orleans, LA 70112.

Repeated DNA sequences comprise almost 50 % of the human genome. These sequences may be divided into those that are tandemly arrayed and those which are interspersed. *Alu* sequences, the most common form of Short INterspersed Elements (SINEs) found within the human genome, are present at a copy number in excess of 500,000/haploid genome. Tandemly arrayed repetitive sequences may be divided into two classes based upon the size of each repeat unit and are generally referred to as Simple Sequence Repeats (SSRs). Minisatellites are composed of repeat units that are greater than 9 bp each while microsatellites have basic repeat units of 9 bp or less. The size and high degree of heterozygosity of microsatellites and minisatellites has made them one of the most important DNA based markers for genetic linkage mapping and forensic identity testing. *Alu* repeats integrate into AT rich regions of the genome, are flanked by short intact direct repeats (formed from the preintegration site) and also contain a characteristic 3' oligo-dA rich tail. In recent years, a number of SSRs have been found adjacent to the 3' end of *Alu* family members. The association between SSRs and *Alu* sequences may occur through the insertion of one repeated DNA sequence (*Alu* element) near a SSR, or be a direct result of mutations which occur in the 3' oligo-dA rich tail after the insertion of the *Alu* SINE followed by the expansion/contraction of these regions.

In order to resolve between these two models for the association of *Alu* sequences and SSRs we performed a DNA sequence analysis of several recently inserted human-specific (HS) *Alu* repeats with SSRs located in the 3' flanking region from both human and non-human primate (chimpanzee) genomes. DNA sequence analysis of the chimpanzee orthologues demonstrated that they did not contain a SSR. These data indicate that the SSRs adjacent to these HS *Alu* sequences arose as a result of mutations within the oligo-dA rich tails of the *Alu* repeats. Therefore, the 500,000 *Alu* repeats dispersed throughout the human genome represent a novel source for the generation of new SSRs.

This work was supported in part by grants from the U.S. Department of Energy (LDRD 94-LW-103) to M.A.B. and National Institutes of Health (RO1 HG 00770) to P.L.D and by Lawrence Livemore National Laboratory under the auspices of the U.S. Department of Energy under contract no.W-7405-ENG-48.

Alu Repeats: Tools for Genomics and Modes of Evolution

Mark A. Batzer [1], David H. Kass [2], Michelle Alegria-Hartman [1], and Prescott L. Deininger [2]

[1]Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550. [2]Department of Biochemistry, Louisiana State University Medical Center, 1901 Perdido St., New Orleans, LA 70112.

There are over 500,000 *Alu* repeats dispersed throughout the human genome in a semi-random manner. *Alu* elements serve as priming sites for the amplification of unique DNA sequences located between *Alu* repeats that reside in relatively close proximity in a process termed inter-*Alu* Polymerase Chain Reaction (PCR). Physical mapping of the human genome involves a variety of complex hybridization based procedures. Some of these procedures rely upon the ability to separate human clones derived from human/rodent hybrid cell lines from those that contain background rodent derived DNA sequences. The ability to block the repetitive element (*Alu* repeat) portion of inter-*Alu* PCR products derived from a variety of complex sources is also crucial for the isolation of unique DNA sequences. We have constructed a new consensus *Alu* repeat probe (pPD39) designed for these purposes.

The *Alu* family of repeats can be divided into distinct subfamilies based upon specific diagnostic mutations. Older subfamilies of *Alu* repeats are more prevalent, while the more recent subfamilies have fewer copies. Many of the younger *Alu* elements are absent from the orthologous loci of non-human primates indicating that they arose in recent evolutionary history by retroposition, the predominant mode of Short INterspersed Element (SINE) amplification. PCR analysis of one young *Alu* subfamily (Sb2) member located on human chromosome 19 within the Low Density Lipoprotein Receptor (LDLR) gene revealed the presence of this element at orthologous positions within the genomes of a number of non-human primates. Analysis of the 5' and 3' flanking DNA sequences surrounding this *Alu* repeat showed that mutations had occurred at a neutral rate, in contrast to a higher degree of variation within the *Alu* sequence. The nucleotide sequence of the *Alu* repeat corresponded to an older Primate-Specific (PS) subfamily member within all of the non-human primate genomes. The alteration of this *Alu* sequence from one of the oldest to one of the youngest *Alu* subfamilies apparently occurred by a gene conversion event. Although gene conversions of *Alu* repeats are rare events, these data suggest that such events do occur, and contribute to the evolution of SINEs.

This work was supported in part by grants from the U.S. Department of Energy (LDRD 94-LW-103) to M.A.B. and National Institutes of Health (RO1 HG 00770) to P.L.D.and by Lawrence Livermore National Laboratory under the auspices of the U.S. Department of Energy contract no.W-7405-ENG-48.

YACs with Unique Capture Handles

Lucy L. Ling, Jonathan Norcross, Elizabeth Borsody and Donald T. Moir
Department of Human Genetics & Molecular Biology, Collaborative Research, Inc., Waltham, MA 02154

Yeast artificial chromosomes have been used extensively for physical mapping. However to obtain sequence ready clones, current approaches are to anchor cosmids or P1s representing an entire YAC. Because STS densities even as high as one per 100 kb may be inadequate, this approach will demand considerable effort which could be avoided if YACs could be sequenced directly or subcloned for sequencing. Current YACs are not suitable sequencing templates, mainly due to (1) limited DNA yield because YACs are typically single copy; (2) YAC DNA cannot be separated easily from yeast DNA; and (3) YACs are frequently chimeric. We have focussed on solving these problems.

Recombination within the yeast host strain has been postulated to contribute to the mechanism of chimerism. The contribution of host genetic background to chimera frequency was determined by introducing the same ligation mixture (greater than 400kb) into three isogenic yeast hosts that differ only in genotype at the RAD52 and RAD1 loci. The frequency of chimeras was measured by fluorescent *in situ* hybridization (FISH) to human prometaphase spreads. Fewer chimeric YACs were found in recombination-deficient hosts, in all size ranges (10% in recombination-deficient hosts compared to 47% in wild-type hosts).

Building on the amplifiable YAC copy number system of Smith et al. (1990, PNAS 87:8242) a new YAC vector, pCGS1000, has been constructed with an EcoR1 cloning site flanked by unique restriction sites and polypurine tag sites for capture by triple helix interaction. Resulting YACs can be digested with *SceI* to remove the vector arm releasing only the insert. The copy number of these YACs can be amplified by growth in the presence of methotrexate, sulfanilamide and thymidine. DNA fragments up to 20kb containing these polypurine sequences were captured specifically by triplex affinity to polypyrimidine third strands attached to a solid support. The effect of the location of the polypurine tag on the quality and quantity of capture was analysed. Vector-YAC PCR products representing both ends of the YAC can be captured selectively by using either the left or right tag polypyrimidine sequences and sequenced directly. We are currently investigating methods for capturing the entire YAC insert.

Supported by DOE Grant DE-FG02-92ER61399

Mapping Markers in Genomic Library Clones Using Oligonucleotide Arrays

R. Lipshutz, R.J. Sapolsky, S.P.A. Fodor

Affymetrix, Santa Clara, CA 95051

We have developed DNA chips and accompanying biochemical and informatic methods for detecting markers to order clones from genomic libraries. Through a series of enzymatic reactions, we selectively amplify and fluorescently label only those portions of the high molecular weight DNA which correspond to elements in a set of oligonucleotide markers. The complementary set of oligonucleotides to these markers is synthesized in a high-density array using light-directed combinatorial chemical synthesis. When the labeled target from a clone is hybridized to the spatially arranged probes, a pattern corresponding to the unique subset of markers contained in that clone is generated. The correlation of the hybridization patterns between the marker sets for two given inserts is an indication of their physical overlap. This methodology for screening and physically mapping clones into contigs is relatively fast and straight-forward and can be easily adapted for automation.

This page intentionally left blank.

Informatics

The GDB Human Genome Data Base Version 6.0

Kenneth H. Fasman, A. Jamie Cuticchia, and David T. Kingsbury

Division of Biomedical Information Sciences, Johns Hopkins University School of Medicine, 2024 E. Monument St., Baltimore MD 21205-2100

The Genome Data Base is currently undergoing a major restructuring. These changes are intended to address a number of long-standing inadequacies in the design and policies of the database. In addition, they will better prepare GDB for participation in the planned federation of biological databases.

Chief among the planned improvements are changes to the representation of mapping data in GDB. The database is being reorganized around a core of derived order and distance data from "first order" genetic and physical maps and "second order" integrated (consensus) maps of the human genome. This information will be reinforced by additional raw mapping data where appropriate. For example, while the CEPH and CHLC databases provide raw linkage data to the public, no equivalent resource exists for YAC/STS content data. GDB V6 will provide one.

The new database will be organized on a revised editorial model which differentiates between original and consensus data. GDB V6 will allow direct user submission and updates on those submissions, along with third-party annotation of those data. In addition, consensus data such as maps and nomenclature and summaries of (possibly conflicting) user data will be maintained by the HUGO editorial committees.

Version 6 of the Genome Data Base is using object-oriented data modeling and software development on both the client and server sides. An object-oriented data model is being implemented on top of a Sybase relational database using the Object Protocol Model (OPM) developed by Victor Markowitz and his colleagues at the Lawrence Berkeley Laboratory. Client software is being written using C++ and a platform-independent development library (Galaxy, Visix Software Inc.) which will allow one set of programs to access GDB data from Windows PC's, Macs, and UNIX X Windows systems.

Additional design improvements will better allow the Genome Data Base to participate in the proposed federation of genomic databases. GDB V6 will feature increased modularity of both applications and databases. The current monolithic database is being replaced by a "mini-federation" of more normalized databases, including separate databases for mapping data, polymorphisms, phenotypes, genomic literature, and contact information. The current monolithic client application will be replaced by a collection of specialized modules that can interact using standard interprocess communication protocols. These modules will then be more easily incorporated into third party software systems. Some isolation of the application software from the schema and database management system will be realized through the implementation of a GDB "object broker." The object broker will provide both query and edit functionality, and can be used to work with the database in both interactive and batch modes.

Data Acquisition and Curation Operations for the GDB Human Genome Data Base

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

Division of Biomedical Information Sciences, Johns Hopkins University School of Medicine, 2024 E. Monument St., Baltimore MD 21205-2100

Just as the techniques for the discovery of mapping information have evolved with the Human Genome Project, so have the methods for acquiring and curating such data. Historically, the responsibility for collating and presenting the data on the human gene map rested with the HUGO chromosome committees and the Human Gene Mapping Workshops. It became clear that as the Human Genome Project grew it would not be possible to consolidate the data on an annual basis. It would rest upon the public mapping databases to continuously collect and curate these data.

The data acquisition strategy of the Genome Data Base has evolved from one of assisting the community in the preparation of data submissions (usually on paper forms) to one which now includes actively scanning the scientific literature for data which were not previously submitted to the database. Since 1992 when the first system of electronic forms was introduced to the community, 77% of the data collected by the Genome Data Base has arrived electronically. However, nearly 95% of all submissions continue to utilize the paper submission forms. One reason for this is the nontrivial investment of resources presently needed to complete an electronic submission. Nearly all electronic submissions have been the product of laboratory databases which produce reports directly in the format of the electronic submission forms. We continue to work with collaborators such as Manfred Zorn of the Lawrence Berkeley Laboratory to produce software so that researchers can submit their data in electronic form jointly to the Genome Data Base and Genome Sequence DataBase. With the increase in the number of tools provided for the creation of electronic submissions, fewer resources should be expended in collecting data from anonymous FTP (file transfer protocol) sites and paper forms.

The HUGO chromosome committees will continue to be responsible for defining the consensus localizations for markers and the consensus maps for each chromosome, as well as oversee human gene nomenclature. To assist them in their interaction with the database, several organizations (the U.K. Medical Research Council, the U.S. Department of Energy, France's INSERM, and the European Union) have funded editorial assistant positions. These individuals assist the Genome Data Base by working with the HUGO editors in their geographic region to update and maintain the database. Additionally, several of the editorial assistants play a role in the scanning of the scientific literature, allowing the Genome Data Base to keep pace with the growing amount of data in the literature. It is hoped that the need for literature scanning will be greatly reduced as the community moves to require researchers to submit mapping data to the database, as they already have for sequence data.

Architecture of the Genome Data Base Version 6.0

Peter Li, David D. Marquette, Christopher W. Brunn, and Kenneth H. Fasman

Division of Biomedical Information Sciences, Johns Hopkins University School of Medicine, 2024 E. Monument St., Baltimore MD 21205-2100

The existing architecture of the Genome Data Base has a number of limitations. The first limitation is its evolving monolithic design -- all data exist in one physical database. This increases maintenance, limits growth, slows deployment, and reduces performance of GDB. The second limitation is its awkward character-based user interface. The lack of a graphical user interface (GUI) makes it awkward for the many GDB users that are familiar with Windows or Mac interfaces. The third limitation is the requirement of SQL for application development. All applications that access GDB must operate at the SQL level and therefore need a comprehensive understanding of the complex relational schema. In addition, run-time copies of SQL-based applications require a license fee.

The architecture of GDB 6.0 addresses the above limitations through the following requirements. The first is an open, federated architecture -- the monolithic database will be partitioned vertically into individual databases such that each can operate independently of the others. This will provide a framework for simpler management, better growth, and faster deployment of GDB. The second requirement is a conceptual modeling of the databases. As DBMS vendors come and go and physical architectures wax and wane in popularity, we need to protect our investment in the domain semantics captured in the database schema. Therefore, we need a conceptual modeling environment that allows migration between DBMS vendors and physical architectures, and allows future adaptation to nonrelational systems. The third requirement is a graphical interface. The interface to GDB V6 should support X Windows, PC, and Mac platforms. To minimize platform-specific code development and maintenance, we also need a multiplatform software development system. The fourth requirement is a method for conceptual data access. In order to reduce the complexity of the schema for application developers and end-users, we want to isolate the physical design from the much simpler conceptual design. Therefore, the interaction between clients (applications) and servers (databases) will be through a layer that translates the physical data to and from the conceptual data. With this approach, a database can be nontraditional but the client can still interact with the database. Furthermore, if the translation layer is kept on the server side, then the end users will not need to pay for application licensing fees.

To fulfill these requirements we are developing GDB 6.0 with a mixture of commercial software, public domain tools, and custom development. The back end issue of conceptual modeling is addressed by the Object Protocol Model tools from Victor Markowitz of Lawrence Berkeley Laboratory. It provides a method for specifying an object-oriented schema and provides a conceptual level API to the resultant physical database. The front end issue of a multiplatform GUI is addressed by the Galaxy programming suite from Visix, Inc. It is a C++ environment that allows one set of source code to be compiled for many computer platforms. The conceptual communication between the front end and the back end is addressed by building an "object broker" that will perform the translations between the physical data and the conceptual data.

THE GENOME SEQUENCE DATA BASE (GSDB)

Jillian Burton, Michael J. Cinkosky, David Crowley, Ada Espinosa-Lujan, James W. Fickett, Timothy Gray, Carol Harger, Mohamad Ijadi, Gifford Keen, Michelle March, Mia McLeod, John O'Neill, Alicia Power, Maria Pumilia, David Rider, Jolene Schwertfeger, Nina Thayer, Jennifer Tipton, Charles D. Troup and Shahar Tsadeek

National Center for Genome Resources (NCGR)
1800 Old Pecos Trail
Santa Fe, NM 87505
(505) 982-7840

The Genome Sequence Data Base (GSDB) is dedicated to supporting scientific research and development by creating, maintaining and distributing a complete, timely, accurate and useful collection of DNA sequences and related information. As an on-line, client-server, relational database, GSDB operates as part of the DOE federated information infrastructure and focuses on meeting the needs of the major genome sequencing laboratories. GSDB is a direct outgrowth of the Los Alamos National Laboratory component of GenBank.

In cooperation with the other major DNA sequence databases (DDBJ, EMBL and GenBank), GSDB collects data directly from authors in many forms, including Authorin submissions. GSDB also supports two new methods of data collection, tailored to the needs of large-scale genome sequencing:

- *Off-site user program.* Anyone connected to the Internet can obtain an account on the GSDB database and run the Annotator's WorkBench (AWB) from their own computer to enter and edit data.
- *Direct database updates.* Centers with in-house Sybase expertise can write special applications that perform updates directly on the master database using client-server access.

Data entered using either of these methods remain invisible to the public until they have passed the GSDB suite of quality control checks.

GSDB may be accessed in the following ways:

- *World Wide Web.* The GSDB Web server (http://www.ncgr.org/gsdb) provides several different access mechanisms, including hyperlinked entry retrieval plus pre-written and *ad hoc* SQL queries.
- *Client-server relational access.* Anyone with a Sybase front-end license may access a read-only copy of the database at the NCGR using either generic database access tools or special-purpose programs.
- *Relational satellites.* For sites with more demanding requirements, local copies of the complete relational database may be installed and maintained automatically using the GSDB database replication system.

Additional information on the Genome Sequence Data Base may be obtained by sending email to gsdb@ncgr.org. GSDB software and documentation, including the complete relational schema manual, may be obtained by anonymous ftp from ftp.ncgr.org or through the Web server.

Algorithms in Support of the Human Genome Project

Dan Gusfield, John Kececioglu, R. Ravi, Jim Knight

Department of Computer Science, University of California, Davis, CA 95616. gusfield@cs.ucdavis.edu
And
Gene Lawler, Archie Cobbs

Department of Computer Science, University of California, Berkeley, CA

Our research covers a wide variety of algorithmic and data structure issues involved in obtaining and analyzing sequence data, in searching databases, in reconstructing evolutionary history from sequence data or from genome rearrangements. The work is both theoretical and applied, and is divided into the following thirteen papers and three programs that we worked on in the past year. Below are the titles of these efforts, which give a fair indication of the thrust and content of our work.

Sequence analysis and database searching:

Fast identification of approximately matching substrings - Cobbs

Improved approximate matching over suffix trees - Cobbs

Faster implementation of a common superstring heuristic - Gusfield

Uniform preprocessing for linear time string matching - Gusfield

Approximate algorithms for multiple sequence alignment - Bafna, Lawler, Pevzner

Computational experience with a branch-and-bound algorithm for maximum-trace multiple sequence alignment - Kececioglu

Genome Rearangements:

Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement - Kececioglu, Sankoff

Efficient Bound for oriented chromosome-inversion distance - Kececioglu, Sankoff

Of Mice and Men: Algorithms for evolutionary distances between genomes with translocation and inversion - Kececioglu and Ravi

Reconstructing a history of recombinations from a set of sequences - Kececioglu and Gusfield

Sequence Reconstruction

Approximate algorithms for multiple alignment to a phylogenetic tree - Jian, Lawler, Wang

Approximation algorithms for tree alignment - Kececioglu, Ravi

Algorithms for laboratory support

A problem in the design of oligo probes for disease detection and in primer selection for PCR - Ravi, Gusfield

Software Development

FLASH: A software tool to find secondary structures and motifs - Knight

A visual tool for construction of secondary structure - Knight

Wparal - parametric alignment in MS-Windows.

Software Tools in Support of Physical Mapping efforts at LANL

The following software tools have been developed to support physical mapping efforts at Los Alamos National Laboratory, Center for Human Genome Studies:

GRAM (Genomic Restriction map AsseMbly)

Soderlund, C., and L. P. MacGavran.

GRAM provides a visualization interface for the alignment of fingerprinted restriction fragments in cosmid clone contigs. GRAM employs a clustering algorithm to merge RFs to a set of unique fragments, and a shuffling algorithm to assemble these fragments into a permutation consistent with the RF content of and implied RF adjacency for the clones of a contig.

MAP

Mundt, M.

MAP depicts contigs of pairwise overlapping clones as an undirected graph and allows the user to edit these depictions. Using interval graph algorithms, MAP highlights forbidden structures in a graph and gives suggestions for repairs. MAP is fully integrated with the Chromosome 16 Physical Map database.

SIGMA (System for Integrated Map Assembly)

Cinkosky, M. J., M. A. Bridgers, W. M. Barber, C. D. Troup, M. Ijadi and J. W. Fickett

X Window-based software tool for creating, editing and viewing integrated genome maps. Combines data from physical, cytogenetic and linkage maps in a single, unified map. (See separate SIGMA abstract and posted Chromosome 16 Physical Map.)

Chromosome 16 Map Browser

Sutherland, R. D.

Map Browser is a point-and-click front-end to the LANL Chromosome 16 database. It presents chromosome 16 breakpoint intervals and all DNA segments lying between them, organized by type. Selecting a segment expands it to provide appropriate detail: Contig membership, YAC hits, breakpoint localization, PCR primers, positioning relations, source data, fingerprint data, gel images, etc.

gridHyb

Beauheim, C., Pecherer, R. M., and Kelly, P.

A software package developed for interpreting hybridization autorads from high-density grids, gridHyb is used to improve the quality of existing maps, and also in the distribution of Chromosome Library clones based on probe data.

These tools will be demonstrated during the Workshop.

Development of a World Wide Web (WWW) HyperText Markup Language (*html+*)Interface to the Lawrence Livermore National Laboratory (LLNL) Human Genome Center

Annette Swartz (swartz3@llnl.gov), Linda Ashworth, and Tom Slezak

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory; Livermore, California 94550.

The World Wide Web (WWW), a network of files connected by hypertext hotlinks, is rapidly becoming the preferred method of navigating the Internet and accessing information stored on remote computers. WWW traffic increased 2500% from June of 1993 to June of 1994. Thus, access to LLNL Human Genome Center data and research via the WWW would facilitate a more rapid dissemination of chromosome 19 mapping information and resources into the scientific community and also to the general public.

The LLNL Human Genome Center is connected to the WWW through a NCSA httpd1.3 server running on a UNIX machine. The Center's web pages provide both an overview of the research activity being done at the LLNL Human Genome Center and a detail view of specific genomics projects. The current version of the metric physical map of chromosome 19 is available through *html+* form-based queries and *cgi* server scripts. This map consists of a set of ordered clones, many of which have FISH distance estimates between them, to which over 300 attributes have been assigned via hybridization, STS mapping and PCR screening. Also online is an ideogram image depicting the band localizations of all chromosome 19 genes in the LLNL database. Genes present on the ideogram are hypertext hotlinked to their Genome Data Base (GDB) entry. GDB gene entries with GenBank and/or MIM numbers are hypertext hotlinked to their corresponding GenBank and/or OMIM entries. Thus the LLNL *html+* interface is a useful complement to GDB since using a WWW client (e.g. Mosaic, MacWeb, WinWeb, Lynx), it is possible to rapidly access LLNL Genome Center data and retrieve GDB, GenBank and OMIM entries. In addition, *html+* image maps (ISMAP) and *cgi* server scripts provide textual descriptions of objects and activities depicted in composite inline graphic images. Direct WWW correspondence with Human Genome Center scientists is supported by an *html+* email form and *cgi* server script.

Future versions of the LLNL Human Genome Center *html+* WWW interface will be based on a redesign of the current chromosome 19 database schema able to store data from multiple genomes. The *html+* interface will support form-based queries of the redesigned database. For example, given an attribute of interest (e.g. gene, D-segment number, clone id), contig and/or restriction map information related to the attribute will be returned. A prototype of the LLNL Human Genome Center *html+*interface will be available for demonstration purposes at the DOE Human Genome Project Contractor and Grantee Workshop.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

The Expressed Gene Anatomy Database (EGAD) project

Chris Fields, Owen White, Carol Bult, Mark D. Adams, and Anthony R. Kerlavage

The Institute for Genomic Research, 932 Clopper Road, Gaithersburg, MD 20878

The Expressed Gene Anatomy Database (EGAD) is a relational database that links genes to mRNA transcripts and protein products, and to data on gene expression, the biochemical activities and functional roles of proteins, and sequence alignments defining isology classes. EGAD is implemented in SYBASE. Views of EGAD have been populated with data on published human transcripts and their products, and with data supporting metazoan and microbial phylogeny projects. User interfaces to EGAD include a sequence viewer and annotation editor, and a flexible graphic query generator and browser capable of linking any SYBASE databases.

EGAD is designed to provide highly-annotated links between multiple databases in a relational federation. The EGAD development effort includes refinement of the existing EGAD structure, semantics, and vocabularies, and the development of cross-references, cross-database join capabilities, and other interoperability procedures with other biological databases.

A Prototype 2nd Generation Database schema for large-scale Physical Mapping

Tom Slezak (slezak@llnl.gov), Mark Wagner, and T. Mimi Yeh

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550

The Human Genome Center at LLNL has been developing a relational database for the last 5 years to support our work on chromosome 19. We had deliberately chosen to make this database specific to the immediate task at hand since we knew that we would gain much experience before we needed to expand to process other genetic real estate.

Closure of the physical map of ch19 is now in sight and we are preparing to expand our mapping work in a wide variety of directions, including (portions of) human ch2, certain gene families across the entire human genome, regions of interest in other genomes with conserved synteny with respect to humans, and possible (eventual) work on various bacterial, plant, and other animal genomes. Separate databases for each chromosome or genome are not well-suited for the comparative biology that lies in our future. We need to scale up our database to be able to handle queries on physical mapping data that span all our objects and their inter-relationships regardless of species.

To accomplish this we are drawing on concepts employed in our #generic map object" database for automated map integration on ch19 as well as ideas taken from the excellent GSDB schema (Michael Cinkosky, et. al.). Major concepts include: All objects will have a unique, permanent identifier; entries will be labeled with owner(s) and date(s); objects and relations will be highly abstracted (single base-class tables for all clones and probes, a single hybridization results table, etc.); laboratory notebook details will be cleanly separated from results, storage, maps, citations, taxonomy, and attributes. This approach will allow us to reference external databases when adequate ones are available (e.g., citations, taxonomy, sequence) for on-line access. It also allows us to make major changes in one component (say, storage) without causing collateral damage to unrelated portions of the schema.

A key point of our design has been to focus on capturing appropriate abstract atomic relationships, which allow subsequent creation of multiple (possibly conflicting) definitions of higher-level map objects. Unlike traditional object-oriented approaches, this relation-oriented approach has proven in our hands in our ch19 database to be highly flexible with respect to changes in biology, user needs and the desire for multiple simultaneous viewpoints. Our early experience with a prototype of this schema will be detailed.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory contract no. W-7405-ENG-48.

Digiboard - A General Purpose Measurement System Used for FISH Mapping

David J. Ow (ow1@llnl.gov), Arthur Kobayashi, Mark C. Wagner, and Tom Slezak

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

We have developed a software program, called Digiboard, which is being used for measurement and ordering in Fluorescence in situ Hybridization (FISH) mapping efforts supporting LLNL's chromosome 19 project. This data has provided the backbone of the physical map by providing both order and distance (see posters by L. Ashworth, et. al., B. Brandriff, et. al., and M. Wagner, et. al. for details).

Slide images are projected onto a 3' by 4' digitizing tablet. The system is calibrated, and then distance measurements from the projected image are collected by selecting end points or by tracing segments using a hand cursor with crosshairs. The measurements are transferred from the tablet to the computer via a serial interface. Sets of measurements, in a user-defined format, can be generated and stored for each experiment. The data sets are organized in a hierarchical manner, and measurement information may be exported to data files for hardcopy output or for post processing by other programs running locally or remotely over the network.

This program is written in C and runs under the X Window system on a Sun Sparc workstation using a Motif based interface. The user interface portion was created using a commercial Graphical User Interface (GUI) builder. By design, the program can be modified to accept measurements or data from other types of input devices. This system was built to solve immediate needs for data measurement until an adequate multi-color digital image processing system can be completed to handle all the demanding requirements of our FISH work.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory contract no. W-7405-ENG-48.

Recent Enhancements to the LLNL Genome Browser: Automated Construction & Display of Physical Maps

Mark Wagner (mwagner@kooler.llnl.gov), Tom Slezak, and Elbert Branscomb

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

The LLNL Genome Browser has undergone extensive enhancement since the last Workshop in February 1993. Our design seeks to provide access and visual display of all data generated by our Human Genome mapping project free of any prejudgment as to what questions are of interest or what objects, properties, and relationships should be displayed at any time. We use a flexible object-based mechanism to track over 600,000 relationships on over 185,000 potentially-mappable objects on human chromosome 19. Rather than forcing all mapping data into a single picture we use a linked collection of displays, each tuned to the needs of a specific level of detail.

New displays have been developed that enable us to manipulate integrated maps of chromosome 19. The partial order map portion of the browser relates our cosmid clones, some of which are partially ordered via FISH and genetic means, to larger entities (YACs, BACs, PACs, cosmid contigs, and Restriction Fragment maps). Although this map does not take into account distance and length information, a "metric map" module adds this extra and often conflicting data. The first generation assembly algorithms used for this construction utilize an ordered greedy approach to position the objects relative to each other. We will continue to refine and expand our capabilities in this area in the coming year.

Restriction Fragment Maps are also generated and displayed, with fragments containing attributes, such as gene exons, indicated. Well over half of ch19 is now mapped at this level of detail.

Relationships between types of maps are indicated through the use of a display that permits the user to see all possible map types that a selected object appears in, and permits the user to view that object instantly in those other maps. For example, this permits the user to select a cosmid in a contig, note that the same cosmid is in a restriction fragment map, and see that cosmid in the map with a single mouse click. Our current version of the browser is in daily use by over 40 biologists at LLNL and is also being used by our collaborators at ORNL, exploiting the client/server design.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory contract no. W-7405-ENG-48.

Integrated Informatics Support for Large-Scale Sequencing at the LBL Human Genome Center

E. Theil, A. Aggarwal, D. Davy, F. Eeckman, T. Fleming, V. Markowitz, J. McCarthy, S. Pitluck, E. Veklerov, M. Zorn

Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720

After approximately two years of effort, the LBL Human Genome Center has demonstrated that it can sequence DNA at a sustained rate of approximately 750kb per each six person sequencing team. The Informatics Group has supported this project with a variety of software programs and databases, most of which have been described at earlier conferences or are detailed in separate abstracts for this meeting.

LBL now is planning to scale up its sequencing effort significantly. In order to accomplish this while continuing to lower the cost per base, it will not be enough to merely hire additional sequencing teams and support them with existing software. What is required instead is a more fully integrated and automated system in which the data produced by the directed strategy developed at LBL is both captured and modeled in software, so that all the information generated prior to sequencing can be exploited subsequently, in assembly and dissemination. Furthermore, this more highly integrated system will help to increase the level of quality control on data and operations by identifying errors and providing computer assistance in troubleshooting. We do not believe that there are general solutions available, but we are able to use a number of robust software components that have been developed elsewhere in conjunction with our own software to form an integrated system tailored to our own particular needs.

We discuss a system architecture that is designed to capture data either automatically or with manual input and which is gradually replacing personal laboratory notebooks with a unified view of the data at any moment. The first pieces of this system will consist of modules for dealing with automatic inspection of ABI sequencing runs, automatic trace cutting and browsing. Other modules recently introduced are automatic generation of transposon maps and the ability to perform post mortems by comparing maps based on actual sizes of sequenced clones with those based on estimated sizes from electrophoresis. This helps to reduce the number of gaps encountered when assembling the sequence.

Another important component now under development is the introduction of a figure of merit associated with each base call. This will be a significant aid as we move to more automatic editing of sequences and the systematic demonstration of quality in sequenced data.

In order to integrate information from and for these modules, our Syndb database will store operational data, finished sequence, and up-to-date maps, all linked to each other. Application modules will communicate with the database to both access and update data as it is produced. Syndb will also support more than one analysis of the basic data (typically gels) in order to troubleshoot inconsistencies (typically false positives) as required.

ON GENERATING ACCURATE TRANSPOSON MAPS

E. Veklerov, C. Martin, E. Theil
Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720

Generation of transposon maps is an important step in the directed sequencing strategy used at LBL. The accuracy of the map affects the success of the assembly process, and thus the per-base cost of the final sequence. We have designed and developed a software package called TRAMP that generates such maps. It also helps researchers study the algorithms involved in this process in order to explore and optimize them.

TRAMP is a highly interactive software package written in Smalltalk,which is an object oriented language that has convenient facilities for handling graphical interface applications. The central part of the package is its selection algorithm which finds a subset of the inserts that, being subsequently used as the priming sites, has a high probability of covering the clone without gaps. At present, manually generated maps alone result in an average coverage (doublestranded) of 93% of an individual 3kb clone. Automatically generated maps take much less time to generate, are more accurate and should reduce both the number of gaps and their average length. We have developed an algorithm that finds a minimum number of inserts which cover the clone and is optimal in the sense that these inserts are distributed as uniformly as possible within the clone.

Another function facilitated by TRAMP is "post-mortem" studies of the transposon maps. After a clone is fully assembled, the positions of the inserts are known exactly and can be compared with the orioginal positions that were estimates based on PCR reations..A preliminary study has shown that the discrepancies between the estimated and actual positions may be on the order of about 10%. A comparison of "before" and "after" maps will help to identify sources of systematic errors, eliminate at least some of them and eventually reduce the cost of sequencing.

Syndb: A Multi-Species Database For Mapping and Sequence Information

A. Aggarwal, D. Davy, J. McCarthy, E. Theil

Human Genome Informatics Group, Lawrence Berkeley Laboratory, Berkeley, CA 94720

Syndb is a variant of ACeDB, a suite of database and display software originally developed by Richard Durbin and Jean Thierry-Mieg to meet the needs of the C. elegans genome research community. Syndb extends the functionality of ACeDB and builds on 21Bdb, another LBL variant of ACEDB used to maintain a public version of LBL's Chromosome 21 Project data.

Currently under active development, Syndb's purpose is to provide additional tools to meet the evolving requirements of LBL's mapping and sequencing projects. It differs from its predecessors primarily in the addition of a new display module which allows the simultaneous viewing of multiple maps of different species from within the same database or of different levels of physical detail of a single map. Also, new data structures have been added to display direct homology comparison in the multiple species display.

Syndb uses ACeDB's basic data structures, which allow all parts of the database to be easily cross-referenced, and ACeDB's user interface, which permits exploration of data via "point and click" with the computer mouse. Other ACeDB modules allow users to display and search DNA sequences for open reading frames, genes, and other features. To these core ACeDB features, LBL has contributed a versatile query-by-example facility, a mechanism for providing on-line descriptions of data fields, and the new multi-level, multi-map display module. Comments and suggestions on all aspects of these new map display capabilities, from the underlying data structures to the visual appearance of the interface are earnestly solicited.

Syndb has replaced 21Bdb as the primary mechanism for providing data from LBL's physical mapping and sequencing projects to other researchers throughout the world. Like 21Bdb, remote access to the full Unix version of Syndb will be available to anyone on the Internet running X-windows software, or via anonymous FTP for remote installation, and as part of the World Wide Web (WWW) via "front-end" ACEDB gateway software developed by Guy Decoux at INRA in France.

Since 1992, LBL has been a major contributor to the ongoing development of ACeDB and related tools, along with its original authors, and a growing user community. Support for this fruitful collaboration comes from the Department of Energy's Human Genome Project; the U.C. Berkeley Drosophila Mapping Project sponsored by NIH; and the U.S. Department of Agriculture's Plant Genome Database Project at the National Agricultural Library. External data sources include Genbank, the Genome Data Base (GDB), the International Genome Database (IGD), Genethon, CEPH, and Jackson Laboratory's Encyclopedia of the Mouse.

Developing a Graphical SQL Editor for the Genomic Database Federation

Dong-Guk Shin

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

We consider that the genome database federation is a fairly straightforward engineering exercise providing (i) appropriate inter-database reference pointers are well maintained, and (ii) all the participating DBMSs are relational and support appropriate APIs. One potential problem would still persist, however. A user who is not familiar with a third party database in the federation would have difficulty understanding the schema of that database. Consequently, he may not be able to import relevant database schemas among many available ones and may have difficulty forming correct SQL expressions. In a federated database environment, this problem of dealing with unfamiliar third party database schemas becomes much more severe due to the manifold added complexity.

The unfamiliar schema problem is caused by the inherent semantic discrepancies that exist between the way the user perceives real world objects and the way the DBA models them into relational forms. We do not currently know how to build an intelligent query interface that can automatically transform the user's need for data into a structural database query expression. The burden for this conceptual transformation is still on the user. The best, short-term alternative solution would be to develop tools and environment that ease the user's query formulation process.

Our approach is to develop tools and an environment in which users can learn and/or examine the third-party database schema in a relatively short time and can produce a correct SQL expression easily. Specifically, our goal is to allow a user to form a distributed SQL query in a graphical fashion within 5 - 10 minutes' time frame despite his unfamiliarities with third-party database schemas of the federation.

(This work was performed during the author's sabbatical leave at Genome Data Base using computing resources provided by Genome Data Base. This work was supported by National Center for Human Genome Research, National Institute of Health, HG00772-01 and a grant from the University of Connecticut Research Foundation.)

An Integrated Approach to Genomic Sequence Analysis

Terry Gaasterland
Natalia Maltsev
Ross Overbeek

Mathematics and Computer Science Division
Argonne National Laboratory

We are working on various projects with the objective of developing an integrated framework for genomic sequence analysis. One effort has involved supporting the analysis of *Mycoplasma capricolum* sequence data (produced by Pat Gillevet of NIH and Walter Gilbert of Harvard). This data was given to us in the form of 372 contigs representing 214kb of sequence. Our analysis produced a set of detailed results (prediction of over 160 specific coding sequences) and a general-purpose tool which we are now using to analyze microbial sequence data.

Other projects include (1) construction of a phylogenetic tree based on rRNA, containing over 2700 taxa, using maximum likelihood (working with Gary Olsen and Carl Woese of the Ribosomal Database Project); (2) development of an integrated database containing sequence data from over 600 microbial organisms, protein sequence and motif data, phylogenetic trees, and alignments; (3) design of an automatic system for analyzing microbial sequence data; and (4) production of a framework to support analysis of organism metabolism (working closely with E. Selkov, developer of the EMP database on enzymes and metabolic pathways).

New Techniques for Integration of Biological Data Sources

Peter Buneman[1], Susan Davidson[1], Chris Overton[2]

[1]Department of Computer and Information Science, University of Pennsylvania, [2] Department of Genetics, University of Pennsylvania School of Medicine

Much biological data resides in sources that are not conventional databases. Examples include NCBI's ASN.1 database, ACEDB, numerous text files, and the output of sequence matching algorithms (FASTA and BLAST). Such sources cannot be queried by conventional database languages nor are there systems that facilitate the restructuring from one of these formats to another. To this end, we are developing related systems. One is a general-purpose query language, CPL, that subsumes existing database languages and provides interfaces for these varied sources. The second is a transformation tool, TSL, that allows the declarative specification of how one data source may be mapped onto a second.

CPL is a language for data access that has evolved from some basic ideas in category theory. It allows us to generalize relational query languages to a much wider range of data types, including those used in the sources mentioned above and to object-oriented database management systems. For example, our system can answer queries such as: "For sequences in the interval p11.1-q13.2 of human chromosome 22, find all alu elements internal to a gene domain". This query is answered first by retrieving map information from the Genome DataBase, and sequence annotation from GenBank (relational or ASN.1). Extraction of sequence corresponding to the primary transcript of each gene is done by the application program QGB (developed locally). And finally, FASTA is used to compare these sequences with a database of alu elements. These programs were easily added to our system, and are called in the same way as queries to data sources.

Data restructuring is one of the most common and difficult tasks in the current biological database environment. Databases are usually designed through a "conceptual modeling" tool and then translated into some practical database management system. The problem is that people want to reason about, transform, and query the conceptual structure, while actually manipulating the physical representation. Current techniques for doing this are largely based on variants of the entity-relationship model, a model open to semantic misinterpretation and one that fails to take account of an adequate variety of data types. We have developed transformation techniques and prototype tool, TSL, for schema and data transformation that, like CPL, is based on the natural semantics of the underlying data types.

These tools greatly enhance our ability to transform, integrate and access heterogeneous data sources. While they may be used for the physical construction of a monolithic database from the existing data sources, we believe this new approach to database languages calls into doubt the advisability of constructing large, inflexible databases.

[1] S.B. Davidson, A.S. Kosky and B. Eckman, "Facilitating Transformations in a Human Genome Project Database". To appear at Conference on Information and Knowledge Management, 1994.

[2] P. Buneman, L. Libkin, D. Suciu, V. Tannen and L. Wong. "Comprehension Syntax", *Sigmod Record* 23(1):87-96, March 1994.

[3] G.C. Overton, J.S. Aaronson, J. Haas and J. Adams, "QGB: A System for Querying Sequence Database Fields and Features." *Journal of Computational Biology*, **Vol 1(1)**, 3-13 1994.

Object-Protocol Model Tools: Overview and Applications

Victor M. Markowitz, I-Min A. Chen, and Ernest Szeto

Data Management Research and Development Group, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, California 94720

Genomic applications require facilities for keeping track and manipulating data. Data need to be described, presented, browsed and queried in a way that scientists can understand. Commercial database management systems (DBMSs) provide data management facilities, but do not provide capabilities for describing, browsing, and querying data in terms familiar to scientists. Thus, both relational and object-oriented databases can hardly be comprehended by scientists and entail tedious, error-prone, and time-consuming database development and maintenance. Consequently, there is a need for data management tools that would allow scientists interact efficiently with both relational and object-oriented DBMSs using clear and concise constructs and operations.

We have developed data management tools that facilitate the development of genomic databases and that allow scientists to define, query, and browse genomic databases in terms of application-specific objects and protocols. These tools are based on a data model called the Object-Protocol Model (OPM) developed by us [1]. OPM provides constructs for directly modeling objects and protocols (laboratory experiments) specific to genomic database applications, for specifying protocols in terms of alternative and sequences of component protocol steps, and for specifying different database views. Furthermore, OPM supports modeling inter-database references that allow developing multidatabase systems and provides object versioning constructs that allow keeping track of the history of objects in a genomic database.

The suite of OPM data management tools currently target relational DBMSs, and include a reverse engineering tool for determining the OPM description of existing relational genomic databases, a graphical editor for specifying the structure of genomic databases, a translator that maps OPM specifications into DBMS specifications, a graphical browsing and data entry interface for browsing and updating genomic databases. The OPM tools provide version management and support for database replication. In future, we plan to port the OPM tools to an object-oriented or object-relational DBMS.

OPM and the OPM data management tools have been used for developing a prototype database for the large scale sequencing project in Dr. Lee Hood's laboratory at University of Washington, Seattle, and are currently used for developing version 6 of Genome Data Base (GDB) at Johns Hopkins School of Medicine, Baltimore.

OPM and OPM data management tools are overviewed in [2]. This overview, other OPM papers, OPM documents, and the OPM tools are available via World Wide Web using URL: ftp://gizmo.lbl.gov/pub/DM_TOOLS/OPM/opm.html.

This work was funded by the Human Genome Program of the Office of Health and Environmental Research, U.S. Department of Energy under Contract DE-AC03-76SF00098.

[1] Chen, I.A., and Markowitz, V.M., Modeling Scientific Applications with an Object Data Model, to appear in Proceedings of the 11th International Conference on Data Engineering, 1995.

[2] Chen, I-Min, and Markowitz, V.M., An Overview of the Object Protocol Model (OPM) and OPM Data Management Tools, TR LBL-33706, Lawrence Berkeley Laboratory, 1994.

SubmitData: Data Submission To Public Genomic Databases*

Manfred D. Zorn

Software Technologies and Applications Group, Information and Computing Sciences Division,
Lawrence Berkeley Laboratory, Berkeley CA 94720

Making information generated by the various genome projects available to the community is very important for the researcher submitting data and for the overall project to justify the expenses and resources. Public genome databases generally provide a protocol that defines the required data formats and details how they accept data, e.g., sequences, mapping information. These protocols have to strike a balance between ease of use for the user and operational considerations of the database provider, but are in most cases rather complex and subject to change to accommodate modifications in the database.

SubmitData is a user interface that formats data for submission to GSDB or GDB. The user interface serves data entry purposes, checking each field for data types, allowed ranges and controlled values, and gives the user feedback on any problems. Besides one-time submissions, templates can be created that can later be merged with TAB-delimited data files, e.g., as produced by common spreadsheet programs. Variables in the template are then replaced by values in defined columns of the input data file. Thus submitting large amounts of related data becomes as easy as selecting a format and supplying an input filename. This allows easy integration of data submission into the data generation process.

The interface is generated directly from the protocol specifications. A specific parser/compiler interprets the protocol definitions and creates internal objects that form the basis of the user interface. Thus a working user interface, i.e., static layout of buttons and fields, data validation, is automatically generated from the protocol definitions. Protocol modifications are propagated by simply regenerating the interface.

The program has been developed using ParcPlace VisualWorks and currently supports GSDB and GDB data submissions.

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

Data Management for the Molecular Cytogenetic Resource*

Jenny E. Marstaller and Manfred D. Zorn

Software Technologies and Applications Group, Information and Computing Sciences Division,
Lawrence Berkeley Laboratory, Berkeley CA 94720

The LBL/UCSF Resource for Molecular Cytogenetics has been created to facilitate the application of molecular cytogenetics in clinical and biological studies. Work is being pursued in three areas: Development and application of improved hybridization technology, selection of probes optimized for use in fluorescence in situ hybridization (FISH) and development of digital imaging microscopy. All of these areas entail creation and manipulation of large images and other laboratory data. Our group is focussed to provide data management support for all the activities in the Resource.

To facilitate the free data exchange between researchers at UCSF and LBL which are a few miles apart we developed a Mosaic interface to access and modify information using the World Wide Web. The data are located on a central database. The Mosaic client allows to formulate retrieval and edit operations that are sent to the database. Results are filtered through a Perl script which generates HTML documents with Hypertext links that are sent back to the Mosaic client. We plan to make data from the Resource available using a similar mechanism that is open to outside access.

Probe information and mapping data from the Resource will be submitted to the public databases, i.e., GDB. In a collaboration with GDB we are developing a data submission tool (see separate abstract by Manfred Zorn) to facilitate the distribution of our research results.

In order to handle large amounts of images we are developing an image annotation database. The images themselves are automatically transferred to the LBL Mass Storage System. The annotation will be reformatted and loaded into a relational database to allow efficient query processing.

We will present an overview and the current status of our work.

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

Automatic Generation of User Interfaces for Genomic Databases*

Manfred D. Zorn, Shirdi R. Prem, Ofer Ben-Shachar[1]

Software Technologies and Applications Group, Information and Computing Sciences Division,
Lawrence Berkeley Laboratory, Berkeley CA 94720
[1]Software Xcellence, Palo Alto, CA

Databases for genomic data are subject to continuing evolution to cope with scientific advances. Modifications in the database definition invariably trigger changes in the user interface. Thus a significant effort is spent in constantly adapting the database to new requirements and the user interface to the modified database definition. To break this vicious cycle we have developed a user interface that is automatically generated from the database definition, i.e., the metadata.

A generic user interface guides the user through a standard flow of actions, from object selection and query formation to viewing the details of an object and following connections to linked objects. A plain text configuration file read upon program start-up provides information for a specific database, e.g., names of objects, attributes, labels, thus customizing the generic interface for a particular application. This creates an object-oriented view of a relational database implemented using an Extended Entity Relationship model.

The configuration file has been automatically created from the metadata. It defines object appearance in the user interface, defines mappers that translate between the database representation of objects and the interface representation, database specific help, and provides for extensive user customizing. In addition, the configuration file defines stored procedures that retrieve data from the database. These procedures are generated from metadata using a query language, Complex Object Query Language (COQL), based on the same Extended Entity Relationship model that has been used in the database design. COQL queries are subsequently translated into SQL procedures.

Thus a working user interface, i.e., static layout of buttons and fields, data retrieval from the database, data conversion between the database output and the user interface data structures, is automatically generated from the database definitions. Modifications are propagated by simply regenerating the interface. This approach, originally developed for the Chromosome Information System, has been applied to an image database, and several other databases.

We will present the current state of the user interface and tools to aide in defining and manipulating the configuration file.

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

Identification, organization and analysis of mammalian repetitive DNA

Jerzy Jurka
Linus Pauling Institute of Science and Medicine, 440 Page Mill Rd, Palo Alto, CA 94306.

There are three major components of this project: organization of databases of mammalian repetitive elements; development of specialized software for analysis of repetitive DNA and discovery and comparative studies of mammalian repeats. Our project has been advanced in all the three areas.

We have expanded previously published reference collection of human repeats. We have also compiled and analyzed a collection of the most abundant simple repeats from primates [1]. Both collections are available electronically via the NCBI server. Reference collections for other mammals continue to be developed and we will review their current status during the workshop.

Identification, elimination and basic studies of repetitive DNA at the sequence level may be very labor-intensive without specialized computer software. We have developed a software called 'CENSOR' [2], based on recently described principles for identification and analysis of repetitive DNA [3]. This software will be communicated and released to the public domain during the Human Genome Workshop.

We have discovered several new families of MER repeats from primates and other mammals. We will update this ongoing work during the workshop. Some of the previously reported MER repeats led us to discovery of unusual subfamilies in the L1 family of repeats [4]. We have also characterized the oldest known and one of the largest families of repeats in mammals (MIR family) [5]. Furthermore, we have discovered the youngest known subfamily of human Alu sequences [6]. These findings will be reported and discussed in the context of genomic studies.

This work was supported by the U.S. Department of Energy, Office of Health and Environmental Research, Grant No. DE-FG03-91ER61152.

[1] Jurka, J., and Pethiyagoda, C. (1994) Simple repetitive DNA sequences from primates: compilation and analysis. *J. Mol. Evol.* (in press).

[2] Jurka, J., Klonowski, P., Dagman, V., and Pelton, P. (1994) CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. Submitted manuscript.

[3] Jurka, J. (1994) Approaches to identification and analysis of interspersed repetitive DNA sequences. In: *Automated DNA Sequencing and Analysis* (Adams, M.D., Fields, C. and Venter, J.C., eds), Academic Press, pp. 294-298

[4] Smit, A.F.A., Tóth, G., Riggs, A.D., and Jurka, J. (1994) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. Submitted manuscript.

[5] Jurka, J., and Batzer, M. (in preparation)

Recursive Relational Representation for DNA and Attribute-Value Lists: Techniques for Reducing Schema Modifications

Pecherer, Robert M.[1], Joe M. Gatewood[2] and Robert D. Sutherland[1]

[1]Theoretical Biology and Biophysics Group; T-10, MS K710; Los Alamos National Laboratory; Los Alamos, New Mexico 87545. [2]Genomics and Structural Biology Group; LS-2, MS 880, LANL.

Database representations to support genome physical mapping and sequencing necessarily change in response to new protocols, new kinds of data, and changes in data requirements for tools developed to analyze this data. Schemas for new entities and changes to existing schemas may induce changes to browsers, interfaces and existing tools if they are to continue to operate correctly. We propose a Recursive Relational representation for DNA Segments which reduces the impact of introducing new DNA Segment types, and the use of an Attribute-Value List field to eliminate schema modification when new data items are needed in existing schemas.

Genome physical mapping and DNA sequencing are typically performed by recursive decomposition. In a typical physical mapping project, chromosomes are cut into clonable units which are subject to characterization by fingerprinting and/or hybridization probing. These smaller DNA segments (which may themselves be further decomposed to even smaller DNA segments) are collectively analyzed and organized by their characteristics into regionalized and/or overlapping sets (contigs). DNA Sequencing is similar: An interesting segment of DNA is sequenced in a series of experiments which generate nucleotide sequence data for consecutive "steps" obtained by "walking" along the segment with the pcr reaction. A consensus sequence is obtained by assembling the sequences obtained for the individual pieces.

With respect to a database representation for modeling the physical decomposition and logical reassembly for both physical mapping and sequencing, the traditional approach (used for example in the original, Los Alamos "Lab NoteBook" and currently in GDB) has been to design one database table for each type of DNA segment: chromosome, yac, cosmid, contig, restriction fragment, sequencing experiment, etc. These individual tables are then connected by specialized "linking" tables which associate (for example) the cosmid clones in one table to the containing contig in another. However, except for size ranges, each of these units is fundamentally the same: A segment of DNA. We propose a generalized relational schema for "DNA Segment", illustrate its use for several sequencing applications, and consider the database consequences for users and systems.

A related database problem concerns the introduction of new data items (or modification or deletion of existing data items) for existing database entities. Schema modification typically has a rippling effect on all software that accesses the affected tables. These may include standardized reports, browsers (that are not automatically generated from schemas) and tools developed for data analysis. This may happen even if the software has no "interest" in the new or modified data items. To address this problem and minimize the impact of these relatively minor schema changes, we borrow the technique of "Attribute-Value Lists" and implement them as a schema-independent, extensible table field. We illustrate several examples, describe the implementation and discuss some implications of the approach.

Research funded by U.S. Department of Energy under Contract W-7405-ENG-36.

Estimation of DNA Substitution Matrices and a Generalized Measure of Evolutionary Distance

W.J. Bruno and L. Arvestad

Theoretical Biology and Biophysics Group; T10, MS K710; Los Alamos National Laboratory; Los Alamos, New Mexico 87545.

When attempting to analyze the evolutionary history of a large number of DNA sequences, the speed and complexity of the algorithm used becomes important. Thus, distance-based algorithms of phylogeny reconsturction are of interest. However, the ability to discriminate among similar trees will sometimes depend on having a very accurate and sensitive distance measure. A distance measure based on an incorrect model for nucleotide substitution can lead to misleading results.

Yang [J. Mol. Evol. 39:105–111 (1994)] showed using a maximum likelihood analysis that currently used methods for distance estimation are based on incorrect models for nucleotide substitution. In particular, he noted that transverstion rates are different for different bases, beyond effects due to base frequency alone. Goldstein and Pollock [Theoretical Population Biology 45:219–226 (1994)] demonstrated that carefully combining Kimura-type measures [Kimura, J. Mol. Evol., 16:111-120 (1980)] of sequence distance based on transversions and transitions gives a robust distance measure that rivals maximum likelihood on simulated data.

We show that it is possible to estimate a substitution matrix without using likelihood methods on real data sets. The method works with small and with large data sets and executes in seconds. Once the substitution matrix is determined, its eigenvectors can be used to calculate a distance measure between pairs of sequences, consistent with the observed pattern of substitution. This method is shown to have good linearity and discrimination properties, and it requires minimal computation.

This work was funded by the DOE Human Genome Program (ERW-F137, R. Moyzis, P.I.). W.J.B. was supported by a DOE Human Genome Postdoctoral Fellowship, and is grateful to the Santa Fe Institute for its hospitality.

Combinatorial algorithms for genome sequence analysis

John D. Kececioglu

Department of Computer Science, University of California, Davis, CA 95616. kece@cs.ucdavis.edu

Our research has focused on algorithms for genome rearrangements, multiple sequence alignment, and DNA sequence assembly. In genome rearrangements we have developed new algorithms for comparing genomes in terms of *inversions* [1] and *translocations* [2], and for reconstructing an evolutionary history for a set of sequences that have evolved by point mutation and *recombination* [3]. Given the order and orientation of genes on a chromosome from two related organisms, we can efficiently compute upper and lower bounds on the minimum number of inversions to transform one gene order into the other, and determine the minimum by branch-and-bound for problems involving up to 250 genes. Given gene orders from genomes of two multichromosomal organisms, we can efficiently find an evolutionary history that has a near-minimum number of inversions and translocations; when gene orientation is known, the history is guaranteed to be within a factor of $\frac{3}{2}$ of optimal; when orientation is unknown, the history is within a factor of 2 of optimal. Given a set of sequences that have evolved by point mutation and recombination, we can efficiently identify within the set a pair of protosequences and a history of recombinations that produces the remaining sequences from the protopair such that the maximum cost of any recombination in the history is minimized; when the set has evolved from more than two protosequences, we can efficiently identify a protoset and a recombination history such that the size of the complement of the protoset is within a factor of $\frac{1}{2}$ of optimal.

We are also completing two large implementation projects: a branch-and-bound algorithm for maximum-trace *multiple sequence alignment*, and a three-phase algorithm for *DNA sequence assembly*. Maximum trace alignment [4] finds a multiple alignment that is in maximum agreement with a set of candidate pairwise alignments described by dot plots; our algorithm branches on a small set of alignment columns obtained from minimum cuts of a graph, and prunes suboptimal alignments with bounds from a new heuristic. Our three-phase algorithm for sequence assembly uses a decomposition similar to Kececioglu and Myers [5], but solves the fragment orientation and layout problems simultaneously using a relaxation to nonbipartite matchings.

This work was supported by a DOE Human Genome Postdoctoral Fellowship.

[1] Kececioglu, John and David Sankoff. "Efficient bounds for oriented chromosome-inversion distance." In Proceedings of the 5th *Symposium on Combinatorial Pattern Matching*, Asilomar, California, Springer-Verlag Lecture Notes in Computer Science, Volume 807, 307–325, June 1994.

[2] Kececioglu, John D. and R. Ravi. "Of mice and men: Algorithms for evolutionary distances between genomes with translocation and inversion." Submitted to the 6th ACM-SIAM *Symposium on Discrete Algorithms*, July 1994.

[3] Kececioglu, John and Dan Gusfield. "Reconstructing a history of recombinations from a set of sequences." In Proceedings of the 5th ACM-SIAM *Symposium on Discrete Algorithms*, Arlington, Virginia, 471–480, January 1994.

[4] Kececioglu, John. "The maximum trace problem in multiple sequence alignment." In Proceedings of the 4th *Symposium on Combinatorial Pattern Matching*, Padova, Italy, Springer-Verlag Lecture Notes in Computer Science, Volume 684, 106–119, June 1993.

[5] Kececioglu, John D. and Eugene W. Myers. "Combinatorial algorithms for DNA sequence assembly." To appear in *Algorithmica*, 1993.

A Syntactic Pattern Recognition System for DNA Sequences

David Searls

Department of Genetics, Univeristy of Pennsylvania School of Medicine, Room 475 CRB, 422 Curie Blvd, Philadelphia PA 19104-6145.

Formal language theory views *languages* as sets of strings over some *alphabet*, and specifies potentially infinite languages with concise sets of rules called *grammars*. Grammars are an exceptionally well-studied methodology, familiar to all computer scientists, for the description of complex, higher-order structures embodied in strings of symbols. Moreover, they can be given as input to general-purpose programs called *parsers* capable of determining whether a given string satisfies the rules of the grammar. Parser technology is also extensively developed, and has been applied as well to the problem of *searching* for complex patterns specified by grammars in large amounts of data, in a technique known as *syntactic pattern recognition*.

We have studied DNA sequences from the perspectives of both formal language theory and practical pattern recognition tasks using linguistic tools. On the formal side we have presented a number of results concerning the mathematical linguistic "complexity" of the language of DNA, e.g. it's position on the Chomsky hierarchy of languages, and the relationship between syntactic structure and secondary structure. We have also defined and characterized a novel grammar formalism, called String Variable Grammar, that is particularly well-suited to the representational needs of DNA. The practical side entails the development and use of a syntactic pattern recognition system for DNA sequences, called GenLang, that takes advantage of structural and/or hierarchical aspects of a domain by using rule-based methods to describe and discriminate such structures. The GenLang system has been used successfully to specify and search for tRNA genes, group I introns, and most recently, protein-encoding genes, achieving results comparable to other, procedural systems.

This work was funded by the DOE Genome Program (DE-FG02-92ER61371).

[1] Searls, D.B. (1988) "Representing Genetic Information with Formal Grammars" *Proceedings of the Seventh National Conference of the American Association for Artificial Intelligence*, AAAI/Morgan Kaufman, pp. 386–391.

[2] Searls, D.B. (1989) "Investigating the Linguistics of DNA with Definite Clause Grammars" In *Logic Programming: Proceedings of the North American Conference* (E. Lusk and R. Overbeek, eds.), MIT Press, pp. 189–208.

[3] Searls, D.B. and Noordewier, M.O. (1991) "Pattern-Matching Search of DNA Sequences Using Logic Grammars" *Proceedings of the Seventh Annual Conference on Artificial Intelligence Applications*, IEEE Computer Society, pp. 3–10.

[4] Searls, D.B. (1992) "The Linguistics of DNA" *American Scientist* **80**: 579–591.

[5] Searls, D.B. (1993) "The Computational Linguistics of Biological Sequences" In *Artificial Intelligence and Molecular Biology* (L. Hunter, ed.), AAAI Press, chapter 2, pp. 47–120.

[6] Searls, D.B. and Dong, S. (1993) "A Syntactic Pattern Recognition System for DNA Sequences" In *Proceedings of the Second International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis* (H.A. Lim, J. Fickett, C.R. Cantor, and R.J. Robbins, eds.). World Scientific Publishing Co., pp. 89–101.

[7] Dong, S. and Searls, D.B. "Gene Structure Prediction by Linguistic Methods" *Genomics*, in press.

Prediction of Coding Regions in Genomic DNA: Optimal and Suboptimal Parses

Eric E. Snyder and Gary D. Stormo

Department of Molecular, Cellular and Developmental Biology
University of Colorado, Boulder, CO 80309

We have developed an approach for predicting coding regions in genomic DNA that utilizes multiple types of evidence, combines those into a single scoring function and then returns both optimal and ranked suboptimal solutions using that scoring function. The current version of the program predicts four classes of sequence: introns and three types of exons, first, last and internal. It uses a variety of statistical tests for these different classes, including those for the signals that define their ends and for baises in their contained sequences. A neural network is used to weight the different types of statistical tests to optimize performance, which we find to be as good or better than other published methods when tested on new examples. However, we find one of the most important features of this system is the ability to examine multiple solutions which is provided by the dynamic programming approach. These multiple, ranked solutions often provide indications of which portions of the predictions are most reliable, and in cases where the highest scoring prediction is not correct it can often be found in a high ranking supoptimal solution. Furthermore alternative splicing patterns can often be found among the high ranking suboptimal solutions. We have performed tests of the robustness of the method when there are sequencing errors in the data, and shown that the system can be trained to optimize performance for data with specified error rates. We are now exploring methods for reliably predicting other classes of sequence regions, especially promoters. These include approaches based on minimal length encoding algorithms and on Sequence Landscape methods. Recent results from these approaches will be described.

Gene Recognition, Modeling, and Homology Search in the GRAIL-genQuest System

Manesh Shah[1], J. Ralph Einstein[1], Sherri Matis[1], Ying Xu[1], Xiaojun Guan[1], Donna Buley[2], Sergey Petrov[1], Loren Hauser[1], Richard J. Mural[2], and Edward C. Uberbacher[1]

[1]Engineering Physics and Mathematics, and [2]Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6364. e-mail:GRAILMAIL@ornl.gov

GRAIL-genQuest is a modular expert system being constructed to analyze and characterize genomic and cDNA sequences. Recognition of gene features, gene modeling, and DNA, protein, and motif databases searches are supported by an e-mail server system and a graphical client-server version. Within the last year significant improvements have been made in the sensitivity and accuracy of coding region recognition and gene modeling.

GRAIL E-mail Server and Feature Recognition: GRAIL provides an on-line e-mail service for locating the protein coding regions of DNA sequences. This interface utilizes a multiple sensor-neural network (Uberbacher and Mural, 1991, PNAS 88:11261-11265) to find coding regions and a rule based interpreter to reduce this output to a table. A new version of the coding recognition portion of this systems is capable of finding 94% of all exons, and 80% of exons less than 100 bases in size. Improvements have also been made in the accuracy with which a given exon is described - the edges are more accurate with 54% of exons predicted with both edges correct to the base, and an additional 40% with one edge exactly correct.

E-mail **GRAIL** can analyze up to 100 kbp of sequence at a time and several sequences may be included in a single e-mail message. The amino acid sequence of predicted coding regions can be searched against the SwissProt database using the Intel iPSC/860 parallel computer, and be searched for functional motifs using the PROSITE database. These search capabilities are part of the new *genQuest* sequence comparison server system (described below) which can be accessed directly or through **GRAIL**.

Gene Modeling and Client-Server GRAIL: In addition to the current coding region recognition capabilities based on a multiple sensor-neural network and rule base, modules for the recognition of features such as splice junctions, transcription and translation start and stop, and other control regions have been constructed and incorporated into an expert system for reliable computer-based modeling of genes. The gene modeling version of GRAIL is available through an X-window-based client-server system. The client-server system allows the user to try a number of different scenarios for a given sequence region and ask "what if" type questions interactively.

A gene assembly program (GAP) which combines the outputs from the various feature recognition modules and attempts to predict the sequence of the spliced mRNA from the genomic DNA sequence is part of this system. Heuristic methods and dynamic programming are used to construct first pass gene models which include the potential for insertions and deletions of initially predicted exons. These actions result in a net improvement in gene characterization, particularly in the recognition of very short coding regions. Genes modeled by this system have an average correlation coefficient compared to the actual gene of 0.93. The models contain 94% of all exons regardless of size and only 3% false positive information. After model construction 79% of exons have both edges correctly defined to the base and an additional 19% have one edge correct. In addition, other features of interest such as poly-A addition sites and repetitive DNA elements can be located using this program. Translation of gene models and database searches are also supported through access to the *genQuest* server (described below). Generation of an annotation report in "feature table" format is supported. Client software can be downloaded by anonymous ftp from arthur.epm.ornl.gov, and help information can be obtained by sending a message with the word help in the first line to GRAIL@ornl.gov e-mail address.

The GenQuest Sequence Comparison Server: The **genQuest** server is an integrated sequence comparison server which can be accessed via e-mail and through a X-windows graphical client-server system. The basic purpose of the server system is to facilitate rapid and sensitive comparison of DNA and protein sequences to existing DNA, protein, and motif databases. Databases accessed by this system include the LANL daily updated DNA sequence database (GSDB), SwissProt, the dbEST expressed sequence tag database, protein motif libraries and motif analysis systems (Prosite, BLOCKS), a repetitive DNA library (from J. Jurka), and sequences in the PDB protein structural database. These options are designed to provide a comprehensive description of newly obtained sequences through homology methods, and can also be accessed from the **XGRAIL** graphical client tool. The system uses a specialized parallel computing environment at the Oak Ridge National Laboratory and is supported and curated by research teams in the genome community.

The **genQuest** e-mail server supports a variety of sequence query types. For searching protein databases, queries may be sent as amino acid or DNA sequence. DNA sequence can be translated in a user specified frame or in all 6 frames. DNA-DNA searches are also supported. User selectable methods for comparison include the Smith-Waterman dynamic programming algorithm, FastA, versions of BLAST, and the IBM dFLASH protein sequence comparison algorithm. A variety of options for search can be specified including gap penalties and option switches for Smith-Waterman, FastA, and BLAST, the number of alignments and scores to be reported, desired target databases for query, choice of PAM and Blosum matrices, and an option for masking out repetitive elements. Multiple target databases can be accessed within a single query.

E-mail turn-around times for the system are quite rapid, less than 1 minute for protein searches, about 1 to 2 minutes for protein-DNA, and several minutes for reasonable length DNA-DNA searches. **GenQuest** can be accessed by e-mail at the Q@ornl.gov e-mail address, and instructions can be obtained by sending a message to that address with the word help in the first line. Further help for **GRAIL** or **genQuest** can be obtained by sending e-mail to the GRAILMAIL@ornl.gov address, and the graphical client tool can be downloaded by anonymous ftp from arthur.epm.ornl.gov.

(This research was 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.)

Biological evaluation of d^2: an algorithm for high performance novel DNA sequence functional assignment.

Winston Hide, John Burke and Daniel Davison
MasPar Computer Corporation and University of Houston.

A number of algorithms exist for searching sequence databases for biologically significant similarities based on primary sequence similarity of aligned sequences. We present here the biological sensitivity and selectivity of d^2, a high performance comparison algorithm that rapidly determines the relative dissimilarity of large datasets of DNA/Protein sequences. We have determined that d^2 is uniquely capable of detecting significant functional similarities between sequences that have no measurable alignable sequence similarity. These relationships remain undetectable by alternate methodologies.

Querying with a lipoprotein lipase DNA sequence results in hits that share functional similarity, such as DNA sequences coding for proteins that regulate lipid metabolism, membrane and fat interaction and lipogenesis. No other algorithm can provide such important pointers to sequence functionality. d^2 uses sequence-word multiplicity as a simple measure of dissimilarity. It is not constrained by the comparison of direct sequence alignments and so can use word contexts to yield new functional information on relationships. It is extremely efficient: comparing a query of length 884 bases (INS1ECLAC) with 19,540,603 bases of the bacterial division of Genbank 76.0 in 52 CPUseconds on a Cray Y/MP-48 supercomputer. A parallel version (in development) is projected to run 100 times faster on a MasPar MP-2216. d^2 is unique in that subsequences of biological interest can be weighted to improve sensitivity and selectivity of a search over existing methods.

We have determined the ability of d^2 to detect biologically significant matches between a query and large datasets of DNA sequences while varying parameters such as word-length and window size. We have optimized parameters to present maximal sensitivity and selectivity relative to FASTA.

Not funded by DOE.

Analysis and Annotation of Nucleic Acid Sequence

David J. States, Ron Cytron, Pankaj Agarwal and Hugh Chou
Institute for Biomedical Computing, Washington University in St. Louis

We are developing improved methods for analyzing nucleic acid sequences based on sequence similarity and very large scale classification techniques. Our previously developed methods for sequence classification provide a basis for defining families of protein sequence and modular domains preserved during evolution. Functional modules of gene and gene product structure and regulatory signals within the genome can be recognized as recurrent patterns in anonymous nucleic acid sequence using large scale classification techniques.

As a first step to extending our methods to a general nucleic acid annotation database, consensus sequences have been derived for each family in the protein sequence classification. This is basically a multiple sequence alignment problem. Some families are quite large (>1000 members) so a computationally efficient algorithm is needed to implement this. We have chosen ClustalW as a tool for this task. Our groups are defined using a minimal spanning tree representation that identifies the most similar members in each family. This tree can be imported directly into ClustalW or a fast heuristic comparison of all family members can be recalculated internally (where tested, the two have been equivalent). A hierarchical multiple sequence alignment scored with the BLOSUM62 matrix is then performed and a consensus sequence based on the whole family or any subtree of the family is then dynamically derived. An important advantage of this approach is that it allows the user to define the granularity of classification interactively. In some cases a very broad classification (e.g. grouping all serine proteases together) may be desired while in other cases a much finer granularity (tracing species variation within the trypsin subfamily) may be needed.

Improved methods for nucleic acid sequence comparison are being developed and a repeat analysis software toolkit has been written. The Dayhoff PAM formalism has been extended to codon based sequence comparisons and scoring systems have been developed for sequences related as protein coding regions, non-coding, transcribed sequences, regions of untranscribed sequence, and regions of similar predicted three-dimensional structure. These methods are being tested on C. elegans and human sequence.

An annotated, intelligently non-redundant sequence database, is being built. This database will complement existing public databases using automated classification technology and manual review. An associated database of all pairwise sequence similarities is also being maintained.

An improved user interface for the classification analysis tool has been developed using WWW, perl and the html protocol. This approach has allowed us to readily link our classification data to the NCBI, EMBL and other network accessible databases. The use of perl allows computationally efficient retrieval systems to be rapidly prototyped and implemented.

Classification data and source code are being distributed on an anonymous FTP site (ibc.wustl.edu) in addition to the WWW interface.

A Conceptual Model for Genome Mapping Data

Mark Graves and Charles B. Lawrence
Department of Cell Biology, Baylor College of Medicine, Houston, TX 77030.

Human genome mapping projects are generating large amounts of data which must be stored in databases to be easily accessible. A good first step in designing any database is to develop a conceptual schema which captures the concepts and relations of the domain. We have developed a simple conceptual model based on graphs which has proven itself useful in the design of several mapping databases. We present the conceptual model and some of the mapping schemas we have developed within the model.

Conceptual modeling is the process of describing the concepts and relationships of a domain that are to be stored in a database [1]. The process takes place within a theoretical framework called a conceptual model. A *conceptual model* is a data model which formalizes the representation and manipulation of concepts and relationships. The conceptual model defines the language used to describe the domain. A conceptual model is used by a database developer to describe the aspects of a domain which are to be captured by a database. The description of a domain is called a *conceptual schema*.

One foundation of conceptual models is graphs. Graphs have proven themselves useful for representing complex domains, may be stored and queried in a database [2], and have a strong mathematical foundation in graph theory. A *graph conceptual model* is a data model which represents the connections between the concepts and relationships in a domain as a graph.

Graph models are useful for representing mapping data. We have developed conceptual schemas for several laboratory databases, including gene (cDNA) mapping, discovery of dinucleotide repeat markers, and YAC-STS hybridization experiments. Graphs can represent the location of genes and markers, order and distance between them [3], and containment relationships. Graph models may also be decomposed into binary relationships which are compatible with physical map viewing tools, such as SIGMA [4].

We have found that conceptual models based on graphs are useful tools for developing mapping databases. They are also a convenient mechanism for communication between biologists and database developers, which helps guide the database development toward a system which meets the needs of genome mappers.

This work was supported by the W.M. Keck Center for Computational Biology and the Baylor Human Genome Center funded by the NIH National Center for Human Genome Research. In addition, C.B.L was supported by a grant from the Department of Energy, and M.G. was supported by a fellowship from the National Library of Medicine and a Department of Energy Human Genome Postdoctoral Fellowship.

[1] Brodie, Michael, John Mylopoulos, and Joachim Schmidt, editors (1984). *On conceptual modelling: Perspectives from artificial intelligence, databases, and programming languages* Springer-Verlag, New York.

[2] Graves, Mark (1994). Querying a Genome Database Using Graphs. In H. Lim, ed., *Proceedings of The Second International Conference on Bioinformatics and Genome Research.* World Scientific Publishing.

[3] Graves, Mark (1993). Integrating order and distance relationships from heterogeneous maps. *In* L. Hunter, D. Searls and J. Shavlik, editors, *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology (ISMB-93)*, Menlo Park, CA. AAAI/MIT Press.

[4] Cinkosky, M.J., Fickett, J.W., Barber, W.M., Bridgers, M.A., and Troup, C.D. (1992) SIGMA: A system for integrated genome map assembly. *Los Alamos Science* 20: 267-269.

Bayes Decoding of Pooling Experiments Using Monte Carlo Methods

W.J. Bruno[1], E. Knill[1,2], A. Schliep[1], D.C. Torney[1]

[1] Theoretical Biology and Biophysics Group; T10, MS K710; Los Alamos National Laboratory; Los Alamos, New Mexico 87545. [2] Computer Research and Applications; CIC3, MS B265; Los Alamos National Laboratory; Los Alamos, New Mexico 87545.

One of the most important experimental procedures used in physical mapping is to screen a library of clones for clones which are positive for a given probe. It is usually not feasible to screen each clone individually. To reduce the number of screening tests, one can form pools from sets of clones and attempt to determine the positive clones by screening the pools. This is the procedure used by most laboratories.

Pooling strategies are designed so that the positive clones can be determined from the screening results for the pools, provided that the number of positive clones is small enough and the pool screenings have no errors. However, we have found that in practice, most pool results are ambiguous due to larger numbers of positive clones and errors. In order to find the positive clones, additional screenings of individual clones are required. To minimize the number of additional screenings requires interpreting the pool results. For simple pooling strategies such as the row-and-column strategies, interpretation of the pool results can be done by hand with reasonable success. These strategies are however far from optimally efficient in terms of the number of pools used.

The ideal method for interpreting pool results involves computing the posterior probabilities of each set of clones being the set of positive clones. We can then determine each clone's posterior probability of being positive. This can be achieved for screening clone libraries because there is a reasonable model for the distribution of sets of positive clones and experimental errors. The main difficulty in implementing this method is that the exact computation of the posterior probabilites is not feasible. However, one can use one of several Monte Carlo methods for estimating these probabilites. The two most promising methods are the Gibbs and the Metropolis-Hastings methods. In these methods, one establishes a Markov process, whose nodes are possible sets of positive clones and whose transition probabilities are determined by the relative posterior probabilites of "adjacent" states. Adjacency is determined by removing or adding a single clone. The posterior probabilty of a clone's being positive is estimated by simulating the Markov process and determining the fraction of the time that a clone is a member of the current state.

We are integrating these algorithms in a general decoding system we have developed and implemented in software. Currently this decoding system uses one of several heuristics to prerank the clones. The heuristics are based on the number of positive pools that a clone belongs to, and whether a positive pool is uniquely explained by a given clone's being positive.

Interactive Algorithms for Rapid Chromosome Copy Number Enumeration of Individual, Whole Cell Nuclei Inside Intact Tissue Specimens

Stephen Lockett[1], Curtis Thompson[2], Damir Sudar[1], James Mullikin[1], Daniel Pinkel[1], Joe Gray[1]

[1]Resource for Molecular Cytogenetics, MS 74-157, Lawrence-Berkeley Laboratory, Berkeley, CA 94720. [2]Collaborating author, The University of California, San Francisco.

Fluorescence in situ hybridization (FISH) is a major tool for analyzing specific nucleic acid sequences in individual cells. Application of this technique to tissue sections is complicated by the presence of fractional cells produced by the sectioning process. In many cases the sections are too thin to contain any intact cells. The continuing improvement in hybridization techniques has permitted obtaining hybridization throughout the volume of formalin fixed/paraffin-embedded tissue over 20 μm thick (Thompson et. al.). These sections contain a central layer of intact nuclei. Optical sectioning with confocal or conventional microscopy provides a stack of two dimensional (x,y) images in successive focal planes (z) from which it should be possible to make accurate genetic measurements. However the complex three dimensional tissue architecture complicates analysis.

To assist with analysis of this data we have developed simple, fast algorithms for displaying and interactively analyzing the 3D images. The display algorithm makes multiple, 2D, maximum-intensity, projection images through the original 3D image. Each projection image is made at a different angle relative to the 3D image and when the projections are viewed in sequence, the 3D image appears semi-transparent and rotating to the viewer. The first interactive analysis algorithm enables the user to mark punctate FISH signals (e.g. those from chromosome specific centromeric probes) using the computer's mouse, in either the original 3D image or the projection images. After marking all the FISH signals in the 3D image, a second analysis algorithm is invoked. For each pair of marked signals (closer than a user-defined distance to each other), this algorithm extracts from the 3D image, a 2D slice image containing the line between both marked signals and parallel to the z (depth) axis. By visual examination of this slice image, the user decides if the signals are in the same nucleus, in different nuclei, or if one or both should be rejected (for example because it is in an incomplete nucleus). The output from the algorithm is the FISH signal copy number distribution for the population of nuclei analyzed and the locations of the analyzed FISH signals and intact nuclei within the 3D image.

The algorithms enable the convenient and rapid determination of signal copy number in individual intact cells in histologically-defined regions of tissue specimens. This will enable analysis of small, premalignant lesions that are not suitable for analysis by other molecular techniques, and permit the direct correlation of molecular cytogenetic information with traditional pathologic grading systems.

This work was funded by the US DOE contract DEAC0376SF00098 and a grant from the Whitaker Foundation.

Thompson, C.T., LeBoit, P.E., Nederlof, P.M., Gray, J.W. (1994) Thick-Section Fluorescence *in Situ* Hybridization on Formalin-Fixed, Paraffin-Embedded Archival Tissue Provides a Histogenetic Profile. *Am J Pathol*, 144, 237-243.

SIGMA: System for Integrated Genome Map Assembly

Michael J. Cinkosky, Shahar Tsadeek, and James W. Fickett

National Center for Genome Resources
1800 Old Pecos Trail
Santa Fe, NM 87505
(505) 982-7840

SIGMA (System for Integrated Genome Map Assembly) is an interactive, graphical genome map editor. SIGMA enables maps to be built from many different types of mapping data including: YAC/STS content, genetic linkage, breakpoint, *in situ* hybridization, radiation hybrid, etc. All of the data are integrated into a single map and SIGMA is able to identify discrepancies between the map and the data on which it is based.

SIGMA has the following major features:

- Graphical, mouse-based genome map editing;
- Supports creation of multiple, user-defined "views" on a single map;
- Enables integration of data at all appropriate levels of resolution, from banded ideograms to restriction fragments;
- Allows users to add new kinds of cloning vectors, chromosome landmarks, or other map objects, without changing the software;
- Split screen allows simultaneous viewing of maps at multiple levels of magnification;
- Able to print maps of any size on any PostScript^{tm} printer;
- Supports integration of data from many different types of physical and linkage experiments;
- Keeps the underlying data as part of the map, allowing users to easily access the data supporting any given map;
- Automatically evaluates the map against the underlying data, pointing out places where the two disagree;
- Available for multiple platforms, including:

  Sun SparcStation, under Sun/OS 4.X

  Apple Macintosh

  DEC Alpha, under OSF-1

SIGMA is currently being used to assemble maps of a number of human chromosomes, including 5, 9, 16, X and Y.

SIGMA is available by anonymous ftp from ftp.ncgr.org. Documentation on the software, as well as a number of sample maps, can be accessed through the World Wide Web (http://www.ncgr.org/sigma).

A Software Tool to Determine Overlaps Between Clones and Existing Restriction Fragment Maps

T. Mimi Yeh (myeh@llnl.gov), and Tom Slezak

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

The LLNL Human Genome Center currently has over 50% of chromosome 19 covered with high-resolution EcoRI restriction fragment maps. As the number of the restriction fragment maps in our database nears 200, an automatic clone matching software tool is needed to help the chromosome 19 closure effort. Such a tool would have two functions: 1) Given a clone (or a list of clones) find the "best" matches against all the restriction fragment maps within our database, 2) Find pairs of restriction maps that possibly overlap on their ends. This will assist us in the gap closing process by freeing the map builders from tedious manual clone matching and by minimizing human errors as much as possible. We note that with such a large amount of chromosonal real estate covered it is quite likely that there may be multiple candidate "fits" for any given clone. We need to find all likely sites and then use other mapping data (insitu or other hybridization, FISH, statistical overlap, etc.) and human judgement to determine the most likely placement.

We use a sliding window method for the comparison process. A window with the size of the input clone length slides along the current restriction map to perform the comparison between the fragments in the sliding window and the input clone fragment list. The sliding window and the input clone fragment lists are both sorted for an efficient comparison. Scores are calculated based on the comparison of matching results for both end windows or interior windows. (End windows are defined as cases where the input clone would extend one or both ends of the existing map.)

The result scores are presented to the user showing the most likely location(s) for both end-extension to existing maps and placement totally with existing maps. By running the program on lists of end clones of existing maps we can dredge for overlaps that have not yet been determined by other means. A number of parameters can be chosen by the user at run time to set the stringency of the comparison. This tool will see wide use in our closure of chromosome 19 and in high-resolution mapping on other genomic targets.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory contract no. W-7405-ENG-48.

Gel Analysis Programs for 'DOG' and STS Content Mapping

S. Pitluck, K. Gong, S. Lewis

Human Genome Informatics Group, Lawrence Berkeley Laboratory, Berkeley, CA 94720

A gel analysis program (Angel) originally developed to analyze restriction gels, has been adapted for the production of maps related to the LBL Directed Sequencing Strategy. This program automatically detects gel lanes and locates and sizes bands in the lanes. In order to adapt Angel for other tasks and simplify even the current minimal user interaction, more specific versions of Angel were created. These special versions also contain several enhancements.

One of the first modifications to Angel allows the operator to use a slide bar to vary the contrast of the displayed image. Another enhancement is the automatic display of both the normal gel lanes as well as the special marker lanes. Another slide bar is available to adjust the location of these overlayed lane markings. Moving the mouse up and down within a marker lane also moves a horizontal line up and down across the image. This is useful for aligning various bands. If the gel image is being used for STS mapping,then clicking the mouse button on any bands outside the marker lanes will result in the recording of the original microtiter plate or row/column locations for that clone. This information is used to determine which P1 was hit by the particular STS.Thus by just pointing and clicking, many images can be rapidly analyzed. In addition, the relevant derived information is saved as a ".ace" file for inclusion in our database. Images can be reexamined at a later date while using the database. Band selections made earlier will then be automatically redisplayed, and revisions to the analysis are supported.

Another version of the software (DAngel) has been created to assist in our Directed Sequencing strategy. For the specific purposes of mapping the 3kb fragments called 'DOGS', the program records the band sizes based on the reference markers, as well as "de-pooling" the clone to identify its microtiter plate location.

In terms of performance, Research Technicians using these programs estimate that processing time has dropped by a factor of about 3 to 5, compared with the largely manual way in which these analyses were performed previously. More important, perhaps, is the reduction in the number of human errors, due to automatic data capture.

Informatics Support for Mapping in Mouse-Human Homology Regions

Sergey Petrov[1], Manesh Shah[1], Loren Hauser[1], Richard Mural[2], and Edward Uberbacher[1]

Engineering Physics and Mathematics Division[1] and Biology Division[2], Oak Ridge National Laboratory, Oak Ridge, TN 37831-6364 615/574-6134, Fax 615/574-7860, Internet:"UBE@ornl.gov".

The purpose of this project is to develop databases and tools for the ORNL Mouse-Human Mapping Project, including the construction of a mapping database for the project, tools for management and archiving of cDNAs and other probes used in the laboratory, and analysis tools for mapping, inter-specific backcross, and other needs. Our initial effort has involved installing and developing a relational SYBASE database for tracking samples and probes, experimental results, analysis, etc. Recent work has focused on a corresponding ACeDB implementation containing mouse mapping data and which provides numerous graphical views of this data. The initial relational database has been constructed with SYBASE using a schema modeled from the one implemented at the LLNL center because of the available documentation for the LLNL system, and to maximize compatibility with the human chromosome 19 mapping effort (major homologies exist between human chromosome 19 and mouse chromosome 7 - the initial focus of the ORNL work). Our ACeDB implementation has been modeled somewhat from the chromosome 21 ACeDB system at LBL (with some model modification) and is designed to contain genetic and physical mouse map data as well as homologous human chromosome data. The utility of exchanging map information with LLNL (human chromsome 19) and potentially other centers has lead to the implementation of procedures for data export, and import of human mapping data into the ORNL databases.

User access to the system is being provided by workstation forms-based data entry and ACeDB graphical data browsing. We have also implemented the LLNL databases browser to view the human chromosome 19 data maintained at LLNL, and arrangements are being made to incorporate mouse mapping information into the browser. Other applications such as the "Encyclopedia of the Mouse", specific tools for archiving and tracking cDNAs and other mapping probes, and analysis of inter-specific backcross data and restriction mapping of YACs have been implemented.

We would like to acknowledge use of ideas from the LLNL and LBL Human Genome Centers. (Research sponsored by the Office of Health and Environmental Research, U.S. Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.)

A Battery of Programs for STS Content Mapping

Farid Alizadeh[2], Richard M. Karp[1], Deborah Weisser[1], Geoffrey Zweig[1]

[1]International Computer Science Institute, Berkeley, CA and University of California at Berkeley. [2]Rutgers University.

The goal of STS content mapping is to infer the linear ordering of a set of Sequence Tagged Sites, given data about the incidence of these sites to the clones in a library. The task is complicated by the presence of errors in the incidence data and clone abnormalities such as chimerism. We have derived a battery of computer programs for performing this task. The programs have performed successfully both on synthetic data and on data from chromosome 21.

Our algorithms are presented with a zero-one matrix giving the measured incidences between probes (STSs) and clones. The generation of our synthetic data is controlled by several parameters, including the numbers of probes and clones, the distribution of clone lengths, the length of the DNA strand being mapped, the false positive and false negative rates, and the fraction of clones that are chimeric.

The main engine of our approach is a simulated annealing algorithm which, given estimates of the false negative and false positive rates and the chimerism rate, seeks to construct the most likely ordering of the STSs. We use several methods of enhancing the performance of the basic simulated annealing algorithm. A preprocessing routine is used to screen out false positives and to identify and split chimeric clones. Because the screening routine occasionally errs by identifying a true probe-clone incidence as a false positive, we couple it with a postprocessing routine which inspects the solution produced by simulated annealing, identifies anomalies traceable to errors in screening, and corrects these errors. To further speed up simulated annealing we also employ routines for generating a good starting solution.

We have also formulated the probe ordering problem as a traveling-salesman problem in which the cities are zero-one vectors and the distance between cities is the number of positions in which the vectors differ. This method is much faster than simulated annealing and performs nearly as well. It is also applicable to a wide range of other physical mapping strategies such as radiation hybrid mapping.

We are presently extending our programs to deal with pooled incidence data rather than incidences between probes and individual clones, to accept *in situ* hybridization data and genetic mapping data, and to work in a feedback loop with experimenters to identify and correct errors in the probe-clone incidence data.

[1] Alizadeh,F., Karp, R.M., Newberg, L. and Weisser,D. (1993) Physical Mapping of Chromosomes: A Combinatorial Problem in Molecular Biology, *Algorithmica*, to appear.

[2] Alizadeh, F., Karp, R.M., Weisser, D., Zweig, G. (1994) Physical Mapping Using Unique Probes, Proc. ACM/SIAM Symp. on Discrete Algorithms.

Software Support for High-Throughput DNA Sequencing

Charles Lawrence[1,3], Victor Solovyev[1] and Eugene Myers[2]

[1]Departments of Cell Biology and Human & Molecular Genetics, Baylor College of Medicine, Houston, TX 77030. [2]Department of Computer Science, University of Arizona, Tucson, AZ. [3]Corresponding author.

One of the barriers to achieving high-throughput DNA sequencing is the processing and management of the sequence data from the time of its generation by the sequencing hardware to its emergence as finished, edited and annotated DNA sequence. The bottleneck is not due to any major unsolved technical challenge, but to the lack of effective, integrated software support for the process.

Over the past 3 years, we have developed an integrated software system whose main goal is to automate the management of sequence data for high-throughput projects, and to provide effective interactive tools for use when human interaction with the data is necessary. The development of the system has been facilitated by using an object-oriented approach for the design of the system, and using tools that support object-oriented system development for its implementation. The ***Genome Reconstruction Manager*** (***GRM***) provides several advances in software support for high-throughput DNA sequencing: support for random, directed, and mixed sequencing strategies; a novel subsystem for fragment assembly (developed by E. Myers, Univ. Arizona); a commercial object database management system for data storage; a client/server architecture for using network computational servers; and an underlying data model that can evolve to support fully-automatic sequence reconstruction.

In a related research project, we have studied the association of error in sequence data with the quality of the underlying primary data. DNA sequence predicted from polyacrylamide gel-based technologies is inaccurate because of variations in the quality of the primary data due to limitations of the technology, and sequence-specific variations due to nucleotide interactions within the DNA molecule and with the gel. The ability to recognize the probability of error in the primary data is useful in reconstructing the target sequence for a DNA sequencing project, and in estimating the accuracy of the final sequence.

We used linear discriminant analysis to assign position-specific probabilities of incorrect, over- and under- prediction of nucleotides for each predicted nucleotide position in the primary sequence data generated by a gel-based DNA sequencing technology. Using this method, correct base predictions can be separated from incorrect, over- and under- predictions with Mahalanobis distances of 7.4, 11.8 and 6.7 respectively. Applying the discriminant to a test data set, a table associating discriminant scores with the probability of a prediction error is easily calculated. This information can then be used to assign the probability of error for each of the three error types to each base position in new sequence data with values between <0.0001 to 0.68.

Measurements of accuracy associated with the primary sequence data can be used as the basis for eliminating or minimizing the human editing of assembled sequence data and to automatically generate consensus sequence with confidence estimates at each position in the final sequence. In future work, we will integrate the results of the error analysis research in ***GRM*** with the goal of automating the generation of consensus sequence.

Lawrence, C.B, Honda, S., Parrott, N.W., Flood, T.C., Gu, L., Zhang, L., Jain, M., Larson, S., and Myers, E.W. 1994. The Genome Reconstruction Manager: A software environment for supporting high-throughput DNA sequencing. *Genomics* (in press).

Lawrence, C.B. and Solovyev, V.V. 1994. Assigning position-specific error probabilities to primary DNA sequence data. *Nucl. Acids Res.* 22:1272-1280.

Software Support for Large Scale Sequencing

Gatewood, Joe,[1] Robert M. Pecherer,[2] and Elaine Best[3]

[1]Genomics and Structural Biology Group; LS-2, MS 880; Los Alamos National Laboratory; Los Alamos, New Mexico 87545. [2]Theoretical Biology and Biophysics Group; T-10, MS K710, LANL. [3]Applications Programming Group; CIC-12, MS B295; LANL.

Current and projected DNA sequencing rates effectively prohibit direct human interaction on experimentally derived raw sequence data -- i.e., inspection, elimination of cloning artifacts, and editing in general. The investigator-as-data-processor bottleneck is further compounded during sequence analysis where DNA homology comparisons against public sequence databases result in redundant or extraneous information: Relevant homology is diluted by the irrelevant.

Our goals in developing software to support large scale sequencing are:

1. High speed sequence data entry where routine computer processing is the rule and human intervention the exception;
2. Sequence analysis where homology comparison and reporting are interactively customizable to enduser needs;
3. For STS generation and sequencing, primer selection is automated and customizable;
4. Sequence order relationships and base confidence information are captured during sequencing and exploited in consensus sequence assembly.

The first three goals have been addressed. For primer directed sequencing, we have developed database representations and are exploring assembly algorithms.

Our system architecture includes four components: Enduser interface, database, context management, and analytical tools. The enduser interface is implemented using Gain Momentum (Sybase) and programmed in GEL, a proprietary scripting language specialized for interactive I/O and task management. Database functionality is provided by the Relational DBMS Sybase using recursive DNA representations. (See "Recursive Relational Representation for DNA and Attribute-Value Lists: Techniques for Reducing Schema Modifications", Pecherer et al., DOE Human Genome Contractors Workshop IV, Santa Fe, NM, November 13-17, 1994.) Context management and analytical tools are written in C for performance and flexibility. Context management provides data management capability for the objects and collections obtained from the database and/or operated upon by analysis software. The analytical tools include homology comparison, feature selection, and primer selection algorithms.

All analytical tools are designed as integral system components to avoid file parsing. We have implemented BLAST with a global alignment capability (BLASTga) to avoid segmented DNA homologies and have incorporated post screening of homology results to eliminate redundant extraneous information resulting from repetitive elements.

Research funded by U.S. Department of Energy under Contract W-7405-ENG-36.

BioPOET: Large Scale Sequence Analysis On Workstation Farms*

Manfred D. Zorn, Jane F. Macfarlane, Rob Armstrong‡, Michael H. Cooper, Nicholas C. Weaver

Software Technologies and Applications Group, Information and Computing Sciences Division,
Lawrence Berkeley Laboratory, Berkeley CA 94720
‡Sandia National Laboratories, Livermore, CA 94550

The rate at which new sequences are being generated has dramatically increased. A standard procedure to analyze sequences is to compare them with already known sequences. Thus longer sequences are matched against increasingly larger databases of sequences.

The available sophisticated computing technology to tackle such problems, e.g., faster machines, parallel processing, distributed computing, exists already. However, the use of these resources requires detailed knowledge of the particular resources to optimally access them.

We developed a framework that allows to partition the necessary tasks and execute them on a workstation farm. A *master* reads the database and creates a task for each sequence. The *workers* request a new task, compare the database with the query sequence, and report the results back to the *master*. A graphical user interface allows easy input and parameter specifications, interacts with a network server to launch the program, and displays the final results graphically. The tasks themselves use existing software, e.g., filter [1] to search the database efficiently and align [2] to generate the final alignment.

The framework makes use of the Parallel Object-oriented Environment and Toolkit, POET, that is modeled after the X11 toolkit and enables both high and low level control of the computational methods. The object-oriented programming paradigm allows data encapsulation and methods to hide implementation details so as to present a unified object view to the user. Existing software can be adapted to exploit the power of parallel processing. Thus sequence analysis can be performed transparently to the user in reasonable time, where POET divides either the query sequence or the database in multiple pieces to run on parallel computers or a number of workstations in a distributed environment.

We will present a prototype system that integrates sequence analysis into the sequencing protocol and performs comparisons of sequences on a workstation farm. The framework has been implemented using the C++ language and uses PVM as communication package. The graphical user interface is implemented in VisualWorks\Smalltalk from ParcPlace Systems.

[1] Chang, W. and Marr T., Approximate String Matching and Local Similarity, in *Combinatorial Pattern Matching*, Springer Verlag, 1994.

[2] Huang X. and Miller W., A space-efficient algorithm for local similarities. *Comput. Appl. Biosci.* **6**:373-381(1990)

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

An Integrated Data Management System for DNA Sequencing

Arthur Kobayashi (kobayashi1@llnl.gov), David J. Ow, Mark C. Wagner, T. Mimi Yeh, and Tom Slezak

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

We are implementing an integrated data management system to support our local DNA sequencing efforts. The system consists of a loosely-coupled set of tools which allows users to define, track, process, and analyze sequence data. We have performed system analysis, design, and prototyping for the new system, and are currently implementing the various components. Our goals include minimizing data entry, consistent naming conventions for the various sample types, online interactive data access, automatic generation of sample sheets or setup files where possible, tracking replications (reprocessing of samples at any level), and ultimately, an integrated view of source, processing, analysis, and mapping data.

We have characterized the system based on the functions necessary to complete sequencing tasks. Functions may then be implemented using the most practical and cost-effective means available--commercial packages, custom software, etc. This approach gives a loose coupling where functions may be stand-alone single-purpose programs, or larger, multi-function programs. These programs may be modified or replaced with minimal impact to other parts of the system. The integration of the system is accomplished by our high-level model of sequencing tasks and sequencing data transformations, and the use of a common schema and relational database as our underlying data repository.

Sequence information consists of source, processing, analysis, and physical mapping data. We have developed a sequence source hierarchy which defines clone libraries using library, clone, prep, DNA, and sequence reaction levels. We have also defined "batch" tables in our database schema (where a batch is a user-defined set of similar items). Users may define batches and then use these batches as a way to conveniently reference related items for display or processing.

We have completed a core set of functions which allows rapid definition and editing of sequence source and processing information. A graphical user interface (GUI) front end provides interactive "dialogs" to select libraries and add, display, edit, or print entries at any level. Sequence reactions can be assigned to a labelling run, added to a sample sheet editor, and output to a printer. The sample sheet assignments are then used to set up the labelling run. Once labelled, the sequence reactions can then be assigned to a sequencing run, and output as a setup file for electrophoresis using a sequencing run editor. The setup file is transferred over the network and used to configure the sequencer. Users then use commercial programs or Unix scripts to edit, assemble or analyze sequence data. Results of these functions are stored in the database and associated with sequence sources.

Our strategy allows us to use a heterogeneous mix of software tools to implement the desired system functions, and relies on a high-level system view for integration of functions and data into a coherent system.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory contract no. W-7405-ENG-48.

Statistical Methods to Improve DNA Sequencing Accuracy

David O. Nelson[1,2], and Terence P. Speed[2]

[1]Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550. [2]Statistics Department, University of California, Berkeley, California 92041.

LLNL is investigating statistical approaches to the problem of determining the DNA sequence underlying data obtained from fluorescence-based gel electrophoresis. Several features of electrophoresis make it interesting to statisticians and probabilists:

- the physical, chemical, and stochastic behavior of the process is complex and still not completely understood
- the yield of fragments of any given size can be quite small and variable
- the mobility of fragments of a given size can depend in predictable ways on the terminating base

In addition, the data generation process in fluorescence-based sequencing poses interesting statistical problems:

- the data consists of samples from one or more continuous, non-stationary signals
- boundaries between segments generated by distinct elements of the underlying sequence are ill-defined or nonexistent in the signal
- the sampling rate of the signal greatly exceeds the transition rate of the underlying discrete sequence

Recently published approaches to base calling, such as Giddings et al. [1] and Tibbetts et al. [2], address some of these issues using elementary statistics and heuristic decision procedures. While such approaches do tend to out perform the native software, the level of improvement in four dye-per-lane systems appears to diminish rapidly beyond 350-400 bases. Further improvements through software will have to come from a more sophisticated approach to recovering sequence from signal.

Our approach to signal recovery and base calling involves combining a stochastic model of the electrophoresis process, which describes the diffusion of DNA through a gel, with adaptive equalization techniques from digital communications theory to recover the underlying sequence. We will present the initial results of our investigation of the extent to which this approach enables us to increase base calling accuracy by providing a rational, statistical foundation to the process of deducing sequence from signal.

Research by D. O. Nelson was performed under the auspices of the U. S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48, with additional support from NSF grant DMS-91-13527. Research by T. P. Speed was partially supported by NSF grant DMS-91-13527.

[1] Giddings, M. C., R. L. Brumley, M. Haker, and L. M. Smith (1993). An adaptive, object-oriented strategy for base calling in DNA sequence analysis. *Nucleic Acids Research* , **21**(19), 4530-4540.

[2] Tibbetts, C., J. M. Bowling, and J. B. Golden III, (1994). Neural networks for automated base calling of gel-based DNA sequencing ladders. In J. C. Venter (Ed.), *Automated DNA Sequencing and Analysis Techniques*, Chapter 31, 219-229. Academic Press.

Towards Simplifying Fragment Assembly

Eugene Myers

Department of Computer Science, University of Arizona, Tucson, AZ 85721

The key computational problem in assembling fragments in a project involving some amount of shotgun sequencing is to determine which overlaps between fragments to use in forming contigs. Particularly when the underlying target sequence is repetitive, there are often several possible overlaps and different choices lead to distinct and often suboptimal reconstructions of the target. We have discovered a very simple preprocess which identifies and melds all fragments that must overlap in any optimal solution, leaving us with an assembly problem that is small and for which the key combinatorial choices are clear. For example, on sequence that is non-repetitive, the process usually melds all fragments providing the one, unique layout for the fragments. This confirms an observation by many that the simple "greedy" algorithm used in most fragment assembly systems is adequate for such problems. The real potential of our simplification arises in problems involving repetitive sequence. For example, given a sequence with 3 exact copies of a 2kb repeat, our process results in 5 contigs that involve 6 overlaps and for which the distinct arrangements are easy to enumerate and the optimal one is self-evident.

We will present our simplification strategy (easily understood by non-mathematicians) along with empirical results showing the reductions gained for various types of sequence. We will then discuss some interesting aspects of the reduced problems and how we can lever them to rigorously solve for the target sequence. What is particularly exciting here is that the combinatorial reduction raises our expectations of being able to solve problems involving *constraints* and *repetitive* sequence in a rigorous, accurate way. It also permits us to correctly formulate the assembly problem in terms of the most probable solution using the one-sided Kolmogorov-Smirnov statistic.

If time permits, we will also report on our progress to assembly 50,000 fragments covering a 700kb stretch of E. Coli with no augmenting or supporting information (data supplied by Fred Blattner's lab). This problem involves 13mb of raw data and is of a scale and order of magnitude larger than any tried before. We are using a combination of supercomputers, state-of-the-art algorithms, and of course, our new simplificatio technique.

Bayesian Restoration of a Hidden Markov Chain with Applications to Sequence Alignment

Gary A. Churchill
Cornell University
Ithaca, NY

Abstract

The problem of assembling DNA sequence fragments generated by a shotgun sequencing project is addressed. In general, the multiple alignment of sequences is recognized to be a difficult and computationally intensive problem. However in the shotgun sequencing context the problem can be reduced to a series of pairwise alignments.

A hidden Markov model describes the process of generating a fragment sequence **f** by copying a subsequence of the bases in a clone sequence **s**. When the sequence **s** and the error rates of the copying process θ are known, each fragment is a conditionally independent realization of this process. We derive the probability distribution of the pairwise sequence alignments between **s** and individual fragment sequences and describe an algorithm that samples alignments according to this distribution.

This sampling from the distribution of assemblies is one step in a Monte Carlo algorithm for consensus estimation. It is intended to address the problem of ambiguities that will inevitably arise in any assembly and present difficulties for consensus and error rate estimation.

We will indicate how the hidden Markov chain model can be adapted to account for empirical error characteristics of the sequencing process. These extensions of the basic model include context dependent errors (e.g. compressions and homopolymer runs) as well variations in accuracy along the length of individual fragments.

Estimating Consensus DNA Sequences

Tom Blackwell, Howard Hughes Medical Institute and Department of Genetics,
Harvard Medical School, 200 Longwood Ave. Boston, Ma 02115
617/432-0503, Fax:-7266, Internet: *blackwel@twod.med.harvard.edu*

Individual DNA sequence reads are commonly truncated after 350 – 500 base pairs to exclude less accurate data at the tail of each read. A data analysis originally developed in a different context may help take advantage of this lower quality DNA sequence information.

The analysis forms a consensus from any number of inaccurate observations of a single DNA sequence; it allows for insertions and deletions in each observed sequence, as well as substitutions. This analysis is more tolerant of insertions and deletions than algorithms currently used, because it averages over all possible multiple alignments of the observed sequence data, rather than fixing on one alignment. Thus the consensus does not depend on details of one choice of alignment. The averaging process relies on estimates of the base calling error rate at each site in the observed sequences. Others have recently described several ways to make such estimates. Consensus sequence formation is only one component in an integrated approach to sequence assembly.

The analysis was developed in the context of proposed 'single molecule' DNA sequencing methods. In this context, one may suppose that error rates are constant throughout the observed sequences, and that errors occur independently from one base to another within each observed sequence, and independently from one observed sequence to the next. It is of interest to know how much observational error can be tolerated and still produce a consensus which is close to the underlying DNA sequence.

2.5 Mb of simulations show that the consensus formed with this approach will reach an accuracy of 10^{-4} errors per nucleotide when it is constructed either using seven observed sequences with 6% error rate in each (insertions, deletions and substitutions combined) or using twenty observed sequences with 20% error rate. It reaches an accuracy of 10^{-3} errors per nucleotide when constructed with sixty observed sequences at 50% error rate. These results are obtained with equal proportions of insertions, deletions and substitutions among the base calling errors, and equal chance of error at every site.

This work was begun at the Center for Human Genome Studies, Los Alamos National Laboratory under U.S. DOE grant B 04861/F118 and continued under Office of Naval Research grant N00014-86K-0246 and National Science Foundation grant DMS-91-04990 to Herman Chernoff at Harvard University. Current support includes U.S. DOE grant FG02-87-ER60565 to George Church at Harvard Medical School.

GnomeView: User Test Results and Creation of a Client-Server System

Richard J. Douthart and Gregory Thomas

Pacific Northwest Laboratory; Battelle Blvd. Richland Washington, 99352

GnomeView is a graphics oriented, database integrated, interface to the Human Genome. A test version has been available to the scientific community via FTP[1] for over a year. The test version includes database abstractions, a users manual, and the GnomeView executable.

GnomeView queries have been monitored electronically and user opinion polled by survey. Each query is tallied and monitored with respect to its specific nature. Results are summarized weekly at the Pacific Northwest Laboratory. These inquiries reveal important information about usage trends of Graphics Users Interfaces in general and the GnomeView interface in particular.

The results of the usage and user opinion surveys have been the basis for formulating specifications and requirements for GnomeView acting in a client-server mode. Present plans are for the server to be located at Oak Ridge National Laboratory integrated with other DOE developed software tools.

This work is supported by the DOE Genome Program (project 13511) under contract DE-AC06-76RLO 1830. PNL is operated for the DOE by the Battelle Memorial Institute

[1]To obtain information about GnomeView: gv-help@gnome.pnl.gov
To obtain instructions on system requirements and downloading: gv-request@gnome. pnl. gov

Sequencing

This page intentionally left blank.

Preparation of Oligonucleotide Arrays for Hybridization Studies

Michael C. Pirrung, Steven W. Shuey, David C. Lever, Lara Fallon, J.-C. Bradley, and William P. Hawe

P. M. Gross Chemical Laboratory, Department of Chemistry, Box 90346, Duke University, Durham, NC 27708.

This project is aimed at developing reliable, high-quality chemical synthetic methods to prepare high-density arrays containing thousands of short DNA sequences. Such arrays can be used to sequence DNA by hybridization using principles enumerated by several groups.[1] Arrays consisting of complete sets of DNA of a given length can be prepared by the novel technique of light-directed synthesis,[2] and photoremovable groups are key to this method. We have developed a superior new photoremovable group for light-directed DNA synthesis that gives a byproduct that is chemically inert and readily measured by optical and fluorescence methods, permitting the yield in each of the photochemical deprotection steps to be verified. The 3',5'-dimethoxybenzoin (DMB) protecting group has been developed for the phosphotriester method of DNA synthesis.[3] It has been used to prepare by solution synthesis several trinucleotides that include sequences containing 5MeC and I. The unnatural nucleotides are intended to 1) increase the melting temperature of hybrids and 2) increase the discrimination (ΔT_m) between one-base mismatches end at the ends of the oligomers (fraying) and perfect hybrids by flanking the reading sequences with a "universal" base that will non-specifically increase base stacking and hydrogen bonding and thereby the T_m. DNA prepared with this method is tethered to the solid phase through its 5' end. The 3',5'-dimethoxybenzoincarbonate photoremovable protecting group has been developed for the phosphoramidite method of DNA synthesis. It has been used to prepare a mixed-sequence decanucleotide by solid-phase synthesis. The purity and sequence of this compound was verified by removal from the support followed HPLC analysis, end-labeling and PAGE analysis, and snake venom phosphodiesterase-catalyzed degradation to mononucleotides followed by HPLC analysis. DNA prepared with this method is tethered to the solid phase through its 3' end. Finally, the 3',5'-dimethoxybenzoincarbonate method has been applied to phosphoramidite synthesis of DNA tethered to the solid phase through its 3' end, permitting the use of arrays in enzymatic transformations such as primer extensions and ligations.

References

1. Bains, W.; Smith, G. C. *J. Theor. Biol.* **1988**, *135*, 303-307. Drmanac, R.; Labat, I.; Brukner, I.; Crkvenjakov, R. *Genomics* **1989**, *4*, 114-128. Khrapko, K. R.; Lysov, Y. P.; Khorlyn, A. A.; Shick, V. V.; Florent'ev, V. L.; Mirzabekov, A. D. *FEBS Lett.* **1989**, *256*, 118-122. Lysov, Y. P.; Florent'ev, V. L.; Khorlyn, A. A.; Khrapko, K. R.; Shick, V. V.; Mirzabekov, A. D. *Dokl. Akad. Nauk. SSSR* **1989**, *303*, 1508-11.

2. Fodor, S.P.A.; Read, J.L.; Pirrung, M.C.; Stryer, L.; Liu, A.T.; Solas, D. *Science* **1991**, *251*, 767.

3. Pirrung, M. C.; Shuey, S. W. *J. Org. Chem.* **1994**, *59*, 0000.

Oligonucleotide Microchip is a Versatile Tool for DNA Analysis: Diagnostic Applications

Igor Ivanov, Eugene Kirillov, Victor E.Barsky, Edward Kreindlin, Sergei Parinov, Edward Timofeev, Gennady Yershov, Vladimir L. Florentiev, and Andrei D. Mirzabekov.

V.A Engelhardt Institute of Molecular Biology, 32 Vavilov Str., B-334, Moscow, 117984, Russia

Sequencing by Hybridization (SBH) is a kind of DNA technology based on the assumption to extract information about DNA sequence by specific hybridization with a set of oligonucleotides [1]. The architecture and complexity of the set defines the information capability. For example, hundreds oligonucleotides could be enough for diagnostics purposes and polymorphism analysis. These oligonucleotides can be fixed or even synthesized on the matrix and then hybridized with DNA.

We continue to develop the microchip DNA technology characterized by immobilization of oligonucleotides into small "cells" (upto 30x30 μm) of gel-support fixed on the glass plate [2]. After hybridization with DNA, fluorescent signals from each "cell" are detected and analyzed by specially designed microscope with CCD camera. Porous 3D structure of the polyacrylamide gel ensures beneficial properties for hybridization and washing out kinetics. The high capacity of immobilization oligonucleotides on gel-support and high hybridization capacity allow also using intercalating dyes for detecting perfect duplexes. In parallel with "standard" hybridization DNA to short oligonucleotides, "contiguous stacking" hybridization has been developed. In the latter case, fluorescently labeled pentamers are hybridized in juxtaposed position to the duplex formed between DNA and immobilized oligos. It has been shown that only perfect "contiguous stacked duplexes" are detected.

We have examined microchip technology in the model diagnostic experiments to identify mutations in blood samples of β-thalassemia patients. The reliable discrimination of point mutations has been achieved by hybridization fluorescently tagged DNA with a set of octamers and decamers. We are showing the possibilities to apply this technology for diagnostics of any mutations along the whole gene length and for studying gene polymorphism.

This work was supported by grants from Russian Human Genome Project, U.S. Department of Energy and Affimax Research Institute.

1. A.D.Mirzabekov, DNA sequencing by hybridization - a megasequencing method and a diagnostic tool? Trends in Biotechnology (1994), vol.12, pp.27-32.

2. K.R.Khrapko et al., A method for DNA sequencing by hybridization with oligonucleotide matrix, J.Sequencing and Mapping (1991), vol.1, pp.375-388.

Electronically controlled DNA hybridization on semi-conductor microchips: A miniaturized DNA chip format for rapid medical diagnostics and DNA sequencing.

Glen A. Evans[1,2], Harold R. Garner[1], Eugene Tu[2], William F. Butler[2], Ronald G. Sosnowski[2], Donald D. Montgomery[2] and Michael J. Heller[2]. [1]McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, 6000 Harry Hines Blvd., Dallas, Texas 75235; and [2]Nanogen, Incorporated, 11588 Sorrento Valley Road, San Diego, California 92121.

DNA hybridization is an important basic technique for molecular biology and medical diagnostics. We have developed a novel technology which allows multiplex DNA hybridization reactions to be carried out in microarray formats or the surface of a unique semiconductor device, or DNA microchip. This technology, termed APEX (for Active Programmable Electronic Matrix) utilizes microdevices containing arrays of electronically addressable microlocations on a silicon surface fabricated using microlithography. Arrays are designed such that direct electronic currents can be independently applied and controlled at each element on the array. Specific oligonucleotide probe sequences are covalently attached to the surface of metal microeletrodes, or to polymer chip coatings, and self-assembly can be directed by the microchip. A prototype APEX device (Nanogen 8850 Chip) has been manufactured and contains 64 addressable microlocations (50 μm x 50 μm) in a space of less than 1 mm^2 on a silicon/silicon dioxide substrate and metal electrodes that can be operated in a direct current mode. The device operates by placing a DNA sample over the array and applies positive bias to the underlying microelectrodes. Target DNA becomes immediately concentrated at each microlocation (by a factor of $> 10^6$) through free field electrophoresis and, as a result of the electrophoretic concentration, the hybridization rate for the target probe and test probe is greatly increased. Reversing the polarity of the electric field allows the removal of non-hybridized DNA, while quantitative control of the reverse bias allows independent adjustment of hybridization stringency at each test site. Exquisite control of hybridization stringency conditions can be achieved at room temperature through electronic control. The hybridization signal of fluorescently-labeled target DNA is detected using CCD arrays or external cooled CCD cameras. This technique has a number of powerful and novel features: 1) it allows hybridization reactions to be programmed, carried out automatically, and results obtained within seconds; 2) it allows precise and independent control of hybridization stringency over each individual sequence in a hybridization array; 3) electronic control over each hybridization site allows highly selective single base miss-match discrimination to be achieved with probes ranging from 7 to 22 nucleotides. We have utilized APEX devices for clinical HLA typing as well as for other diagnostic and second generation chips with a larger number of elements are being developed for massive parallel DNA analysis for human genotyping and for sequencing by hybridization applications. This novel technique may have a variety of other research and commercial applications as well as special applications in the Human Genome Project.

Positional Effects of Nearest-Neighbor Base Pairs and Mismatched Base Pairs in Short DNA Duplexes

Mitchel J. Doktycz, [1]Maxwell Morris, K. Bruce Jacobson, [3]Kenneth L. Beattie, and [2]Robert S. Foote
Health Sciences Research Division, [1]Math and Computer Sciences Division, and [2]Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-8080
[3]DNA Technology Laboratory, H.A.R.C., The Woodlands, Texas 77381

The careful characterization of the hybrid stability of short DNA molecules will be necessary for the success of SBH as a sequence diagnostic method. Knowing the conditions to achieve maximal hybrid stability will allow design of protocols to achieve optimal discrimination against mismatched hybrids. Furthermore, knowledge of the hybrid stability will be a guide in selection of sample preparation procedures, in the selection and positioning of the immobilized DNAs and in effective labeling and detection schemes.

Prediction of the duplex stability of molecules of the type to be used in SBH can be based on thermodynamic stability data. Such data have been obtained primarily from experiments on longer DNAs. For shorter oligomers, end effects, which refer to the initiation of melting from the solvent exposed ends, complicate the application of thermodynamic data collected on long DNAs to the types of molecules used in SBH. This effect may be propagated several base pairs in from the end and can contribute to destabilization of the end base pairs and reduced diagnostic accuracy by increasing the relative stability of mismatches. To evaluate this interaction and others which may be peculiar to small DNA molecules, a systematic evaluation of solution hybridizations has been conducted. The molecule sets studied are based on octamers of the sequences $^{5'}$NNNTGGAC$^{3'}$ (set 1), $^{5'}$GCNNNGAC$^{3'}$ (set 2) and the respective complements $^{5'}$GTCCANNN$^{3'}$ and $^{5'}$GTCNNNGC$^{3'}$, where N = A, T, G, or C for a total of 256 molecules. Modeling of melting temperatures of the 128 perfectly matching duplex molecules have allowed evaluation of the stability of A•T and G•C base pairs and the 10 nearest neighbor stacking interactions as a function of base pair position. These same molecule sets have also been used for the evaluation of the 8 mismatched base pairs as a function of position.

Concurrent with the optical melting studies is an examination of the hybridization of these octamers to immobilized 8-mers. DNA sequences corresponding to either one half of the set 1 or set 2 oligomers have been synthesized, using photodeprotection, onto a glass surface. Representative, complementary molecules were radioactively labeled and hybridized to determine the hybridization specificity.

(Research sponsored by the Office of Health and Energy Research, United States Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, additional support was received from NIH Grant 1 P20 HG00666.)

Accessing Genetic Information with DNA Arrays

Stephen P.A. Fodor and Robert Lipshutz

Affymetrix, 3380 Central Expressway, Santa Clara, CA 95051

Recent advances in DNA chip technology have produced a robust platform for accessing genetic information. This platform consists of DNA chips, instrumentation to read the chips and software to interpret the data. DNA oligonucleotide arrays are fabricated using high-resolution photolithography in combination with solid-phase oligonucleotide synthesis. Fluorescent nucleic acid targets are hybridized to the array. The hybridization pattern, as determined by epifluorescence microscopy, reveals details of the target sequence. Applications of this technology have been demonstrated including comparative gene sequencing, mutation detection, clone mapping and sequencing by hybridization. These applications are poised to address many of the goals of the Human Genome Project.

Sequencing by Hybridization: Towards an Efficient Large-Scale DNA Methodology

Radomir Crkvenjakov

Hyseq, Inc., 670 Almanor Ave., Sunnyvale, CA 94086

Hyseq has acquired the patent on SBH, a sequencing method in which computerized sequence assembly by overlap is not preceded by sequential monomer by monomer determination of restricted portions of sequence. Instead, determination of partial to complete lists of constituent oligomers in 1-2 kb DNA fragments by oligonucleotide hybridization provides the shortest possible stretches for sequence assembly.

In a technology transfer from Argonne National Laboratory, Hyseq obtained the rights on DOE sponsored research of an SBH format 1 production line capable of collecting ten million hybridization scores per day and to a novel variant of SBH - format 3 or "super chip." The large-scale methodology optimally combining SBH format 1 with automated single pass gel sequencing in a production-line setting can factor out the disadvantages of sequence assembly inherent to SBH while retaining its potential for acceleration of the sequencing process by several orders of magnitude. For example, sequencing a 4.5 Mb bacterial genome requires 10,000 gel sequencing reactions and 25 million hybridization scores with 3,500 heptamer probes; both tasks easily accomplished within a 6 month time-frame with accuracy of 0.001 at an initial cost of 20 cents per base.

Building on format 1 development, Hyseq will also focus on the unique niche in DNA diagnostics: screening a large number of multi-locus, multi-patient 200-2000 bp PCR samples by resequencing with a small number of probes in centralized facilities thus providing gene sequencing at a fraction of the present cost.

Development of Instrumentation for DNA Sequencing at a Rate of 40 Million Bases per Day

Edward S. Yeung,[1] Huan-Tsung Chang, Qingbo Li, Xiandan Lu

Ames Laboratory-USDOE and Department of Chemistry, Iowa State University, Ames, IA 50011.
[1]Corresponding author.

One of the many bottlenecks in high-speed DNA sequencing using available instrumentation is the electrophoretic separation and the subsequent identification of DNA fragments derived from the Sanger reaction. A combination of fast separation by capillary electrophoresis (30 bases per minute per channel) and simultaneous monitoring (100 channels in parallel) has led to a raw sequencing rate of 3,000 bases per minute in a single instrument. This technology is readily scalable to 1,000 capillaries to achieve a gross sequencing rate of 40 million bases per day.

The substantial increase in sequencing rate is a result of several technical advances in our laboratory. (1) The use of commercial linear polymers for sieving allows replaceable yet reproducible matrices to be prepared that have lower viscosity (thus faster migration rates) compared to polyacrylamide. (2) The use of a charge-injection device camera allows random data acquisition to decrease data storage and data transfer time. (3) The use of distinct excitation wavelengths and cut-off emission filters allows maximum light throughput for efficient excitation and sensitive detection employing the standard 4-dye coding. (4) The use of index-matching and 1:1 imaging reduces stray light without sacrificing the convenience of on-column detection.

Capillary Array Electrophoresis for High Throughput DNA Analysis

John Bashkin, David Barker, Dave Roach, Matt Bartosiewicz, Joe Leong, Tom Zarella, Rick Johnston

Molecular Dynamics, Sunnyvale, CA 94086

Capillary electrophoresis can be used effectively for rapid DNA analysis. Applications include DNA sequencing, diseases diagnostics, genetic typing, and genome mapping. However, experimental throughput has been limited by the requirement of running samples sequentially down a single capillary. Expanded use of the technique has also been hindered by the cost and reliability of polyacrylamide-filled capillaries, and the poor resolution achieved by currently available replaceable matrices. We have extended the work of R. Mathies [1] and developed a scanning instrument capable of electrophoresis and analyte detection in 48 capillaries simultaneously.

Our capillary array electrophoresis (CAE) instrument detects DNA through Laser Induced Fluorescence (LIF), employing an Ar^+ laser at 488nm. Hydroxyethyl cellulose of various molecular weights and percentages are used as the separation matrix, or sieving buffer. The system is connected to either a vacuum or pressure source to introduce the matrix into the capillaries rapidly and conveniently. The total turnaround time for replacement of the sieving buffer between electrophoretic separations is approximately 20 minutes. We present data on separations of dsDNA restriction fragments and PCR products, with single base-pair resolution over 100-400bp, and a total size range of 50bp-12kbp. The fragments are fluorescently labeled by inclusion of thiazol orange in the sieving buffer. Total experimental run time from sample injection to resolution of fragments up to 2kbp is 20-30 minutes. The electrical current is monitored in all capillaries during an experiment. This data is stored and can be used to correct for slight differences in the electrical power delivered to each capillary across the array.

Data analysis software, ArrayQuant™, has also been developed under the WindowsNT™ environment. This package accomplishes automatic peak detection, fragment sizing, and error analysis for the electropherograms. Data reports for all 48 capillaries are generated semi-automatically and conveniently transferred to Excel™ spreadsheets.

This work was funded by a Department of Energy SBIR Phase II Grant (DE-FG03-92 ER81299, D. Barker, P.I.).

[1] Huang, X.C., Quesada, M.A., Mathies, R.A. (1992) DNA Sequencing Using Capillary Array Electrophoresis. *Anal. Chem.*, **64**, 2149-2154.

MULTIPLE CAPILLARY DNA SEQUENCING

Norman J. Dovichi[1,2,3], Jian-Zhong Zhang[1], Jian-Ying Zhao[1], Jiang Rong[1], Rong Liu[1], John Elliott[2], Sue Bay[1], Pieter Roos[1], and Larry Coulson[1]

[1]Department of Chemistry and [2]Department of Medical Microbiology, University of Alberta, Edmonton, Alberta CANADA T6G 2G2. [3]corresponding author.

We have developed a multiple capillary DNA sequencer. This instrument has several important attributes. First, by operation at an electric field of 200 V/cm, we are able to separate DNA sequencing fragments rapidly and efficiently. Typically, 1.5 hours are required to separate fragments up to 550 bases in length. Second, the separation is performed with 5%T 0%C polyacrylamide. This non-crosslinked matrix has sufficiently low viscosity that it can be pumped from the capillary and replaced with fresh material when required. Third, the fluorescence detection cuvette is manufactured locally by means of microlithography technology. This detection cuvette provides robust and precise alignment of the optical system. Fourth, the system is based on Applied Biosystems' four color sequencing technology. We routinely separate both PRISM-labeled dideoxynucleotide terminated fragments and conventional primer-labeled sequencing fragments. The majority of molecular biologists are familiar and comfortable with this labeling technology, which should enhance the acceptance of the sequencer by the user community. Fifth, we use fiber-optic coupled avalanche photodiode photodetectors, which provide low noise and high sensitivity detection. We have obtained detection limits of 120 fluorescein molecules injected onto the capillaries. We operate a number of capillaries simultaneously, increasing the throughput of the system. The system operates with 5 to 32 capillaries; the five capillary device is ideally suited for small-scale sequencing projects, while the larger instruments are appropriate for genomic scale sequencing. Finally, we have developed an automated DNA base-calling algorithm.

This combination of speed, flexibility, and labeling technology provides the first example of a practical and useful DNA sequencer based on capillary electrophoresis. The instrument is very robust, and has been routinely used for six months. We have begun to sequence a number of templates, and we will report sequencing accuracy and sequence read length at the conference. We are currently using an exonuclease method to synthesize a set of nested deletions in a directed sequencing strategy. This approach greatly simplifies sequence assembly problems compared with a shot-gun sequencing strategy.

A next-generation system is under construction that will operate 96 capillaries simultaneously. This system will eventually be expanded to 864 capillaries. An 864 capillary instrument will produce nearly half a million bases of raw sequence data in a 1.5 hour run. Our automated base-calling algorithm is being developed for the 864-capillary instrument; the algorithm is very fast, requiring less than a second for a typical sequencing run.

DNA ANALYSIS BY CAPILLARY ELECTROPHORESIS: SEQUENCING AND MUTANT DETECTION

Barry L. Karger, Jan Berka, Marie C. Ruiz-Martinez, Steve Carson, Appasani S.M. Krishnarao, Kirstin Hebenbrock and Frantisek Foret

Barnett Institute, Northeastern University, Boston, MA 02115

A core technology based on capillary electrophoresis with a replaceable linear polyacrylamide sieving matrix and laser-induced fluorescence detection has been developed for DNA analysis. We will first report on the application of this technology to DNA sequencing. Specifically, we are using the modular primer-walking strategy and dye-labelled terminators for sequencing stretches of DNA at a minimum rate of 500 bases/hr. The design of a fully automated multiple capillary instrument which can replace the polymer matrix after each run will be shown. In this design, emphasis has been placed on flexibility, for use with a variety of excitation and emission wavelengths, while maintaining low detection limits and high spectral resolution. The front-end automation for primer selection, Sanger reaction, sample clean-up and injection will also be described.

A second topic will deal with high resolution separation and analysis of point mutations, using the above core technology. Both SSCP and CDCE (constant denaturant capillary electrophoresis) have been developed utilizing multiple dye-labelling strategies. The excellent resolution of heteroduplexes from homoduplexes in CDCE permits detection of mutants in low frequency relative to wild type (< 1 part per thousand). With appropriate post-column collection, it is possible to sequence individual species for mutant structure determination.

NEW DIRECTIONS IN HIGH-SENSITIVITY FLUORESCENCE DETECTION OF DNA AND CAPILLARY ARRAY ELECTROPHORESIS*

Richard A. Mathies, Jingyue Ju[1], Huiping Zhu, Steven M. Clark, Adam T. Woolley, Yiwen Wang, Scott C. Benson and Alexander N. Glazer, Chemistry Department and Department of Cell and Molecular Biology, University of California, Berkeley CA 94720.

Capillary Array Electrophoresis, coupled with confocal fluorescence detection,[2] is a valuable new technique for performing high-speed, high-throughput DNA sequencing[3,4] and fragment sizing.[5] In current implementations, up to 50 capillaries can be run in parallel and separations are complete ~10 times faster than slab gels. Our recent efforts have focused on (i) the development of improved fluorescence reagents for DNA sequencing, (ii) the development of methods for using intercalating dyes for double-stranded DNA fragment detection in CE, (iii) and the development of methods for microfabricating capillary arrays.

To improve the spectroscopic properties of the labels used in DNA sequencing, we have synthesized sequencing primers labeled with pairs of dyes that are coupled by fluorescence energy transfer (ET).[6] The donor is chosen to provide intense absorption at the laser wavelength (488 nm) while the acceptors are chosen to provide large Stokes shifts and distinctive emission spectra. The spacing of the dyes along the primer is selected to provide efficient energy transfer with no quenching. The mobility shifts of the ET primers are less than those observed using current commercial dye-labeled primers and the fluorescence intensities are as much as 5 times stronger. Results using these ET dye-labeled primers for 4-color DNA sequencing (collaboration with Dr. Carl Fuller at U. S. Biochemical) and for 2-color allelic fragment detection will be presented.

A variety of monomeric and dimeric intercalating dyes have been used to improve the sensitivity and versatility of DNA fragment detection in CE.[7] The monomeric intercalating dyes TO, TO6, YO, Propidium 2 and Propidium 3 have been used for on-column staining. TO and TO6 are able to detect as little as 1 fg/μL of a 600 bp fragment in the initial sample. Separations of DNA precomplexed with the dimeric intercalating dyes EthD, TOTAB, and YOYO have also been successful, although it is important to work at low dye:DNA ratios and to use 9-aminoacridine in the running buffer to achieve high-resolution separations.

In a continuing effort to miniaturize DNA analysis systems, we have used photolithographic techniques to microfabricate capillary arrays.[8] These arrays were fabricated on planar glass slides by first etching channel patterns, and then forming the capillaries by thermally bonding the etched substrate to a top glass plate. The channels have an effective length of 3.5 cm and the 10 micron deep channels ranged from 30-120 μm in width. Using these CAE chips, high-resolution separations of double-stranded ΦX/HaeIII DNA have been performed from 70-1000 bp in only 120 seconds! Since up to 100 such channels can be fabricated on a single glass slide, this work establishes the feasibility of developing miniaturized DNA analysis chips.

*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-FG-91-61125.

[1]DOE Human Genome Distinguished Postdoctoral Fellow

[2]R.A. Mathies et al., *Rev. Sci. Instrum.* **65**, 807-812 (1994).
[3]R.A. Mathies and X.C. Huang, *Nature (London)* **359,** 167-168 (1992).
[4]R.A. Mathies and X.C. Huang, *Automated DNA Sequencing and Analysis*, eds. M.D. Adams, C. Fields & J. C. Venter, Academic Press, pp. 17-28 (1994).
[5]S.M. Clark and R.A. Mathies, *Anal. Biochem.* **215**, 163-170 (1993).
[6]J. Ju, C. Fuller, C. Ruan, A.N. Glazer and R.A. Mathies, Fluorescence Energy Transfer Primers for DNA Sequence Analysis, in preparation.
[7]H. Zhu, S.M. Clark, S.C. Benson, H.S. Rye, A.N. Glazer and R.A. Mathies, *Analytical Chemistry* **66,** 1941-1948 (1994).
[8]A.T. Woolley and R.A. Mathies, Ultra-High-Speed DNA Fragment Separations Using Microfabricated Capillary Array Electrophoresis Chips, *Proc. Natl. Acad. Sci. U.S.A.*, in press.

Sequencing of DNA by Gel Electrophoresis in Micromachined Channels

Joe Balch[1], Courtney Davidson[1], Jeff Gingrich[1], Muhammad Sharaf[2], Larry Brewer[1], Jackson Koo[1] Doug Smith[2], Michael Albin[2], and Anthony Carrano[1]

[1]Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore California. 94550, [2]Perkin Elmer Corporation, Applied Biosystems Division, Foster City, California 94404.

Sequencing of DNA by gel electrophoresis is typically performed in slab gel systems. Efforts to increase sequencing rates have generally relied on increasing sample load capacity (higher lane density or multiple capillary systems) and increasing the electric field to obtain faster fragment separation (requiring the use of thinner gels or capillaries to reduce heat dissipation). As an alternative to slab and capillary systems, we are investigating a hybrid technique based upon a high density array of electrophoresis channels fabricated using micromachining technologies on a single, large substrate at fixed locations. Furthermore, standard polyacrylamide gel (PAG) compositions can be poured into the channels using conventional techniques without the problem of bubble formation and other defects commonly incurred in PAG filled capillaries.

We have found that electrophoretic resolution is dependent upon the surface finish of the micromachined channels. We have developed and refined fabrication techniques that result in electrophoretic resolution in microchannels comparable to the electrophoresis resolution measured in standard slab gels formed between two flat glass plates. We have obtained resolution that results in accurate base calling to greater than 500 DNA bases per channel (200 micrometer deep by 1 mm wide by 25 cm long microchannels filled with 6% PAG). Present efforts are underway to develop large electrophoresis channel arrays on a single glass substrate for high throughput DNA sequencing.

This work was performed under a Cooperative Research and Development Agreement (CRADA) between Perkin-Elmer Corporation, Applied Biosystems Division and by Lawrence Livermore National Laboratory under the auspices of the U.S. Department of Energy contract no. W-7405-Eng-48.

High-Speed Automated DNA Sequencer Utilizing From-the-Side Laser Excitation

Michael S. Westphall, Robert L. Brumley Jr., Eric C. Buxton, and Lloyd M. Smith

Department of Chemistry, University of Wisconsin, Madison, WI 53706

The Human Genome Initiative is an ambitious international effort to map and sequence the three billion bases of DNA encoded in the human genome. If successfully completed, the resultant sequence database will be a tool of unparalleled power for biomedical research. One of the major challenges of this project is in the area of DNA sequencing technology. At this time, virtually all DNA sequencing is based upon the separation of DNA fragments in high resolution polyacrylamide gels. This method, as generally practiced, is one to two orders of magnitude too slow and expensive for the successful completion of the Human Genome project. One reasonable approach to improved sequencing of DNA fragments is to increase the performance of such gel-based sequencing methods.

Decreased sequencing times may be obtained by increasing the magnitude of the electric field employed. This is not possible with conventional sequencing, due to the fact that the additional heat associated with the increased electric field cannot be adequately dissipated. Recent developments in the use of thin gels have addressed this problem. Performing electrophoresis in ultrathin (50 to 100 microns) gels greatly increases the heat transfer efficiency, thus allowing the benefits of larger electric fields to be obtained. An increase in separation speed of about an order of magnitude is readily achieved. Thin gels have successfully been used in capillary [1] and slab formats [2,3].

A detection system has been designed for use with a multiple fluorophore sequencing strategy in horizontal ultra-thin slab gels. The system employs laser through-the-side excitation and a cooled CCD detector; this allows for the parallel detection of up to 24 sets of four fluorescently labeled DNA sequencing reactions during their electrophoretic separation in ultrathin (115μm) denaturing polyacrylamide gels. Four hundred bases of sequence information is obtained from 100ng of M13 template DNA in an hour, corresponding to an overall instrument throughput of over 9600 bases/hr. A detailed description and the operating characteristics of this system are presented.

[1] Lucky, J. A., Drossman, H., Kostichka, A. J., Mead, D. A., D'Cunha, J., Norris, T. B., and Smith, L. M., *Nucleic Acids Research*, **18**, 4417, (1990)
[2] Brumley, R. L., Jr., and Smith L. M., *Nucleic Acids Research*, **19**, 4121 (1991).
[3] Kostichka, A. J., Marchbanks, M., Brumley, R. L., Drossman, H., and Smith, L. M., *Bio/Technology*, **10**, 78 (1991).

Fluorescent Sequencer Development

William F. Kolbe and Jocelyn C. Schultz

Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720

In order to address the need for increased DNA sequencing capability, we have initiated a program of gel-based fluorescent sequencer development. Experiments are being performed on two separate platforms. A conventional sized (30 cm x 25 cm x 350 mm) gel apparatus is used to develop improved fluorescence detection methods in order to obtain both high spatial resolution and fast response. This will permit a significant increase in the number of samples which can be loaded on a given gel. At the same time, an ultra-thin gel apparatus (15 cm x 25 cm x 50–100 mm) is being constructed to provide high-speed electrophoretic separations. Detection methods developed in the conventional sized gel apparatus will be incorporated in the ultra-thin gel apparatus as they become available.

The detection system uses a laser beam passing transversely through the gel to excite all of the DNA lanes simultaneously and a fiber optic array to collect the fluorescence produced. A lens array is employed to image the fluorescence onto the ends of the fibers. The output end of the fiber array is formed into a compact rectangular shape compatible with a cooled charge-coupled device camera used to detect the light. The alignment of the laser beam passing through the gel is stabilized by means of a feedback system employing a motorized steering mirror and photodiode detectors positioned at each end of the fiberoptic array. The spatial resolution of the system is sufficient to permit the detection of 100 lanes in a gel width of 25 cm. The system is currently operational with single color detection (one DNA sequence ladder per lane) and is being extended to a four color configuration.

The ultra-thin gel apparatus has been constructed and is undergoing preliminary testing. It should be possible to incorporate the detection system described above in this apparatus in the near future. This combination should lead to greatly enhanced sequencing capabilities.

DESIGN, SYNTHESIS AND CHARACTERIZATION OF ENERGY TRANSFER FLUORESCENT DYE TAGGED OLIGONUCLEOTIDE PRIMERS FOR DNA SEQUENCE ANALYSIS*

Jingyue Ju, Chihchuan Ruan[1], Carl W. Fuller[1], Alexander N. Glazer and Richard A. Mathies
Department of Chemistry and Department of Cell and Molecular Biology, University of California, Berkeley, CA 94720-1460. [1]United States Biochemical Corporation, P.O. Box 22400, Cleveland, Ohio 44122

The objective of our work is to develop novel fluorescently labeled primers for DNA sequencing and other types of multicolor genetic analysis that allow higher levels of detection sensitivity.[2] We describe here: (1) the design and synthesis of energy transfer (ET) fluorescent dye-labeled oligonucleotide primers with improved spectroscopic and electrophoretic properties; (2) The evaluation of these primers in DNA sequencing on an ABI 373A DNA sequencer and a capillary array electrophoresis (CAE) system. Specifically, energy transfer fluorescent dyes (donor and acceptor) are incorporated on the M13 (– 40) universal primer. The primers are denoted as D-N-A, where D is the donor, A is the acceptor, and N is the number of bases between D and A. In all the primers prepared, the donor is the fluorescein derivative FAM (F) incorporated at the 5' end of the oligonucleotide. The acceptors are the commercially available dyes JOE (J), TAMRA (T) or ROX (R) incorporated at the location of a modified T in the primer sequence. Twenty ET primers with the same donor chromophore and different acceptors having different donor-acceptor separations were synthesized and purified. The evaluation of the spectroscopic and electrophoretic properties of these primers led to an elegant way of tuning the mobility shift of the primers on electrophoresis as well as achieving optimum fluorescence intensities. This is achieved by adjusting the spacing of the donor and acceptors to get a similar mobility match and still maintain good energy transfer. Four of the primers (F-10-F, F-10-J, F-3-T and F-3-R) with optimal properties were selected for four-color DNA sequencing on an ABI 373A sequencer. The fluorescence intensity of the DNA sequencing fragments generated with ET primers is 2-5 times higher than that of the corresponding fragments generated with the conventional single dye-labeled primers. In a typical sequencing run, a 500-base sequence is determined with an accuracy of >99%. The design, synthesis, purification and spectroscopic properties of ET primers, the way of adjusting the electrophoretic properties of the ET primers, and the DNA sequencing protocol with ET primers will be presented.

*This research was 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-FG-91-61125. J. J. was supported by a Human Genome Distinguished Postdoctoral Fellowship sponsored by the U.S. Department of Energy, Office of Health and Environmental Research, and administered by the Oak Ridge Institute for Science and Education. Support from United States Biochemical Corporation is also gratefully acknowledged.

[2]J. Ju, C. W. Fuller, C. Ruan, A. N. Glazer and R. A. Mathies, Fluorescence Energy Transfer Primers for DNA Sequence Analysis, in preparation.

Luminescent Lanthanide Ions as Labels for DNA Sequencing and Mapping

Gilbert M. Brown[1], Jeffrey E. Elbert,[1] Frederick V. Sloop,[2] Mitchel J. Doktycz,[3] Richard A. Sachleben,[1] and K. Bruce Jacobson[3]

Chemical and Analytical Sciences Division,[1] Biology Division,[2] and Health Sciences Research Division,[3] Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37831-6119

The goal of this research program is the development of a new luminescent labeling system for DNA that will be a successor to the fluorescent organic dyes currently used for sequencing. This labeling system can also be used with hybridization probes for DNAs attached to nylon or nitrocellulose membranes (Southern Blot). Luminescence of the lanthanide ions Sm(III), Eu(III), Tb(III), and Dy(III) has narrower bandwidths than the fluorescent organic compounds currently used to label DNA, and these narrow linewidths will allow multiple probes to be detected simultaneously with little overlap of emission and potentially to get greater resolution in the bands. Furthermore, the long lifetimes allow detection with a lower background. A derivative of the macrocyclic chelating agent, 1,4,7,10-tetraazacyclododecane-1,4,7,10-tetraacetic acid (DOTA), is used to attach the lanthanide [Ln(III)] ions to oligonucleotides, and this ligand forms stable, kinetically inert complexes with these metal ions. The ligand, as an isothiocyanate derivative, is reacted with a hexylamine linker arm on an oligonucleotide to generate a labeled primer for the Sanger sequencing procedure or PCR. Detection sensitivity for luminescence from Ln(III) ions can be greatly enhanced if excitation is to a ligand based state having a long lived triplet electronic state. A naphthylmethyl derivative of DOTA and an acetophenone derivative of DO3A (tetraazacyclododecanetriacetic acid) were prepared and the sensitizer groups served as antennas to funnel excitation energy to complexed Ln(III) ions. The benzene ring of the acetophenone derivative is being nitrated so that upon reduction the resulting amine will provide an attachment point to DNA. A third derivative, the benzylisothiocyano derivative of tetraazacyclododecane with three acetate arms and an acetophenone antenna has been prepared and characterization is in progress. The chemistry for attachment of the ligand to oligonucleotides is being developed so that it can be added as the last step in the automated synthesis of the oligonucleotide while still attached to a solid support. Detection limits for oligonucleotides, labeled with the Ln(III) reagents, were determined following capillary electrophoresis with on-column detection using a pulsed laser and time-gated luminescence detection. Work is in progress toward preparing and testing an M13 primer, labeled with an Eu(III) complex having a sensitizing antenna group, for use in the Sanger sequencing procedure. Initial sequencing experiments will be carried out using the well studied M13mp18 system. We will utilize two-color labeling using Eu(III) and Tb(III) complexes with capillary electrophoresis to separate the DNA fragments and two filter-photomultiplier tube assemblies for luminescence detection. These results will guide development of a second generation instrument which will have a polychromator/CCD detector for simultaneous detection of all four Ln(III) ions.

This research was sponsored by the Office of Health and Environmental Research, U. S. Department of Energy, under contract No. DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc. This research was sponsored in part by the appointment of J. E. Elbert to a Human Genome Distinguished Postdoctoral Fellowship sponsored by the U. S. Department of Energy, Office of Health and Environmental Research, and administered by the Oak Ridge Institute for Science and Energy.

Laser Desorption Mass Spectrometry for Fast DNA Sequencing

C. H. Winston Chen,[1] Kai Tang, Nelli I. Taranenko, and Steve L. Allman

Photophysics Group, Chemical Physics Section, Health Sciences Research Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831. [1]Corresponding author.

During the past year, we have achieved some major breakthroughs for using mass spectrometry for DNA analysis. They are (1) successful detection of single-stranded DNA the size of 500 nucleotides and double-stranded DNA the size of 500 base pairs; (2) the detection sensitivity of large DNA segments reaches to the femtomole region, and (3) the first demonstration of using laser desorption mass spectrometry for cystic fibrosis diagnostics. We believe these achievements should become important milestones toward the use of mass spectrometry for fast DNA sequencing. There is a good possibility that fast mass spectrometric DNA sequencing can be demonstrated in the very near future.

Laser desorption mass spectrometry has been considered as a potential new method for fast DNA sequencing. Our approach is to use matrix-assisted laser desorption to produce parent ions of DNA segments and a time-of-flight mass spectrometer to identify the sizes of DNA segments. Thus, the approach is similar to gel electrophoresis sequencing using Sanger's enzymatic method. However, no gel, no radioactive tagging, and no dye labeling are required. In addition, the sequencing process can possibly be finished within a few hundred microseconds instead of hours and days. In order to use mass spectrometry for fast DNA sequencing, the following three criteria need to be satisfied. They are (1) detection of large DNA segments, (2) sensitivity reaches the femtomole region, and (3) mass resolution good enough to separate DNA segments of a single nucleotide difference. It has been very difficult to detect large DNA segments by mass spectrometry before due to the fragile chemical properties of DNA and low detection sensitivity of DNA ions. We discovered several new matrices to increase the production of DNA ions. By innovative design of a mass spectrometer, we can increase the ion energy up to 75 KeV to enhance the detection sensitivity. At present, we have fulfilled two key criteria for using mass spectrometry for fast DNA sequencing. The major effort in the near future is to improve the resolution. Different approaches are being pursued. When high resolution of mass spectrometry can be achieved and automation of sample preparation is developed, the sequencing speed to reach 500 megabases per year can certainly be feasible.

In addition to the three major achievements described above, other important accomplished works are:

(1) Success of detecting DNA samples in PCR solutions and Sanger's solution without the need of purification.
(2) Achieve quantitative measurements of DNA by matrix-assisted laser desorption method.
(3) Discovery of the existance of metastable DNA ions.
(4) Detection of restriction enzyme digested DNA.
(5) Detection of point mutation in P-53 gene.
(6) Successful detection of 30 blind samples from cystic fibrosis patients.

Details will be presented in the meeting.

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

Progress in DNA sequence mixture readout by mass spectrometry

Chau-Wen Chou, David Dogruel, Jennifer Krone, Randall Nelson, David Schieltz and Peter Williams

Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287-1604

Mass spectrometric readout of Sanger sequencing mixtures not only has the potential to accelerate ladder readout in large-scale sequencing, but also may offer greatly reduced error rates. *Faster readout* may or may not accelerate the overall sequencing process, but the potential exists to keep pace with improvements in other areas, e.g. rapid end-sequence feedback for primer walking strategies. However, *error elimination* would produce a directly predictable rate improvement of up to a factor of 10 by eliminating the redundancy of repeated sequencing. Requirements for mass spectrometric readout include the capability to generate gas-phase molecular ions of intact DNA strands, with uniform sensitivity, mass resolution compatible with the readout length required, i.e. 1:400 for a 400 base read length, and mass spectral simplicity, i.e. one peak per DNA fragment. Large molecules can be vaporized by pulsed laser ablation from a volatile or photodegradable matrix. We are currently pursuing two approaches to this goal: (i) a search for photodegradable matrices which do not photolytically or otherwise fragment the DNA and (ii) optimization of ablation from aqueous media (ice). The best mixture spectra to date have been obtained by ablation of thin films of frozen aqueous DNA solutions, but with very poor reproducibility. Fig. 1 shows ice ablation mass spectra of the C- and G-terminated "lanes" of a synthetic sequence mixture. Favorable features of these spectra for sequencing applications are: (a) only 1 peak per DNA strand apart from some possible impurities at low mass, (b) good signal-to-noise, (c) rather uniform sensitivity (peak integrals constant within ~ 30% from 10--mer up to the 89-mer) (d) good registration, even though the spectra were taken some weeks apart. Missing peaks from the A and T "lanes" are immediately apparent, so that the absolute mass scale allows unambiguous *error flagging*, and precise mass measurement (to ± 3 Da) of the 2-nucleotide gaps up to about the 40-mer allows *error recovery*: the identity of the missing nucleotide (A or T) can be deduced even though the A-T mass difference is only 9 Da.

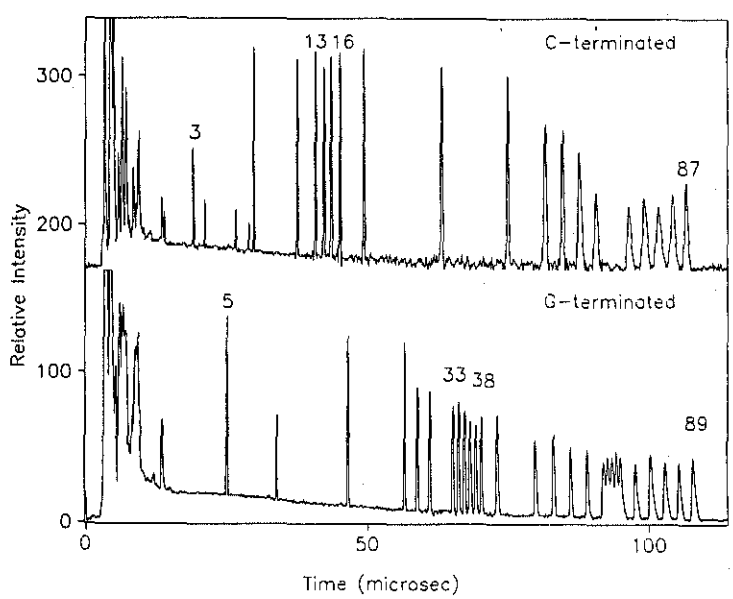

Fig. 1. Time-of-flight mass spectra of synthetic ss-DNA sequence mixtures ablated from frozen aqueous solution films on oxidized copper substrate. Laser: 589 nm, ~ 8 ns pulse length, ~ 10^9 W/cm^2. Spectra are sums from 12 consecutive laser shots.

Upper: C-terminated, 19 components (71-mer - 87-mer concentrations increased x 2).

Lower: G-terminated, 27 components, equimolar mixture

Fragmentation Studies of Oligonucleotides Using Matrix-Assisted Laser Desorption Mass Spectrometry

Christine M. Nelson, Lin Zhu, Stephane Mouradian, Wei Tang and Lloyd M. Smith
Department of Chemistry, University of Wisconsin, Madison, WI 53706.

The development of matrix-assisted laser desorption and its demonstrated performance with large proteins, up to 300,000 daltons in size, has generated substantial interest in utilizing this technique as a replacement for the gel electrophoretic separation step of the four Sanger fragment mixtures. If successful, the main advantages of this method over traditional gel-based methods are that polyacrylamide gels are no longer necessary and that the speed of separation, detection and data acquisition can be performed in seconds compared to the hours or so required in the ultrafast gel electrophoretic formats. This alternative approach is still in the preliminary stages of development. If we are able to prove its feasibility, we expect to develop an instrument with a theoretical throughput of 360,000 bases per day.

Our group has demonstrated the possibility to obtain sequencing information of oligonucleotides with 17-40 bases using the matrix 3-hydroxypicolinic acid (3HPA). In this experiment, the mass range is limited by a decrease in peak intensity for the larger components of the mixture. The current limitation of size in the MALDI analysis of oligonucleotides has generated significant interest in characterizing the problems associated with the application of this technique to nucleic acids. Early results in our laboratory and in others have shown that fragmentation is an important issue in the MALDI analysis of nucleic acids and is possibly responsible for the current limitations of size and base composition in oligonucleotide analysis.

Fragmentation studies of small oligodeoxynucleotides using MALDI-MS at 355 nm radiation from the matrix 2,5-dihydroxybenzoic acid were performed. Homopolymers of deoxythymidine are readily analyzed, however, other homopolymers and mixed sequence oligomers have been more recalcitrant. A variety of asymmetric oligonucleotides ($d(T_4N_4T_7)$ where N= A, C or G) were synthesized to study this fragmentation in greater detail. Similar fragmentation patterns were observed for the three samples which indicated that the primary fragmentation pathway is loss of a base followed by backbone cleavage at the 3' C-O bond of the corresponding deoxyribose. A statistical cleavage model accurately described the observed patterns of fragmentation; the model yields backbone cleavage probabilities at A, C and G of 0.13, 0.26 and 0.27, respectively.

Additional studies to further our understanding of the fragmentation mechanism have also been undertaken in which we studied the fragmentation behavior of several normal and modified DNA and RNA oligomers using MALDI, including methylated and brominated cytosine bases, abasic residues, and RNA incorporated into DNA.

Analysis of Oligonucleotide Mixtures by Electrospray Ionization-Mass Spectrometry

Xueheng Cheng, David C. Gale, Harold R. Udseth and Richard D. Smith*

Chemical Sciences Department, Pacific Northwest Laboratory, Richland, WA 99352

Our aim is to develop electrospray ionization mass spectrometry (ESI-MS) methods for high speed DNA sequencing of oligonucleotide mixtures, that can be integrated into an effective overall sequencing strategy. ESI produces intact molecular ions from DNA fragments of different size and sequence with high efficiency [1]. Our aim is to determine mass spectrometric conditions that are compatible with biological sample preparation and that avoid problems due to dissociation, aggregation, or adduction of the ionized DNA fragments. Oligonucleotide ions are typically produced from ESI with a broad distribution of net charge states for each molecular species (i.e., $(M\text{-}nH)^{n-}$, where n is a series of integers), and thus leading to difficulties in analysis of complex mixtures [1]. To make identification of each component in a sequencing mixture possible, the charge states of molecular ions can be reduced by manipulating the ESI process and/or by using gas-phase reactions. The charge-state reduction methods being examined include: (1) reactions with organic acids (in the solution to be electrosprayed, the ESI-MS interface or the gas phase); (2) the labeling of the oligonucleotides with a designed functional group for production of molecular ions of very low charge states; and (3) the shielding of potential charge sites on the oligonucleotide *phosphate/phosphodiester* groups with polyamines (and the subsequent gas-phase removal of the neutral amines). In initial studies two methods for charge state reduction of gas phase oligonucleotide negative ions have been tested: (1) the addition of acids to the oligonucleotide solution and (2) the formation of diamine adducts followed by dissociation in the interface region [2]. In the first method, the efficiency of charge state reduction depends on the pK_a, the concentration and the nature of the acids. Acetic and formic acids were found to be better reagents than HCl, CF_3CO_2H and H_3PO_4. The second method has the advantage that the stability of oligonucleotides is not affected but requires the optimization of the interface dissociation conditions and the amounts of diamine added to the oligonucleotide solution. Both methods show promise for charge state reduction and results have been demonstrated for small oligonucleotides (e.g., $pd(T)_{12}$ and d(AGCT)) [2]. Substantial reduction in spectral complexity was also observed for a four-component mixture of oligonucleotides upon charge state reduction. Our aim is to provide a basis for the development of an overall approach to high speed sequencing so as to provide a basis for the subsequent step of prototyping a cost effective high-throughput instrument for broad application.

* Corresponding author.

[1] "New Developments in Biochemical Mass Spectrometry: Electrospray Ionization", R. D. Smith, J. A. Loo, C. G. Edmonds, C. J. Barinaga, and H. R. Udseth, **Anal. Chem.**, 62, 882-889 (1990).

[2] "Charge State Reduction of Oligonucleotide Negative Ions from Electrospray Ionization", X. Cheng, D. C. Gale, H. R. Udseth, and R. D. Smith, **Anal. Chem.**, submitted.

High Speed Sequencing of Single DNA Molecules in the Gas Phase by FTICR-MS

Richard D. Smith*, Xueheng Cheng, S. A. Hofstadler, J. A. Bruce and Charles G. Edmonds

Chemical Sciences Department and Environmental and Molecular Sciences Laboratory, Pacific Northwest Laboratory, Richland, WA 99352

This project is aimed at the development of a totally new concept for high speed DNA sequencing based upon the analysis of single (i.e., individual) large DNA fragments using electrospray ionization (ESI) combined with Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. In our approach, large single-stranded DNA segments extending to as much as 25 kilobases (and possibly much larger), is first transferred to the gas phase using ESI. The multiply-charged molecular ions is then be trapped in the cell of an FTICR mass spectrometer, where one or more *single ion(s)* can be selected for analysis and its mass-to-charge ratio (*m/z*) measured both rapidly and non-destructively. Single ion detection is achievable due to the high charge state of the electrosprayed ions.

Our efforts under the first two+ years of this project have demonstrated the capability for the formation, extended trapping, isolation, and monitoring of sequential reactions of highly charged DNA molecular ions with molecular weights well into the megadalton range [1-4]. We have shown that large multiply-charged individual ions of both single and *double-stranded* DNA anions can also be efficiently trapped in the FTICR cell, and their mass-to-charge ratios measured with very high accuracy. Thus, it is feasible to quickly determine the mass of each lost unit as the DNA is subjected to rapid reactive degradation steps. Our aim is to now develop methods based upon the use of ion-molecule or photochemical processes that can promote a stepwise reactive degradation of gas-phase DNA anions. Successful development of one of these approaches could greatly reduce the cost and enhance the speed of DNA sequencing, potentially allowing for sequencing *DNA segments of more* than 25 kilobase in length, on a time-scale of minutes with negligible error rates, with the added potential for conducting many such measurements in parallel. *The techniques* being developed promise to lead to a host of new methods for DNA characterization, potentially extending to the size of much larger DNA restriction fragments (>500 kilobases).

* Corresponding author.

[1] "Trapping, Detection and Reaction of Very Large Single Molecular Ions by Mass Spectrometry," R. D. Smith, X. Cheng, J. E. Bruce, S.A. Hofstadler and G.A. Anderson, **Nature**, 369, 137-139 (1994).

[2] "Charge State Shifting of Individual Multiply-Charged Ions of Bovine Albumin Dimer and Molecular Weight Determination Using an Individual-Ion Approach," X. Cheng, R. Bakhtiar, S. Van Orden, and R. D. Smith, **Anal. Chem.**, 66, 2084-2087 (1994).

[3] "Trapping, Detection, and Mass Measurement of Individual Ions in a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer,: J.E. Bruce, X. Cheng, R. Bakhtiar, Q. Wu, S.A. Hofstadler, G.A. Anderson, and R.D. Smith, **J. Amer. Chem. Soc.**, in press.

[4] "Direct Charge Number and Molecular Weight Determination of Large Individual Ions by Electrospray Ionization-Fourier Transform Ion Cyclotron Resonance Mass Spectrometry", R. Chen, Q. Wu, D.W. Mitchell, S.A. Hofstadler, A.L. Rockwood, and R. D. Smith, **Anal. Chem.**, in press.

Electrophore-Labeled DNA for Enhanced Sensitivity in Matrix-Assisted Laser Desorption Mass Spectrometry *

Phillip F. Britt, Gregory B. Hurst, and Michelle V. Buchanan
Chemical and Analytical Sciences Division, Oak Ridge National Laboratory
Oak Ridge TN 37831

In order to speed up sequencing and other DNA analyses that yield information encoded as a series of DNA fragments of different molecular weights, strategies based on matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) are being developed as alternatives to the lengthy gel electrophoresis separation and detection method that is currently used. One disadvantage of MALDI-MS, particularly for larger DNA fragments, is the relatively poor sensitivity that results from the low apparent efficiency for production of ions from the neutral DNA molecules desorbed by the laser [1,2]. We are therefore investigating methods to improve MALDI-MS sensitivity by increasing this ionization efficiency. Our first approach has been to derivatize DNA with an electrophore in order to exploit the presence of electrons in the MALDI plume. The electrophore-tagged DNA should efficiently capture these electrons, potentially resulting in increased ionization efficiency and enhanced sensitivity.

We have attached a series of electrophore tags to the 5' terminus of the M13-40 sequencing primer 5'-GTTTTCCCAGTCACGAC-3'. Electrophore groups tested to date include pentafluorophenyl-, p-nitrophenyl-, m-nitrophenyl-, and 9-fluorenone. Using a methodology similar to that used to attach tags for fluorescence detection of sequencing ladders on gels [3,4], the aminohexyl derivative of the primer is reacted with either the N-hydroxysuccinimide or the isothiocyanate derivative of the electrophore. MALDI-MS analysis of the products shows that the electrophore-labeled primer can be successfully synthesized, and that the label remains attached to the primer throughout the MALDI-MS measurement process. The next phase of the work will be to examine an expanded series of electrophore tags to identify a candidate that is less susceptible to processes leading to loss of signal, such as rapid autodetachment (loss of the captured electron) or dissociative attachment. In addition, modifications to the mass spectrometer that will provide more favorable conditions for electron capture by the labeled oligonucleotides are in the design phase.

*Research sponsored by the National Institutes of Health, National Center for Human Genome Research, Grant No. 1 R55 HG/OD00819-01A1, under Interagency Agreement 1884-F026-A1 with the U.S. Department of Energy under Contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

1. Nelson, R.; Rainbow, D.; Lohr, P.; Williams, P. *Science* **1989**, *246*, 1585.
2. Romano, L.; Levis, R. *J. Am. Chem. Soc.* **1991**, *113*, 9665.
3. Telser, J.; Cruickshank, K.A.; Morrison, L.E.; Netzel, R.L.; Chan, C.-K. *J. Am. Chem. Soc.* **1989**, *111*, 7226-7232.
4. Smith, L.M.; Fung, S.; Hunkapiller, M.W.; Hunkapiller, T.J.; Hood, L.E. *Nucleic Acids Res.* **1985**, *13*, 2399-2412.

Human DNA Sequencing at the LBL Human Genome Center

Michael Palazzolo, Christopher Martin, William Kimmerly, Edward Rubin, J. F. Cheng, Joseph Jaklevic, Edward Theil, and Mohan Narla.
Human Genome Center, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720

The Human Genome Center (HGC) of the Lawrence Berkeley Laboratory is oriented almost exclusively towards developing and implementing directed methodologies for cost-effective and accurate high throughput human DNA sequencing. This work has five components. The first three components of the Center are all involved with new technology development for sequencing and are based on a collaboration between biologists, the automation group, and computer scientists. The fourth component is the sequencing production effort itself. The fifth component of the HGC is the biology effort that interfaces and performs experiments derived from the completed sequence data. The biology component of the new technology group has developed a directed strategy of DNA sequencing in which high resolution physical maps are generated so that a small set of standard primer binding sites are positioned every 300 bp. This mapped set of templates is then sequenced. Using this strategy templates can be selected in a minimally redundant fashion which means that template preparation requirements are reduced ten-fold and sequencing reactions can be reduced five-fold. In addition, sequence assembly is straightforward because all the templates are mapped in relation to each other, with a resolution of about 30 bp, prior to sequencing. The biology group is continuing to optimize the biological procedures of the directed process. The second component of the center is the work of the automation group which is developing instrumentation to support the directed sequencing approach. Some of the modules have been completed and are currently in use: an image station that captures and analyzes the mapping information from agarose gels, a colony picker, a robotic library replicator, and a modified Biomek that sets up PCR assays and sequencing reactions. A water-based thermocycler and a 12-channel oligonucleotide synthesizer are being rigorously tested by the biologists prior to entering production. Novel methods to analyze PCR fragments as well as preliminary plans to integrate some of the initial modules are in the early planning stages. The third component of the HGC is the informatics group. The major goal of the group is to develop software that facilitates the sequencing effort. The developmental effort is aimed at all aspects of the process, beginning with the physical mapping efforts, continuing through the generation of the high resolution map and template selection, followed by sequencing and assembly of templates and concludes with the analysis and dissemination of the sequencing information. The programs that keep track of and display the physical mapping data are nearing completion. The emphasis of the work is shifting now to sequence assembly, editing, and analysis. Another aspect of the work is focused on developing mechanisms to make the data publicly accessible as it is being generated. The fourth component of the center is the production effort. A team of 5-6 FTE's can currently generate 600-700 kb per year. The goal of the production effort is to maintain this rate as well as to add additional teams in the next few years and also increase the productivity of each team. The final component of the HGC is the biology program which has been reconstituted to be closely integrated with the overall sequencing effort. The biology effort will play a role in selecting templates beforehand and then developing biological approaches to interpret such a large amount of data in a meaningful way.

Directed Genomic DNA Sequencing

Christopher H. Martin, Cheryl A. Davis, Carol A. Mayeda, Herb Moise and Michael J. Palazzolo
Human Genome Center
Lawrence Berkeley Laboratory, Berkeley, CA 94720

Our group has developed a novel directed approach to genomic sequencing in which every sequencing template is mapped to a resolution of 30 base pairs prior to being sequenced. This high resolution mapping information yields two important advantages. First, genomic sequence can be determined with far fewer sequencing reactions than approaches which utilize random coverage to obtain the bulk of the complete sequence. Second, the difficulty of the sequence assembly problem is greatly reduced, as the mapping information provides significant assistance to the reconstruction of the original sequence.

This project is supported by a unique collaborative arrangement between NIH and DOE. The Department of Energy is funding organism-independant technology development for the directed sequencing stragtgy, and also production genomic DNA sequencing in humans. NIH is funding Drosophila production sequencing.

Our group has sequenced over 1,500,000 base pairs of human and Drosophila genomic DNA using the directed sequencing approach. We currently have two 6.5 member sequencing teams in operation. The first is DOE funded and is focusing on a 1.2 Mb growth factor cluster surrounding the interleukin-4 gene on human chromosome 5. The second team is NIH funded and is targeted at a genetically well characterized region of approximately 2.2 megabases that surrounds the Adh gene of Drosophila.

Our goal is to use the directed sequencing strategy as a biological platform for the collaborative development of a highly automated system for the sequencing of genomic DNA. We work in close collaboration with the automation and computation groups of the Human Genome Center at LBL in order to pursue this goal of reducing the labor requirement of genomic sequencing. The ability to develop custom hardware and software modules that address the scope of the genomic sequencing production work is essential in our efforts to reduce the cost of genomic sequencing. Several of these automation and informatics modules are already in routine production use and are having a major impact on the efficiency of the directed sequencing process. The functioning of some of these modules in a production sequencing environment and their effect on productivity will be presented. Several additional modules are now being tested or are under construction and are aimed at further increases on the throughput of the process.

The two production teams (one is dedicated to Drosophila sequencing, the other to human sequencing) of the Directed Sequencing Group consist of Marnel Bondoc, Annie Chiang, Thomas Cloutier, Paul Critz, Cheryl Davis, Cheryl Ericsson, Michael Jaklevic, Robert Kim, Michelle Lee, Melvin Li, Carol Mayeda, Afaf Steiert-El Kheir and Millicent Yee.

Large-scale Sequencing of the Human and Mouse T Cell Receptor Loci

Leroy Hood[1], Lee Rowen[1], Kai Wang[1], Inyoul Lee[1], Cecilie Boysen[2], and Ben F. Koop[3].

[1]Department of Molecular Biotechnology, University of Washington, FJ-20, Seattle, Washington 98195. [2]Caltech, 391 South Holiston, #147-75, Pasadena, California 91125. [3]Department of Biology, University of Victoria, Victoria, British Colombia, V8W 2Y2.

T cell receptors play a major role in immunity and autoimmune diseases. For this reason, their genomic sequence has been chosen as a model system for the development of strategies and tools related to the human genome project. The complete genomic sequence of a multigene locus enables a delineation of genes and gene boundaries, and an assessment of the proportion of pseudogenes. Additionally, it provides PCR access to microsatellite markers and other possible sites of polymorphic variance and, therefore, will facilitate efforts to discover mutations related to disease susceptibility. Strong sequence homologies found in a cross-species comparison between human and mouse counterparts will assist in identifying regulatory regions, new genes and alternative functions for DNA sequence information The cross-species comparison will also conduce to an understanding of the evolutionary mechanisms that underlie overall gene organization.

Well over a megabase of the T cell receptor loci from human and mouse have been sequenced using the shotgun strategy. The preponderance of the data comes from the TCR beta loci (~95% complete in human, ~60% complete in mouse). Our current efforts are aimed at completing the beta loci and the human alpha locus.

Major new discoveries from the human and mouse T cell receptor loci sequence include the following:

A cross-species comparison of the human and mouse alpha-delta constant regions reveals ~70% homology across nearly 100 kb of sequence, suggesting that non-coding DNA may have heretofore undiscovered functions.

The T cell receptor beta locus is also the site of the human and mouse pancreatic trypsinogen multigene family, suggesting that genes with apparently unrelated functions can occupy the same genomic space.

Approximately half of the human T cell receptor beta locus is comprised of long homologous repeats in which members of multigene subfamilies are embedded. These repeats suggest a mechanism for the divergence of gene function. Indeed, a portion of the human TCR beta locus has even been translocated to another chromosome. The mouse locus, by way of contrast, contains far less repeated DNA. In this regard, the comparative genomic sequences have provided an explanation for why there are twice as many TCR beta variable gene segments in human as mouse, even though both species have about the same number of subfamilies.

We anticipate that the sequences of the human and mouse TCR alpha loci, when completed, will yield further insights into features of mammalian genomes. We also anticipate improvements in sequencing technology, data management, and organizational strategies will increase the ease and throughput of mammalian genomic sequencing.

Automated Fluorescent Detection for Multiplex DNA Sequencing

Robert B. Weiss[1,3], F. Mark Ferguson[1], Leonard Di Sera[1], Alvin Kimball[1], Josh Cherry[1], Mark Stump[1], Andy Marks[1], Tony Schurtz[1], Diane Dunn[1,2] and Raymond F. Gesteland[1,2].

[1]Department of Human Genetics and the [2]Howard Hughes Medical Institute, University of Utah, Salt Lake City, UT 84112. [3] Corresponding author.

Instrumentation for automated hybridization and detection of DNA hybrids on nylon membranes is a focal point of our research. Recently, we have devised a method for amplifying fluorescent light output on nylon membranes by using an alkaline phosphatase-conjugated probe system combined with a fluorogenic alkaline phosphatase substrate [1]. The amplified signal allows sensitive detection of DNA hybrids in the sub-femtomole/band range.

Development of integrated instruments capable of multiple hybridization/detection cycles is underway. The hybridization apparatus contains a set of nested Plexiglas cylinders: a heated inner drum with nylon membranes fixed to its outer surface rotates through a fluid puddle formed by an eccentric outer drum enclosure. The inner drum has a viewable surface area of ~ 3000 cm^2. Fluid delivery and drain are achieved using solenoid valve blocks, with segregated blocks for wash fluids, probe fluids, enzyme fluids and substrate fluids. A stepper motor supplies the drive system for drum rotation. This precision drive allows synchronization of the drum rotation with the charge transfer across one dimension of a two dimensional CCD camera during the detection process. The peltier-cooled 2048 x 96 pixel CCD camera collects a continuous image of the inner drum's surface by operating in the Time Delay and Integration mode.

Several hybridization/detection instruments have been constructed and are now undergoing testing with a variety of hybridization formats. Several probe-enzyme conjugates, both direct oligonucleotide-alkaline phosphatase conjugates and indirect biotin-streptavidin-AP or digoxigenin-anti-digoxigenin-AP, are being tested. Results will be shown from automated hybridization and detection of high-density colony grids of cosmid libraries using STS primers as probes. A major use of these instruments is in both multiplex mapping and sequencing of transposon inserts in large plasmid (~20 Kb) templates. Transposon inserts are mapped by probing Southern blots of restriction digests from multiplexed plasmids. Mapping membranes contain 96 lanes, and five of these membranes can be mounted on a single drum. A ten vector multiplex family allows mapping of 4,800 transposon inserts with a single drum load of the instrument, followed by 21 cycles of hybridization and detection. Multiplex sequence ladders in both 16 lane set and 32 lane set formats are being tested. Hybridization and detection of multiplex genotyping markers (PCR products of simple sequence repeats) is also under development. Continued refinement and development of these instruments will provide a key segment for the automation and integration of large-scale multiplex sequencing.

This work was funded by DOE grant DE-FG03-94ER-61817 (R.F.Gesteland, P.I.)

[1] Cherry, J.L., Young, H., Di Sera, L.J., Ferguson, F.M., Kimball, A.W., Dunn, D.M., Gesteland, R.F., and Weiss, R.B. (1994). Enzyme-linked fluorescent detection for automated multiplex DNA sequencing. *Genomics* **20**, 68-74

The *C. elegans* Genome Sequencing Project.

Richard K. Wilson[1] and the *C. elegans* Genome Consortium[1,2].

[1]Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO, USA, and
[2]The Sanger Center, Hinxton, Cambridgeshire, UK.

The nematode worm, *C. elegans*, has proved particularly amenable to genetic analysis of development and neurobiology. Its haploid genome, distributed over six chromosomes, contains about 100 megabases (Mb) nearly all of which have been mapped to 15 large contigs. Following an initial pilot phase sequencing project of a 3 Mb region on chromosome III, our laboratories in Cambridge (UK) and St. Louis (USA) plan to complete the sequence of the entire genome within the next five years.

Our approach to genomic sequencing is to first derive random M13 (1-2 kb) or phagemid (6-9 kb) subclones from cosmids, although where there is no cosmid coverage, libraries have been produced from whole or partial YACs. This "shotgun" phase is followed by a directed "walking" phase to completion. Subclones are sequenced and analyzed on ABI 373A fluorescent gel readers with an overall coverage redundancy of around six-fold. Recent improvements to the ABI 373A gel readers have substantially increased the amount of data which can be obtained from a single run.

Software for assembly and finishing are continually improving: a recent version of the Staden assembly program incorporates several useful features within the contig editor to aid finishing. In addition, novel automation which speeds template preparation and DNA sequencing has been developed at both sites.

We have now sequenced over 7 Mb. Analysis of this region by GENEFINDER suggests an average gene density of 1 gene per 5 kb. The percentage of ESTs (derived from cDNAs) identified in the genome sequenced so far suggests that the gene density in this region is close to average for the whole genome. Approximately 45% of the predicted genes show homology to previously identified genes from all organisms.

With the sequence of the central region of chromosome III essentially complete, we are now following the same strategy for sequencing chromosome II and the X chromosome.

Not funded by DOE.

WHOLE GENOME SHOTGUN SEQUENCING OF THE 2.0 Mb GENOME OF *HAEMOPHILUS INFLUENZAE*

Carol J. Bult, Robert D. Fleischmann, Mark D. Adams, Joseph M. Merrick, Jeannine Gocayne, Li-ing Liu, Anthony Kerlavage, Hamilton O. Smith, and J. Craig Venter
The Institute for Genomic Research, Gaithersburg, MD, and Johns Hopkins University, Baltimore, MD.

The accepted approach for sequencing large segments of DNA (>100 kb) has been to spend substantial effort in the development of lambda or cosmid libraries and their subsequent mapping. Developing approaches for rapid and efficient sequencing and assembly of large segments of DNA is critical for genome sequencing projects. We have developed a whole genome sequencing approach that eliminates "top down" efforts. A random shotgun approach to whole genome sequencing of the approximately 2.0 Mb genome of *Haemophilus influenzae* was undertaken by creating a sheared random genomic library with an average insert size of 1.5 - 2.0 kb cloned into pUC18. Using the high throughput capacity of the TIGR DNA sequencing facility we have prepared approximately 20,000 randomly selected *H. influenzae* double stranded templates and analyzed them on ABI 373A DNA sequencers with the 48 cm gel plate modification. Initial sequencing and assembly data from our random genomic library fit the Lander and Waterman model for predicting the rate of assembly of a 2 Mb genome.

We have developed a number of software programs which allow us to manage large scale genome sequencing projects All data and analysis results are written directly into our SYBASE database. A number of methods are being evaluated for assembling the sequence data into contigs and generating a consensus sequence. One of the more successful approaches involves clustering all the available sequences by doing pairwise Smith Waterman comparisons. The cluster are then assembled by a multiple sequence alignment program written at TIGR, and which runs on our MASPAR computer. Large clusters are assembled using the ABI AutoAssembler program. To date, greater than 10 Mb of raw sequence data has been generated. This accounts for approximately 1.6 Mb of the *H. influenzae* genome which we currently have ordered into 61 groups containing182 contigs. The remainder of the genome can be accounted for in the 6 rRNA repeats, 3 additional repeats and gaps. Closure is being accomplished by using our database information to target the sequencing of specific templates which are likely to close gaps.

Microbial Genome Sequencing

Douglas R. Smith and John Reeve[1]
Genome Therapeutics Corporation
[1]Ohio State University

The goal of this project is to determine the sequence of several microbial genomes of particular importance from the perspectives of energy production and bioremediation. The target organisms are *Methanobacterium thermoautotrophicum* (1.7 Mb), chloroplast (100 kb), *Rhodococcus rhodochrous* (2.5 Mb), *Methanopyrus kandleri* (1.7 Mb), *Synechococcus sp.* (2.7 Mb), and *Haloferax volcanii* (4.2 Mb). The sequencing will be done using state-of-the-art multiplex sequencing technology and will build on our current success in generating and analyzing around 1 Mb of finished and fully annotated genomic sequence from the *Mycobacterium leprae* and *M. tuberculosis* genomes.

Each genome will be shotgun sequenced in its entirety at about eightfold random coverage. Clone pooling, DNA preparation, and cycled dideoxy sequencing reactions will be performed using established procedures and robotic instrumentation wherever applicable. Finish sequencing will be performed by cycled sequencing methods from subclone pools. Data acquisition will be accomplished by digital film or infrared fluorescence scanning. Sequences will be read on computer workstations using an automated image-analysis program, REPLICA™ [developed by L. Mintz and G. Church (HHMI at Harvard Medical School)]. The sequences will be assembled into contigs, and the contigs will be proofread using a modified GelAssemble platform and REPLICA. Analysis for genes and structural features will be carried out with a variety of programs including a specialized set we are currently using for the analysis of large bacterial genomic regions (Large Sequence Analysis Suite).

Mathematical modeling of shotgun sequencing was done using the mathematically rigorous equations of Lander and Waterman (1988) assuming a read length of 400 bases and an 80% overall success rate. These calculations predict that, at eightfold coverage, a 1.7-Mb genome is expected to assemble into 19 contigs averaging 90 kb in size. The gaps between contigs are predicted to be small (2.3×10^{-4} chance that a gap will be greater than 400 bp) and a bridgeable (1.6×10^{-9} chance that a gap could not be bridged by walking on one of the existing 2-kb shotgun sequencing templates).

All sequences will be made available with full annotation for gene locations as soon as possible (and certainly within 6 months) after completion. Sequences will be submitted using automated ASN1 routines, which we are currently using to annotate mycobacterial cosmid sequences. To facilitate biochemical studies, the Ohio State group will generate and distribute a resource of sequence-mapped clones (lambda or cosmid) for each genome.

This work is funded by a cooperative agreement between DOE and Genome Therapeutics Corp. (DE-FC02-95ER61967).

High-Throughput DNA Sequencing and Characterization of Diverse Microbial Genomes

J. Craig Venter and Carl Woese[1]
The Institute for Genomic Research
[1]University of Illinois

The broad objective of the Microbial Genome Consortium (MGC) is to characterize the genomes of diverse microorganisms and to use the resulting genome information to further understand microbial phylogeny, physiology, structural biology, and ecology.

Specific Aims

1. Implement MGC as a functioning collaboration and establish a Microbial Genome Advisory Board (MGAB) that will oversee the program and recommend candidate genomes for sequencing. Organisms selected will be phylogenetically diverse and have genomes whose sizes are in the 1.5- to 2.8-Mbp range (or larger in the out years) and are of value in both applied and basic biology. In the first year, this will include the complete sequence of *Methanococcus jannaschii,* which is a barophilic, hyperthermophilic methanogen with a genome size of 2 Mbp.

2. Sequence one or more complete microbial genomes per year and randomly sample a large number of genomes.

3. Analyze the data generated. Screen all new data for related sequences in the databases to identify homologues and gene products. Use comparative analysis to identify reading frames (and other features) in regions without identified products. Annotate the sequences with respect to likely significant features, including known homologues.

4. Deposit the data in the sequence data banks and publish the results in a timely manner. Improve access by establishing a public-access Microbial Genome Data Base (MGDB) containing genome sequence data and related data in an integrated, queryable form (MGDB will be provided in the first year as a cost-share by the laboratories in the consortium).

5. Investigate candidate small eukaryotic genomes for possible later sampling or complete sequencing.

6. Coordinate our efforts and organism selections with ongoing studies of the microbial diversity in natural populations, with an initial focus on subsurface and selected high-temperature ecosystems.

This work is funded by two cooperative agreements with DOE (DE-FC02-95ER61962 to The Institute for Genomic Research and DE-FC02-95ER61963 to the University of Illinois).

Automated Multiplex Microbial Genome Sequencing

Joshua L. Cherry[1], Debi Nelson[1], Peter Cartwright[1], Mark Stump[1], Diane Dunn[1] and Robert B. Weiss[1,2].

[1]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112. [2] Corresponding author.

We are initiating large-scale genomic DNA sequencing of microorganisms of industrial and biological interest. Our first project will be the complete sequence of the genome of *Pyrococcus furiosus* (DSM 3638), a hyperthermophilic member of the "Archaea." This organism was isolated from hot marine sediments with temperatures near 100°C. The Archaea constitute one of the three great phylogenetic branches of living things. They are the least well studied of these three groups. The complete sequencing of an archael genome will yield a tremendous quantity of information about this largely unexplored territory. The data will have much use in phylogenetic analyses and sequence comparison involving widely diverged taxa. Many archaeons, especially methanogens and thermophiles, have potential economic importance. Interest in thermophilic organisms has been growing in recent years, and thermostabile enzymes have a wide variety of industrial uses. The sequencing of the entire genome of a thermophile would amount to the cloning of a large number of such enzymes. It has been observed by X-ray crystallographers that proteins from thermophiles generally form usable crystals more easily than their mesophilic homologues. Many structure determinations have depended upon the cloning of a thermophilic homologue. Our sequencing efforts will enable us to provide such clones easily and rapidly. Comparison of thermophilic protein sequences to homologous mesophilic sequences will provide much information concerning thermal stability, and protein stability in general.

Novel methods and instrumentation for enzyme-linked fluorescent multiplex sequencing will be utilized [1]. The contiguous sequence of individual 20 kb inserts will be determined by a transposon-based directed strategy. A *P. furiosus* library of 20 kb insert size is being built in 10 multiplex vectors. These vectors provide multiplex tags for both the end sequencing and transposon mapping phases of the process. These fragments will be fed into the transposon system in an ordered fashion via an end sequencing strategy which obviates the need for physical mapping. The first step in the process will be end sequencing of the *P. furiosus* genomic library This end sequence will consist of ~2000 single-pass sequence ladders from the ends of the 20 kb inserts. This sequence will itself yield much information and will result in the discovery of many genes for hyperthermophilic proteins. Sequence matching between newly completed insert sequences and the database of end sequences will allow selection of new inserts to be sequenced, leading ultimately to completion of the genome.

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

[1] Cherry, J.L., Young, H., Di Sera, L.J., Ferguson, F.M., Kimball, A.W., Dunn, D.M., Gesteland, R.F., and Weiss, R.B. (1994). Enzyme-linked fluorescent detection for automated multiplex DNA sequencing. *Genomics* **20**, 68-74

Genomic Sequencing of Human and Rodent DNA Repair Gene Regions

Jane E. Lamerdin, Mishelle A. Montgomery, Stephanie A. Stilwagen, Jakob M. Kirchner, Edmund P. Salazar, Christine A. Weber, Robert S. Tebbs, Kerry W. Brookman, Larry H. Thompson, and Anthony V. Carrano

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

The ability to repair damaged DNA in cells is critical to the survival and reproduction of all eucaryotes. Deficiency of specific DNA repair processes can manifest as cancers or other diseases. The process of DNA repair occurs through two main pathways, nucleotide excision repair (NER) and mismatch repair, and each of these pathways involves the interaction of numerous proteins. Several of the NER proteins might also play a role in transcription. Three DNA repair genes, ERCC1, ERCC2, and XRCC1, have been mapped to human chromosome 19q13.2-13.3. The genomic sequence of the ERCC1 gene has been reported previously (Martin-Gallardo et al., Nature Genetics 1:34-39, 1992). We describe here the sequencing of approximately 130 kbp of non-contiguous genomic DNA containing the mouse and human XRCC1 genes, the human and hamster ERCC2 genes, and the human ERCC4 gene (which maps to chromosome 16).

Cosmids containing the XRCC1 and ERCC2 genes were identified by hybridization with cDNA probes to the appropriate cosmid library (eg. mouse, hamster, or human chromosome 19). The human ERCC4 cosmid was identified as a secondary transformant that provided functional correction of a repair-deficient hamster cell line (UV41). Four of the five cosmids were sonicated and subcloned into M13 or phagemid vectors. The essential ERCC2 gene region in the hamster cosmid was defined and subcloned into plasmid and M13 vectors, which were subjected to exonuclease III to generate nested deletions. All templates were sequenced using Taq dye primer cycle sequencing kits (Applied Biosystems Division of Perkin Elmer, ABI) on an ABI 800 Catalyst Workstation. Resultant sequencing ladders were loaded on 4.75% or 6% sequencing gels and data collected on ABI 373A DNA sequencers. Sequence chromatograms were imported into GENeration (Intelligenetics, IG), where editing and assembly were performed.

Comparative genomic sequencing of the human and mouse XRCC1 gene regions reveals the presence of 26 elements that are at least 65% homologous. Seventeen of these elements correspond to the exons of XRCC1 (the human and mouse coding regions are 84% identical), and two correspond to introns whose lengths are completely conserved. The human ERCC2 gene is comprised of 23 exons and is 98% identical to the hamster gene at the protein level. Exon lengths between species are completely conserved, and some intron lengths are conserved for this gene as well. The human XRCC1 cosmid has an average density of 1.1 Alu per kbp, but due to clustering, the local density is as high as 1.8 Alu per kbp. The human ERCC2 cosmid appears to have a similar density, and the clustering effect is even more pronounced. In one 1.4 kbp region of intron 12, there are 3.5 tandemly arrayed Alu elements, and a monomer on the opposite strand; a density of 2.8 Alu/ kbp. In contrast, the SINEs B1 and B2 are present in the mouse XRCC1 cosmid at a density of 0.4 per kbp, and in the hamster ERCC2 gene region at a density of 0.3 SINEs per kbp.

The human ERCC2 cosmid contains an additional 8 ORFs with excellent coding potential (as identified by XGRAIL 1.1) on the opposite strand and at the 3'-end of the ERCC2 gene. At this time, its function is unknown. The presence of a homologous region in the hamster cosmid is under investigation. The human ERCC4 gene has been recently identified and cloned by complementation analysis, but no sequence data has been reported. We have begun sequencing the cosmid containing the gene, and several putative cDNAs.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48

Evaluation of High Throughput M13 DNA Isolation Protocols

Maria de Jesus, Alex Copeland, Jane Lamerdin, Anthony V. Carrano, and Ray Mariella

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

As a preliminary to implementing an automated system for the isolation of DNA from M13 cosmid subclones for our sequencing core facility, we have undertaken an evaluation of several DNA isolation protocols. We describe the results of direct comparisons between four protocols: two commercially available magnetic bead based protocols, a Triton/thermal extraction protocol [1], and a protocol that relies on a PEG precipitation, followed by phenol/chloroform extraction [2].

Currently, we can isolate DNA from approximately 100 subclones per day per person manually using the Qiagen kit and 80 subclones overnight using an Autogen 740 robot. Read lengths of 500-550 bases are obtained on 4.75% polyacrylamide gels with a 34cm well-to-read distance on the ABI 373A DNA Sequencer. Our goal is to increase DNA isolation capacity to approximately 1000 samples per day per person, while maintaining or extending current read lengths.

Template quality and quantity were assayed by agarose gel electrophoresis. Samples were sequenced by dye primer cycle sequencing either manually or using an ABI Catalyst with either Taq Polymerase or Sequitherm (Epicentre Technologies), according to the manufacturers' instructions. Samples were analyzed on 4.75% polyacrylamide gels on an ABI 373A. Protocols were evaluated for sequence quality, read length, robustness, ease of use, automatability, cost, and time.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48.

[1] Mardis, E.R., (1994) *Nucleic Acids Research*, **22**(11), 2173.
[2] Zollo, M. and Chen, E., (1994) *BioTechniques*, **16**(3), 370.

PROGRESS TOWARD AUTOMATION OF THE FRONT-END OF DNA SEQUENCING

Richard A. Guilfoyle, Jim Uzgiris, Huamin Ji, Qinghua Liu, Dan Chen, Michael Westphall, Robert Brumley, Brian Boville, Jessica Hayden, Zhen Guo, Andy Thiel, Arthur Johnson, Todd Francisco, Ricardo Gee, and Lloyd M. Smith

University of Wisconsin, Dept. of Chemistry, Madison, WI 53706

Results will be presented addressing recent progress made in our laboratory regarding the continued development of our front-end strategy for high-throughput M13-based shotgun cloning and sequencing. Our "front-end" will be discussed in terms of ongoing technology development and strategy designs as pertaining to new methods for: (1) purification and restriction mapping of cosmid inserts, (2) random fragmentation of purified cosmid inserts, (3) M13 library constructions, (4) isolation of M13 clones, (5) ordering of M13 shotgun clones and selection of the minimally overlapping inserts, (6) purification of M13 templates, (7) quantitation of M13 templates, and (8) 4-dye fluorescent dideoxy sequencing reactions. Also, an initiated plan for robotic integration of these technologies will be discussed in terms of achieving robust compatibility with the ABI373A and horizontal ultra-thin gel electrophoresis (HUGE).

Specifically, and as directly related to the above, ongoing and planned experimentation to be discussed will include: (1) a new vector facilitating triple-helix-affinity-capture (TAC) purification and restriction mapping of cosmid inserts, (2) CviJ1 digestion vs. nebulization for random fragmentation of target DNA, (3) M13 library constructions in the direct selection vector, M13-100, now equipped with univeral primers and a TAC sequence for purification of ssDNA, (4) isolation of M13-100 clones by flow cytometric sorting vs. plaque-picking and limiting dilution of transfected cells, (5) SBH format 2-related ordering of M13 clones by hybridization to oligo arrays on glass supports, (6) purification of ssDNA M13 templates by TAC vs. glass-fiber bottom microtiter wells, (7) YOYO cyanine dye-based quantitation of ssDNA template concentrations in microtiter plates using the Molecular Dynamics Fluorimager 575, and (8) direct (eg. non-cycle sequencing) Bst DNA polymerase catalyzed sequencing reactions performed on the Gilson 215 robotic liquid handling station. ABI373A and HUGE base-calling data will also be presented which examine an automated solid-phase scheme (biotin-streptavidin/magnetic bead) for the purification of DNA sequencing reactions. Robotic integration will be discussed in terms of initial and long-term objectives utilizing the ROBOLAB 9600 System (Robocon, Austria).

Studies of T7 DNA Polymerase to Improve the Properties of DNA Sequencing Enzymes

Stanley Tabor, Jeff Himawan, and Charles C. Richardson. Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115

Bacteriophage T7 DNA polymerase modified to remove its 3' to 5' exonuclease activity has properties that are advantageous for DNA sequence analysis. Two of the most important are low discrimination against chain terminating dideoxynucleotides and high processivity in polymerizing DNA. Our studies have been directed towards understanding the mechanism by which T7 DNA polymerase achieves these properties, with the goal of improving it as well as other DNA polymerases for use in DNA sequencing.

T7 DNA polymerase incorporates dideoxynucleotides more efficiently than all known DNA polymerases, not discriminating against them in the presence of manganese. A consequence of this are bands of uniform intensities after gel electrophoresis, a major advantage for DNA sequence analysis. In contrast, two other polymerases used for DNA sequencing, Klenow fragment of E. coli DNA polymerase I and Taq DNA polymerase, discriminate against dideoxynucleotides up to one thousand fold. This results in marked variability in band intensities. This dramatic difference is surprising in light of the fact that all three belong to the same related family of polymerases. We have been using a number of approaches in an attempt to understand the relevant differences that account for this. In one approach we have made use of the fact that T7 DNA polymerase is required for the replication of phage T7 to isolate mutant phage that grow in the presence of dideoxynucleosides. The location of mutations in T7 DNA polymerase responsible for this phenotype reveals the domain where the polymerase interacts with the ribose moiety of the dNTP. In another approach, we are interchanging domains of T7 DNA polymerase with those of Klenow fragment and Taq DNA polymerase, and determining the effect each change has on the ability to incorporate dideoxynucleotides. These mutant polymerases have allowed us to define the region responsible for binding the ribose moiety of dNTPs, and have produced a series of potentially interesting hybrid polymerase molecules with altered specificities for dNTPs.

The high processivity of T7 DNA polymerase is achieved by its interaction with accessory proteins, the predominant one being a tight association with E. coli thioredoxin. We are taking a number of approaches to studying the mechanism by which thioredoxin confers high processivity on T7 DNA polymerase. Our results suggest that thioredoxin acts as a "clamp" which together with T7 DNA polymerase encircles the duplex region of a primer-template. These experiments include the effect of thioredoxin on the footprint of T7 DNA polymerase binding to a primer-template, and the effect of mutant thioredoxins on the polymerase activity of wild-type and mutant T7 DNA polymerases. The goal of these studies is to develop mutant thioredoxins and/or T7 DNA polymerases that result in yet stronger interactions with the DNA, that should improve the properties of the enzyme for DNA sequence analysis. Two other T7 accessory proteins, the T7 helicase/primase and DNA binding protein, also interact with T7 DNA polymerase and dramatically improve its processivity under specific conditions.

Knowledge of the structure of T7 DNA polymerase would be extremely helpful for these studies. Towards this end we, in collaboration with Dr. Thomas Ellenberger (Harvard Medical School) have obtained well ordered crystals of the complex of T7 DNA polymerase and thioredoxin. A complete native data set to 2.8 Å has been determined using the Cornell High Energy Synchrotron Source. We are currently undergoing a screen of heavy atom derivatives in order to solve the structure. In addition to providing the structure of the DNA polymerase widely used for DNA sequencing projects today, this would also represent the first structure obtained of a DNA polymerase with its processivity factor, and would provide important clues in understanding the mechanism by which accessory proteins increase the processivity of DNA polymerases.

Design of a Parallel Array Oligonucleotide Synthesizer

Thomas Brennan, Scott Hunicke-Smith, Deval Lashkari, and Ronald Davis
Stanford DNA Sequencing and Technology Center
Departments of Genetics and Biochemistry
Stanford University
Stanford, CA 94305

An automated oligonucleotide synthesizer has been developed which can simultaneously and rapidly synthesize up to 96 different oligonucleotides in a standard 96-well format. The machine is capable of both mixed length and mixed scale synthesis, and can accommodate all of the standard internal base and 5'-labeled modifications. Conventional phosphoramidite chemistry is used. The standard synthesis scale is 20 nmol. The appropriate reagents are delivered by banks of valves into the individual wells containing the growing oligonucleotide chain which is bound to a solid support. Each well has a filter bottom to enable the removal of spent reagents. On-line trityl analysis is employed to monitor the coupling efficiency in each well, which is typically > 98%. Oligonucleotides up to 90-mers have been successfully prepared.

With the development of this machine, it has now become practical and cost effective to synthesize thousands to tens of thousands of "custom" oligonucleotides for directed primer walking strategies, as well as PCR mapping, gene and vector construction projects.

This work has been supported by DOE DEFG0393ER6155 and NIH HG00205.

Using Expressed Sequences as Nucleation Points for Genomic Sequencing

Michael R. Altherr, Amanda Ford, Cleo Naranjo, Judy Buckingham, Chris Munk and Robert K. Moyzis.

Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545.

Individual chromosomes provide the skeletal framework on which genomic data is organized. As the physical map and an ordered assemblage of molecular clones for chromosome 16 nears completion, we have embarked on the construction of an 'expressed sequence map' of this chromosome. Assuming that there are 100,000 human genes, approximately 3,000 should be encoded by chromosome 16. To this end, we have chosen the strategy of exon amplification to identify expressed sequences on chromosome 16. Our strategy employs 96-well plate pools of DNA from the flow sorted and arrayed chromosome 16 cosmid library as the substrate for exon trapping. At present we have generated 1800 exon clones from 75 of the 150 plates in the chromosome 16 library. We have sequenced over 500 of these clones and determined that approximately 60% appear to be independent exons. These sequences are being mapped to specific locations on chromosome 16 using a panel of somatic cell hybrids, YACs and cosmids that have previously been integrated into our high resolution physical map. In addition, these sequences are being subjected to database analysis to determine whether they represent previously characterized genes or contained conserved motifs that might provide some insight as to their biological function. The exonic sequences obtained in this way can be used as nodes from which to initiate genomic sequencing efforts. We have used a 230 kbp cosmid contig from the 7q telomere to model this integrated exon amplification/genomic sequencing approach.

Application of Single Molecule Detection to DNA Sequencing and Sizing

W. Patrick Ambrose, Peter M. Goodwin, James H. Jett, Mitchell E. Johnson, John C. Martin, Babetta L. Marrone, Jeffery T. Petty, Jay A. Schecker, Ming Wu, Richard A. Keller

Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545

Alberto Haces, Po-Jen Shih, and John D. Harding

Corporate Research and Molecular Biology Research and Development, Life Technologies Inc. (GIBCO BRL), 8717 Grovemont Circle, Gaithersburg, MD 20898

We are developing a laser-based technique for the rapid sequencing of 40-kb or larger fragments of DNA at a rate of 100 to 1000 bases per second. Our approach relies on fluorescent labeling of the bases in a single fragment of DNA, attachment of this labeled DNA fragment to a support, movement of the supported DNA into a flowing sample stream, and detection of the individual fluorescently labeled bases by laser-induced fluorescence as they are cleaved from the DNA fragment by an exonuclease. The ability to sequence large fragments of DNA will reduce significantly the amount of subcloning and the number of overlapping sequences required to assemble megabase segments of sequence information. Current status will be presented.

We are also applying our sensitive fluorescence detection to sizing of DNA fragments. Large, fluorescently stained restriction fragments of lambda phage DNA are sized by passing individual fragments through a focused, continuous- wave laser beam in an ultrasensitive flow cytometer at a rate of ~60 fragments per second. The size of the fluorescence burst emitted by each stained fragment as it passes through the laser beam, is measured in one millisecond. We have demonstrated flow cytometric sizing of DNA fragments in a ~0.1-pg sample of a restriction digest of lambda DNA in 164 seconds with sizing accuracy better than 98%. Our current sizing range is 2 kb to 150 kb.

Toward an Automated System for High-Throughput DNA Sequencing: 1. Primer Walking from a Hexamer Library

Jan Kieleczawa, Shiping Zhang, John J. Dunn and F. William Studier

Biology Department, Brookhaven National Laboratory, Upton, New York 11973

Strings of three adjacent hexamers can prime DNA sequencing reactions specifically and efficiently when the template DNA is saturated with a single-stranded DNA-binding protein (SSB) [1]. The SSB prevents individual hexamers from priming at isolated complementary sites, but base stacking between hexamers bound at adjacent complementary sites drives formation of a priming complex. A library of all 4096 hexamers provides ready access to primers, and priming is effective using Sequenase and *E. coli* SSB at 0 °C with templates at least as large as 40 kbp.

Priming by strings of three hexamers has been tested at more than 2000 sites in M13 single-stranded DNAs of 6.4 kb or 7.3 kb, and at more than 1000 sites in T7 DNA, a double-stranded DNA of 40 kbp, in radioactively labeled sequencing reactions. More than half of the 4096 possible hexamers have been sampled, and it appears that almost all hexamers will be able to participate effectively in these priming reactions, although hexamers containing only A and T seem to be less effective at an outside position than internally in a priming string. Priming is prevented within template hairpin structures that are too stable to be disrupted by SSB, but such structures are expected to be recognizable and infrequent. Outside of such hairpins, about three-fourths of the hexamer strings tested on M13 DNA and more than half of those tested on T7 DNA primed easily readable sequence. The ability to select hexamer strings that prime effectively should improve as the factors that affect priming are defined and controlled.

Instant access to primers at negligible cost opens the prospect of being able to sequence by primer walking on multiple templates as fast as sequencing reactions can be assembled. Taking full advantage of such capacity requires a matching capacity to read the sequence from the products of the sequencing reactions. The four-color fluorescent terminators supplied by ABI allow rapid, automated readout of the sequence, and we find that priming by strings of three hexamers can be as effective as priming by long primers in generating sequence information with these terminators. This poster will present results of fluorescent sequencing with an ABI Sequencer, where reactions were primed with strings of three hexamers at many different sites in M13 templates. Progress toward developing a multiple capillary electrophoresis system using a replaceable matrix and full spectral detection, and in developing vectors for supplying 40-kbp templates in quantities suitable for primer walking, will be presented in accompanying posters. The ultimate goal is to develop a fully automated system for high-throughput DNA sequencing that uses primers supplied from a hexamer library for primer walking on multiple templates in parallel.

[1] Kieleczawa, J., Dunn, J. J., and Studier, F. W. (1992) DNA sequencing by primer walking with strings of contiguous hexamers. *Science*, **258**, 1787-1791.

Toward an Automated System for High-Throughput DNA Sequencing: 2. Multiple Capillary Electrophoresis with a Replaceable Matrix

Mark A. Quesada, Jan Kieleczawa, Shiping Zhang, John J. Dunn and F. William Studier

Biology Department, Brookhaven National Laboratory, Upton, New York 11973

Primer walking on multiple templates, using primers supplied from a hexamer library, has the potential for sustained, high-throughput production of sequencing reactions [1]. A capillary electrophoresis and fluorescence detection system is being developed to provide the capacity and sensitivity needed to take advantage of this potential. Our initial goal is to read several hundred bases in less than an hour from sequencing reactions primed by hexamer strings, using the ABI four-color fluorescent terminators and a replaceable resolving matrix of linear polyacrylamide [2]. Ultimately, we plan to analyze many capillaries in parallel, using complete fluorescence spectra to determine the base sequence and assign confidence levels.

A prototype single-capillary electrophoresis system with four-color detection was constructed and used to assess requirements for analyzing the products of sequencing reactions primed by hexamer strings. We are collaborating with Karger's group at Northeastern University to implement capillary electrophoresis with replaceable linear polyacrylamide matrices. Results will be presented detecting the products of sequencing reactions primed by hexamer strings and terminated by ABI four-color fluorescent terminators, priming at different sites in M13 templates.

Preliminary work toward developing a multiple capillary system will also be presented. Fiber-optic illumination and detection is being tested relative to a confocal system. Multiple fiber-optic output to a spectrograph would allow full spectral detection of many capillaries in parallel on a CCD, collecting data fast enough to read 400 or more bases an hour from each capillary. A potential configuration for a fully automated multiple capillary system will be described.

[1] Kieleczawa, J., Dunn, J. J., and Studier, F. W. (1992) DNA sequencing by primer walking with strings of contiguous hexamers. *Science*, **258**, 1787-1791.

[2] Ruiz-Martinez, M. C., Berka, J., Belenkii, A., Foret, F., Miller, A. W., and Karger, B. L. (1993) DNA sequencing by capillary electrophoresis with replaceable linear polyacrylamide and laser-induced fluorescence detection. *Anal. Chem.* **65**, 2851-2858.

Toward an Automated System for High-Throughput DNA Sequencing: 3. Fesmid Vectors for Sequencing by Primer Walking

John J. Dunn and F. William Studier

Biology Department, Brookhaven National Laboratory, Upton, New York 11973

Reliable and plentiful sources of template DNAs will be required to realize the potential for automated DNA sequencing by primer walking, using primers supplied from a hexamer library [1]. We hope to use a strategy that involves sequencing directly on 40-kbp templates. The ends of a random set of 40-kbp clones from a YAC, bacterial genome, or other large DNA would provide multiple sites to begin primer walking. With sufficient coverage, merging walks from different clones could produce the sequence of the original DNA and an ordered set of clones without the need for prior mapping steps. The multiple start sites in the original DNA should also help to overcome possible problems with repeated sequences.

We are developing new vectors for preparing 40-kbp DNAs as templates for DNA sequencing. These vectors are derived from the fosmid vectors developed in Simon's laboratory [2] and are referred to as fesmids. As with fosmids, they are designed to package 40-kbp fragments of genomic DNA into λ phage particles, which allows the high-efficiency λ packaging system to be used for generating libraries of clones. And likewise, upon injection into the host cell, the fesmids are maintained as single copies under control of the replication and partitioning functions of the F factor, which helps to stabilize potentially toxic clones.

The unique feature of the fesmids is that the cloned DNA fragment is flanked by replication and packaging signals recognized by bacteriophage T7. Upon infection by T7, the cloned fragment is amplified and packaged into phage particles, leaving most of the vector sequence behind. The size of the vector sequence is such that any genomic fragment able to be packaged in λ (capacity 48.5 kbp) will also be packaged in T7 (capacity 40 kbp). Phage particles containing the genomic DNA are easily isolated from the lysate and high quality template prepared from them.

A disappointment with the first fesmid vectors was that only perhaps 5-20% of the T7 phage particles contained the cloned DNA, the remainder containing T7 DNA. To increase the fraction of particles that contained the cloned DNA, the vectors were modified to contain a copy of the lytic replicon of bacteriophage P1 under control of the *lac* repressor. Amplification of the plasmid by induction of this replicon a few hours before T7 infection can produce lysates where perhaps 90% of the phage particles contain the cloned DNA.

[1] Kieleczawa, J., Dunn, J. J., and Studier, F. W. (1992) DNA sequencing by primer walking with strings of contiguous hexamers. *Science*, **258**, 1787-1791.

[2] Kim, U.-J., Shizuya, H., de Jong, P. J., Birren, B., and Simon, M. I. (1992) Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res.* **20**, 1083-1085.

Modular Primers in Automated DNA Sequencing

Levy Ulanovsky, Irina Sobolev & Lev Kotler
Dept. of Structural Biology, Weizmann Institute of Science, Rehovot, 76100 Israel.

Recently, in two independent findings, hexamers and even pentamers were shown to assemble into long unique primers upon their contiguous annealing to the template with no ligation required (1, 2). Studier and colleagues used SSB to suppress alternative priming (1), while we found that adjacently annealed pentamers, hexamers and heptamers could uniquely prime sequencing reactions without the difficult-to-remove SSB (2).

Here we describe our progress in the development of the technology of DNA sequencing by modular primer walking, eliminating the primer synthesis bottleneck. Modular primers are assembled from three 5-mer, 6-mer or 7-mer modules selected from a presynthesized library of as few as 1000 oligonucleotides. The three modules anneal contiguously at the selected site in the template, and, in a striking way, prime there uniquely, while each of them is not unique for the most part, when used separately. This technology is expected to reduce the time per walk by a factor of 20 to 50, and the cost of DNA sequencing 5 to 15 fold. Both time and expense will be saved not only on the primer synthesis *per se* but, more importantly, as a result of the automation of the complete cycle of walking sequencing, made possible by the instant availability of the primers.

In our most recent advance we show that modular primers can now be used with dye-terminators on the ABI 373A automated sequencer (3). The single most important requirement is that of the three modules, the two upstream ones should have their 3'-ends modified to prevent their extension by the polymerase. The success rate and the quality of the automated sequencing with modular primers are similar to those with the conventional 17-20 base long primers, and for the most part few if any base-calling errors are found within the first 400 bases of the sequence run, with no stretch-liner, even though little optimization has yet been done for the reaction conditions. The protocol is only improved in that no precipitation or phenol extraction (obstacle to closed-end automation) is required. We currently use the pentamer-based primers of the 5+7+7 structure with Pu-Pu base-stacking between the 5-mer (to be extended) and the adjacent 7-mer. Both 3'-blocked heptamers have two degenerate positions each, and thus the same size library as the pentamer (512 sequences).

1. Kieleczawa, J, Dunn,J.&Studier, F. (1992), Science 258: 1789-1791.
2. Kotler, L., Zevin-Sonkin, D., Sobolev, I., Beskin, A.&Ulanovsky, L. (1993), Proc. Natl. Acad. Sci. USA 90: 4241-4245.
3. Kotler,L.,Sobolev,I.&Ulanovsky,L. (1994), BioTechniques (in press).

ONE-STEP PCR SEQUENCING

Ken Porter, David Briley, and Barbara Ramsay Shaw*
Department of Chemistry, Duke University, Durham, NC 27708

A method is described to simultaneously amplify and sequence single- or double-stranded DNA. The method, which we call One-Step PCR Sequencing, is unique in that it employs a new class of αP-borane 2'-deoxynucleoside 5'-triphosphates (dNT^bPs) first synthesized in our laboratory. These boronated triphosphates exhibit useful properties: (a) they are heat stable, (b) they can be incorporated, base-specifically, into DNA during the polymerase chain reaction, and (c) once incorporated, the boranophosphate nucleotides block the action of exonuclease III. For One-Step Sequencing, a small percentage of boranophosphates are incorporated into DNA during PCR, and the positions of the stably-incorporated boronated dNMPs can be revealed by a simple exonuclease digestion, thereby defining the sequence of the PCR product. The One-Step method should eliminate two costly and time-consuming steps associated with current PCR sequencing techniques: (a) DNA purification following amplification and (b) single-sided primer extension with dideoxynucleotide chain terminators. As a consequence, the One-Step method should both decrease the time required to sequence PCR products, and render PCR sequencing completely automatable. Further, since the boranophosphate sequence delimiters are incorporated into the DNA during PCR, the method is readily amenable to bidirectional sequencing.

We have performed several preliminary experiments which are described below.

A. PCR is not inhibited by the presence of a small percentage of dNT^bPs. In fact, the dNT^bPs are incorporated extremely well into the PCR products and render the PCR products resistant to exonuclease III.

B. Primers can be extended in the presence of 100% of one boronated dNTP (A, T, G, or C) plus the three remaining normal dNTPs. These extension products are resistant base-specifically to exonuclease III under conditions in which all-normal extension products are degraded completely. Also, the presence of boronated nucleotides in these oligomers does not affect their electrophoretic mobility in 16% polyacrylamide gels.

C. One-Step PCR Sequencing was demonstrated with an end-labeled primer. The 509 bp region between phage T7 positions 34534 and 35042 was amplified by PCR using one labeled and one unlabeled PCR primer in the presence of a small percentage of each dNT^bP. The PCR products were digested with exonuclease III and separated by PAGE. Approximately 350 bases could be read from each strand.

D. Bidirectional One-Step PCR Sequencing was demonstrated with a biotinylated primer. The 629 bp region between T7 positions 21786 and 22414 was amplified by PCR using one biotinylated and one unmodified PCR primer in the presence of α-^{33}P-dATP and each dNT^bP. The products were incubated with streptavidin-linked magnetic beads, immobilized with a magnet, and digested with exonuclease III. Fragments from both strands were isolated and separated by PAGE. Sequencing data over 200-300 bases with boranophosphates compared favorably with dideoxy cycle sequencing.

Our method is rapidly approaching the quality of data which can be produced by cycle sequencing, yet requires much less DNA than the standard cycle-sequencing reaction. More DNA is required for cycle sequencing because the dideoxy-NTP chain truncators limit product formation to a linear accumulation. In contrast, One-Step PCR sequencing requires minimal starting template DNA because the boranophosphate sequence delimiters permit exponential PCR amplification. **The key advantages of boranophosphates as sequence delimiters in PCR are that they (1) delineate the DNA sequence and (2) do not obstruct exponential amplification.**

Quantitative Comparison of the Incorporation of Borano- and Thio- Deoxynucleoside Triphosphate Analogs by Klenow Polymerase

Kenneth Porter, Katie Ealey, J. David Briley, Faqing Huang and Barbara Ramsay Shaw

Department of Chemistry, Duke University, Durham, N.C. 27708

A new class of phosphate-substituted nucleotides, the 2'-deoxynucleoside 5'-α-borano triphosphates, has been shown to be good substrates for DNA polymerases. These boranophosphates form the basis of a sequencing method wherein DNA is subjected to repeated amplification/denaturation in the presence of both normal and α-boranophosphates, and then treated with exonuclease to produce fragments truncated at the nearest boranophosphate nucleotide (see accompanying abstract by Porter, Briley and Shaw).

The kinetic parameters for the incorporation of borano-analogs were determined by steady-state kinetic analysis and compared to the corresponding thio- and normal dNTP analogs. Labeled primer was extended by exonuclease-free Klenow in the presence of various concentrations of the appropriate dNTP or dNTP analog and the percent extension was quantitated and converted to a measurement of the initial reaction velocity. Values for k_{cat} and K_m were calculated from a Michaelis-Menten plot by nonlinear regression. The k_{cat} values for both the borano- and the thio- nucleotide analogs were about 40% that of normal dNTPs. However, the K_m values for boronated nucleotides (except for T) were less than the K_m values of normal dNTPs, while the K_m values for thiophosphates were significantly higher than borano and normal dNTPs. The efficiency of incorporation (k_{cat}/K_m) for the boranophosphates was approximately 30% that of normal dNTPs, whereas the efficiency of the thiophosphates was approximately 15% that of normal dNTPs. Therefore, the boranophosphates, although less efficient substrates than normal dNTPs, were shown to be twice as efficient as the thiophosphates.

Improvement of Ligation-Mediated PCR for DNA Sequence Determination

Kenneth.S. Graham, Arian F.A. Smit, and Arthur D. Riggs

Division of Biology, Beckman Research Institute of the City of Hope, Duarte CA 91010

We have been investigating the use of ligation-mediated PCR (LMPCR) as an approach to sequence determination of DNA that is difficult or impossible to clone, since LMPCR enables direct sequence determination of total genomic mammalian DNA. The basic method consists of 1) treatment of the genomic DNA with sequence specific cleavage agents, 2) primer extension with a gene specific oligonucleotide 3) ligation of an oligonucleotide to the blunt ends, 4) exponential PCR using a gene specific primer and a linker specific primer, and 5) sequencing analysis of the PCR products. All areas of the technique have been examined and improved, resulting in a 30% increase in overall speed, improved specificity and use of non-radiolabeled detection. In the area of sequence specific cleavage agents, a major improvement was achieved by incorporating an A specific cleavage protocol. Both primer extension and PCR were improved through the use of thermal stable polymerases such as Vent (exo-) for first primer extension and then Vent for PCR. The greater thermal stability of these enzymes allowed the use of higher Tm primers which give increased specificity and lower background. Since these enzymes have greater processivity than the previously used Sequenase and Taq, longer reads have been achieved and the sensitivity to G-C rich sequences has been reduced. Unambiguous determination of previously unknown sequences was accomplished. Moreover, the increased specificity and lower background also allowed the use dye conjugated primers for nonradioactive product visualization by use of an ABI automated sequencer.

Investigation was begun of a potentially major improvement using ligation of linker to the 3′ end of primer extended molecules instead of, as in standard LMPCR, to the 5′ end of molecules rendered blunt-ended by primer extension. Initial experiments were done using T4 ligase and were promising, but efficiencies were low, so we have been investigating a novel chemically mediated ligation procedure which should increase the efficiency of 3′ ligation.

Sequence determination by LMPCR is essentially a primer-walking method, so it ultimately depends on the specificity of oligonucleotide primers. We observed, initially at the DHFR locus, that some primers gave smears even though they were expected to be for single-copy target sequences. This work led to the discovery of a new, abundant family of mammalian apparent LTR retroposons (MaLRs) [1]. This analysis of repetitive sequences has now been expanded to cover both SINES and LINES. Consensus sequences for several repetitive element families have been derived and have been contributed to the repetitive sequence database being compiled by J. Jurka.

This work was funded by the DOE Genome Program (ER6-1137, A.D. Riggs P.I.)

[1] Smit, A.F.A. (1993) Identification of a new, abundant superfamily of mammalian LTR-transposons, *Nucleic Acids Res.* **21**: 1863-1872.

This page intentionally left blank.

Instrumentation

Automated Methods for Large-Scale Physical Mapping and Sequencing

Patricia A. Medvick[1], Tony J. Beugelsdijk, Jerome T. Chen, Bobi K. Den Hartog

Systems and Robotics Group, ESA-6, MS J580, and Center for Human Genome Studies; Los Alamos National Laboratory, Los Alamos, NM 87545. [1]Corresponding author.

As a major contribution to the Los Alamos mapping and library distribution effort, we have successfully created a prototype gridding system. This system reduces user interaction to initial setup, is flexible, and can produce 8 duplicate membrane arrays from 16 microtiter plates in two hours [1,2]. System development provided insight into the requirements of molecular biologists in the laboratory. With prolonged use of the prototype gridding system, analysis of the probed membranes was rate limiting. Our software for automating this analysis procedure addressed the data glut and data-tracking problems that result from isolated automation of individual steps in a process. These problems are a familiar occurrence in analytical chemistry labs. To combat the analytical-chemistry throughput problem for the Department of Energy cleanup effort, the Robotic Technology Development Program has supported the Contaminant Analysis Automation (CAA) project, an integrated effort involving five national laboratories [3]. The CAA concept, in which samples are directed through intelligent submodules by a central controller, consisting of a user interface, task sequence controller, database and expert system, has been developed and implemented. This paradigm is directly applicable to the molecular biology laboratory, particularly to a high-throughput effort such as the mapping and sequencing of the human genome.

Our next generation gridding system conforms to this model. The software interfaces and controller are directly applicable to our next task of constructing a high-throughput sequencer. The gridding system contains four hardware modules consisting of a robot, tool-washer, plate stacker/restacker, and barcode reader. Controllers for these modules include an Adept controller, two single-board IBM-type PCs and a Sun workstation that contains the central control software. The central control software consists of a user interface, database, task sequencer, and communication interface to each of the hardware controllers. Configuration information for each hardware module is stored in the database and provides a list of tasks performed by each module. Scripts provide the pattern of tasks for the system and can be modified by the biologist. This model for hardware control provides the flexibility to incorporate changes to the process and to upgrade hardware as new equipment becomes available. The control software is directly applicable to the integration of a high-throughput sequencing system and is suitable for facilitating cooperation between centers of automation development.

This work was funded by the DOE Genome Program (ERW-F142, P. Medvick, P.I.).

[1] Medvick, P. A., R. M. Hollen, R. S. Roberts, D. Trimmer, and T. J. Beugelsdijk. (1992) Automated DNA Hybridization Array Construction and Database Design for Robotic Control and for Source Determination of Hybridization Responses. *International Journal of Genome Research* **1**, 1, 17-23.

[2] Medvick, P. A., R. M. Hollen, and R. S. Roberts. (1991) Development of an Automated Workcell for DNA Hybridization Array Construction, *Journal of Laboratory Robotics and Automation.*, **3**, 4/5, 169-173.

[3] Whole issue of *Laboratory Robotics and Automation*, **6**, 2, 55-104.

A Rapid, Microtiter-Plate, Robotics-Compatible, Multi-Station Thermal Cycler

A.D.A. Hansen, J.M. Jaklevic, V.M. Stevko, W.F. Kimmerly, *et al.*

Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720

We have developed a thermal cycler based on the circulation of temperature-controlled water directly to the underside of thin-walled polycarbonate microtiter plates. The water is rapidly switched from a set of reservoirs. The plate wells are loaded with typically 15–20 μl of reagent mix for the PCR process. Heat transfer through the thin polycarbonate is sufficiently rapid that the contents reach thermal equilibrium with the water in less than 15 seconds, although unexpected improvements in performance are observed in non-equilibrium runs in which the temperature switching is more rapid. Complete PCR amplification runs of 40 three-step cycles have been performed in as little as 14.5 minutes, with the results showing substantially enhanced specificity compared to conventional technology requiring run times in excess of 100 minutes. 96-, 192- and 384-well plates can be used. The plate clamping station includes a heated lid, eliminating the need for mineral oil overlay of the reactants, and was specifically designed to be accessible for robotic loading and unloading of the plates. The present apparatus has three plate stations, fed from common reservoirs but operating with independent switching cycles. The high overall throughput is required to meet the needs of the LBL Human Genome Center effort. The enhanced performance and specificity observed in some cases may allow for extension of PCR methods in mapping and sequencing strategies. Experience developed in using the system in a production environment will be described together with future design modifications.

TURBO PCR—An Integrated Robotic System for Continuously Setting Up and Running Multiple PCR Reactions

John Meng, Don Uber, Joe Jaklevic

Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720

The Three Unit RoBOtic thermal cycler—TURBO PCR for short—is a multi-station thermal-cycler unit serviced by an ORCA robot unit which also services a Biomek robot unit. The ORCA robot loads the Biomek robot from an extensive hotel array containing tip racks, microtiter plates and titer-tube racks of input material. It reads bar codes from input plates and output plates as they are used, permitting material to be tracked as it passes through the system. The Biomek robot is programmed to perform the chemistry necessary to set up 96 PCR reactions in a microtiter plate. Once the reactions are set up, the ORCA passes loaded plates from the Biomek robot into the thermal cycler. It removes plates from the thermal cycler at the end of PCR runs. The ORCA hotels can hold supply quantities sufficient to service nine input plates, which may consist of either titer-tube racks or thin wall microtiter plates. The hotels may be loaded with additional inputs and supplies at any time, making it practical for the system to be run continuously. Setup and orchestration of the TURBO PCR's operations is done from an independent computer running a state-machine interpreter which communicates via shared files in the ORCA computer, the Biomek computer and the Thermal Cycler computer. Other units, such as an imaging station and a gel load-and-run station, easily connect to the interpreter, and additional units may be connected as they come on-line. The interpreter also performs data base operations and monitors the system for unit failures, keeping track of hardware operations and experimental materials.

Lawrence Berkeley Laboratory DNA Preparation Machine

Martin J. Pollard

Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720

A DNA preparation machine is in a late stage of construction. It is designed to execute a modified boil prep to extract plasmid DNA from *E. coli* bacteria. The machine consists of a gantry style robot (827 W x 1715 L x 340 H mm) coupled with a modified IEC 7000 centrifuge capable of spinning microtiter plates to 5000 g's. The four fixed robot tools consist of two 8 channel Biomek MP200 pipettors, 5 eight channel dispensing manifolds, and a pneumatic gripper for moving the microtiter plates within the work envelope and into and out of the centrifuge. The working surface is an optical table with locating fixtures for the microtiter plates, the pipet tip racks and other miscellaneous hardware. The hot plate and the cold plate each hold two microtiter plates. The microtiter plates are pressed against the constant temperature surfaces pneumatically. The IEC 7000 centrifuge has been modified with a computer interface, an automatically lifting cover, an indexing rotor, and robot compatible sample buckets.

The instrument is controlled with a 486 Gateway 2000 PC with a Visual Basic software interface. The software interface allows both operational and teach modes. The teach mode allows the operator to develop protocol modules by guiding the robot through each step of the process and recording a description of each step in a text file. During the operational mode a collection of text files, describing the entire protocol, is read, parsed, and executed on the machine. The protocol files are easily edited in the teach mode or on a text editor.

The LBL DNA Prep Machine is designed to process samples in two microtiter plates (192 samples) in 3–4 hours. The instrument will run unattended while operating. One run will result in enough samples to produce 35 kb of finished sequence in LBL's directed sequencing strategy. This instrument will complete the series of automation modules required for obtaining sequence from plasmid inserts. The large work envelope of the machine will accommodate a variety of other applications as well.

Automated Pooling of Large DNA Libraries

Donald C. Uber

Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720

Our mapping strategies require screening large libraries (e.g. 100,000 clones) with multiple probes to find the occurrences of specific DNA sequences. Since the expected number of hits in a library with a particular probe is relatively few (e.g. five), the effort in this task can be greatly reduced by first combining the clones into pools according to a prescribed pattern, and screening just the pools. By observing which pools "light up", it is then straightforward to identify which clones match the probe.

Using a Hewlett-Packard ORCA robot, we have automated a 3D pooling strategy in which 960 clones in 96-well format are combined into 30 pools: 8 row pools, 12 column pools, and 10 plate pools. This scheme reduces the number of screening assays by a factor of 32. The ORCA uses a custom 12-channel pipet tool with standard 9 mm spacing, and makes pools in troughs that accept 12 samples simultaneously. A single disposable pipet tip per clone is used, as compared to three tips when the process is done manually. A run of 10 source plates is completed in less than three hours, including making a copy of each source plate, as compared to five hours manually.

Automated Colony Picking Machine

Martin J. Pollard, Donald C. Uber, Jack S. Zelver

Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720

The LBL automated colony picking machine uses a room temperature CCD camera to image 100 x 100 mm colony plates. The Optimas software automatically selects colonies using criteria such as size, circularity, aspect ratio, nearest neighbor distance, etc. The colony image coordinates are then translated into mechanical colony picking coordinates.

The mechanical colony picker positions the source colony plate and the destination microtiter plate under a carousel of needles which pick and place the colonies. Each needle dips into the source plate onto a colony and then rotates until it is over the appropriate well in the 96-well destination microtiter plate. It then dips into the growth medium in that well to transfer picked cells. The needles move on to a semicircular cam that dips them into an ultrasonic cleaning bath. Picking, placing, and cleaning are performed simultaneously.

Usage of the colony picker has shifted from occasional picking of large libraries to more frequent picking of smaller libraries used as part of LBL's directed sequencing strategy. This mode requires picking 2000 colonies, five times every two weeks. To facilitate this, a new Windows-based graphical user interface was developed for the control software. The software now operates in two modes depending on how you start the application. The first picks a continuous stream of colonies into microtiter plates to generate large colony libraries. The second picks colonies into a single microtiter plate. Each microtiter plate constitutes a separate "experiment". Typically colonies are picked into 20 microtiter plates during a session. The new software makes it easier to use and more intuitive even for relatively inexperienced users.

The original programs were written in DOS-based QuickBasic and have been converted to Visual Basic. The code now takes advantage of the services provided by the object-oriented features of Visual Basic. Most modules were re-designed to be event-driven and parallel rather than in-line, and much of the code was replaced by object methods and property manipulations. Ease of maintenance was preserved because the two BASIC languages are mostly either identical or similar. An exception handling scheme was created to allow recovery from a variety of mishaps. This allows the user to restart picking into the middle of a destination plate, and allows a different sequence of input files to be chosen.

We have also upgraded the procedure which calibrates the imager and picker to each other. The picker pierces a sheet of aluminum foil with a predetermined array of holes, using the same needle. This array is imaged, and a nonlinear fit is performed between the locations of the holes in the image and their nominal locations when punched. This computation yields a set of polynomial coefficients that are used to transform colony centroids into picker coordinates. This transformation simultaneously handles the necessary scaling, translation, and rotation between imager and picker coordinates, and corrects for lens distortion and camera positioning errors.

To account for variations in straightness among the needles, the foil is punched with another grid of holes using all the needles. Using the coefficients described above, the hole for each needle is mapped from its image location back into picker coordinates, and compared with its nominal punched location. The difference, or offset, is applied to the needle when it picks a colony. The new calibration procedure has proven to be extremely convenient and accurate.

Automation of a Directed Sequencing Strategy

Joseph M. Jaklevic

Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720

The Lawrence Berkeley Laboratory Human Genome Center is currently expanding a large-scale directed sequencing project toward a multi-megabase per year capability. The LBL directed sequencing strategy is based on STS content mapping of P-1 inserts followed by high-resolution mapping of 3 kb fragments and transposon assisted sequencing. This approach results in a ten-fold reduction in template preparation and a five-fold reduction in sequencing relative to random strategies. Automation of this directed approach presents many unique challenges and opportunities in the area of laboratory instrumentation. In particular, the extensive high-resolution mapping which precedes template sequencing places large demands upon PCR gel assays and related protocols.

Our approach to automation of this strategy includes the use of specific robotics-based tools for colony picking, library replication, pooling and for performing PCR reaction preps and sequencing reactions. In addition, a number of specialized instrumentation modules have been developed. An automated image acquisition and analysis system employs a digital camera to acquire images from ethidium bromide stained gels. Image processing and analysis software is then used to automatically locate bands and assign sizes for mapping purposes. A robotics-compatible, multi-station thermal cycler is currently capable of performing 600 amplifications per hour in a three-plate, 96-well format. The next version will have a larger number of individual stations and will be capable of using 384-well plates. An automated 12-channel oligosynthesizer produces custom oligos at a rate and cost significantly improved with respect to existing commercial systems. It is designed to be expandable in multiples of 12 channels up to a 96-well format. Progress in all of the areas will be reviewed.

Recent efforts to integrate these specialized modules into functional units with robotic materials handling and data tracking will also be presented.

Low Cost Automated Preparation of Plasmid, Cosmid And Yeast DNA

Tuyen Nguyen, Randy F. Sivila, Joshua P. Dyer, and William P. MacConnell

MacConnell Research Corporation, San Diego, California

MacConnell Research currently manufactures and sells a low cost automated bench-top instrument that can purify up to 24 samples of plasmid DNA simultaneously in one hour at a cost of $0.50 per sample and under $8000 for the instrument. The patented instrument uses a form of agarose gel electrophoresis to purify the plasmid DNA and electroelutes in into approximately a 20 µl volume. The instrument has many advantages of other robotic and manual methods including that fact that it is: two times faster, at least six times less expensive, much smaller in size, easier to operate, less cost per sample, and results in DNA pure enough for direct use in fluorescent automated sequencing. In addition, the instrument has only one moving part associated with its purification process and does not have internal reagent reservoirs. The instrument process begins with bacterial culture which is loaded directly into a disposable cassette in the machine.

In the proposed SBIR phase I work we will develop a version of the above instrument which will simultaneously process 96 bacterial samples in 1.5 hours and will prepare cosmid, yeast and mammalian DNA in quantities of 1-5 microgram per cassette lane. The goal will be to obtain DNA of sufficient purity for direct use in automated fluorescent and manual sequencing as well as other molecular biology protocols. The proposed 96-channel instrument will purify over 1000 plasmid DNA preps per eight hour day.

This work is being supported in part by DOE SBIR Phase I Grant # DE FG 03 - 94ER81802 / A000, W. MacConnell P.I.

Molecular Biology Laboratories on Microchips

Stephen C. Jacobson, Alvin W. Moore, Jr., and J. Michael Ramsey
Chemical & Analytical Sciences Division
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge, TN 37831-6142

We have recently demonstrated the ability to perform a number of chemical analysis manipulations on microfabricated devices that have areas on the order of 1 cm^2 and volumes of a few tenths of a cubic centimeter. These devices are fabricated using conventional micromachining techniques on glass and fused silica substrates. Microchannels and microreactors are fabricated by chemically etching features in a planar substrate. The channels and reaction chambers are closed by bonding a cover plate over the machined areas. Microfluidic manipulation is effected through the use of electroosmotic flow. We have demonstrated the ability to manipulate fluid volumes as small as 100 picoliters with the precision of less than 1% rsd. Several chemical separation techniques have been demonstrated in our laboratory on microfabricated devices including free solution electrophoresis, open tubular electrochromatography, micellar electrokinetic capillary chromatography, and capillary gel electrophoresis. Moreover, the ability to perform chemical reactions "on chip" under computer control with picoliter and nanoliter volumes has been demonstrated. Performance of all these separation techniques have been either equivalent to or greater than conventional laboratory devices. We are now applying these new capabilities to molecular biology problems.

"The submitted manuscript has been authored by a contractor of the U.S. Government under contract No. DE-AC05-84OR21400. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes."

Research sponsored by U.S. Department of Energy, Office of Basic Energy Sciences, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

Chromosome Mapping by Quantitative Image Analysis: New Tools and Applications

Babetta L. Marrone, Gary C. Salzman, Stephanie Pendergrass, Julia Gonzales, Cheri Potter, James H. Jett, and Larry Deaven[1]

Cell Growth Damage and Repair Group, LS-1, M888, Life Sciences Division, and [1]Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545

Several software tools have been developed to facilitate cytogenetic mapping of DNA sequences. The DNA sequences are labeled on human metaphase chromosomes by one- or two-color Fluorescence *In Situ* Hybridization (FISH) and chromosome images are captured by computer-assisted digital microscopy. The tools include: 1) High precision fractional length analysis, for locating FISH-labeled DNA sequences with 1 Mb resolution on metaphase chromosomes; 2) High resolution two-color analysis, for ordering closely spaced or partially overlapping FISH signals; and 3) DAPI-band profiling to facilitate cytogenetic band assignments of FISH-labeled DNA sequences or sequences labeled by Primed *In Situ* elongation (PRINS). Examples of the applications of these tools will be presented, including localization and sub-band ordering of mega-YACs surrounding DNA repair genes on chromosome 2 and chromosome 5; whole genomic screening for clusters of DNA repeat sequences labeled by PRINS; and mapping gaps in the nearly completed LANL chromosome 16 mega-YAC map.

Work performed under U.S. Department of Energy contract W-7405-ENG-36.

A High Speed Method for Chromosome Sorting Using the Photoinduced Cross-linking of Psoralen Derivatives with Chromosomal DNA: Preliminary Studies

M. C. Roslaniec,[1] J. C. Martin,[1] R. J. Reynolds,[1] L. S. Cram[1]

[1]Life Sciences Division - 1; Los Alamos National Laboratory; Los Alamos, New Mexico, 87545

A High Speed Optical Chromosome Sorter based on selective, irreversible photoinactivation of chromosomal DNA is being developed. Chromosomes are analyzed as in traditional FCM, however, rather than relying on droplet formation, (a rate limiting factor in droplet sorting), sorting is achieved by photoinduced cross-linking of undesired chromosomes with a photosensitive compound such as psoralen. Cross-linked DNA cannot be denatured and is rendered unclonable. Desired chromosomes are not irradiated and will remain clonable providing a means of separation. Initial results indicate that treating pBluescript SK(+) phagemid DNA (100 ng) with ≈ 4.0 μg of 8-methoxypsoralen and 35 kJ/m^2 UV lamp irradiation produces ≈ 5 photoadditions (mono- and diadducts)/kbp. We have determined that this level of cross-linking effectively inhibits transfection of XL1-blue *E.* coli cells. Furthermore, since our ultimate goal is the construction of chromosome specific libraries, we are using the s-Cos-1 vector to examine the effects of photo-cross-linking on cosmid cloning. Finally, we report preliminary results of the effect of chromosome photoinactivation in flow.

Supported by the National Flow Cytometry Resource, NIH grant RR01315 and by the U. S. Department of Energy

Sizing DNA Fragments by Flow Cytometry - Extending the Size Range

Jeffrey T. Petty,[†] Mitchell E. Johnson,[†] Peter M. Goodwin,[†] James H. Jett,[‡] John C. Martin,[‡] and Richard A. Keller[†]

[†]Chemical Science and Technology Division, MS M888, Los Alamos National Laboratory, Los Alamos, NM 87545. [‡]Life Science Division, MS M888, Los Alamos National Laboratory, Los Alamos, NM 87545.

Using flow cytometry, sizes of DNA fragments were obtained from the fluorescence intensity of samples stained with a thiazole orange dye. The stained fragments passed through a low power (30 mW) continuous-wave laser beam one at a time using transit times of 1 - 5 ms. As little as 50 fg of DNA was analyzed at a rate of 40 fragments/sec in times ranging from 1 - 15 min. A detectable lower size limit of 1.5 kilobase pairs (kbp) was demonstrated, and a linear relationship between fluorescence intensity and fragment length was observed for the fragments ≥ 4.4 kbp. Issues relating to size resolution in the 2 - 50 kbp range are discussed.

This work is supported by the Los Alamos Center for Human Genome Studies under United States Department of Energy Contract W-7405-ENG-36 and the National Flow Cytometry Resource , NIH RR 01315.

[1] Goodwin, P. M.; Johnson, M. E.; Martin, J. C.; Ambrose, W. P.; Marrone, B. L.; Jett, J. H.; Keller, R. A. (1993) Rapid Sizing of Individual Fluorescently Stained DNA Fragments by Flow Cytometry *Nucleic Acids Res.* **21**, 803.

[2] Johnson, M. E.; Goodwin, P. M.; Ambrose, W. P.; Martin, J. C.; Marrone, B. L.; Jett, J. H.; Keller, R. A. (1993) Sizing of DNA Fragments by Flow Cytometry *SPIE* **1895**, 69.

Three Dimensional Imaging of DNA Fragments During Electrophoresis Using a Confocal Detector

Larry Brewer , Courtney Davidson, Joe Balch, and Anthony Carrano

Human Genome Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore California. 94550

We have measured the three dimensional distribution of individual fragments of DNA during electrophoresis. Separations (65 V/cm) were carried out in a 6% polyacrylamide gel confined by a glass microchannel with cross sectional dimensions of 200 microns deep by 1000 microns wide. The detection system consisted of a confocal microscope with a measured depth of focus of 30 microns (FWHM) plus a photomultiplier to detect the laser induced fluorescence from the dye-labeled DNA. The photomultiplier detector was used in a photon counting mode because of the small signal levels (~ 1000 Hz). The electrophoresis microchannel plates were mechanically scanned using a precision X-Y stage and a Z axis microscope focus controller, both under computer control. Preliminary results indicate that the DNA was confined in the center of the microchannel, with an approximately Gaussian profile along all three axes. DNA fragments ranging from 50 to 300 bases were profiled and found to have similar spatial distributions. These results will be important for designing optimized high-throughput electrophoresis systems.

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

New Method For The Detection Of Right-Angle-Scattered Light In Flow Cytometry

Raymond Mariella Jr., D. Masquelier, Gerald Eveleth, and Richard Langlois

Human Genoime Center, L-452, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550.

We report a new physical configuration for the detection of right-angle-scattered (RAS) light in Flow Cytometry which greatly increases the signal-to-noise ratio, narrows the coefficient of variation for uniformly-sized latex spheres, and greatly eases the alignment requirements, too. The new technique views the scattered light which is trapped within the optical waveguide of the flow stream in air.

In previous Flow Cytometers, the RAS light has been viewed perpendicularly to the liquid flow, typically using a high-numerical-aperture (NA) microscope objective lens or fiber optic. Some of the difficulties associated with this approach are the very limited depth of field of high-NA optics, and the necessity to align precisely the exact focal point of the lens with the point where the excitation light source intersects the sample flow stream. Our invention uses the unconfined aqueous flow stream itself as a 0.75-NA optical waveguide. There is no "focal point" for this configuration. Alignment simply requires aligning the light source onto the flow stream; the liquid optical waveguide is then automatically "aligned". For the collection of elastically-scattered light, another advantage occurs: the background level of scattered light is extremely low when using the flow-stream waveguide (FSW), because the same physical properties which confine the desired light within the stream also keep random scattered light out, no obscuration bar is needed for the collection of RAS light.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no.W-7405-ENG-48.

Large-Scale Oligosynthesis in a Multichannel Format

L.E. Sindelar and J.M. Jaklevic

Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley CA 94720

We describe an approach to large-scale parallel oligosynthesis in which a multi-well format is used. The reactions are carried out in open wells using an argon ambient atmosphere to prevent reagent contamination. The controlled-pore glass beads which form the substrate for synthesis are held in individual wells with high-density polyethylene filter bottoms through which reagents are drawn into a vacuum manifold. The synthesis is carried out using direct reagent dispensing into the individual reaction wells. A computer controls the sequence in which reagents are dispensed and the timing of the periodic vacuum pulses required to synthesize the desired sequence.

Experiments to date have demonstrated the viability of the approach for a variety of test sequences. Results obtained with HPLC analysis demonstrate coupling efficiencies as high as 99.5% under optimized conditions. Use of the oligomers for DNA sequencing templates and as PCR primers has been demonstrated in production applications. The current instrument design consists of a series of discrete, 12-channel reaction chambers capable of multiplexing in a 12 x N format where N can be 1 to 8, i.e., 96 wells. A projected time interval for 12 parallel syntheses is 2:30 hours, with 96 syntheses in 3:30 hours. The current 12-channel system is now being used in a full production environment and is easily capable of producing 120 custom oligos per work week. The proposed extension to the larger 12 x N format will provide proportionately greater output. Because of the reduced volume of reagents required in the open well format, significant cost savings are projected.

Modular Instruments for Genome Mapping and Sequencing

Ger van den Engh, Richard Esposito, Jordan Hopkins, David Basiji, Barb Trask, Maynard Olson

Department of Molecular Biotechnology, University of Washington FJ-20, Seattle, 98195

Progress in the Human Genome effort will depend on the emergence of automated techniques for mapping and sequencing. We envision that the necessary scale enlargement (from a few MBases/year to a Mbase/day) will be achieved by processes of an industrial character. We expect that maximum efficiency will be obtained by organizing the mapping process into a chain of simple steps that are carried out by stations arranged along a linear "assembly" line. Samples will be processed sequentially rather than in synchronous batches. Sequential, pipe-lined operations carried out in parallel maximally utilize the capacity of the individual process elements.

We previously developed a modular, high performance flow cytometer, known as the MoFlo system. This instrument has been designed around a digital bus that accepts electronic modules that carry out the functions of flow cytometry and cell sorting. Event processing and data acquisition occur in a pipe-lined, parallel fashion. The optical system is similarly modularized and highly adaptable to different experimental requiremetns. This modular approach to instrument design resulted in a machine that compares favorably to competing designs with respect to simplicity of design, performance and ease of operation. MoFlo systems are being used with high efficiency and reliability for chromosome analysis, chromosome sorting and rare event analysis.

We are building on the experience gained in designing modular instruments for genome analysis. Modules of the high speed sorter are now being reconfigured for the mapping and sequencing process. Todate, MoFlo modules have been utilized for the construction of a gel scanner for restriction fragment analysis and an instrument for bacterial-clone sorting and expansion. These instruments have common control and data exchange protocols. Our experience as well as designs for future modules will be presented. The prospects for integration into a autonomous mapping and sequencing process will be discussed.

This work was supported by U.S. D.O.E. grant DE-FG06-93ER61662.

This page intentionally left blank.

Ethical, Legal, and Social Issues

Adjudication of Genetic Testing and Gene Therapy Evidence in Federal and State Courts

Franklin M. Zweig[1]

Advances in genetics forecast new challenges for federal and state courts. Supported by the ELSI Program of the DOE Human Genome Project, this project is developing a desk book for judges to assist their management of the complex evidence issues inherent in genetics-related litigation. A companion videotape and CD ROM reference guide is being planned.

Adjudication desk books are traditional tools for judicial use. They typically are resources deployed in judicial chambers upon case assignment. They describe the scope of the litigation issues at hand. They provide glossaries and case law summaries. They suggest tasks to be considered at each procedural stage of lawsuit adjudication. They do not prescribe procedures or outcomes. They provide options for consideration of the presiding judge.

Applied genetics has vigorously been litigated in the forensic field for the past five years. Heated contests can be predicted to occur with respect to the medical technologies derived from molecular biology. This project includes the following activities to prepare courts:

1. Preparation of a scientific primer on genetic testing and the prospects for gene therapy in language and with exhibits that make concepts and empirical findings more accessible to courts.

2. Inventory of the case law involving genetics issues in order to document a precedent foundation that can be consulted in detail.

3. Commentary on the salient procedural implications for genetics information in a prosecution or a civil lawsuit in which genetics information is introduced. These procedural implications include screening evidence proffered pre-trial for its scientific validity in comportment to the rules, statutes, and case law governing the admissibility of scientific evidence and expert witness testimony; qualification of expert witnesses; rulings in response to evidence-related motions; examination of expert witnesses at trial; and issuance of jury instructions.

4. Description of collateral legal actions in which genetic testing and gene therapy evidence may prove pivotal, including but not limited to lawsuits over: medical information privacy infringements; medical practice guidelines issued by an authoritative source for either quality assurance, cost-containment, or liability limitations purposes, or a combination thereof; the pursuit of lawsuits related to payments and rates for health insurance reimbursement; the pursuit of lawsuits related to health care provider, laboratory, or utilization reviewer malpractice.

5. A workshop on genetics background materials and sensitizing considerations to help determine the desk book's content, organization, and presentation mode.

Franklin M. Zweig, PI, is Senior Research Staff Scientist for Law and Judicial Policy at The George Washington University, Washington, D.C., and President, Einstein Institute for Science, Health and the Courts. Correspondence address: 3122 Brooklawn Terrrace, Chevy Chase, Maryland 20815 (301) 913-0448, tel and fax.

The Human Genome Project: Information Management, Access, and Regulation

Development of Educational Materials for High School Biology

(Grant No. DE-FG03-93ER61584)

Joseph D. McInerney and Lynda B. Micikas, Biological Sciences Curriculum Study (BSCS)

Informatics is a central, but sometimes overlooked, aspect of the Human Genome Project (HGP). Management of the vast amounts of data generated by the HGP presents technological and logistical challenges to those who require access to those data. In addition, the electronic storage of genomic information raises important questions of ethics and public policy, many revolving around privacy and confidentiality.

BSCS has addressed the scientific, technological, ethical, and policy aspects of genome informatics in an instructional program titled *The Human Genome Project: Information Management, Access, and Regulation.* The program, intended for use in introductory high school biology, includes a 125-page print module and software. The print materials provide approximately 50 pages of background material for teachers and seven days of classroom instruction. The software includes two model databases: a research database that contains anonymous data (map data, sequence data, and biological/clinical information) and a registry that attaches names (three kindreds, 52 individuals) to genomic data. The students manipulate the databases as they work through the classroom inquiries.

BSCS used the development process that it has refined continually since the inception of the organization in 1958. The materials were designed and written by experts in human and medical genetics, molecular biology, informatics, ethics, public policy, and classroom teaching. The education committees of the American Society of Human Genetics, Council of Regional Networks of Genetic Services, and the National Society of Genetic Counselors reviewed the conceptual framework for the program and the field-test materials. BSCS field tested the program with 1,000 students in high schools across America. Following final revision based on the field-test data and external reviews, BSCS will distribute the print materials and software free of charge to all 50,000 high school biology teachers in the United States.

A Hispanic Educational Program for Scientific, Ethical, Legal and Social Aspects of the Human Genome Project

Margaret C. Jefferson[1,3] and Mary Ann Sesma[2]

[1]Department of Biology & Microbiology; California State University, Los Angeles; 5151 State University Drive; Los Angeles, California 90032. [2]Los Angeles Unified School District; 1621 Sunnyhill Drive; Monterey Park, CA 91754. [3]Corresponding author.

This project is a multidisciplinary, bilingual (Spanish-English) educational program about the Human Genome Project (HGP) and related ethical, legal and social issues (ELSI). The target population is the bilingual Hispanic within the Los Angeles Unified School District, who represents 65% of an approximately 850,000 pre-school to adult enrollment, and their teachers and parents. The major focus is to develop culturally competent, linguistically appropriate, and relevant curriculum that leads to Hispanic student and family interactions.

The current students in the local school population and their parents demonstrate little knowledge of HGP-ELSI when pre-tests were given to these students or when small focus groups of parents met. The basic curriculum to be developed consists of the full text translation into Spanish of existing BSCS HGP curriculum plus supplemental materials for students and parents that focus on the genetics of New World Hispanics and HGP-ELSI. A major goal is to open up a channel of familial dialogue between parents and their students about HGP-ELSI, and to develop linkages for access to health and educational services for individuals seeking assistance. Other activities include the development of assessment criteria for both parent and student curriculum components; development of a tracking system to measure student progress; pretesting the curriculum prior to start of academic year; implementation and testing of curriculum at Bell High School (a year round, four year high school with over 4,000 students, 93% Hispanic) with follow-up modification; dissemination, implementation and testing of curriculum at four additional, predominantly Hispanic high schools; and national dissemination of project objectives and outcomes.

Family progress will be measured by attendance at an annual conference of students, parents, and teachers at California State University, Los Angeles. Two aspects of this conference are of major importance: (1) parent involvement in the educational experiences of their children and (2) display of student work at the conference for positive reinforcement. Throughout the academic school year, teams of students will produce weekly projects related to the science / ELSI concepts they are learning in the classroom which will be displayed at the annual conference. In addition, students will prepare one page newsletters to take home each week for discussion with their parents. This project will be closely linked with several other programs, such as Healthy Start (California Senate Bill 620). This linkage provides opportunities to develop communication systems that are appropriate to the Hispanic population of students and parents.

This work is funded by the DOE HGP-ELSI program (DE-FG03-94ER61797 A000, M. C. Jefferson and M.A. Sesma are co-P.I.).

HUMAN GENETICS EDUCATION FOR MIDDLE AND SECONDARY SCIENCE TEACHERS

Debra L. Collins, M.S.[1], Linda Segebrecht, M.S.[2], R. Neil Schimke, M.D.[1]

[1]University of Kansas Medical Center; 3901 Rainbow Blvd.; Dept. of Endocrinology and Genetics, 4023 Wescoe; Kansas City, KS 66160-7318. [2]Science Pioneers, Inc.; 425 Volker Blvd.; Kansas City, MO 62110.

The goal of this project is to increase public awareness of the Human Genome Project through a series of educational workshops, development of a teacher mentor network, and dissemination activities by teachers. Workshop topics focused on the application of ethical, legal, social, and public policy implications to new genetic technology. All participants complete four phases of the project over a two year period:

Phase I: first one week workshop
Phase II: classroom use of materials and information
Phase III: second one week workshop
Phase IV: peer dissemination of information

Speakers included genetic counselors and clinical geneticists, lawyers and ethicists familiar with HGP/ELSI and public policy issues, researchers using DNA technology, teachers experienced in presenting HGP/ELSI topics in their classrooms, and consumers. Families (consumers) presented information on the impact of genetic conditions on their lives, discussed misconceptions about their condition, and gave educators a humanistic context unavailable from textbook descriptions of genetic conditions.

Uses of biotechnology were demonstrated through tours of laboratory facilities using new DNA technologies in research, clinical care, and forensic science.

Teachers used the DOE / BSCS curriculum in their classroom, presented peer teaching programs, and began disseminating workshop information at local, state, and national teacher meetings. Most peer teaching will be completed in the 1994 - 1995 school year.

To date, 115 teachers from 40 states and the District of Columbia have attended these DOE / HGP workshops. These educators will teach more than 16,000 students each year. Approximately 19% of the students are from minority populations. Participants have disseminated information about the Human Genome Project to more than 2,600 peer teachers in 33 states at more than 150 different educational programs and meetings.

The project is evaluated through teacher assessment of knowledge, preparedness and confidence, as well as a national pre and post-survey of students' knowledge of genetics and Human Genome Project topics. Preliminary data shows a large increase in student knowledge in participants' classrooms compared to a control group.

A database of genetics educational materials has been compiled at our Education Center with annotated information on developed curricula, books, booklets, brochures, computer programs, hands-on materials, newsletters, posters, videotapes, and other resources. This list is available to educators.

Assessment of Genome Education

Cheryl Dell, Ph.D.[1], Diana DeVries, M.A.[2], and Diane Baker, M.S.[1]

[1]Education Program, University of Michigan Genome Center, 2570 MSRB II, Ann Arbor, 48109-0674.
[2]Center for the Study of Higher and Post Secondary Education, School of Education, University of Michigan, 610 E. University, 2117 SEB, Ann Arbor, MI 48109-1238.

The provision of genome-related educational outreach is on the increase and is finding receptive and eager audiences in a variety of fields. Assessment tools are needed which go beyond examining participant satisfaction with instructor defined goals to measuring participant defined needs and actual changes in participant knowledge, behavior and attitudes. We will present an assessment model developed for an audience of genetic counselors who attended a one-week course titled, *Molecular Diagnostics, Genetic Counseling and the Human Genome Project* sponsored by the University of Michigan Human Genome Center in August, 1994. This course was designed for genetic counselors in response to their increasing utilization of DNA-based diagnostic techniques in patient management. The course was attended by 25 individuals and included five days of lectures, discussions, hands-on laboratories and exercises surrounding the technology of DNA-based testing.

The assessment plan was developed around three theoretically based components: subjective (participant self-report), objective (measurable changes in instructor-defined tasks), and participatory (participant-defined goals and needs). This design includes eight distinct assessment activities: (1) a one-page statement identifying why the participant wished to take the course; (2) a participant profile describing training and experience; (3) a pre-test on knowledge about specific genetic techniques and concepts; (4) small group discussions of participant goals, course design and the assessment process; (5) a post-test repetition of the pre-test instrument administered at the close of the course; (6) small group discussions about the impact of the course on participant knowledge base and practice; (7) evaluation of course logistics, speakers, and organization; and (8) a follow-up assessment administered six months after the course to investigate behavior or performance changes resulting from the course experience.

This poster will use the assessment design and the resulting data to discuss implications for genome-related educational outreach and the field of genetic counseling. In particular, these assessment activities highlighted the isolation in which genetic counselors work as well as their felt need to define practice guidelines regarding DNA-based testing. This assessment also provides insights regarding the process of making links between rapidly developing DNA-based technology and practice-based service to individual families.

STUDENT PUGWASH USA

International Conference Working Group:
The Social Costs and Medical Benefits of Human Genetic Information

Betsy Fader, Executive Director
Student Pugwash USA,
1638 R Street, NW • Suite 32
Washington, DC 20009
Tel: (202) 328-6555 Fax: (202) 797-4664

Student Pugwash USA, a national, educational, non-profit organization, helps young people of diverse academic and ethnic backgrounds gain a better understanding of social and ethical issues raised by science and technological advancements in the following key areas: health and medicine, environment and energy resources, peace and global security, population and international development, and information/ computer technologies. Student Pugwash USA's interactive, educational programs and conferences bring international, interdisciplinary groups of motivated students together with scientists, policy makers, members of academe, and industry leaders for examination of critical issues at the juncture of science, technology and society.

In June 1994, Student Pugwash USA conducted a week-long international conference focusing on science, technology and ethical responsibility. The Conference, entitled "Science and Technology for the 21st Century: Meeting the Needs of the Global Community", brought together approximately 90 college and university students and 65 eminent professionals from 25 different countries for intensive discussions on the role of science and technology in world affairs. Six different "working groups" were assembled for the week-long event held at Johns Hopkins University in Baltimore, Maryland, with one working group focusing on the ethical, legal and social implications of the Human Genome Project. The working group raised the following key questions relating to the Human Genome Project, including:

- What effects will the genetic knowledge derived from the Human Genome Project have on access to insurance, jobs, and civil rights?
- What long-term impacts will the technologies associated with genetic research have upon society, public health, the global gene pool, and bioengineering?
- Is the Human Genome Initiative the most effective and equitable use of scarce scientific resources?

The Department of Energy (DoE)-supported working group, "The Social Costs and Medical Benefits of Human Genetic Information," included 13 students and 8 "seniors" (experts/resource people), each representing a range of international and academic backgrounds, experiences and perspectives. Among the "senior" participants of the working group, for example, were genetic researchers, genetic counselors, physicians, and ethicists. In advance of the International Conference, the student participants prepared original research papers on ethics and the use of genetic information, which then served as the focal point for discussion in the working group throughout the conference week. Upon the conclusion of the week, students presented a working group statement to the full conference community, including policy recommendations for the scientific community.

Support from DoE is also being used to support the compilation of a student-authored, expert-edited educational resource, the *Global Issues Guidebook.* The *Guidebook* draws upon outstanding papers prepared for the International Conference, and will feature one section on the ethical, legal and social impacts of the Human Genome Project. Like the International Conference, the *Guidebook* will highlight the interdisciplinary nature of global issues, explore creative solutions for their resolution, and evaluate their impact on both society and the individual. Information contained in the *Guidebook* will be presented in a format which can be used as a guide in classroom or student group discussions. Following its publication in early 1995, over 500 copies will be distributed across the U.S. and internationally, in loose-leaf binder and electronic formats to allow for continual updating.

1638 R Street, NW Suite 32
Washington, D.C. 20009
Tel: 202 328-6555
Fax: 202 797-4664

The Human Genome: Science and the Social Consequences

Charles Carlson, Randy Comer, Randall Fontes, Micah Garb, Glenn Gutleben, Anne Jennings, Mary Miller, Frank Millero, III

Exploratorium 3601 Lyon St, San Francisco, CA 94123

The Exploratorium is embarking on a long-term commitment to increase public awareness of the Human Genome Project (HGP), the basic science of genetics, and related ethical, legal, and social issues. Our comprehensive multi-year proposal includes developing new exhibits, designing demonstrations, and creating a Genetics Pathway. In addition we are planning a lecture series directed specifically at social and ethical issues raised by the HGP and genetic technology. This lecture series will serve both as a public forum for discussion as well as a means of creating video records for inclusion in a free-standing exhibit. All aspects of the project involve collaborations with regional biotechnology firms, universities, and museums.

To date we have created twelve interactive genetics exhibits which are currently on display or in prototype form (*Blood Typing, Cells, Dancing DNA, DNA: The Master Molecule, Genetic Characteristics, Marching Bands, Molecular Library, Protein Production Line, Sickling Cells, Simulated Movement, Tree of Life, Zebrafish Development*). The new exhibits under development include: *Challenges and Choices, Fruit Flies, Fungi, A Graphical Presentation of the Human Genome, Interactive History of the Automated Sequencer, Musical Mutants, Panning for DNA demonstration,* and *Reading DNA.*

The final collection of exhibits and programs will address the concepts of DNA as the molecule of heredity, mutation, variation, and the relationship between genes and proteins. It will also focus on the HGP as an example of how basic research can result in practical applications and raise important social and ethical questions. We are meeting the challenge of presenting this complex and often foreign subject by exposing visitors to the phenomenon itself, through living and non-living biological experiments as well as mechanical and computer models. The exhibits will be enjoyed by more than 625, 000 visitors annually, including 70,000 students on field trips and 500 teachers trained by the Teacher Institute and the School in the Exploratorium.

This work was funded by the DOE Genome Program (DE-FG03-93ER61583)

Human Genetics for Nonscientists:

Practical Workshops for Policy Makers and Opinion Leaders

Mark Bloom[1], David Micklos[1], and Jan Witkowski[2]

[1]DNA Learning Center, [2]Banbury Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724.

Communications and sociological research suggests that information campaigns have an indirect effect on public opinions and behavior. Information often appears to flows to the public in two steps. First, "opinion leaders" assess information from a variety of sources and form attitudes about issues. These well-informed individuals, in turn, influence the opinions and behaviors of people around them. Thus, information campaigns aimed at a relatively small number of opinion leaders may provide a cost-effective means to reach large segments of the public.

With this model in mind, we designed a workshop for influential nonscientists who interface with human genetics research and society. Most participants come to the workshop with extensive "book knowledge" on genetics, but only fragmentary understanding of the technology upon which modern genetic analysis is based. Thus, the workshop aims to deepen insight into the research process and fill in gaps in participant's understanding. The workshop is composed of three components that juxtapose the theory, practice, applications, and implications of human gene manipulation. Concept seminars, presented by project staff, introduce key principles that underpin human genome analysis, including the molecular basis of inheritance, gene mapping and cloning, and DNA diagnosis. These topics were made tangible through laboratory sessions where participants construct a restriction map and make their own DNA fingerprints using PCR. Feature seminars, presented by genetics professionals, provide first-person insight into the research process and the ethical dilemmas of human genetics research.

Four workshops supported by the first round of funding drew together an eclectic group of administrators and communicators from federal and state governments, genetic support groups, foundations, associations, the media, science education, law, and ethics. With second-round funding in 1994, we targeted a new group of opinion leaders with great potential to influence the medical community: education directors at hospitals. The initial workshop in this series, held in April, drew participants from 14 hospitals in the New York metropolitan area. The second workshop will draw medical education directors representing many regions of the country.

Student Sequencing Program

Leroy Hood, Maynard Olson: Department of Molecular Biotechnology, University of Washington, FJ-20, Seattle, WA 91895

This program provides high school teachers with the training, equipment and support to lead students through the exercise of sequencing small portions of the human genome. Results of student analysis will be placed in appropriate national scientific data bases, alongside data contributed by professional scientists. Scientists at the UW's Department of Molecular Biotechnology work with high school biology teachers and genetic counselors to design hands-on laboratories suitable for schools and curricular materials which address social and ethical issues arising from the genome project. Funded by the Department of Energy, the program is being launched in Seattle in 1993-94 and should be available nationally to schools by 1995-96. Program chairs: Maynard Olson, Ph.D., Leroy Hood, Ph.D., and Maureen Munn, Ph.D.

Involvement of High School Students in the Sequencing of the Human Genome

Maureen M. Munn, Maynard V. Olson and Leroy Hood

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

We are developing a program that allows high school students to participate directly in the mapping and sequencing of the human genome. This program provides high school teachers with the proper training, equipment, and support to lead their students through the exercise of sequencing small portions of human DNA. Development of this project has been facilitated by our collaboration with six Washington teachers, who advise on experimental design and curriculum development and pilot the program in their classrooms. Besides the hands-on labs, students and their teachers discuss the social and ethical issues emerging as the human genome is mapped. The development of the ELSI curriculum is guided by the expertise of geneticist/ medical ethicist, Sharon Durphy, and two genetic counselors, Ann Spencer and Deborah Doyle.

Currently, we are testing two experimental modules, DNA synthesis (an introduction to DNA replication and the techniques used to study it) and DNA sequencing. To allow collaboration among the participating classrooms, we are shotgun sequencing a small human gene, with each classroom contributing the sequence of two or more fragments. This collaborative approach emphasizes to the students their responsibility for careful data analysis.

All laboratory procedures are carried out in the classrooms, using equipment and supplies provided by the program. In the first class period, the students prepare their sequencing reactions, using the Sanger method and single-stranded DNA templates. The following day, they resolve their DNA fragments by denaturing gel electrophoresis on a benchtop apparatus. To avoid the use of radioactivity, the DNA is transferred from the gel to a nylon membrane and stained by an enzymatic reaction that creates a colored image. The resulting sequencing ladders are analyzed by the students. The students then enter their data into a computer data base that allows them to align and compare similar sequences and perform simple analyses such as checking for open reading frames and common repetitive sequences. In addition, they are able to scan the existing sequence data bases using Blast and the e-mail server. More advanced classes will help to assemble the fragments and determine the chromosomal location of our gene.

While we hope the human genome sequencing experience will interest some students in science careers, a broader goal is to encourage high school students to think constructively and creatively about the implications of scientific findings so that the coming generation of adults will make judicious decisions affecting public policies.

This work is sponsored by the U.S Department of Energy under grant No. DE-FG06-94ER61798.

DNA Banking and DNA Databanking by Academic and Commercial Laboratories

J.E. McEwen [1,2] and P.R. Reilly,[1].

[1] Eunice Kennedy Shriver Center, Division of Social Science, Ethics, and Law, Waltham, Massachusetts;
[2] Boston College Law School, Newton, Massachusetts.

The advent of DNA-based testing is giving rise to DNA banking (the long-term storage of cells, transformed cell lines, or extracted DNA for subsequent retrieval and analysis) and DNA databanking (the indefinite storage of information derived from DNA analysis, such as the linkage profiles of persons at risk for a particular genetic disease)[1]. The large scale acquisition and storage of DNA and DNA data has important implications for the privacy rights of individuals [2,3].

A survey of 148 academically based and commercial DNA diagnostic laboratories was conducted to determine: (1) the extent of their DNA banking activities; (2) their policies and experiences regarding access to DNA samples and data; (3) the quality assurance measures they employ; and (4) whether they have written policies and/or depositor's agreements addressing specific issues. These issues include: (1) who may have access to DNA samples and data (*e.g.*, the depositor; his or her physician, spouse/partner, adult children, or other relatives; other third parties); (2) whether scientists may have access to anonymous samples or data for research use; (3) whether they have plans to contact depositors or retest samples if improved tests for a disorder become available; (4) disposition of samples at the end of the contract period, if the laboratory ceases operations, if storage fees are unpaid, or after a death or divorce; (5) the consequence of unauthorized release, loss, or accidental destruction of samples; and (6) whether depositors may share in profits from the commercialization of tests or treatments developed in part from studies of stored DNA.

The results suggest that many laboratories are banking DNA, that many have already amassed a large number of samples, and that a significant number plan to further develop DNA banking as a laboratory service over the next two years. Few laboratories have developed written policies governing DNA banking--and fewer still have drafted documents that define the rights and obligations of the parties. There may be a need for increased regulation of DNA banking and DNA databanking and for better defined policies with respect to protecting individual privacy.

This work was funded by the DOE Genome Program (DE-FG02-91ER61237, P.R. Reilly, P.I.) MCH grant MCJ-259151-02-0, and Department of Mental Retardation of the Commonwealth of Massachusetts Contract 1000-10003-SC.

[1] Reilly, P.R. (1992) DNA banking. *Am. J. Hum. Genet.* **51**, 1169-1170.
[2] American Society of Human Genetics, Ad Hoc Committee on DNA Technology. (1988) DNA banking and DNA analysis: Points to consider. *Am. J. Hum. Genet* **42**, 781-83.
[3] Annas, G.J. (1993) Privacy rules for DNA databanks. *JAMA* **270**, 2346-2350

Assessing Genetic Risks: Implications for Health and Social Policy

Lori B. Andrews,[1] Jane E. Fullarton,[2] Neil A. Holtzman,[3] and Arno G. Motulsky[4].

[1] Fellow, American Bar Foundation, Chicago, IL. [2] Study Director, Assessing Genetic Risks, Institute of Medicine, Washington, D.C. [3] Professor of Pediatrics, Health Policy, and Management and Epidemiology, The Johns Hopkins University Hospital, Baltimore, MD. [4] Chair, Committee on Assessing Genetic Risks; Professor of Medicine and Genetics, Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA.

This study of the scientific, ethical, legal, and social issues implicit in the field of genetic diagnosis, testing, and screening was supported jointly by the National Center for Human Genome Research at the National Institutes of Health and the Department of Energy's Health Effects and Life Sciences Research Office. Supplemental funding was also provided by the Markey Charitable Trust and the Institute of Medicine.

The committee took its starting point from the advice of the 1975 National Academy of Sciences study, *Genetic Screening: Programs, Principles, and Research.*[1] The recommendations from that report, written almost 20 years ago, remain valid today. The committee reaffirmed the sentiments expressed in the 1975 report and updated and broadened their application for the 1990s and beyond.

The committee posed its recommendations in terms of general principles for the evaluation of expanded genetic diagnosis, testing, and screening. Although these recommendations reflect what is known today, and what experts foresee for the next few years, the committee had no crystal ball and, therefore, tried to develop criteria and to suggest processes for assessing when new tests are ready for pilot introduction and for widespread application in the population.

The committee's fundamental ethical principles include voluntariness, informed consent, and confidentiality, which in turn derive from respect for autonomy, equity, and privacy. Other committee principles described in this report include: the necessity of high-quality tests (of high specificity and sensitivity) performed with the highest level of proficiency and interpreted correctly; and conveying information to clients–both before and after testing–in an easily understood manner through genetic education and counseling that is relevant to the needs and concerns of the client. These principles are the absolute foundation of genetic testing.

It is the view of the committee that, until benefits and risks have been defined, genetic testing and screening programs remain a form of human investigation. Therefore, routine use of tests should be preceded by pilot studies that demonstrate their safety and effectiveness. Standard safeguards should be applied in conducting these pilot studies, and independent review of the pilot studies should be conducted to determine whether the test should be offered clinically. Publicly supported population-based screening programs are justified only for disorders of significant severity, impact, frequency, and distribution, and when there is consensus that the available interventions warrant the expenditure of funds.

---

1 National Academy of Sciences (1975) *Genetic Screening: Programs, Principles and Research.* Committee for the Study of Inborn Errors of Metabolism. Washington, D.C.:NAS.

Impact of Human Genome-Derived Technology on Genetic Testing, Screening and Counseling: Cultural, Ethical and Legal Issues.

Ralph W. Trottier[1], Lee A. Crandall[2], Faye Cobb Hodgin[1], Mwalimu Imara[1], Ray Moseley[3], and David Phoenix[4]

[1]Morehouse School of Medicine, Atlanta, Georgia, [2]University of Illinois, Champagne, Illinois, [3]University of Florida, Gainesville, Florida, [4]Clemson University, Clemson, South Carolina.

The primary focus of this research involves a detailed analysis comparing and contrasting genetic services programs and delivery systems in the states of Georgia and Florida. Description of findings include a profile of personnel and their responsibilities as functionaries in the delivery of public sector genetic services. An important basic issue considered in this study and which must be addressed through research in a more global context involves the extent of awareness of the potential impact anticipated to result from Human Genome Project research and the degree to which public sector programs are preparing to incorporate, in an equitable and just way, new technological advances in genetic diagnostics and treatment.

The project is conducted through on-site interviews and correspondence with key professionals involved in genetic out-reach service programs. Data and information collected form the fabric of a rich description regarding geographic, organizational and social perspectives of public sector genetics. Newborn screening (NBS) legislation was analyzed in detail to determine whether there is a clear and currently relevant intent and purpose in fulfilling expressed state's interests, whether provisions to educate the public are included, the extent to which privacy (confidentiality-protecting) safeguards have been included, and how service provision is documented and monitored. Administration of genetic services delivery systems was examined from organizational and operational viewpoints. On-site field research provided the opportunity to gain firsthand understanding of operational functions with respect to population demographics on a regional basis, roles of key personnel at the health district/subdistrict level and identification of potential gaps in delivery of and access to services. Leading roles of academically-based tertiary care centers in both states are compared and contrasted.

Early Intervention Programs (EIP) resulting from the Education for the Handicapped Act Amendments of 1986 are now in early stages of implementation. The EIP approach involves multi-agency and family interactions in forming and evaluating goals and objectives designed to facilitate a smooth transition of developmentally delayed children into their school years. The Georgia and Florida EIP were analyzed to determine the extent to which they interact with genetic services in the public sector. Research identifying genetic factors in learning disabilities and causes of mental retardation will be of keen interest to EIP. Whether EIP will lead to setting new standards for incorporating additional NBS tests (e.g., for fragile X) remains to be determined but would appear to be a reasonable possibility. Recommendations for future ELSI research in defining standards for public sector genetic services delivery and policy options conclude this presentation.

This research is funded by DOE/NIH co-sponsorship under the administration of the DOE Genome Program, Grant #DE-FG0292ER61396, Ralph W. Trottier, P.I., Lee A. Crandall, Co-P.I.

Social Science Studies of Privacy and Their Implications for Policy Choices in the Uses of Genetic Information

Alan F. Westin [1]

[1] Center for Social and Legal Research, Suite 414, Two University Plaza, Hackensack, N.J. 07601

This DOE-funded Project updates the theoretical and empirical study of privacy and of new-technology challenges to democratic privacy balances published by Westin in 1967 [1].

The project has collected and examined social science writings and studies since 1967 in light of Westin's generally accepted concepts of the four primary states or conditions of privacy (solitude, intimacy, anonymity, and reserve); four central functions of privacy in modern, democratic society (personal autonomy, emotional release, self-evaluation, and limited and protected communication); the balancing of privacy, disclosure, and surveillance as competing but necessary social values and processes; the impacts of physical, psychological, and data surveillance technologies on pre-1960's balances of privacy in social life and under legal rules; the key elements of a technology-assessment for privacy impacts; an analytic framework for contemporary balancing of privacy, disclosure, and surveillance interests; and a legal, political, and social program of privacy-protection measures.

The Principal Investigator (Westin) has been supported by an expert, inter-disciplinary Senior Advisory Committee (David Flaherty, Lance Hoffman, Neil Holtzman, Kenneth Laudon, Dorothy Nelkin, Kimberly Quaid, and Philip Reilly).

As a central resource, the Project has commissioned reviews of the social science literature relating to privacy since 1967 in anthropology, sociology, political science, economics, psychology and psychiatry, jurisprudence, and empirical legal studies. The reviewers were Colin Bennett, Gail Geller, William Regenold, and Carol Traver. Their drafts were critiqued by the PI and Senior Advisors; read by outside experts in the fields; and are now in final rewrites.

In addition, a comprehensive review of public and leader opinion surveys on privacy since 1967 was conducted and written up by the PI and two research assistants.

Using the original Westin framework and the social science studies since 1967, the Project is preparing a monograph presenting a Privacy-Impact Assessment of the new scientific discoveries and applications emerging from the Human Genome Project (HGP). This assessment describes the uses of genetic science and genetic concepts (and their societal impacts) in the 19th and 20th centuries, prior to the discovery of DNA; the period from DNA discovery to the HGP; and the past decade of activity spurred by the HGP and applications of its findings.

The assessment then lays out the interest groups and competing positions relating to uses of genetic information, genetic testing, and genetic data banks for both medical and social uses of genetic data; compares the ideas and the emerging politics of genetic-information applications to the treatment of information-technology applications (computers and telecommunications) between the mid-1960's and the present; examines the application of fair information practices or privacy-protection standards to the genetic-information field; provides a near and long term forecast of likely applications and uses of genetic information in the U.S.; and poses a series of policy choices relating to privacy and genetic information, with a discussion of their implications.

The Project will produce three products: an "assessment" monograph on privacy and the uses of genetic information; an edited volume containing each of the four social science review papers and the survey-research review; and a definitive bibliography of social science works on privacy.

[1] *Privacy and Freedom* (New York: Atheneum, 1967)

Pathways to Genetic Screening: Molecular Genetics Meets the High-risk Family

By Troy Duster[1] and Diane Beeson[1,2]

This project examines the social processes that occur as families at risk for two of the most common autosomal recessive diseases, sickle cell disease (SC) and cystic fibrosis (CF), encounter genetic testing. Since each of these diseases is found primarily in a different ethnic/racial group (CF in European Americans and SS in African Americans), each with differing economic and political resources, this research will clarify the role of culture in integrating genetic knowledge and interventions into lived experience.

Data are drawn from interviews with members of families in which a gene for CF or SC has been identified. Data collection consists primarily of focused interviews with approximately 300 family members exploring constructions of the following topics: direct personal experience with genetic disorders; the meaning of the disorder for the life of affected individuals and family members; health care and insurance issues; prevention and testing; family communication; communication beyond the family.

A variety of patterns of response to these issues has been identified, including the following:

- Women are the primary communicators of genetic knowledge within the family.
- Communication occurs primarily at the time of diagnosis of a child with the disorder and during medical crises.
- Families frequently establish restrictive communication rules related to genetic issues that inhibit involvement in prevention.
- Testing is rarely sought by high-risk family members of either group and at this time occurs primarily as a result of provider initiative.
- Genotype, even when known, is rarely a factor in choosing a partner or in reproductive decision making.
- The most widely considered and accepted form of genetic testing is prenatal testing.
- Acceptance or interest in prenatal testing is unrelated to willingness to abort an affected fetus.
- Men in all socioeconomic groups are more likely to deny genetic risk of their contribution to a child's disorder.
- Grandparents exhibit high levels of distress related to their potential genetic contribution to their grandchild's disorder.
- Although African-Americans are more likely to have been tested for carrier status, those who have been tested are more reluctant to integrate this information into reproductive decision making than the predominantly European-American known CF carriers, and more critical of biomedical approaches to reproduction.

These patterns are being confirmed and amplified as the research concludes its second of three years. In spite of significant differences, the two cultural groups under study share a number of patterns that contrast sharply with cultural assumptions of the social world of molecular genetics. One key theme that emerges from our ongoing research with these families is that the social worlds of molecular genetics and high-risk families are on a potential collision course on the matter of genetic testing, due to differences in the ways in which genetic information is framed in each setting. These differences have important implications for medical practice, health-seeking behavior, intra-familial communication and health policy.

[1] Institute for the Study of Social Change, 2420 Bowditch Street, University of California, Berkeley, CA 94720
[2] Dept. of Sociology and Social Services, California State University, Hayward, CA 94542

Genetic Privacy & Discrimination: What the Federal and State Governments Are Doing, Aren't Doing, and Should Be Doing

Michael Yesley
Los Alamos National Laboratory; Los Alamos, NM 87545
505/667-3766, Fax: /665-4424, Internet: *msy@lanl.gov*

Advances in genetics are making it possible to generate increasing amounts of sensitive information about individuals. Although the potential benefits of this knowledge are substantial, there are also potential harms from undesired disclosures and unfair uses of genetic information. To protect against these harms, Congress and many state legislatures are beginning to consider and adopt laws that prohibit various practices involving genetic information. Many of these measures are too narrow, too broad, or misdirected, due to jurisdictional limitations and the novelty and complexity of the subject matter. Furthermore, the conflicting demands of those who would be affected by the measures has slowed their adoption or narrowed their reach.

We are reviewing current and proposed laws and regulations aimed at protecting genetic privacy and barring genetic discrimination, to determine what are the issues and most appropriate solutions. In some cases the solutions will be unclear: society must decide if fairness requires equal treatment of genetic difference.

Protecting Genetic Privacy by Regulating the Collection, Analysis, Use, and Storage of DNA and Information Obtained from DNA Analysis

George J. Annas, Leonard H. Glantz, and Patricia A. Roche

Health Law Department, Boston University School of Public Health, Boston, MA

Both DNA samples and information derived from genetic analysis of DNA samples are maintained in various facilities which may be termed "DNA databanks." The use of both the samples and the information obtained from them could be regulated by attempting to regulate the activities of such banks. Our analysis, however, has concluded that there is such a variety of what may be seen as DNA databanks that it makes more sense to directly regulate the activities of all people who collect, analyze, store and use information obtained from DNA samples. Since the activities involved often cross state lines, and uniform regulations are desirable, federal regulation is to be greatly preferred to state or local regulation. We began this research project in early 1993 by focusing on six questions: (1) how is genetic information different from medical information?; (2) when should it be permissible to collect DNA samples from individuals?; (3) when is consent required to store DNA samples and information derived from them?; (4) who owns genetic material and genetic information?; (5) who should have access to stored DNA and genetic information obtained from it; and (6) should time limits be set on the storage of DNA? [1]

The answers to these six questions led us to others, and ultimately led us to draft a "Genetic Privacy Act" the purpose of which is to provide legal rules for the collection, analysis, storage, and use of DNA samples and the genetic information obtained from them. [2] The Act, which complements current federal proposals to protect medical information but moves beyond them, establishes rules to protect genetic privacy by requiring explicit authorization for the collection of DNA samples for genetic analysis, limiting the uses that can be made of DNA samples and the genetic information obtained from them, and setting forth penalties for violation of the rules. The full text of the Act, together with a detailed explanation of its provisions, will be available.

This research was funded by the Office of Health and Environmental Research of the DOE (DE-FG02-93ER61626).

[1] Annas GJ. Privacy rules for DNA databanks: protecting coded 'future diaries' JAMA 1993; 270:2346-50.

[2] Protecting genetic privacy (interview with George Annas) Trial Aug 1994; 30(8):43-46.

The Role of Patents in Technology Transfer in the Human Genome Project

Rebecca S. Eisenberg, J.D.

University of Michigan Law School

Abstract

The Human Genome Project provides government funds for generating vast amounts of information in the hope that the information will ultimately be put to use in developing new products and processes for the diagnosis and treatment of human disease. Much of this information is generated in government and university laboratories that cannot undertake the downstream research and development necessary to translate basic research discoveries into commercial products. Technology transfer to the private sector is thus a prerequisite to the development of genome-related products, but achieving effective technology transfer in such a project is a matter of some complexity.

Federal policy since 1980 has reflected a presumption that the most effective way to promote technology transfer and commercial development in the private sector of discoveries made in the course of government-sponsored research is to patent them. Yet the reactions of industry trade groups to the filing of patent applications by the National Institutes of Health on thousands of partial complementary DNA (cDNA) sequences of unknown function that were identified in government laboratories suggest that the prevailing pro-patent policy may oversimplify the complexities involved in technology transfer. This controversy provides a useful focal point for considering when the results of government-sponsored research should be patented and when they should be dedicated to the public domain.

This page intentionally left blank.

Infrastructure

Human Genome Management Information System

Betty K. Mansfield, Anne E. Adamson, Denise K. Casey, K. Alicia Davidson, Sheryl A. Martin, **John S. Wassom**, Judy M. Wyrick, and Laura N. Yust

Biomedical and Environmental Information Analysis Section; Health and Science Research Division; Oak Ridge National Laboratory; 1060 Commerce Park, MS 6480; Oak Ridge, TN 37830
615/576-6669, Fax: /574-9888, Internet: *bkq@ornl.gov*

The Human Genome Management Information System (HGMIS), which was inaugurated in 1989, provides technical communication and information services for the DOE Office of Health and Environmental Research (OHER) Human Genome Program Task Group. HGMIS is charged with (1) helping to communicate genome-related matters and research to contractors, grantees, and other publications; (2) serving as a clearinghouse for information on the U.S. genome project; and (3) reducing duplication of research efforts by providing a forum for information exchange among Human Genome Project investigators worldwide. HGMIS also occasionally compiles and organizes administrative data for DOE by preparing reports and meeting minutes, conducting information searches, writing and editing, and assisting DOE staff at meetings.

To fulfill its communication goals, HGMIS publishes the bimonthly newsletter *Human Genome News (HGN)*, cosponsored by OHER and the NIH National Center for Human Genome Research. HGMIS also produces a primer on molecular genetics and reports on the DOE Human Genome Program, contractor-grantee workshops, and other related subjects; and makes its publications available via World Wide Web and Gopher (*gopher.gdb.org*). The newsletter and several reports have been recognized with awards by the Society for Technical Communication, East Tennessee Chapter.

HGN features technical and general interest articles, meeting reports, national and international project news, features on informatics and resources for facilitating research, genome event and training calendars, and grant and fellowship announcements. Some 9000 newsletter subscribers include genome and basic researchers at national laboratories, universities, and other research institutions; professors and teachers; industry representatives; legal personnel; ethicists; students; genetic counselors; physicians; science writers; and other interested individuals. To conserve resources and increase cost-effectiveness, HGMIS uses bulk mailing and constantly updates and revises the mailing list, dropping the names of those who fail to respond to subscription-renewal notices.

HGN also serves as a primary source for discipline-specific publications that extract or reprint information on the Human Genome Project. Some of these are *Bioinformatics;* the ethics journal *Eubios;* several university publications; and newsletters of genome centers, state biotechnology organizations (*BT Catalyst),* chromosome-specific support groups (*The Chromosome 18 Communique'*), high school biology teachers (*The Genetic Messenger)*, Student Pugwash USA (*Tough Questions*), and the National Society of Genetic Counselors (*Perspectives in Genetic Counseling*). The U.K. publication *Gnome News* regularly reprints the GDB Forum pages from *HGN* for its readers.

By 1994 over half the subscribers had requested some type of information or document, including program and workshop reports, the DOE-NIH 5-year plan, and the DOE informatics summit report. In addition, numerous copies of full or partial documents have been distributed for educational purposes.

An outstanding example is the *Primer on Molecular Genetics*, expanded and revised by HGMIS from an earlier DOE document. Originally an appendix to the program reports, the separate primer has been reprinted several times because of its demand as a handout for genome centers and as a resource for teachers, genetic counselors, and educational organizations. These organizations include high schools; universities; and medical schools, which use the primer in their continuing-education curriculum. More than 28,500 hard copies have been distributed without charge, and the primer is available on the World Wide Web (URL: *http://www.gdb.org/hopkins.html*).

In addition to their publishing efforts, HGMIS staff answer questions about the Human Genome Project and supply general information by telephone, fax, and e-mail; the most frequent inquirers are graduate students, researchers, medical professionals, and private companies and individuals. For example, experts in biotechnology and other industries use HGMIS as a source of data for identifying goods and services that might be useful to genome researchers. HGMIS staff also have the opportunity to exchange ideas and suggestions with investigators, industry representatives, and others when they display the DOE Human Genome Project traveling exhibit at scientific conferences and genome-related meetings.

HGMIS invites comments and suggestions about its documents and services, which are available upon request and without charge.

This work is sponsored by the Office of Health and Environmental Research, U.S. Department of Energy, under contract No. DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

The Human Genome Distinguished Postdoctoral Fellowships

Linda Holmes and **Al Wohlpart**
Science/Engineering Education Division; Oak Ridge Institute for Science and Education; Oak Ridge, TN 37831-0117
615/576-9934, Fax: /241-5219

The Human Genome Distinguished Postdoctoral Fellowships were initiated in FY 1991 by the DOE Office of Health and Environmental Research (OHER) to support genome research by recent doctoral degree recipients. Fellowships of up to 2 years are tenable at any DOE, university, or private laboratory if the proposed advisor at that laboratory receives at least $150,000 per year in support from OHER for Human Genome Program research. Fellows earn stipends of $37,500 the first year and $40,500 the second. Eligible applicants must be U.S. citizens or permanent resident aliens and must have received their doctoral degrees within 3 years of the earliest possible starting date, which is May 1 of the appointment year.

The Oak Ridge Institute for Science and Education (ORISE), administrator of the fellowships, prepares and distributes program literature to universities and laboratories across the country, accepts applications and convenes a panel to make award recommendations, and issues stipend checks to fellows. The review panel identifies finalists from which DOE chooses the award winners. Up to six awards will be made each year. Deadline for the FY 1995 fellowship cycle is February 1, 1995. For more information or an application packet, contact Linda Holmes at the Science/Engineering Education Division; ORISE, Rm. 45; P.O. Box 117; Oak Ridge, TN 37831-0117 (615/576-9934, Fax: /241-5219).

Support of Human Genome Program Proposal Reviews

Walter Williams, James Wright, and Bryan Coulter
Science/Engineering Education Division; Oak Ridge Institute for Science and Education; Oak Ridge, TN 37831-0117
Williams: 615/576-4811, Fax: /241-2727, Internet: *williamw@orau.gov*
Coulter: 615/574-8633, Fax: /241-2727, Internet: *coulterb@orau.gov*

The Oak Ridge Institute for Science and Education (ORISE), operated by Oak Ridge Associated Universities, provides assistance to the DOE Office of Health and Environmental Research in the technical review of proposals submitted in response to solicitations by the DOE Human Genome Program. ORISE staff members create and maintain a database containing all proposal information, including abstracts, relevant names and addresses, and budget data. This information is compiled and presented to proposal reviewers. Before review meetings ORISE staff members make appropriate hotel and meeting arrangements, provide each reviewer with proposal copies, and coordinate reviewer travel and honoraria payment. Other support includes assistance with program advertising and preparation of reviewer comments following each review.

Appendices

This page intentionally left blank.

Appendix A: Author Index

First authors are in **bold.**

Adams, Mark D. 11, 83, 150
Adamson, Anne E. 202
Agarwal, Pankaj 105
Aggarwal, A. 87, 89
Aksenov, N.D. 13
Albin, Michael 134
Alegria-Hartman, Michelle 73
Alizadeh, Farid 113
Allen, Susan A. 42
Allison, D.P. 16
Allman, Steve L. 139
Altherr, Michael R. 1, 28, 38, 39, 41, 159
Ambrose, W. Patrick 160
Amemiya, Chris 49
Andrews, Lori B. 195
Ankener, Wendy 71
Annas, George J. 200
Apostolou, S. 36, 37
Arcot, Santosh S. 72
Armstrong, Rob 116
Arvestad, L. 99
Ashworth, Linda K. 29, 30, 43–46, 48, 82
Athwal, Raghbir S. 23
Bailey, Susan M. 12
Bailis, Julie M. 58
Baker, Diane 188
Balch, Joe 43, 134, 180
Balding, David J. 27
Ballabio, A. 56
Banks, Theresa A. 32
Barber, W.M. 81
Barker, David 130
Barsky, Victor E. 124
Bartholomew, Christopher 32
Bartosiewicz, Matt 130
Bashirzadeh, Romina 4
Bashkin, John 130
Basiji, David 183
Batzer, Mark A. 43, 72, 73
Bay, Sue 131
Bazan, Hernan 72
Beattie, Kenneth 126
Beauheim, C. 81
Becker, K. 11
Beeson, Diane 198
Benson, Scott C. 133
Ben-Shachar, Ofer 96
Bergmann, Anne 26, 44
Berka, Jan 132
Best, Elaine 115
Beugelsdijk, Tony J. 168
Birren, Bruce 19, 20
Blackwell, Tom 121
Blair, Patrick J. 32
Bloom, Mark 191
Boehrer, Denise 21
Borsani, G. 56
Borsody, Elizabeth 4, 74
Boville, Brian 156
Boysen, Cecilie 147
Bradley, J.-C. 123
Brandriff, Brigitte F. 30, 43, 44, 47
Branscomb, Elbert 31, 43, 86
Brennan, Thomas 158
Brewer, Larry 134, 180
Bridgers, M.A. 81
Briley, J. David 165, 166
Britt, Phillip F. 144
Brody, Tom 63
Bronstein, Irena 68
Brookman, Kerry W. 154
Broude, Natalia 34
Brown, Gilbert M. 138
Brown, Nancy C. 25
Bruce, David C. 27, 39, 41
Bruce, J.A. 143
Brumley, Jr., Robert L. 135, 156
Brunn, Christopher W. 78
Bruno, William J. 27, 39, 41, 99, 107
Buchanan, Michelle V. 144
Buckingham, Judy M. 25, 159
Bukanov, Nikolay 34
Buley, Donna 103
Bult, Carol 83, 150
Buneman, Peter 92
Burbee, David 58

Burgin, Matt 45, 46, 48, 50
Burke, John 104
Burton, Jillian 79
Butler, Laurie 68
Butler, William F. 125
Buxton, Eric C. 135
C. elegans Genome Consortium 149
Callen, D.F. 36–39, 41
Campbell, Evelyn C. 25
Campbell, Mary L. 25
Cantor, Charles R. 34
Capehart, Joe 22
Carlson, Charles 190
Carlson, Christopher 71
Carrano, Anthony V. 26, 42–48, 134, 154, 155, 180
Carson, Steve 132
Cartwright, Peter 153
Carver, Ethan 30
Casey, Denise K. 202
Chandra, Abha 34
Chang, Huan-Tsung 129
Chasteen, L.A. 53
Chen, C.H. Winston 139
Chen, Chira 22
Chen, Dan 156
Chen, I-Min A. 93
Chen, Jerome T. 168
Chen, X.N. 11
Cheney, Brian 45
Cheng, Jan-Fang 24, 54, 55, 62, 145
Cheng, Xueheng 142, 143
Cherry, Joshua L. 148, 153
Chinault, A.C. 56
Chipperfield, Michael A. 77
Chou, Chau-Wen 140
Chou, Hugh 105
Christensen, Mari 44
Churchill, Gary A. 120
Churukian, Allan C. 58
Cinkosky, Michael J. 79, 81, 109
Clark, Stephen P. 58
Clark, Steven M. 133
Clingan, R.L. 56
Cobbs, Archie 80
Collins, Colin 10, 14, 16, 62, 66
Collins, Debra L. 187
Comer, Randy 190
Cooper, Michael H. 116
Copeland, Alex 155
Coulson, Larry 131
Coulter, Bryan 204
Cox, D.R. 35
Cram, L.S. 13, 178
Crandall, Lee A. 196
Crkvenjakov, Radomir 128
Crowley, David 79
Cuticchia, A. Jamie 76, 77
Cytron, Ron 105
Dalan, A. Burak 34
Daneshvar, L. 10, 66
Danganan, Linda 44
Davidson, Courtney 134, 180
Davidson, K. Alicia 202
Davidson, Susan 92
Davis, Cheryl A. 146
Davis, Ronald 158
Davison, Daniel 104
Davy, D. 87, 89
de Fatima Bonaldo, Maria 8
de Jesus, Maria T. 42, 155
de Jong, Pieter J. 22
de Kanter, Mark 15
de la Chapelle, A. 35
Deaven, Larry L. 25, 39, 41, 52, 53, 177
Deininger, Prescott L. 72, 73
Dell, Cheryl 188
Denison, Karen 2, 3, 11
DeVries, Diana 188
Di Sera, Leonard 148
Diggle, Karin 58
Doggett, Norman A. 27, 36, 37, 39–41
Dogruel, David 140
Doktycz, Mitchel J. 126, 138
Douthart, Richard J. 122
Dovichi, Norman J. 131
Doyle, Johannah 30
Drew, P. 11
Dubois, JoAnn 4
Duesing, L.A. 39–41
Dunn, Diane 148, 153
Dunn, John J. 161–163
Dunn, William 30
Duster, Troy 198
Dyer, Joshua P. 175
Ealey, Katie 166

Edmonds, Charles G. 143
Eeckman, F. 87
Efstratiadis, Argiris 8
Eichler, E.E. 56, 57
Einstein, J. Ralph 103
Eisenberg, Rebecca S. 201
Elbert, Jeffrey E. 138
Elliott, Jeffrey M. 45, 46
Elliott, John 131
Espinosa-Lujan, Ada 79
Esposito, Richard 183
Evans, Glen A. 58, 125
Evans, Jim 9
Eveleth, Gerald 26, 181
Fader, Betsy 189
Fallon, Lara 123
Falls, Kathy 4
Fasman, Kenneth H. 76, 78
Fawcett, John J. 25
Feder, M. 53
Ferguson, F. Mark 148
Ferraro, G.B. 56
Fertitta, Anne 44
Fickett, James W. 79, 81, 109
Fields, Chris 83
Firulli, B.A. 56
Fleischmann, Robert D. 150
Fleming, T. 87
Florentiev, Vladimir L. 124
Fodor, Stephen P.A. 75, 127
Fontes, Randall 190
Foote, Robert S. 126
Ford, Amanda A. 1, 39, 41, 159
Foret, Frantisek 132
Francisco, Todd 156
Franco, B. 56
Frengen, Eirik 22
Friedman, Cynthia 9
Froula, Jeff 62
Fullarton, Jane E. 195
Fuller, Carl W. 137
Gaasterland, Terry 91
Gale, David C. 142
Garb, Micah 190
Garcia, Emilio 43, 45, 46
Garner, Harold R. 58, 125
Garnes, Jeffrey A. 26, 43, 51
Gatewood, Joe M. 2, 3, 11, 52, 98, 115
Gee, Ricardo 156
Generoso, Estela 30
Georgescu, Anca M. 42, 50
Gerwehr, S. 11
Gesteland, Raymond F. 148
Gibbs, R.A. 56
Gingrich, Jeffrey C. 21, 26, 43, 51, 62, 134
Giorgi, Dominique 5
Glantz, Karen 31
Glantz, Leonard H. 200
Glazer, Alexander N. 69, 133, 137
Gnirke, Andreas 61
Gocayne, Jeannine 150
Godfrey, Virginia L. 32
Godovikova, T.S. 70
Gong, K. 111
Gonzales, Julia 177
Goodart, S. 53
Goodwin, Edwin H. 12
Goodwin, Peter M. 160, 179
Goold, R. 35
Gordon, Laurie A. 44, 47, 48, 50
Gorvad, Ann 45, 46
Grady, Deborah L. 25, 52, 53
Graham, Kenneth S. 167
Graves, Joan 64
Graves, Mark 106
Gray, Joe 10, 14–16, 62, 66, 108
Gray, Timothy 79
Green, Eric 9
Greulich, Karin 10, 14, 66
Grillo, A. 56
Grosz, Michael 18
Gu, Y. 56, 57
Guan, Xiaojun 103
Guan, Xiaoping 22
Guellaën, G. 11
Guilfoyle, Richard A. 156
Guo, Zhen 156
Gusfield, Dan 80
Gutleben, Glenn 190
Haces, Alberto 160
Hansen, A.D.A. 169
Harding, John D. 160
Harger, Carol 79
Harris, M. 35
Harrison, Rhonda 34

Hartog, Bobi K. 168
Hauser, Loren 103, 112
Haveran, Liam 4
Hawe, William P. 123
Hayden, Jessica 156
Hebenbrock, Kirstin 132
Hebert, S. 35
Heller, Michael J. 125
Hide, Winston 104
Himawan, Jeff 157
Hintz, Mary 62
Hodgin, Faye Cobb 196
Hoffman, Susan M.G. 42, 48, 49
Hofstadler, S.A. 143
Hogan, Brigid L.M. 32
Holmes, Linda 203
Holtzman, Neil A. 195
Hood, Leroy 11, 147, 192, 193
Hopkins, Jordan 183
Houts, Julie 52
Hoyt, Peter R. 32
Huang, Faqing 166
Huang, G.M. 11
Hubert, R. 11
Huiet, Layne 65
Hunicke-Smith, Scott 158
Hurst, Gregory B. 144
Hutchinson, Jane S. 58
Hwang, D. 11
Ihle, James N. 32
Ijadi, Mohamad 79, 81
Imara, Mwalimu 196
Ioannou, Panayotis A. 22
Ivanov, Igor 124
Jackson, Kim 58
Jacobson, K. Bruce 126, 138
Jacobson, Stephen C. 176
Jaklevic, Joseph M. 145, 169, 170, 174, 182
Jefferson, Margaret C. 186
Jelenc, Pierre 8
Jennings, Anne 190
Jessee, Joel 22
Jett, James H. 160, 177, 179
Jewett, Phil 25
Ji, Huamin 156
Johnson, Arthur 156
Johnson, Dabney K. 6
Johnson, Mitchell E. 160, 179
Johnson, Stephanie 48
Johnson, Wanda 26
Johnston, Rick 130
Ju, Jingyue 133, 137
Jurka, Jerzy 97
Kallioniemi, Olli 62
Kanjuparamban, Teresa 4
Kao, Fa-Ten 17
Karger, Barry L. 132
Karp, Richard M. 113
Kass, David H. 73
Kaszuba, David 15
Kaur, G. Pal 23
Kececioglu, John D. 80, 100
Keen, Gifford 79
Keeney, Scott 63
Kelleher, Zachary 18
Keller, Richard A. 160, 179
Kelly, P. 81
Kerlavage, Anthony R. 83, 150
Kerley, Marilyn K. 32
Kestilä, Marjo 50
Khristich, Jason V. 58
Kieleczawa, Jan 161, 162
Kim, Ung-Jin 11, 19, 20
Kimball, Alvin 148
Kimmerly, William 59, 145, 169
Kingsbury, David T. 76
Kirchner, Jakob M. 154
Kirillov, Eugene 124
Knight, Jim 80
Knill, Emanuel 27, 39, 41, 107
Knorre, D.G. 70
Kobayashi, Arthur 85, 117
Kolbe, William F. 136
Koo, Jackson 134
Koop, Ben F. 147
Korenberg, J.R. 11
Kotler, Lev 164
Kouprina, Natalya 64
Kowbel, David 10, 62, 66
Kravatsky, Y.V. 13
Kreindlin, Edward 124
Krishnarao, Appasani S.M. 132
Krone, Jennifer 140
Kuo, Wen-Lin 10, 62, 66
Kuznetsova, A. 13

Kwok, Pui-Yan 71
Kwon, H. 33
Lamerdin, Jane 43, 45, 46, 51, 154, 155
Lane, S. 36, 37
Langlois, Richard 26, 43, 181
Larionov, Vladimir 64
Lashkari, Deval 158
Lawler, Gene 80
Lawrence, Charles B. 106, 114
Lawton, Lee 8
Lee, Denise 44
Lee, Inyoul 147
Lee, William H. 32
Lehesjoki, A.E. 35
Lehrach, Hans 22
Lemanski, Cheryl 2
Lennon, Gregory G. 5, 30, 42, 43, 45, 46
Leong, Joe 130
Lever, David C. 123
Lewis, Keith 59
Lewis, S. 111
Li, Peter 78
Li, Qingbo 129
Lieuallen, Kimberly 5
Liew, C.-C. 11
Lim, Regina 61
Ling, Lucy L. 4, 74
Linn, Stuart 63
Lipshutz, Robert 75, 127
Liu, Li-ing 150
Liu, Qinghua 156
Liu, Rong 131
Lobb, Rebecca 2
Lockett, Stephen 15, 66, 108
Longmire, Jonathan L. 25
Los Alamos Genomics Group 38
Lovett, M. 53
Lowenstein, M.G. 39, 41
Lowry, Jimmie 65
Lowry, Steve 54
Lu, F. 56
Lu, J. 56
Lu, Xiandan 129
Luchina, N.N. 67
Lundstrom, Ron 4
Lustre, Veronica M. 59
MacConnell, William P. 175
Macfarlane, Jane F. 116
MacGavran, L.P. 81
Maltsev, Natalia 91
Mancino, Valeria 19
Männikkä, Minna 50
Mansfield, Betty K. 202
Marcano, Stewart 58
March, Michelle 79
Marchuck, Yelena 58
Mariella, Jr., Raymond 43, 155, 181
Markowitz, Victor M. 87, 93
Marks, Andy 148
Marquette, David D. 78
Marrone, Babetta L. 160, 177
Marstaller, Jenny E. 10, 95
Martin, Chris S. 68
Martin, Christopher H. 59, 62, 88, 145, 146
Martin, John C. 160, 178, 179
Martin, Sheryl A. 202
Martin-Gallardo, Antonia 9
Martinez, Adelmo 25
Masquelier, D. 181
Massa, Hillary 9
Mathies, Richard A. 133, 137
Matis, Sherri 103
Mayeda, Carol A. 146
McCarthy, J. 87, 89
McCormick, Mary K. 25
McEwen, J.E. 194
McInerney, Joseph D. 185
McLeod, Mia 79
McPherson, J.D. 53
Medvick, Patricia A. 168
Meincke, Linda J. 25, 39, 41
Meng, John 170
Mengos, April 19
Merrick, Joseph M. 150
Meyne, Julianne 12
Micikas, Lynda B. 185
Micklos, David 191
Miller, Mary 190
Millero III, Frank 190
Mirzabekov, Andrei D. 124
Mitchell, S. 11
Mohrenweiser, Harvey W. 30, 43, 44, 47, 49
Moir, Donald T. 4, 74
Moise, Herb 146
Montgomery, Donald D. 125

Montgomery, Mishelle A. 154
Moore, Jr., Alvin W. 176
Morris, Maxwell 126
Moseley, Ray 196
Motulsky, Arno G. 195
Mouradian, Stephane 141
Moustakas, Demetri 34
Moyer, J. 33
Moyzis, Robert K. 12, 25, 28, 39, 41, 52, 53, 159
Mucenski, Michael L. 29, 32, 33
Mulley, J.C. 36, 37
Mullikin, James 15, 66, 108
Mundt, M. 39, 41, 81
Munk, Chris 159
Munn, Maureen M. 193
Mural, Richard J. 103, 112
Muzny, D.M. 56
Myers, Eugene 114, 119
Myers, R. 35
Naranjo, Cleo 1, 159
Narla, Mohan 145
Nasedkina, T.V. 70
Neff, Mark W. 60
Nelson, Christine M. 141
Nelson, D.L. 56, 57
Nelson, David O. 44, 118
Nelson, Debi 153
Nelson, Randall 140
Nguyen, Tuyen 175
Nichols, Anne F. 63
Nickerson, Deborah A. 71
Nickerson, E. 53
Norcross, Jonathan 4, 74
O'Connor, K. 35
O'Neill, John 79
Olsen, Anne S. 30, 42–44, 50
Olson, Maynard 61, 183, 192, 193
Oostra, B.A. 57
Overbeek, Ross 91
Overhauser, J. 53
Overton, Chris 92
Ow, David J. 85, 117
Palazzolo, Michael J. 59, 62, 145, 146
Parinov, Sergei 124
Parrish, J.E. 57
Pecherer, Robert M. 40, 81, 98, 115
Pendergrass, Stephanie 177
Pennacchio, Len 5
Perez-Castro, Ana V. 28
Peterson, E.T. 53
Petrov, Sergey 103, 112
Petty, Jeffrey T. 160, 179
Phoenix, David 196
Pietrzak, Eugenia 22
Pinkel, Daniel 10, 14, 15, 62, 66, 108
Piper, Jim 15
Pirrung, Michael C. 123
Pitluck, S. 87, 111
Polanovsky, Oleg L. 67
Poletaev, A.I. 13, 70
Pollard, Martin J. 171, 173
Polymeropoulos, Mihael H. 7
Porter, Christopher J. 77
Porter, Kenneth 165, 166
Potter, Cheri 177
Potter, S. Steven 32
Power, Alicia 79
Prange, Christa 5
Prem, Shirdi R. 96
Probst, Shane 58
Przetakiewicz, Marek 34
Pumilia, Maria 79
Quackenbush, John 58
Quesada, Mark A. 162
Quigley, Denise I. 12
Ramsey, J. Michael 176
Ravi, R. 80
Reeve, John 151
Reilly, P.R. 194
Resnick, Michael 64
Reyes, Tony 65
Reynolds, J. 57
Reynolds, R.J. 178
Richards, C.S. 57
Richards, R.I. 36, 37, 39
Richards, S. 56
Richards, W.G. 33
Richardson, Charles C. 157
Rider, David 79
Riedell, L. 10, 66
Riggs, Arthur D. 167
Rinchik, Eugene 30
Rine, Jasper 60
Roach, Dave 130

Robinson, Donna L. 25, 52, 53
Roche, Patricia A. 200
Rodkins, Annie 58
Romberg, Lori 58
Rommens, Johanna 62
Rong, Jiang 131
Roos, Pieter 131
Roslaniec, M.C. 178
Roth, E.J. 56
Rouquier, Sylvie 5
Rouse, Barry T. 32
Rowen, Lee 147
Ruan, Chihchuan 137
Rubin, Edward M. 24, 54, 145
Ruiz-Martinez, Marie C. 132
Rye, Hays S. 69
Sachleben, Richard A. 138
Salazar, Edmund P. 154
Salzman, Gary C. 177
Sandhu, Arbansjit K. 23
Sano, Takeshi 34
Sapolsky, R.J. 75
Schecker, Jay A. 160
Scheidecker, Lisa K. 51
Schieltz, David 140
Schimke, R. Neil 187
Schliep, A. 107
Schneider, Greg 68
Schor, Pat L. 25
Schultz, Jocelyn C. 136
Schurtz, Tony 148
Schwertfeger, Jolene 79
Scott, Duncan 54
Scott, Ron 58
Searls, David 101
Segebrecht, Linda 187
Selleri, Licia 58
Sesma, Mary Ann 186
Shadravan, Farideh 10, 62, 66
Shaffer, L.G. 57
Shah, Manesh 103, 112
Shannon, Mark 29
Sharaf, Muhammad 134
Shatrova, A.N. 13
Shaw, Barbara Ramsey 165, 166
Shih, Po-Jen 160
Shin, Dong-Guk 90
Shizuya, Hiroaki 11, 19, 20
Shuey, Steven W. 123
Simmons, A. 53
Simon, Melvin I. 11, 19, 20
Sindelar, L.E. 182
Sivila, Randy F. 175
Slepak, Tania 19
Slezak, Tom 43, 82, 84-86, 110, 117
Sloop, Frederick V. 138
Smit, Arian F.A. 167
Smith, Cassandra L. 34
Smith, Desmond J. 24
Smith, Doug 134, 151
Smith, Hamilton O. 150
Smith, Lloyd M. 135, 141, 156
Smith, Loanne R. 12
Smith, Michael W. 58
Smith, Richard D. 142, 143
Snyder, Eric E. 102
Soares, Marcelo Bento 8
Sobolev, Irina 164
Soderlund, C. 81
Soliman, K. 45
Solomon, Jerry 20
Solovyev, Victor 114
Sosnowski, Ronald G. 125
Speed, Terence P. 118
Spikes, A.S. 57
Stanford, Beverly 30
States, David J. 105
Stepchenko, Alexander G. 67
Stevko, V.M. 169
Stilwagen, Stephanie A. 154
Stone, N. 35
Stormo, Gary D. 102
Strathmann, Mike 60
Stubbs, Lisa J. 6, 29-31, 33
Studier, F. William 161-163
Stultz, Karen E. 59
Stump, Mark 148, 153
Su, Long 8
Sudar, Damir 15, 66, 108
Sun, Dazhong 59
Sun, Tian-Qiang 18
Sutherland, G.R. 36, 37, 39, 41
Sutherland, Robert D. 39-41, 81, 98
Swartz, Annette 47, 82
Szafranski, Przemekyslaw 34

Szeto, Ernest 93
Tabor, Stanley 157
Talbot, Jr., C. Conover 77
Tang, Kai 139
Tang, Wei 141
Tanner, Minna 62
Taranenko, Nelli I. 139
Tebbs, Robert S. 154
Tesmer, J.G. 39, 41
Thayer, Nina 79
Theil, Edward 87–89, 145
Thiel, Andy 156
Thomas, Gregory 122
Thomas, Sylvia 58
Thompson, Curtis 14, 108
Thompson, Larry H. 154
Thompson, Sue 52
Thundat, T. 16
Timofeev, Edward 124
Tipton, Jennifer 79
Tobin, Joshua 58
Torney, David C. 27, 39, 41, 107
Trask, Barbara J. 9, 183
Trottier, Ralph W. 196
Troup, Charles D. 79, 81
Tryggvason, Karl 50
Tsadeek, Shahar 79, 109
Tsujimoto, Susan 30, 47
Tsukamoto, Kazuhiro 34
Tu, Eugene 125
Tumolo, Annette 65
Uber, Donald C. 170, 172, 173
Uberbacher, Edward C. 103, 112
Udseth, Harold R. 142
Ugozzoli, Luis 65
Ulanovsky, Levy 164
Uzgiris, Jim 156
van den Engh, Ger 9, 183
van der Feltz, Gus 15
Veklerov, E. 87, 88
Venter, J. Craig 11, 150, 152
Verkerk, A.J.M.H. 57
Vos, Jean-Michel H. 18
Wagner, Mark C. 84–86, 117
Wagner, Robert 52
Wallace, Bruce 65
Wang, Kai 147
Wang, M. 10, 14
Wang, Yiwen 133
Wapenaar, M.C. 56
Warmack, R.J. 16
Warrington, J.A. 35
Wasmuth, J.J. 53
Wassom, John S. 202
Watt, Susan 30
Weaver, Nicholas C. 116
Weber, Christine A. 154
Wei, Yalin H. 58
Weier, H.-U. 14
Weier, Ulli 10, 62, 66
Weiss, Robert B. 148, 153
Weisser, Deborah 113
Wentland, M.A. 56
Westin, Alan F. 197
Westphall, Michael S. 135, 156
White, Owen 83
Whitmore, S.A. 36, 37
Whittaker, Clive C. 27
Wilkinson, J.E. 33
Williams, Peter 140
Williams, Walter 204
Wilson, Julie 28
Wilson, Richard K. 149
Witkowski, Jan 191
Witney, Frank 65
Woese, Carl 152
Wohlpart, Al 203
Wong, Benjamin S. 26
Wong, Gane 61
Woolley, Adam T. 133
Woychik, R.P. 33
Wright, James 204
Wu, Ming 160
Wyrick, Judy M. 202
Xu, Ying 103
Yaar, Ronald 34
Yager, Thomas D. 71
Yamakawa, K. 11
Yeh, T. Mimi 84, 110, 117
Yershov, Gennady 124
Yesley, Michael S. 199
Yeung, Edward S. 129
Yoder, B. 33
Yokota, Hiroki 9
Youngblom, Janey 9
Yu, Jingwei 17

Yu, Jun 61
Yue, P. 10, 66
Yust, Laura N. 202
Zarella, Tom 130
Zelver, Jack S. 173
Zenin, V.V. 13
Zhang, Jian-Zhong 131
Zhang, Shiping 161, 162
Zhao, Jian-Ying 131
Zhu, Huiping 133
Zhu, Lin 141
Zhu, Yiwen 24, 54, 55
Zimmermann, Wolfgang 30
Zoghbi, H.Y. 56
Zorn, Manfred D. 10, 66, 87, 94–96, 116
Zweig, Franklin M. 184
Zweig, Geoffrey 113

Appendix B: National Laboratories Index

U.S. Department of Energy Laboratories

Human Genome Program work at the national laboratories is described in the following abstracts.

Ames Research Center 129

Argonne National Laboratory 91

Brookhaven National Laboratory 161-63

Lawrence Berkeley Laboratory 10, 14-16, 24, 54, 55, 59, 60, 62, 66, 87-89, 93-96, 108, 111, 116, 136, 145, 146, 169-74, 182

Lawrence Livermore National Laboratory 5, 21, 26, 29, 30, 31, 42-51, 62, 72, 73, 82, 84-86, 110, 117, 118, 134, 154, 155, 180, 181

Los Alamos National Laboratory 1-3, 11-13, 25, 27, 28, 36-41, 52, 53, 81, 98, 99, 107, 115, 159, 160, 168, 177-79, 199

Oak Ridge National Laboratory 6, 16, 29, 30-33, 103, 112, 126, 138, 139, 144, 176, 202

Pacific Northwest Laboratory 122, 142, 143

Sandia National Laboratories 116

Appendix C: Anticipated Workshop Attendees

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Chris Abajian
Dept. of Molecular Biotechnology
University of Washington, FJ-20
Seattle, WA 98195
phone: 206-685-7367
fax: 206-685-7301
e-mail: molbiotk@u.washington.edu

Mark D. Adams
The Institute for Genomic Research
932 Clopper Road
Gaithersburg, MD 20878
phone: 301-869-9056
fax: 301-869-9423
e-mail: Mdadams@tigr.org

Arun Aggarwal
Information and Computing Science
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-6821
fax: 510-486-6816
e-mail: akaggarwal@lbl.gov

Mike Albin
Perkin Elmer–ABI Division
850 Lincoln Center Drive
Foster City CA 94404
phone: 415-638-5753
fax:
e-mail:

Michelle Allegria-Hartman
Lawrence Livermore National Lab.
P. O. Box 808, L-452
Livermore, CA 94551-9900
phone: 510-423-3637
fax: 510-423-3608
e-mail:

Susan Allen
Human Genome Center
Lawrence Livermore National Lab.
P.O. Box 808, L-452
Livermore, CA 94551-9900
phone: 510-422-2779
fax: 510-422-2282
e-mail:

David P. Allison
Bldg. 4500 S MS 6123
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, TN 37831
phone: 615-574-6215
fax: 615-574-6210
e-mail: not available

Michael R. Altherr
Genomics and Structural Biology Group
Los Alamos National Laboratory
MS M880
Los Alamos, NM 87545
phone: 505-665-6144
fax: 505-665-3024
e-mail: Altherr@Flovax.lanl.gov

George J. Annas
Law, Medicine and Ethics Program
Boston University School of Public Health
80 E. Concord Street A-509
Boston, MA 02118-2394
phone: 617-638-4626
fax: 617-638-5299
e-mail:

Santosh Arcot
Biology & Biotechnology Division
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
7000 East Avenue
Livermore CA 94551-9000
phone: 510-423-3637
fax: 510-422-2282
e-mail:

Norman Arnheim
Department of Biology
University of Southern California
AHF 103
Los Angeles, CA 90089-0371
phone: 213-740-7675
fax: 213-740-8631
e-mail: arnheim@molbio.usc.edu

Linda Ashworth
Lawrence Livermore National Lab.
P. O. Box 808
7000 East Avenue, L-452
Livermore, CA 94551
phone: 510-422-5665
fax:
e-mail: ashworthl@.llnl.gov

Raghbir S. Athwal
Medical Research Bldg, Room 700
Fels Institute for Cancer Res. and Molec.
Temple University
3420 N. Broad Street
Philadelphia, PA 19140
phone: 215-707-6931 or 4300
fax: 215-707-4318
e-mail:

Diane Baker
Human Genome Center
University of Michigan
Bldg MSRB II, Box 0674, Room 2570
Ann Arbor, MI 48109-0674
phone: 313-763-2933
fax: 313-763-3784
e-mail:

Joseph Balch
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore, CA 94551
phone: 510-422-8643
fax: 510-423-3608
e-mail: balch1@llnl.gov

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

David Barker
Molecular Dynamics
928 E. Arques Avenue
Sunnyvale, CA 94086
phone: 408-737-3003/737-1222
fax: 408-773-8343
e-mail:

John Bashkin
Molecular Dynamics
928 E. Arques Avenue
Sunnyvale, CA 94086
phone: 408-737-33130
fax: 408-773-8343
e-mail:

Mark Batzer
Human Genome Center,Biological &
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
7000 East Avenue
Livermore, CA 94551
phone: 510-422-5721
fax: 510-422-2282
e-mail: batzer2@llnl.gov

Peter D. Bayne
Corporate Development Department
Promega Corporation
2800 Woods Hollow Road
Madison WI 53711
phone: 608-277-2542
fax: 608-277-2601
e-mail: pbayne@promega.com

Diane R. Beeson
Department of Sociology
California State University, Hayward
Hayward, CA 94542
phone: 510-881-4127
fax: 510-727-2276
e-mail:

George I. Bell
Theoretical Division, T-10
Los Alamos National Laboratory
MS K710, T-10
Los Alamos, NM 87545
phone: 505-665-3805
fax: 505-665-3493
e-mail: gib@lanl.gov

Anne Bergmann
Human Genome Center
Lawrence Livermore National Laboratory
P.O. Box 808, MS L-452
Livermore, CA 94551-9900
phone: 510-423-3633
fax:
e-mail:

Anthony J. Beugelsdijk
Los Alamos National Laboratory
P.O. Box 1663, MS J580
Los Alamos, NM 87545
phone: 505-667-3169
fax: 505-665-3911
e-mail: beugelsdijk@esa.lanl.gov

Howard S. Bilofsky
Core Scientific Information Technologies
SmithKline Beecham Pharmaceuticals
P.O. Box 1539
709 Swedeland Road
King of Prussia PA 19406
phone: 215-270-6464
fax: 215-270-5580
e-mail: bilofsky@smithkline.com

Tom Blackwell
Howard Hughes Medical Institute and
Harvard Medical School
200 Longwood Avenue
Boston, MA 02115
phone: 617-432-0503
fax: 617-432-7266
e-mail: blackwel@twod.med.harvard.edu

Frederick R. Blattner
University of Wisconsin
Genetics Bldg
445 Henry Mall
Madison, WI 53706
phone: 608-262-2534
fax: 608-263-7459
e-mail: fred@genetics.wisc.edu

Denise Boehrer
Biology & Biotechnology Division
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore CA 94551-9000
phone: 510-422-8361
fax: 510-422-2282
e-mail:

E. Morton Bradbury
MS M881, Life Sciences Division, Center for
LS-DO Los Alamos National Laboratory
P.O. Box 1663
Los Alamos NM 87545
phone: 505-667-2690
fax: 505-667-2891
e-mail: bradbury_e_morton@lanl.gov

J.-C. Bradley
Department of Chemistry
Duke University
P.O. Box 90346
Durham NC 27708-0349
phone: 919-660-1500
fax: 919-660-1605
e-mail:

Brigitte F. Brandriff
Human Genome Center
Lawrence Livermore National Lab.
P. O. Box 808, L-452
Livermore, CA 94551
phone: 510-423-0758
fax: 510-423-3608
e-mail: brandriff1@llnl.gov

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Elbert W. Branscomb
Biomedical Science Division
Lawrence Livermore National Lab.
P. O. Box 5507, L-452
7000 East Avenue
Livermore, CA 94550
phone: 510-422-5681
fax: 510-423-3608
e-mail: elbert@alu.llnl.gov or

Thomas M. Brennan
Department of Genetics
Stanford University, School of Medicine
Room M305
Stanford, CA 94305-5120
phone: 415-725-7423
fax: 415-723-1534
e-mail: brennan@sumex-aim.stanford.edu

David Brevels
Genomics & Structural Biology Group
Los Alamos National Laboratory
Group LS-2, MS M880
Los Alamos, NM 87545
phone: 505-667-2726
fax: 505-665-3024
e-mail:

Larry Brewer
Biology & Biotechnology Division
Lawrence Livermore National Laboratory
P.O. Box 808, L-156
Livermore CA 94551-9000
phone: 510-422-8361
fax: 510-422-2282
e-mail:

Irena Bronstein
Attn: Anne Marie Youell
Tropix, Inc.
47 Wiggins Avenue
Bedford, MA 01730-2314
phone: 617-271-0045
fax: 617-275-8581
e-mail:

Gilbert M. Brown
Chemistry and Analytical Sciences Division
Oak Ridge National Laboratory
P. O. Box 2008, Bldg 4500S, Rm C-250
MS 6119
Oak Ridge TN 37831-6119
phone: 615-576-2756
fax: 615-574-4939
e-mail: gbn@ornl.gov

Nancy Brown
MS M880, LS-2
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-685-3858
fax: 505-665-
e-mail:

Chris Brunn
Genome Data Base
2024 E Monument St., Suite 1-200
Baltimore MD 21205-2100
phone: 410-955-9705
fax: 410-614-0434
e-mail: brunn@gdb.org

William Bruno
T-10, Mail Stop K-710
Los Alamos National Laboratory
Los Alamos, NM 87545
phone: 505-665-3802
fax: 505-665-3493
e-mail: billb@temin.lanl.gov

Michelle V. Buchanan
Chemistry Department
Oak Ridge National Laboratory
P.O. Box 2008
5510-A, MS 6365
Oak Ridge TN 37831-6365
phone: 615-574-4868
fax: 615-576-8559
e-mail: VBu@ORNLSTC

Judith Buckingham
Genomics & Structural Biology
Los Alamos National Laboratory
LS-2, MS M880
Los Alamos, NM 87545
phone: 505-667-2764
fax:
e-mail:

Peter Buneman
Department Computer and Information
University of Pennsylvania
200 S. 33rd Street
Philadelphia PA 19104-6389
phone: 215-898-7703
fax: 215-898-0587
e-mail: peter@cis.upenn.edu

Matt Burgin
Biology & Biotechnology Division
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore CA 94551-9000
phone: 510-422-8361
fax: 510-422-2282
e-mail:

Christian Burks
MS K710, Group T-10; LANSCE/ER
Los Alamos National Laboratory
Los Alamos, NM 87545
phone: 505-667-6683
fax: 505-665-3493
e-mail: cb@t10.lanl.gov

Eric Buxton
Dept. of Chemistry
University of Wisconsin
1101 University Avenue
Madison, WI 53706
phone: 608-262-2021
fax: 608-262-0381
e-mail: buxton@whitewater.chem.wisc.ed

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

David Callen
Dept. of Cytogenetics & Molecular Genetics
Adelaide Women and Children's Hospital
72 King William Road
North Adelaide S.A. Australia
phone: 618-204-6715/7111
fax: 618-204-7342
e-mail: dcallen@ache.adelaide.edu.au

Graham Cameron
Hinxton Hall
EBI
Hinxton,
Cambridge, Engla CB10 1RQ
phone: 011-44-223-494467
fax: 011-44-223-494968
e-mail:

Evelyn W. Campbell
Center for Human Genome Studies
Los Alamos National Laboratory
MS M880, LS-1
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-667-2696
fax: 505-665-3024
e-mail:

Mary Campbell
Center for Human Genome Studies
Los Alamos National Laboratory
MS M880, LS-1
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-667-2696
fax: 505-665-3024
e-mail:

Charles Cantor
Dir., Center for Advanced Biotechnology
Boston University
36 Cummington Street
Boston, MA 02215
phone: 617-353-8504
fax: 617-353-8501
e-mail: crc@enga.bu.edu

Charles Carlson
Director Life Sciences
The Exploratorium
3601 Lyon Street
San Francisco, CA 94123
phone: 415-561-0341
fax: 415-561-0307
e-mail: charliec@exploratorium.edu

Anthony Carrano
Director, Human Genome Center
Lawrence Livermore National Lab.
BBRP, L-452
P. O. Box 5507
Livermore, CA 94551
phone: 510-422-5698
fax: 510-423-3110
e-mail: avc@sts.llnl.gov

C. Thomas Caskey
Institute for Molecular Genetics
Baylor College of Medicine
Texas Medical Center, T-809
One Baylor Plaza
Houston, TX 77030-3498
phone: 713-798-4774
fax: 713-798-7383
e-mail: CCaskey@bcm.tmc.edu

Leslie A. Chasteen
Genomics & Structural Biology
Los Alamos National Laboratory
LS-2, MS M880
Los Alamos, NM 87545
phone: 505-667-2726
fax:
e-mail:

C. H. Winston Chen
MS 6378; Health Sciences Research
Oak Ridge National Laboratory
P. O. Box 2008
Bldg. 5500,
Oak Ridge TN 37831-6119
phone: 615-574-5895
fax: 615-576-2115
e-mail: chenc@ornl.gov

David Chen
MS A114
Los Alamos National Laboratory
Los Alamos, NM 87545
phone: 505-667-1600
fax: 505-665-3858
e-mail:

I-Min A. Chen
ICSD
Lawrence Berkeley Laboratory
MS 50B-3238
1 Cyclotron Road
Berkeley CA 94720
phone: 510-486-7264
fax: 510-486-4004
e-mail:

Jan-Fang Cheng
Human Genome Center
Lawrence Berkeley Laboratory
1 Cyclotron Road
MS 74-157
Berkeley, CA 94720
phone: 510-486-6590
fax: 510-486-6816
e-mail: JFCheng@lbl.gov

Xueheng Cheng
Life Sciences Center (K1-50)
Pacific Northwest Laboratory
P.O. Box 999
1101 University Avenue
Richland, WA 99352
phone: 509-376-0723/5665/5-3738
fax: 509-376-0418
e-mail:

Han-Chang Chi
Genomics & Structural Biology
Los Alamos National Laboratory
LS-2, MS M880
Los Alamos, NM 87545
phone: 505-665-1274
fax:
e-mail:

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Mari Christensen
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore, CA 94550
phone: 510-422-5657
fax:
e-mail:

George M. Church
Department of Genetics
Harvard Medical School
Warren Alpert Building, Room 513
200 Longwood Avenue
Boston, MA 02115
phone: 617-432-7562
fax: 617-432-7663
e-mail: church@rascal.med.harvard.edu

Gary A. Churchill
Biometrics Unit
Cornell University
337 Warren Hall
Ithaca, NY 14853-7801
phone: 607-255-5488
fax: 607-255-4698
e-mail: gary@amanita.biom.cornell.edu

Michael J. Cinkosky
Director of Information Services
National Center for Genome Resources
1800 Old Pecos Trail, Suite E
Santa Fe, NM 87505
phone: 505-982-7840
fax: 505-982-7690
e-mail: Michael.Cinkosky@ncgr.org

Lynn Clark
Center for Human Genome Studies
Los Alamos National Laboratory
MS M885-CHGS
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-667-3912
fax: 505-665-2891
e-mail: clark@telomere.lanl.gov

Colin Collins
Resource for Molecular Cytogenetics, MS
Lawrence Berkeley Laboratory
1 Cyclotron Rd.
Berkeley CA 94720
phone: 510-486-5810
fax: 510-486-6746
e-mail: collins@white.lbl.gov

Debra L. Collins
Dept. of Endocrinology, metabolism, and
University of Kansas Medical Center
3901 Rainbow Blvd.
Kansas City, KS 66160-7318
phone: 913-588-6043
fax: 913-588-3995
e-mail: ukanvm.cc.ukans.edu

Alex Copeland
Human Genome Center
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore, CA 94550
phone: 510-422-0236
fax: 510-422-3608
e-mail:

L. Corell
Human Genome Center
Lawrence Livermore National Laboratory
P.O. Box 808, L452
Livermore, CA 94551-9900
phone: 510-423-8062
fax: 510-422-2282
e-mail:

Bryan Coulter
Science and Engineering Education Division
Oak Ridge Institute for Science and
P.O. Box 117
Oak Ridge TN 37831-0117
phone: 615-574-8633
fax: 615-241-2727
e-mail: coulterb@orau.gov

L. Scott Cram
Life Sciences Division
Los Alamos National Laboratory
LS-DO MS M881
Los Alamos, NM 87545
phone: 505-667-2890
fax: 505-667-2891
e-mail: lxc@lanl.gov or sec.

Lee A. Crandall
Dept. of Community Health & Family
University of Florida
Box 100222
JHMHC
Gainesville, FL 32610-0222
phone: 904-392-4321
fax: 904-392-7349
e-mail:

Radomir Crkvenjakov
HYSEQ, Inc.
670 Almanor Ave.,
Sunnyvale, CA 94086
phone: 408-524-8100
fax: 408-524-8141
e-mail:

Jamie Cuticchia
Data Manager, Genome DB
John Hopkins University
2024 E Monument Street
Baltimore, MD 21205
phone: 410-614-0438
fax: 410-614-0434
e-mail: jamie@gdb.org

Ron Cytron
Biomedical Computer Lab
Washington University
Campus Box 8068
700 South Euclid Avenue
St. Louis MO 63110
phone: 314-362-2129
fax: 314-362-0234
e-mail:

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Carol Dahl
NCGR, National Center for Human Genome
National Institutes of Health
Office of Scientific Review, Bldg. 38A, Rm.
9000 Rockville Pike
Bethesda MD 20892
phone: 301-496-7531
fax: 301-480-2770
e-mail: xd2@cu.nih.gov

Courtney Davidson
Electronic Engineering; L-222
Lawrence Livermore National Laboratory
P.O. Box 808, Bldg. 153/2021
Livermore, CA 94550
phone: 510-423-7168
fax:
e-mail:

Susan Davidson
Computer Science
University of Pennsylvania
200 S. 33rd Street
Philadelphia PA 19104
phone:
fax: 215-898-0587
e-mail:

Cheryl Davis
Human Genome Center
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5917
fax:
e-mail:

Ronald W. Davis
Department of Biochemistry
Stanford University
School of Medicine
Beckman Center, Room B400
Stanford, CA 94305-5307
phone: 415-723-6277
fax: 415-723-6783
e-mail:

Maria de Jesus
Biology & Biotechnology Division
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore CA 94551-9000
phone: 510-422-8361
fax: 510-422-2282
e-mail:

Pieter de Jong
Human Genetics Department
Roswell Park Cancer Institute
Elm & Carlton Streets
Buffalo, NY 14052
phone: 716-845-3168
fax: 716-845-8449
e-mail: pieter@dejong.med.buffalo.edu

Larry Deaven
Center for Human Genome Studies, Life Sci.
Los Alamos National Laboratory
CHGS, MS M885-CHGS
Group LS-4
Los Alamos, NM 87545
phone: 505-667-3912
fax: 505-667-2891
e-mail: deaven@telomere.lanl.gov

Chery L. Dell
Director, Genome Education Program
University of Michigan Human Genome
2570 MSRB II, Box 0674
1150 West Medical Center Drive
Ann Arbor, MI 48109-0674
phone: 313-747-2738
fax: 313-764-4133
e-mail: cheryl.dell@med.umich.edu

Bobi K. Den Hartog
Engineering Sciences & Applications, ESA-6
Los Alamos National Laboratory
MS J580
Los Alamos NM 87545
phone: 505-665-3231
fax: 505-665-3911
e-mail: bobi@niobrara.esa.lanl.gov

Karen Denison
MS 880
Los Alamos National Laborabory
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-665-0781
fax: 505-665-3024
e-mail:

Norman Doggett
Genetics Group
Los Alamos National Laboratory
MS M880, LS-2
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-665-4007
fax: 505-665-3024
e-mail: Doggett@Flovax.LANL.Gov

Richard J. Douthart
Life Sciences Center (K1-50)
Pacific Northwest Laboratory
Mail Stop K1-50
P. O. Box 999
Richland, WA 99352
phone: 509-375-2653/3738
fax: 509-375-6821/3649?
e-mail: dick@gnome.pnl.gov

Norman J. Dovichi
Department of Chemistry
University of Alberta
E3-44 Chemistry Bldg.
Edmonton, Alberta
CANADA T6G 2G2
phone: 403-492-2845/3254
fax: 403-492-8231
e-mail: norm_dovichi@dept.chem.ualbert

Johannah Doyle
Biology Division
Oak Ridge National Laboratory
Bldg 9210, MS 8077
Oak Ridge TN 37831-8077
phone: 615-574-0848
fax: 615-574-1283
e-mail: stubbs@biovx1.bio.ornl.gov

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

John J. Dunn
Biology Department
Brookhaven National Laboratory
P.O. Box 5000
Upton, NY 11973-5000
phone: 516-282-3012
fax: 516-282-3407
e-mail: dunn@genome1.bio.bnl.gov

Michele Durand
Scientific Attache
French Consulate
540 Bush Street
San Francisco, CA 94108
phone: 415-397-0596
fax: 415-397-9947
e-mail:

Troy Duster
Institute for the Study of Social Change
University of California, Berkeley
2420 Bowditch Street
Berkeley, CA 94720
phone: 510-642-0813
fax: 510-642-8674
e-mail:

Charles G. Edmonds
Life Sciences Center (K1-50)
Pacific Northwest Laboratory
P.O. Box 999
Richland, WA 99352
phone: 509-375-3738
fax:
e-mail:

Carol Edwards
Genomics & Structural Biology Group
Los Alamos National Laboratory
Group LS-2, MS M880
Los Alamos, NM 87545
phone: 505-667-2726
fax: 505-665-3024
e-mail:

Frank H. Eeckman
Engineering
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-7384
fax: 510-486-6816
e-mail: fheeckman@lbl.gov

Rebecca S. Eisenberg
Law School
University of Michigan, Ann Arbor
625 S. State Street, 439 Hutchins Hall
Ann Arbor Michig 48109-1215
phone: 313-763-1372
fax: 313-763-9375
e-mail: usergde1@um.cc.umich.edu

Jeffrey M. Elliott
Human Genome Center
Lawrence Livermore National Laboratory
MS L-452
P.O. Box 808
Livermore, CA 94551-9900
phone: 510-422-6517
fax:
e-mail:

Michael L. Engle
Group T-10, MS K710
Los Alamos National Laboratory
Los Alamos, NM
phone: 505-665-2598
fax: 505-665-3493
e-mail:

Cheryl Ericsson
Human Genome Center
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5917
fax:
e-mail:

Glen A. Evans
McDermott Center for Human Growth and
University of Texas Southwestern Medical
6000 Harry Hines Blvd.
Dallas, TX 75235-8591
phone: 214-648-1660
fax: 214-648-1666
e-mail: gevans@SWMED.EDU

Jerry Eveleth
Human Genome Center
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Bldg 361, Room 1953
Livermore, CA 94551
phone: 510-294-5180
fax: 510-422-2282
e-mail: Jerry@flowcentral.llnl.gov

Betsy Fader
Student Pugwash USA
1638 R Street, NW - Suite 32
Washington, DC 20009
phone: 202-328-6555
fax: 202-797-4664
e-mail:

Lara Fallon
Department of Chemistry
Duke University
P.O. Box 90346
Durham NC 27708-0349
phone: 919-660-1500
fax: 919-660-1605
e-mail:

Kenneth H. Fasman
Genome Data Base
Johns Hopkins University
Welch Medical Library
2024 E. Monument Street
Baltimore, MD 21205
phone: 410-614-0439
fax: 410-614-0434
e-mail: ken@gdb.org

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Christopher Fields
National Center for Genome Resources
1800 Old Pecos Trail
Santa Fe NM 87505
phone: 505-982-7840
fax: 505-982-7690
e-mail: chris.fields@ncgr.org

Terri K. Fleming
Engineering
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5733
fax: 510-486-6816
e-mail: tkfleming@lbl.gov

Stephen Fodor
President & Scientific Director
Affymetrix Research Institute
3380 Central Expressway
Santa Clara, CA 95051
phone: 408-522-6010
fax: 408-481-0920
e-mail: steve_fodor@affymetrix.com

Robert S. Foote
Biology Division
Oak Ridge National Laboratory
Bldg. 9207, MS 8077
P.O. Box 2009
Oak Ridge, TN 37831-8077
phone: 615-574-0801
fax: 615-574-1274
e-mail:

Amanda A. Ford
Genomics & Structural Biology
Los Alamos National Laboratory
LS-2, MS M880
Los Alamos, NM 87545
phone: 505-665-6144
fax:
e-mail:

Ken Fordyce
Princeton Separations, Inc.
P.O. Box 300
Adelphia, NJ 07710
phone: 908-431-3338
fax: 908-431-3768
e-mail:

David M Fram
Belmont Research
84 Sherman Street
Cambridge, MA 02140
phone: 617-868-6878, ext. 216
fax: 617-868-2654
e-mail:

Eirik Frengen
Human Genetics Department
Roswell Park Cancer Institute
Elm & Carlton Streets
Buffalo, NY 14052
phone: 716-845-3168
fax: 716-845-8449
e-mail: eirik@dejong.med.buffalo.edu

Carl W. Fuller
United States Biochemical
26111 Miles Road
Cleveland OH 44128
phone: 216-765-5000
fax: 216-360-0975/800-535-0898
e-mail: bz476@Cleveland.freenet.edu

David Galas
Vice President of Research and
Darwin Molecular, Inc.
22025 20th Avenue SE, Suite 1000
Bothell, WA 98021
phone: 206-489-8011
fax: 206-489-8020
e-mail: Galas@Darwin.com

Emilio Garcia
Human Genome Center
Lawrence Livermore National Lab.
P. O. Box 808, L-452
Livermore, CA 94551-9900
phone: 510-422-8002
fax: 510-423-2282
e-mail: garcia@pvs.llnl.gov

Skip Garner
Automation Lab, at GESTEC
Southwestern Medical Center, University of
6000 Harry Hines Boulevard
Dallas TX 75235
phone: 214-648-1600
fax: 214-648-1666
e-mail: garner@swmed.edu

Jeffrey A. Garnes
Human Genome Center
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore, CA 94551-9900
phone: 510-423-3637
fax:
e-mail:

Joe M. Gatewood
Center for Human Genome Studies
Los Alamos National Laboratory
LS-2, MS M880
Los Alamos, NM 87545
phone: 505-665-2906/667-2726
fax: 505-665-3024
e-mail:

Anca M. Georgescu
MS L-452
Lawrence Livermore National Laboratory
P.O. Box 808, Bldg. 3701/1664
Livermore, CA 94551
phone: 510-423-3633
fax:
e-mail:

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Raymond F. Gesteland
Department of Human Genetics
University of Utah
6160 Eccles Genetics Bldg.
Salt Lake City, UT 84112
phone: 801-581-5190
fax: 801-585-3910
e-mail: rayg@genetics.med.utah.edu

Jeff Gingrich
Human Genome Center
Lawrence Livermore Laboratory
P.O. Box 808, L-452
Livermore CA 94551
phone: 510-423-8145
fax: 510-422-2282
e-mail: gingrich1@llnl.gov

Leonard H. Glantz
Health Law Dept.
Boston Univ. School of Public Health
80 E. Concord Street
Boston MA 02118
phone: 617-638-4626
fax: 617-638-5299
e-mail:

Alexander N. Glazer
Stanley/Donner ASU
University of California, Berkeley
230A Stanley Hall
Berkeley, CA 94720
phone: 510-642-3126/643-6302
fax: 510-643-9290
e-mail: ALEXANDER_GLAZER@MAILLINK.

Julia Gonzales
Genomics & Structural Biology
Los Alamos National Laboratory
LS-1, MS M888
Los Alamos, NM 87545
phone: 505-665-3859
fax:
e-mail:

Lynne A. Goodwin
Genomics & Structural Biology
Los Alamos National Laboratory
LS-2, MS M880
Los Alamos, NM 87545
phone: 505-665-3763
fax:
e-mail:

Steven J. Gordon
Intelligent Automation Systems
142 Rogers Street
Cambridge, MA 02142
phone: 617-354-3830
fax: 617-547-9727
e-mail:

Deborah L. Grady
Genetics Group
Los Alamos National Laboratory
MS M886
Los Alamos, NM 87545
phone: 505-667-2695
fax:
e-mail:

Kenneth Graham
Biology Department
Beckman Research Institute of the City of
1450 East Duarte Road
Duarte, CA 91010-0269
phone: 818-301-8352
fax: 818-358-7703
e-mail:

Mark Graves
Department of Cell Biology
Baylor College of Medicine
One Baylor Plaza, Rm. M525
Houston, TX 77030
phone: 713-798-8271
fax: 713-798-3759
e-mail: mgraves@bcm.tmc.edu

Joe W. Gray
Resource for Molecular Cytogenetics
University of California, San Francisco
P.O. Box 0808
San Francisco CA 94143-0808
phone: 415-476-3559
fax: 415-476-8218
e-mail: jwgray@lbl.gov

Darrell Jay Grimes
G-110/GTN ER-74
DOE, Environmental Sciences Division
19901 Germantown Rd.
Germantown MD 20874
phone: 301-903-4183
fax: 301-903-8519
e-mail:

Michael Grosz
Campus Box 7295, School of Medicine
University of North Carolina at Chapel Hill
UNC Lineberger Comprehensive Cancer
Chapel Hill, NC 27599-7295
phone: 919-966-3036
fax: 919-966-3015
e-mail:

Richard A. Guilfoyle
Dept. of Chemistry
University of Wisconsin
1101 University Avenue
Madison, WI 53706
phone: 608-263-2594
fax: 608-262-0381
e-mail:

Daniel Gusfield
Dept. of Computer Science
University of California, Davis
3051 EU II
Davis, CA 95616
phone: 916-752-7131
fax: 916-752-4767
e-mail: gusfield@cs.ucdavis.edu

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Anthony Hansen
Lawrence Berkeley Laboratory
MS 70A-3363
1 Cylcotron Road
Berkeley, CA 94720
phone: 510-486-7158
fax: 510-486-5857
e-mail: adhansen@lbl.gov

John Harding
Corporate Research
Life Technologies, Inc. , Bethesda Research
P.O. Box 6009
8717 Grovemont Circle
Gaithersburg, MD 20877
phone: 301-670-7649
fax: 301-670-7727
e-mail:

Carol Harger
Group T-10, MS K710
Los Alamos National Laboratory
Los Alamos, NM 87545
phone: 505-665-1923
fax:
e-mail: cah@hormone.lanl.gov

Mike Harrell
MS 204
ESL, Inc.
495 Java Drive
Sunnyvale, CA 94088-3510
phone: 408-738-2888 X6615
fax: 408-743-6425
e-mail: Mike_HARRELL@ESL.COM

William P. Hawe
Department of Chemistry
Duke University
P.O. Box 90349
Durham NC 27708-0349
phone: 919-660-1500
fax: 919-660-1605
e-mail:

Damar Hawkins
The Institute for Genomic Research
932 Clopper Road
Gaithersburg MD 20878
phone: 301-869-9056
fax: 301-869-9423
e-mail: dhawkins@tigr.org

Stephen R. Heller
ARS, BARC-West, Bldg. 005, Room 337
United States Department of Agriculture
10300 Baltimore Blvd.
Beltsville, MD 20705-2350
phone: 301-504-6055
fax: 301-504-6231
e-mail: srheller@asrr.arsusda.gov

Winston Hide
MasPar Computer Corporation
749 North Mary Avenue
Sunnyvale CA 94086
phone: 408-736-3300
fax: 408-736-9560
e-mail:

Jeff Himawan
Dept. of Biological Chem. & Mol. Phar.
Harvard Medical School
240 Longwood Avenue
Boston, MA 02115
phone:
fax:
e-mail:

Kathleen Hodgkins
100 N. Center St., Apt. 4
Westminster, MD 21157
phone: 410-876-4463 (home)
fax:
e-mail:

Susan M. G. Hoffman
Human Genome Center
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore, CA 94551-9900
phone: 510-423-7687
fax: 510-422-2282
e-mail: hoffman@cea.llnl.gov

Linda Holmes
Science and Engineering Education Division
Oak Ridge Institute for Science and
P.O. Box 117
Oak Ridge TN 37831-0117
phone: 615-576-3192
fax: 615-241-5219
e-mail: holmesl@orau.gov

Leroy H. Hood
Department of Molecular Biotechnology
University of Washington
FJ-20
4909 25th Avenue N.E.
Seattle, WA 98195
phone: 206-685-4340
fax: 206-685-7301
e-mail: molbiotk@u.washington.edu

Herbert Hooper
Soane Technologies, Inc.
3916 Trust Way
Hayward CA 94545
phone: 510-293-1850
fax: 510-293-1860
e-mail:

Eliezer Huberman
Center for Mechagnistic Biology and
Argonne National Laboratory
Bldg. 202, 9700 South Cass Avenue,
Argonne, IL 60439-4833
phone: 708-252-3819
fax: 708-252-3853
e-mail:

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Suzanne Huebner
National Center for Genome Resources
1800 Old Pecos Trail, Suite E
Santa Fe, NM 87505
phone: 505-982-7840
fax: 505-982-7690
e-mail:

Layne Huiet
Life Science Group
Bio-Rad Laboratories
2000 Alfred Nobel Drive
Hercules, CA 94547
phone: 510-741-1000
fax: 510-741-1060
e-mail:

Tim Hunkapiller
Dept. of Molecular Biotechnology
University of Washington
MS GJ-10
4909 25th Avenue N.E.
Seattle, WA 98195
phone: 206-685-7365
fax: 206-685-7363
e-mail: tim@mudhoney.mbt.washington.e

Greg Hurst
Analytical Chemistry Div.
Oak Ridge National Laboratory
P.O. Box 2008
Bldg. 4500S, MS 6120
Oak Ridge TN 37831-6365
phone: 615-574-4968
fax: 615-576-7956
e-mail: h9g@stc20.ctd.ornl.gov

Diane Isonaka
Darwin Molecular Corporation
22025 20th Avenue SE, Suite 1000
Bothell, WA 98021
phone: 206-489-8000
fax: 206-489-8019
e-mail:

Sorin Istrail
Dept. of Algorithms and Discrete Math.
Sandia National Labs
Albuquerque, NM 87185-1110
phone: 505-845-7612
fax:
e-mail: scistra@cs.sandia.gov

Igor B. Ivanov
Institute of Molecular Biology
Engelhardt Institute of Molecular Biology
Russian Academy of Sciences
32 Vavilov Str., B-334
Moscow 117984 Russia
phone: 011-7-095-135-9846
fax: 011-7-095-135-1405
e-mail: iigor@imb.msk.su

Dan Jacobson
Genome Data Base GDB
2024 E Monument
Baltimore MD 21205
phone:
fax:
e-mail: danj@gdb.org

K. Bruce Jacobson
Biology Division
Oak Ridge National Laboratory
P.O. Box 2009
Bldg 9211, MS 8080
Oak Ridge, TN 37831-8080
phone: 615-574-1204
fax: 615-574-1274
e-mail: JacobsonKB@ornl.gov or

Stephen C. Jacobson
Analytical Chemistry Division
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge TN 37831-6142
phone: 615-574-8002
fax: 615-574-8363
e-mail: z4j@ornl.gov

Joseph M. Jaklevic
Human Genome Center and Engineering
Lawrence Berkeley Laboratory
MS 70A-3363 F
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5647
fax: 510-486-5857
e-mail: JMJaklevic@lbl.gov

Margaret Jefferson
Department of Biology and Microbiology
California State University, Los Angeles
5151 State University Drive,
Los Angeles CA 90032-8201
phone: 213-343-2059
fax: 213-343-2095
e-mail: mjeffer@flytrap.calstatela.edu

James H. Jett
Center for Human Genome Studies, Life Sci.
Los Alamos National Laboratory
MS M888, LS-1
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-667-3843
fax: 505-665-3024
e-mail: jett@flovax.lanl.gov

Dabney K. Johnson
Biology Division
Oak Ridge National Laboratory
P.O. Box 2009
Bldg 9210, MS 8077
Oak Ridge TN 37831
phone: 615-574-0953
fax: 615-574-1283
e-mail: johnsondk@biovx1.bio.ornl.gov

Richard Johnston
Molecular Dynamics
928 E. Arques Avenue
Sunnyvale, CA 94086
phone: 408-737-3142
fax: 408-773-8343
e-mail:

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Robert Jones
Darwin Molecular
22025 20th Ave., Suite 1000
Bothell WA 98021
phone: 206-489-0983
fax:
e-mail: jones@darwin.com

Jingyue Ju
Depts. of Chemistry and Molecular and Cell
University of California, Berkeley
310 Hildebrand
Berkeley, CA 94720-1460
phone: 510-642-4192
fax: 510-642-3599
e-mail:

Jerzy W Jurka
Linus Pauling Institute of Science and
440 Page Mill Road
Palo Alto, CA 94306-2025
phone: 415-327-4064
fax: 415-327-8564
e-mail: jurka@jmullins.stanford.edu

Fa-Ten Kao
Biophysics and Genetics
Eleanor Roosevelt Institute
University of Colorado Health Sciences
1899 Gaylord Street
Denver, CO 80262
phone: 303-333-4515
fax: 303-333-8423
e-mail: kao@druid.hsc.colorado.edu

Barry L. Karger
Barnett Institute of Chemical Analysis and
Northeastern University
360 Huntington Avenue, 341 Mugar Ave.
Boston, MA 02115
phone: 617-373-2867/2868
fax: 617-373-2855
e-mail:

Richard M. Karp
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley CA 94704
phone: 510-642-4274
fax: 510-643-7684
e-mail: KARP@CS.Berkeley.EDU

Kanchi "K" Karunaratne
MS 70-3363
Lawrence Berkeley Laboratory
One Cyclotron Road
Berkeley CA 94720
phone: 510-486-6736
fax: 510-486-5857
e-mail: KSKarunaratne@lbl.gov

John Kececioglu
Divsion of Computer Science
University of California at Davis
Davis, CA 95616
phone:
fax:
e-mail:

Gifford Keen
National Center for Genome Resources
1800 Old Pecos Trail, Suite E
Santa Fe, NM 87505
phone: 505-982-7840
fax: 505-982-7690
e-mail:

Jan Kieleczawa
Biology Department
Brookhaven National Laboratory
P.O. Box 5000, Bldg. 463
Upton, NY 11973-5000
phone: 516-282-2133
fax: 516-282-3407
e-mail: KIEL@BNL.BIO.GOV

Ung-Jin Kim
Division of Biology, 147-75
California Instititute of Technology
MS 147-75
Pasadena, CA 91125
phone: 818-395-4154
fax: 818-796-7066
e-mail: ung@ash.tree.caltech.edu

William J. Kimmerly
Human Genome Center
Lawrence Berkeley Lab.
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-7330
fax: 510-486-6816
e-mail: kimmerly@genome.lbl.gov

David Kingsbury
Genome Data Base
Johns Hopkins University
2024 E. Monument Street
Baltimore MD 21205
phone: 410-955-9705
fax: 410-955-0985 / 410-614-0434
e-mail: dkingsbu@gdb.org

Emanuel Knill
Theoretical Biology & Biophysics
Los Alamos National Laboratory
T-10, K987
Los Alamos, NM 87545
phone: 505-665-8283
fax: 505-665-3493
e-mail: knill@lanl.gov

Arthur Kobayashi
Biology & Biotechnology Division
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore CA 94551-9000
phone: 510-422-8361
fax: 510-422-2282
e-mail:

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Bill Kolbe
Lawrence Berkeley Laboratory
MS 70A-3363
1 Cylcotron Road
Berkeley, CA 94720
phone: 510-486-7199 or 4650
fax: 510-486-5857
e-mail: WFKolbe@lbl.gov

Julie R. Korenberg
Department of Medical Genetics
Cedars-Sinai Medical Center, ASB 378
University of California
110 George Burns Road, Davis Bldg., Suite
Los Angeles, CA 90048-1869
phone: 310-855-7627
fax: 310-652-8010
e-mail: Korenberg@csmcmvax.bitnet

Paul M. Kraemer
Genomics & Structural Biology
Los Alamos National Laboratory
LS-2, MS M888
Los Alamos, NM 87545
phone: 505-667-2794
fax: 505-665-3024
e-mail: Kraemer@telomere.lanl.gov

Wen-Lin Kuo
Division of Molecular Cytology
University of California, San Francisco
Department of Laboratory Medicine
Box 0808, MCB 230
San Francisco, CA 94143-0808
phone: 415-476-3559
fax: 415-476-8218
e-mail:

Pui-Yan Kwok
Division of Dermatology
University of Washington, School of
Box 8123
660 S. Euclid Street
St. Louis MO 63110
phone: 314-362-8236
fax: 314-362-8159
e-mail: kwok@psts.wustl.edu

Jane E. Lamerdin
Human Genome Center
Lawrence Livermore National Lab.
P. O. Box 808, MS L-452
Livermore, CA 94550
phone:
fax:
e-mail:

Rich G. Langlois
Human Genome Center
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore, CA
phone: 510-422-5616
fax: 510-422-2282
e-mail: langlois1@llnl.gov

Vladimir Larionov
Laboratory of Molecular Genetics
National Institute of Health
P.O. Box 12233
Research Triangle NC 27709
phone:
fax:
e-mail:

Charles B. Lawrence
Department of Cell Biology
Baylor College of Medicine
One Baylor Plaza, Rm. M525
Houston, TX 77030
phone: 713-798-6226
fax: 713-798-3759
e-mail: chas@bcm.tmc.edu

Cheryl Lemanski
MS M880
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-667-1600
fax: 505-665-3858
e-mail:

Gregory G. Lennon
Biology and Biotechnology Research
Lawrence Livermore National Laboratory
P.O. Box 808, MS L-452
Livermore, CA 94551
phone: 510-422-5711
fax: 510-423-3608
e-mail: greg@mendel.llnl.gov

Stanley Letovsky
Genome Data Base at John Hopkins Univ.
2024 E. Monument St., Suite 1-200
Baltimore MD 21205-2100
phone: 410-955-9705
fax: 410-614-0434
e-mail: letovsky@gdb.org

Allison Levy
AT Biochem
30 Spring Mill Drive
Malvern PA 19355
phone: 800-282-4626
fax: 610-889-9304
e-mail:

Peter Li
Genome Data Base at Johns Hopkins U.
2024 E. Monument Street
Baltimore MD 21205
phone: 410-614-0439
fax: 410-614-0434
e-mail: peterli@gdb.org

Kimberly Lieuallen
L-452
Lawrence Livermore National Laboratory
P.O. Box 808, Bldg. 361/1850
Livermore, CA 94550
phone: 510-422-5651
fax: 510-422-2282
e-mail: Kim@mendel.llnl.gov

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Lo-See (Lucy) Ling
Collaborative Research, Inc.
1365 Main Street
Waltham, MA 02154
phone: 617-487-7979
fax: 617-891-5062
e-mail: LING@CRI.COM

Robert J. Lipshutz
Affymetrix
3380 Central Expressway
Santa Clara, CA 95051
phone: 408-522-6010
fax: 408-481-0422
e-mail:

Rebecca Lobb
MS M880
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-665-0781
fax: 505-665-3858
e-mail:

Valery Logan
Dept. of Molecular Biotechnology
University of Washington,
FJ-20
Seattle, WA 98195
phone: 206-685-7367
fax: 206-685-7301
e-mail: molbiotk@u.washington.edu

Jonathan Longmire
Center for Human Genome Studies, Life Sci.
Los Alamos National Laboratory
MS M886
Los Alamos, NM 87545
phone: 505-667-8208
fax: 505-665-3024
e-mail:

Michael Lowenstein
Life Sciences, Genomic & Structural Biology
Los Alamos National Laboratory,
LS-2, MS M880
Los Alamos, NM 87645
phone: 505-667-2726
fax: 505-665-3024
e-mail:

Ron Lundstrom
Collaborative Research, Inc.
1365 Main Street
Waltham, MA 02154
phone: 617-487-7979
fax: 617-891-5062
e-mail: ron@cric.com

Veronica M. Lustre
Human Genome Center
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5915
fax:
e-mail:

William P. MacConnell
MacConnell Research Corporation
11408 Sorrento Vallley Road, Suite 202
San Diego CA 92121
phone: 619-452-2603
fax: 619-452-6753/2603
e-mail:

Catherine A. Macken
Theoretical Biology and Biophysics Group
Los Alamos National Laboratory
Group T-10, MS K710
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-665-1970
fax: 505-665-3493
e-mail:

Betty K. Mansfield
Human Genome Management Information
Oak Ridge National Laboratory
1060 Commerce Park, MS 6480
Oak Ridge TN 37830
phone: 615-576-6669
fax: 615-574-9888
e-mail: bkq@ornl.gov

Elaine R. Mardis
Genome Sequencing Center
Washington University School of Medicine
Box 8501
4444 Forest Park Ave.
St. Louis MO 63108
phone: 314-286-1805
fax: 314-286-1810
e-mail: emardis@watson.wustl.edu

Ray Mariella
Biology & Biotechnology Division
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore CA 94551-9000
phone: 510-422-8361
fax: 510-422-2282
e-mail:

Victor M. Markowitz
Information Services
Lawrence Berkeley Laboratory
MS 50B-3238
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-6835
fax: 510-486-4004
e-mail: VMMarkowitz@lbl.gov

David Marquette
Genome Data Base at John Hopkins Univ.
2024 E. Monument St., Suite 1-200
Baltimore MD 21205-2100
phone: 410-955-9705
fax: 410-614-0434
e-mail: ddm@gdb.org

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Thomas G. Marr
Cold Spring Harbor Laboratory
P.O. Box 100
100 Bungtown Road
Cold Spring NY 11724
phone: 516-367-8393
fax: 516-367-8461
e-mail: marr@cshl.org

Babetta L. Marrone
LS-1, M888
Los Alamos National Laboratory
Los Alamos, NM 87545
phone: 505-667-3279
fax: 505-665-3024
e-mail: blm@lanl.gov

Christopher H. Martin
Human Genome Center
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5654
fax: 510-486-7282
e-mail: chrism@genome.lbl.gov

Christopher S. Martin
Attn: Anne Marie Youell
Tropix, Inc.
47 Wiggins Avenue
Bedford, MA 01730-2314
phone: 617-271-0045
fax: 617-275-8581
e-mail:

John C. Martin
Center for Human Genome Studies
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545
phone:
fax:
e-mail:

Richard A. Mathies
Depts. of Chemistry and Molecular and Cell
University of California, Berkeley
310 Hildebrand
Berkeley, CA 94720
phone: 510-642-4192
fax: 510-642-3599
e-mail: rich@zinc.cchem.berkeley.edu

Carol Mayeda
Lawrence Berkeley Laboratory
1 Cyclotron Road
MS 74-157
Berkeley, CA 94720
phone: 510-486-5915
fax:
e-mail:

Linda McCamant
Science and Engineering Education Division
Oak Ridge Institute for Science and
P.O. Box 117
Oak Ridge, TN 37831-0117
phone: 615-576-1089
fax: 615-241-5219
e-mail: mccamanl@orau.gov

John McCarthy
Informatics and Computing Science Dept.
Lawrence Berkeley Laboratory
MS 50B-3238
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5307
fax: 510-486-4004
e-mail: jlmccarthy@lbl.gov

Jean E. McEwen
Shriver Center for Mental Retardation, Inc.
200 Trapelo Road
Waltham, MA 02254
phone: 617-552-4381
fax:
e-mail:

Joseph D. McInerney
Biological Sciences Curriculum Study
Colorado College
830 North Tejon, Suite 405
Colorado Springs, CO 80903-4720
phone: 719-578-1136
fax: 719-578-9126
e-mail:

David A. Mead
Life Science Group
Bio-Rad Laboratories
2000 Alfred Nobel Drive
Hercules, CA 94547
phone: 510-741-1000/ ext.6753
fax: 510-741-1060
e-mail: dmead@haley.genetics.bio-rad.co

Patricia A. Medvick
Engineering Sciences & Applications, ESA-6
Los Alamos National Laboratory
P.O. Box 1663
LANL MS J 580
Los Alamos, NM 87545
phone: 505-667-2676
fax: 505-665-3911
e-mail: pm@lanl.gov

Linda J. Meincke
Center for Human Genome Studies
Los Alamos National Laboratory
LS-2, MS M880
Los Alamos, NM 87545
phone: 505-665-1274
fax: 505-665-3024
e-mail:

Deirdre Meldrum
Electrical Engineering, FT-10
University of Washington
Seattle WA 98195
phone: 206-685-7639
fax: 206-543-3842
e-mail: deedee@uw-isdl.ee.washington.e

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

John D. Meng
Human Genome Center/Engineering
Lawrence Berkeley Laboratory
MS 70A-3363
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5117
fax: 510-486-5857
e-mail: JDMeng@lbl.gov

Julianne Meyne
Center for Human Genome Studies
Los Alamos National Laboratory
LS-2, MS M880
Los Alamos, NM 87545
phone: 505-667-2697
fax: 505-665-3024
e-mail:

Jerome P. Miksche
ARS, Plant Genome Project
United States Department of Agriculture
Bldg. 005, Rm. 331C, BARC-West
10300 Baltimore Blvd.
Beltsville, MD 20705-2350
phone: 301-504-6029
fax: 301-504-6231
e-mail: jmiksche@asrr_arsusda.gov

Andrei D. Mirzabekov
Director
Engelhardt Institute of Molecular Biology
Russian Academy of Sciences
Vavilov Str. 32, B-334
Moscow 117984 Russia
phone: 011-7-095-135-0559
fax: 011-7-095-135-1405
e-mail: amir@imb.msk.su

Harvey Mohrenweiser
Human Genome Center
Lawrence Livermore National Lab.
P. O. Box 808, L-452
7000 East Avenue
Livermore, CA 94551
phone: 510-423-0534
fax: 510-422-2282
e-mail: harvey@cea.llnl.gov

Don Moir
Collaborative Research, Inc.
1365 Main Street
Waltham, MA 02154
phone: 617-487-7979
fax: 617-891-5062
e-mail:

Robert Molinari
Protogene Laboratory
1615 Plymouth St.
Mountain View CA 94043
phone: 415-428-3810
fax: 415-428-1600
e-mail: protogen@netcom.com

Robert K. Moyzis
Director, Human Genome Center
Los Alamos National Laboratory
CHGS - MS M885
Los Alamos, NM 87545
phone: 505-667-3912
fax: 505-667-2891
e-mail: moyzis@telomere.lanl.gov

Mike Mucenski
Biology Division
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge TN 37831-8077
phone: 615-574-0703
fax: 615-574-1283
e-mail:

Michael Mueller
Molecular Applications Group
445 Sherman Avenue, Suite T
Palo Alto, CA 94306
phone: 415-473-3030
fax: 415-473-1795
e-mail: mueller@mag.com

Jim Mullikin
MS 74-157
Lawrence Berkeley Laboratory
1 Cyclotron Rd.
Berkeley CA 94720
phone: 510-486-5197
fax: 510-486-5343
e-mail: JCMullikin@lbl.gov

Christine Munk
MS A114
Los Alamos National Laboratory
Los Alamos, NM 87545
phone: 505-667-1600
fax: 505-665-3858
e-mail:

Maureen Munn
Dept. of Molecular Biotechnology
University of Washington, FJ-20
Seattle, WA 98195
phone: 206-685-7367
fax: 206-685-7301
e-mail: molbiotk@u.washington.edu

Richard Mural
Biology Division
Oak Ridge National Laboratory
P.O. Box 2008
Bldg.9211, MS 8080
Oak Ridge TN 37830-8077
phone: 615-576-2938
fax: 615-574-1274
e-mail: M9L@biovx1.bio.ornl.gov

Eugene W. Myers
Department of Computer Science
University of Arizona
Gould-Simpson Building,
Tucson, AZ 85721
phone: 602-621-6612
fax: 602-621-4246
e-mail: gene@cs.arizona.edu

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Cleo Naranjo
Center for Human Genome Studies
Los Alamos National Laboratory
Los Alamos, NM 87545
phone: 505-665-4438
fax: 505-665-3024
e-mail:

Mohandas Narla
Head, Human Genome Center
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-7029
fax: 510-486-6746
e-mail:

Mark Neff
Barker Hall, Sachs Laboratory
University of California, Berkeley
Berkeley, CA 94720
phone: 510-643-7908
fax: 510-643-5035
e-mail: neff@mendel.berkeley.edu

Christine Nelson
Dept. of Chemistry
University of Wisconsin
1101 University Avenue
Madison, WI 53706
phone: 608-263-2594
fax: 608-262-0381
e-mail:

David L. Nelson
Co-director, BCM Human Genome Center
Baylor College of Medicine
Dept. of Molecular & Human Genetics, Rm
One Baylor Plaza
Houston, TX 77030
phone: 713-798-4787
fax: 713-798-5386 or 6370
e-mail: nelson@bcm.tmc.edu

David O. Nelson
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
Livermore, CA 94551
phone: 510-423-8898
fax: 510-423-3608
e-mail: daven@stille.llnl.gov

Anne Nichols
Department of Molecular and Cell Biology
University of California, Berkeley
Linn Laboratory
401 Barker Hall
Berkeley, CA 94720
phone: 510-642-7522
fax: 510-643-5035
e-mail:

Deborah Nickerson
Dept. of Molecular Biotechnology, FJ-20
University of Washington
4909 25th Avenue, NE.
Seattle, WA 98195
phone: 206-685-7387
fax: 206-685-7344
e-mail: debnick@u.washington.edu

Rachel Oberai-Soltz
Digital Equipment Corp.
50 Nagog Park, AKO2-2/D8
Acton, MA 01720-3499
phone: 508-264-7844(in Acton)
fax: 508-264-7133
e-mail: roberai@akocoa.enet.dec.com

Anne Olsen
Biomedical Sciences Division
Lawrence Livermore National Lab.
P. O. Box 5507, L-452
Livermore, CA 94550
phone: 510-423-4927
fax: 510-423-3608
e-mail: olsen@ecor1.llnl.gov

Maynard V. Olson
Dept. of Molecular Biotechnology
Universtiy of Washington
MS GJ-10
Seattle, WA 98195
phone: 206-685-7366/46
fax: 206-685-7344
e-mail: mvo@u.washington.edu

Christian Oste
MS D-33-E
Beckman Instruments, Inc.
P.O. Box 3100
2500 Harbor Blvd.
Fullerton, CA 92634-3100
phone: 714-773-8481
fax: 714-773-7600
e-mail: ccoste@ccgate.dp.beckman.com

Ross Overbeek
Math and Computer Science Division
Argonne National Laboratory
MCS 221/D236
9700 S. Cass Avenue
Argonne, IL 60439
phone: 708-252-7856
fax: 708-972-5986
e-mail: overbeek@mcs.anl.gov

Chris Overton
Department of Genetics
University of Pennsylvania School of
CRB 475
422 Curie Blvd., Ste 1-200
Philadelphia, PA 19104-6145
phone: 215-573-3105
fax: 215-573-3111
e-mail: coverton@cbil.humgen.upen.edu

David J. Ow
Human Genome Center
Lawrence Livermore National Laboratory
P.O. Box 808, L452
Livermore, CA 94551-9900
phone: 510-423-8062
fax: 510-422-2282
e-mail: ow1@llnl.gov

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Michael Palazzolo
Human Genome Center
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5915
fax: 510-486-6816
e-mail: michaelp@genome.lbl.gov

Julia E. Parrish
Department of Molecular and Human
Baylor College of Medicine
One Baylor Plaza
Houston, TX 77030
phone: 713-798-7898
fax: 713-798-5386
e-mail: jparrish@bcm.tmc.edu

John Peeters
Environmental Safety and Health
Department of Energy
EH 43–GTN
Washington, DC 20585
phone: 301-903-5902
fax: 301-903-5072
e-mail: peeters@ux5.lbl.gov

Stepanie Pendergrass
Genomics & Structural Biology
Los Alamos National Laboratory
LS-1, MS M888
Los Alamos, NM 87545
phone: 505-667-2499
fax:
e-mail:

Ana Perez-Castro
Genomics & Structural Biology
Los Alamos National Laboratory
LS-2, MS M888
Los Alamos, NM 87545
phone: 505-667-0894
fax:
e-mail:

Jackie Perry
Research Assistant, Molecular Biology
INCYTE Pharmaceuticals, Inc.
3330 Hillview Avenue
Palo Alto, CA 94304
phone: 415-855-0555
fax: 415-855-0572
e-mail: jackie@qmgate.incyte.com

Ellen Peterson
Genomics & Structural Biology
Los Alamos National Laboratory
LS-2, MS M880
Los Alamos, NM 87545
phone: 505-665-5039
fax:
e-mail: etp@lanl.gov

Sergey Petrov
Martin Marietta Energy Systems, Inc.
Oak Ridge National Laboratory
P.O. Box 2008
Bldg 6025, MS-6364
Oak Ridge TN 37831-6364
phone: 615-574-8934
fax: 615-574-7860
e-mail: petrovs@ornl.gov

Jeffrey T. Petty
Genomics & Structural Biology
Los Alamos National Laboratory
CST-2, MS M888
Los Alamos, NM 87545
phone: 505-665-2092
fax: 505-665-3024
e-mail:

Pavel Pevzner
Department of Computer Science &
Pennsylvania State University
University Park PA 16803
phone: 814-863-3599
fax: 814-865-3176
e-mail: pevzner@cse.psu.edu

Liz Pham
Robbins Scientific
814 San Aliso Ave.
Sunnyvale CA 94086
phone: 408-734-8400
fax:
e-mail:

Daniel Pinkel
Department of Laboratory Medicine,
University of California, San Francisco
513 Parnassus
San Francisco CA 94143-0808
phone: 415-476-3659
fax: 415-476-8218
e-mail: pinkel@dine.ucsf.edu

Michael C. Pirrung
Department of chemistry
Duke University
P.M. Gross Chemical Lab
Box 90346
Durham, NC 27708-3046
phone: 919-660-1556
fax: 919-660-1591
e-mail: chem@chem.duke.edu

Samuel (Sam) Pitluck
Engineering
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-6317
fax: 510-486-6816
e-mail: s_pitluck@lbl.gov

Oleg L. Polanovsky
Laboratory of Molecular Genetic
Engelhardt Institute of Molecular Biology
Russian Academy of Sciences
Vavilov str 32
Moscow 117984 Russia
phone: 011-7-095-135-2311
fax: 011-7-095-135-14-05
e-mail: pol@imb.msk.su

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Marty Pollard
Human Genome Center
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-4561
fax: 510-486-6816
e-mail: mjpollard

Mihael Polymeropoulos
National Center for Human Genome
National Institute of Health
49 Convent Dr. MSC 4470, Room 4A68
Bethesda, MD 20892-4470
phone: 301-402-2035 (lab)
fax: 301-402-2170
e-mail:

Ken Porter
Departement of Chemistry
Duke University
Box 90349
101 Gross Chemistry Labs
Durham NC 27708
phone: 919-660-1500
fax:
e-mail:

Dan Prestridge
Dir., Molecular Biology Computer Center
University of Minnesota
1479 Gortner Ave.
St. Paul MN 55108
phone: 612-625-3744
fax: 612-625-5780
e-mail: danp@molbio.imn.edu

Theodore T. Puck
Director and Senior Fellow
Eleanor Roosevelt Inst. for Cancer Research
University of Colorado, Health Sciences
1899 Gaylord Street
Denver CO 80206-1210
phone: 303-333-4515
fax: 303-333-8423
e-mail:

Mark A. Quesada
Brookhaven National Laboratory
P.O. Box 5000, Bldg. 463
Upton, NY 11973-5000
phone: 516-282-3390
fax: 516-282-3407
e-mail:

Barbara Ramsay Shaw
P.M. Gross Chemical Laboratory
Duke University
Box 90346
Durham, NC 27708-0346
phone: 919-660-1551
fax: 919-660-1605
e-mail:

J. Michael Ramsey
Analytical Chemistry Division, Optical
Oak Ridge National Laboratory
P. O. Box 2008
Bldg. 4500S, C-158
Oak Ridge TN 37831-6142
phone: 615-574-5662
fax: 615-574-8363
e-mail: ramseyjm@ornl.gov

Richard Rava
Affymetrix
3380 Central Expressway
Santa Clara, CA 95051
phone: 408-522-6010
fax: 408-481-0920
e-mail:

David E. Reichle
Environmental, Life & Social Sciences
Oak Ridge National Laboratory
P. O. Box 2008, MS-6253
Bldg. 4500N, MS 6253
Oak Ridge TN 37831-6253
phone: 615-574-4333
fax: 615-574-9869
e-mail: der@ornl.gov/ reichlede@ornl.gov

Philip R. Reilly, JD
President, Chief Executive Officer
Shriver Center
Center for Mental Retardation
200 Trapelo Road
Waltham, MA 02254
phone: 617-642-0230
fax: 617-893-5340
e-mail:

Peter Richterich
Collaborative Research, Inc.
100 Beaver Street
Waltham, MA 02154
phone: 617-487-7979/ ext.220
fax: 617-487-7960
e-mail: peter@neptun.cric.com

Arthur D. Riggs
Biology Department
Beckman Research Institute
City of Hope
1450 East Duarte Road
Duarte, CA 91010-0269
phone: 818-301-8352
fax: 818-358-7703
e-mail:

Donna Robinson
Center for Human Genome Studies
Los Alamos National Laboratory
Los Alamos, NM 87545
phone: 505-665-4438
fax: 505-665-3024
e-mail:

Patricia A. Roche
Health Law Department
Boston University
80 East Concord Street
Boston MA 02118
phone: 617-638-4626
fax: 617-638-5299
e-mail:

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Lori Romberg
Sequencing Lab. at McDermott Center
University of Texas Southwestern Medical
6000 Harry Hines Blvd.
Dallas, TX 75235-8591
phone: 214-648-1622
fax: 214-648-1666
e-mail: romberg@swmed.edu

Mary C. Roslaniec
Genomics & Structural Biology
Los Alamos National Laboratory
LS-1, MS M888
Los Alamos, NM 87545
phone: 505-667-2788
fax:
e-mail:

Darlene B. Roszak
Ransom Hill Bioscience
P.O. Box 219
Ramona, CA 92065
phone: 619-789-9483
fax: 619-789-6902
e-mail:

Lee Rowen
Dept. of Molecular Biotechnology, GJ-10
University of Washington
Seattle, WA 98195
phone: 206-685-7337/7367
fax: 206-685-7344/7301
e-mail: molbiotk@u.washington.edu

Eddy Rubin
Human Genome Center
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5072 or 6714
fax: 510-486-6746
e-mail:

Gerald Rubin
Department of Molecular and Cell Biology
University of California, Berkeley
539 Life Sciences Addition Building
Berkeley, CA 94720
phone: 510-643-9945
fax: 510-643-9947
e-mail:

Jeffrey D. Saffer
Life Sciences Center (K1-50)
Pacific Northwest Laboratory
Mail Stop K1-50
P. O. Box 999
Richland, WA 99352
phone: 509-375-3738
fax: 509-375-6821/3649?
e-mail:

Arbansjit K. Sandhu
School of Medicine, Fels Institute
Temple University
Medical Research Building
3420 N. Broad Street
Philadelphia, PA 19140
phone: 215-707-4300
fax: 215-707-4318
e-mail:

Bruce Sawhill
Theoretical Biology & Biophysics
Los Alamos National Laboratory
T-10, K710
Los Alamos, NM 87545
phone: 505-665-
fax: 505-665-3493
e-mail: sawhill@lanl.gov

Ira Schildkraut
New England BioLabs
32 Tozer Road
Beverly, MA 01915-5510
phone: 508-927-5054/ext. 214
fax: 508-921-1527
e-mail: schildkr@neb.com

Richard D. Schneider
Life Sciences Center (K1-50)
Pacific Northwest Laboratory
Mail Stop K1-50
P. O. Box 999
Richland, WA 99352
phone: 509-375-3738
fax: 509-375-6821/3649?
e-mail:

Jocelyn Schultz
Human Genome Center
Lawrence Berkeley Laboratory
MS 70A-3363
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5324 or 4112
fax:
e-mail:

Annette Schwartz
Biology & Biotechnology Division
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore CA 94551-9000
phone: 510-422-8361
fax: 510-422-2282
e-mail:

David B. Searls
Department of Genetics, 475 CRB
University of Pennsylvania School of
422 Curie Boulevard
Philadelphia, PA 19104-6145
phone: 215-573-3107
fax: 215-573-3111
e-mail: dsearls@cbil.humgen.upenn.edu

Jeffrey J. Seilhamer
Vice President,
INCYTE Pharmaceuticals, Inc.
3330 Hillview Avenue
Palo Alto, CA 94304
phone: 415-855-0555
fax: 415-855-0572
e-mail: jeff@incyte.com

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Mary Ann Sesma
Los Angeles Unified School District
1621 Sunnyhill Drive
Monterey Park, CA 91754
phone: 213-261-5860
fax: 213-261-8692
e-mail: msesma@marlborough.la.ca.us

Valerie P. Setlow
Health Sciences Policy Division
Institute of Medicine, National Academy of
2101 Constitution Avenue
Washington DC 20418
phone: 202-334-2351
fax: 202-334-1385
e-mail: vsetlow@nas.edu

Lowell E. Sever
Life Sciences Center (K1-50)
Pacific Northwest Laboratory
Mail Stop K1-50
P. O. Box 999
Richland, WA 99352
phone: 509-375-3738
fax: 509-375-6821/3649?
e-mail:

Farideh Shadravan
Human Genome Center
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 97420
phone: 5963
fax: 5343
e-mail:

Manesh Shah
Engineering Physics & Math Division
Oak Ridge National Laboratory
P.O. Box 2008
Bldg.6025, MS 6364
Oak Ridge TN 37831-6050
phone: 615-574-6134
fax: 615-574-7860
e-mail: X9M@ornl.gov

Mark Shannon
Biology Division
Oak Ridge National Laboratory
P.O. Box 2009
Bldg 9210, MS 8077
Oak Ridge TN 37831-8077
phone: 615-574-1237
fax: 615-574-1283
e-mail:

Peggy Sharp
Human Genome Center
Lawrence Livermore National Laboratory
P.O. Box 808, L452
Livermore, CA 94551-9900
phone: 510-423-8062
fax: 510-422-2282
e-mail:

Kathy Shera
Center for Human Genome Studies
Los Alamos National Laboratory
MS-434
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-667-3228
fax: 505-665-3644
e-mail:

Dong-Guk Shin
Computer Science & Engineering Dept.
University of Connecticut
260 Glenbrook Rd., U-155
Storrs CT 06269-3155
phone: 203-486-3719
fax:
e-mail: shin@cse.uconn.edu

Hiroaki Shizuya
Division of Biology
California Institute of Technology
147-75
Pasadena, CA 91125
phone: 818-395-4154
fax: 818-796-7066
e-mail: shizuya@caltech.edu

Chris Shumate
Hamilton Company
4970 Energy Way
Reno NV 89502
phone: 800-648-5950/ext.255
fax: 702-856-7259
e-mail:

Melvin I. Simon
Biology Division
California Institute of Technology
147-75
1201 East California Boulevard
Pasadena, CA 91125
phone: 818-395-3944
fax: 818-796-7066
e-mail:

Linda Sindelar
Human Genome Center
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-4817
fax: 510-486-6816
e-mail: LESindelar@lbl.goc

Thomas R. Slezak
Human Genome Center
Lawrence Livermore National Laboratory
P. O. Box 5507, L-452
7000 East Avenue
Livermore, CA 94550
phone: 510-422-5746
fax: 510-423-3608
e-mail: slezak@llnl.gov or

Frederick V. Sloop
Biology Division
Oak Ridge National Laboratory
Bldg. 4500S, MS 6119
P.O. Box 2008
Oak Ridge TN 37831-6119
phone: 615-574-6716
fax: 615-574-4939
e-mail: sloopfvjr@ornl.gov

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

David Smith
Director, Health Effects & Life Sciences
U.S. Department of Energy
ER-72
Washington, DC 20585
phone: 301-903-5468
fax: 301-903-8521
e-mail:

Desmond J. Smith
Life Sciences Division
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5090
fax: No fax
e-mail:

Douglas Smith
Collaborative Research Inc.
1365 Main Street
Waltham MA 02154
phone: 617-487-7979
fax:
e-mail:

Lloyd M. Smith
Department of Chemistry; Analytical
University of Wisconsin-Madison
1101 University Avenue
Madison, WI 53706-1396
phone: 608-263-2594
fax: 608-262-0453
e-mail: smith@bert.chem.wisc.edu

Randall Smith
Dept. of Molecular and Human Genetics;
Baylor College of Medicine
One Baylor Plaza, Room T921
Houston, TX 77030
phone: 713-798-4735
fax: 713-798-6521
e-mail: rsmith@bcm.tmc.edu

Richard D. Smith
Life Sciences Center, P8-19/326 Bld/300
Pacific Northwest Laboratory
MS K1-50
P.O. Box 999
Richland, WA 99352
phone: 509-376-0723/5665
fax: 509-376-0418
e-mail: rd_smith@pnl.gov

Tracy Smull
MasPar Computer Corporation
749 North Mary Avenue
Sunnyvale CA 94086
phone: 408-736-3300
fax: 408-736-2942
e-mail: smull@maspar.com

David Soane
Soane Technologies, Inc.
3916 Trust Way
Hayward, CA 94545
phone: 510-293-1850
fax: 510-293-1860
e-mail:

Marcelo Bento Soares
Dept. of Psychiatry, College of Physicians &
Columbia University
722 West 168th Street, Box #41
New York, NY 10032
phone: 212-960-2313
fax: 212-781-3577
e-mail: cuc@cuccfa.ccc.columbia.edu

Terence Speed
Statistics Dept., Chair
University of California, Berkeley
327 Evans Hall
Berkeley, CA 94720
phone: 510-642-0613/2781
fax: 510-642-7892
e-mail: TERRY@STAT.Berkeley.EDU

Sylvia Spengler
Lawrence Berkeley Laboratory
1 Cyclotron Road
MS Donner 459
Berkeley CA 94720
phone: 510-486-5874
fax: 510-486-5717
e-mail: SJSpengler@lbl.gov

David J. States
Biomedical Computer Lab
Washington University
Campus Box 8068
700 South Euclid Avenue
St. Louis Misso 63110
phone: 314-362-2135
fax: 314-362-0234
e-mail: states@ibc.wustl.edu

John Steinkamp
MS A114
Los Alamos National Laboratory
Los Alamos, NM 87545
phone: 505-667-1600
fax: 505-665-3858
e-mail:

A.G. Stepchenko
Institute of Molecular Biology
Engelhardt Institute of Molecular Biology
Russian Academy of Sciences
Vavilov Str. 32, B-334
Moscow 117984 Russia
phone: 011-7-095-135-0559
fax: 011-7-095-135-1405
e-mail:

Marvin Stodolsky
Human Genome Task Group
HELSRD, U.S. Department of Energy
ER-72 GTN
Washington, DC 20585
phone: 301-903-4475
fax: 301-903-8521
e-mail: Marvin.Stodolsky%er@mailgw.er.d

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Gary D. Stormo
Dept. of Molecular, Cellular & Develop.
University of Colorado
Campus Box 347
Boulder, CO 80309-0347
phone: 303-492-1476
fax: 303-492-7744
e-mail: stormo@boulder.colorado.edu

Robert Strausberg
Assistant to the Dir. for Tech.
National Center for Human Genome
Bldg. 38A Room 610
9000 Rockville Pike
Bethesda, MD 20892
phone: 301-496-7531
fax: 301-480-2770
e-mail: cxr@cu.nih.gov

Gary Striniste
Los Alamos National Laboratory
LS1 MS M888
Los Alamos, NM 87545
phone: 505-667-2737
fax: 505-665-3024
e-mail:

Lisa Stubbs
Biology Division
Oak Ridge National Laboratory
Bldg 9210, MS 8077
P.O. Box 2009
Oak Ridge TN 37831-8077
phone: 615-574-0848
fax: 615-574-1283
e-mail: stubbs@biovx1.bio.ornl.gov/

F. William Studier
Biology Department
Brookhaven National Laboratory
P.O. Box 5000
Upton, NY 11973-5000
phone: 516-282-3390 or 3012
fax: 516-282-3407
e-mail: studier@genome1.bio.bnl.gov

Mark Stump
Human Genetics Dept.
University of Utah
2100 Eccles Genetics Bldg.
Salt Lake City, UT 84112
phone: 801-585-5173
fax: 801-585-3910 (weiss)
e-mail:

Dazhong Sun
MS 74-157
Lawrence Berkeley Laboratory
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5915
fax: 510-486-7282
e-mail:

Ernest Szeto
ICSD
Lawrence Berkeley Laboratory
MS 50B-3238
1 Cyclotron Road
Berkeley CA 94720
phone: 510-486-6411
fax: 510-486-4004
e-mail: E_Szeto@lbl.gov

Stanley Tabor
Dept. of Biological Chem. & Mol. Phar
Harvard Medical School
240 Longwood Avenue
Boston, MA 02115
phone: 617-432-3128
fax: 617-738-0516
e-mail:

Judy Tesmer
MS A114
Los Alamos National Laboratory
Los Alamos, NM 87545
phone: 505-667-1600
fax: 505-665-3858
e-mail:

Edward H. Theil
MS 46A-1120
Lawrence Berkeley Laboratory
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-7501
fax: 510-486-5926
e-mail: EHTheil@lbl.gov

Clark Tibbetts
Microbiology
Vanderbilt University School of Medicine
Nashville TN 37232-2363
phone: 615-322-3375
fax: 615-343-7392
e-mail: tibbetts@ctrvax.vanderbilt.edu

David Torney
Theoretical Biology & Biophysics
Los Alamos National Laboratory
T-10, K710
Los Alamos, NM 87545
phone: 505-667-9452
fax: 505-665-3493
e-mail: dct@ipmati.lanl.gov

Barbara J. Trask
Dept. of Molecular Biotechnology, FJ-20
University of Washington School Medicine
Fluke Hall Room 235
Seattle, WA 98195
phone: 206-685-7347
fax: 206-685-7354
e-mail: trask@biotech.washington.edu

Ralph W. Trottier
Dept. of Pharmacology and Toxicology
Morehouse School of Medicine
720 Westview Drive, S.W.
Atlanta, GA 30310-1495
phone: 404-752-1710
fax: 404-756-8675
e-mail:

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Susan Tsujimoto
Human Genome Center
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore, CA 94551-9900
phone: 510-422-1934
fax: 510-422-2282
e-mail: susant@kafka.llnl.gov

Annette Tumolo
Genetic Systems Division
Bio-Rad Laboratories, Inc.
2000 Alfred Nobel Drive
Hercules, CA 94547
phone: 510-741-1000
fax: 510-741-1060
e-mail:

Waneta Tuttle
Southwest Medical Ventures
2309 Renard Pl. SE # 204
Albuquerque NM 87106
phone: 505-764-0174
fax: 505-764-0074
e-mail:

Donald C. Uber
Human Genome Center
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-6378
fax: 510-486-6816
e-mail: DCUber@lbl.gov

Edward Uberbacher
Engineering Physics & Mathematics Div.
Oak Ridge National Laboratory
Bldg 6025, MS 6364
P.O. Box 2008, Bldg. 6025
Oak Ridge TN 37831-6364
phone: 615-574-6134
fax: 615-574-7860
e-mail: ube@ornl.gov

Levy Ulanovsky
Dept. of Structural Biology
Weizmann Institute
Rehovot Israel 76100
phone: 972-8-343-547
fax: 972-8-344-105
e-mail: bplevy@weizmann.weizmann.ac.il

Gerrit J. van den Engh
Dept. of Molecular Biotechnology, FJ-20
University of Washington School of
Fluke Hall
4909 25th Avenue N.E.
Seattle, WA 98195
phone: 206-685-7345
fax: 206-685-7354
e-mail: engh@biotech.washington.edu

J. Craig Venter
Institute for Genomic Research
932 Clopper Road
Gaithersburg, MD 20878
phone: 301-869-9056
fax: 301-869-9423/977-7286
e-mail:

Amy Voltz
Genome Data Base at John Hopkins Univ.
2024 E. Monument St., Suite 1-200
Baltimore MD 21205-2100
phone: 410-955-9705
fax: 410-614-0434
e-mail: areiner@welchlink.welch.jhu.edu

Jean-Michel Vos
Campus Box 7295, School of Medicine
University of North Carolina at Chapel Hill
UNC Lineberger Comprehensive Cancer
Chapel Hill, NC 27599-7295
phone: 919-966-3036
fax: 919-966-3015
e-mail: vos@med.unc.edu

Mark Wagner
L-156
Lawrence Livermore National Lab.
P. O. Box 808
Livermore, CA 94551
phone: 510-422-2866
fax:
e-mail: mwagner@llnl.gov

Robert P. Wagner
MS M880
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87544
phone:
fax:
e-mail: wagner@flovax.lanl.gov

Carol M. Warner
Department of Biology
Northeastern University
Boston, MA 02115
phone: 617-373-4036
fax:
e-mail:

Janet A. Warrington
Department of Genetics
Stanford Medical Center, Stanford
300 Pasteur Rd. M314
Stanford, CA 94305-5120
phone: 415-725-8034
fax: 415-725-1534
e-mail: jaw@camis.stanford.edu

Heinz-Ulrich Weier
Life Sciences Division
Lawrence Berkeley Laboratory
MS 74-157
1 Cyclotron Road
Berkeley CA 94720
phone: 510-486-5347
fax: 510-486-5343
e-mail: weier@white.lbl.gov

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Robert Weiss
Human Genetics Dept.
University of Utah
2100 Eccles Genetics Bldg.
Salt Lake City, UT 84112
phone: 801-585-3435
fax: 801-585-3910
e-mail: weiss@gene1.med.utah.edu

Alan F. Westin
Center for Social and Legal Research
2 University Plaza, Suite 414
Hackensack, NJ 07601
phone: 201-996-1154
fax: 201-996-1883
e-mail: alanrp@aol.com

Michael Westphall
Dept. of Chemistry
University of Wisconsin
1101 University Avenue
Madison, WI 53706
phone: 608-263-2594
fax: 608-262-0381
e-mail:

Thomas Whaley
Genomics & Structural Biology Group
Los Alamos National Laboratory
Group LS-2, MS M880
Los Alamos, NM 87545
phone: 505-667-2726
fax: 505-665-3024
e-mail:

Peter Williams
Dept. of Chemistry and Biochemistry
Arizona State University
Tempe, AZ 85287-1604
phone: 602-965-4107
fax: 602-965-2747
e-mail: pw@asu.edu

Richard Wilson
Genetics
Washington University School of Medicine
Genome Sequencing Center, Box 8501
4444 Forest Park Ave.
St. Louis MO 63108
phone: 314-286-1804
fax: 314-286-1810
e-mail: rick@geneman.wustl.edu

Frank Witney
Life Science Group
Bio-Rad Laboratories
2000 Alfred Nobel Drive
Hercules, CA 94547
phone: 510-741-1000
fax: 510-741-1060
e-mail:

Sheldon Wolff
Dir., Lab. of Radiobiology & Environ. Health
University of California, San Francisco
LR102, Box 0750
3rd & Parnassus Avenues
San Francisco, CA 94143
phone: 415-476-1636
fax: 415-476-0721
e-mail: shelly@radlab.ucsf.edu

Ben Wong
Biology & Biotechnology Division
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
Livermore CA 94551-9000
phone: 510-422-8361
fax: 510-422-2282
e-mail:

John Wooley
OHER, Director of Information & Resources
U.S. Department of Energy
ER71, GTN
Washington, DC 20585
phone: 301-903-3153
fax: 301-903-5051
e-mail:

Richard P. Woychik
Mammalian Genetics, Biology Division
Oak Ridge National Laboratory
P.O. Box 2009, MS 8077
Bear Creek Road, Y-12
Oak Ridge TN 37831-8077
phone: 615-574-3966
fax: 615-574-1271
e-mail: woychik@biovx1.bio.ornl.gov

James Wright
Science and Engineering Education Division
Oak Ridge Institute for Science and
P.O. Box 117
Oak Ridge TN 37831-6117
phone: 615-576-1716
fax: 615-241-2727
e-mail:

Jung-Rung Wu
Genomics & Structural Biology Group
Los Alamos National Laboratory
Group LS-2, MS M880
Los Alamos, NM 87545
phone: 505-667-2726
fax: 505-665-3024
e-mail:

T. Mimi Yeh
Human Genome Center
Lawrence Livermore National Laboratory
P.O. Box 808, L-452
7000 East Avenue
Livermore, CA 94550
phone: 510-422-2315
fax:
e-mail: myeh@llnl.gov

Michael S. Yesley
Center for Human Genome Studies
Los Alamos National Lab.
MS A187
P.O. Box 1663
Los Alamos, NM 87545
phone: 505-665-2523
fax: 505-665-4424
e-mail: yesley_michael_s@ofvax.lanl.gov

List of Potential Attendees
DOE Contractor–Grantee Meeting
November 13-17, 1994

Edward S. Yeung
Department of Chemistry
Ames Laboratory
Iowa State University
Ames, IA 50011-3020
phone: 515-294-8062
fax: 515-294-0266
e-mail: yeung@ameslab.gov

Jingwei Yu
Eleanor Roosevelt Institute
1899 Gaylord Street
Denver, CO 80206
phone: 303-333-4515
fax: 303-333-8423
e-mail: yuj@essex.hsc.colorado.edu

Jung Yue
Depts. of Chemistry and Molecular and Cell
University of California, Berkeley
310 Hildebrand
Berkeley, CA 94720
phone: 510-642-4192
fax: 510-642-3599
e-mail: rich@zinc.cchem.berkeley.edu

Jack S. Zelver
Lawrence Berkeley Laboratory
MS 70A/3363
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-7522
fax: 510-486-5857
e-mail: JSZelver@lbl.gov

Shiping Zhang
Biology Department, Bldg 463
Brookhaven National Laboratory
P.O. Box 5000
Upton, NY 11973-5000
phone: 516-282-3390
fax: 516-282-3407
e-mail:

Manfred Zorn
MS 50B-3238
Lawrence Berkeley Laboratory
1 Cyclotron Road
Berkeley, CA 94720
phone: 510-486-5041
fax: 510-486-4004
e-mail: MDZorn@lbl.gov

Franklin M. Zweig
Center for Health Policy Research, Law and
George Washington University
3122 Brooklawn Terrace
Chevy Chase, MD 20815
phone: 301-913-0448
fax: 301-913-5739
e-mail: