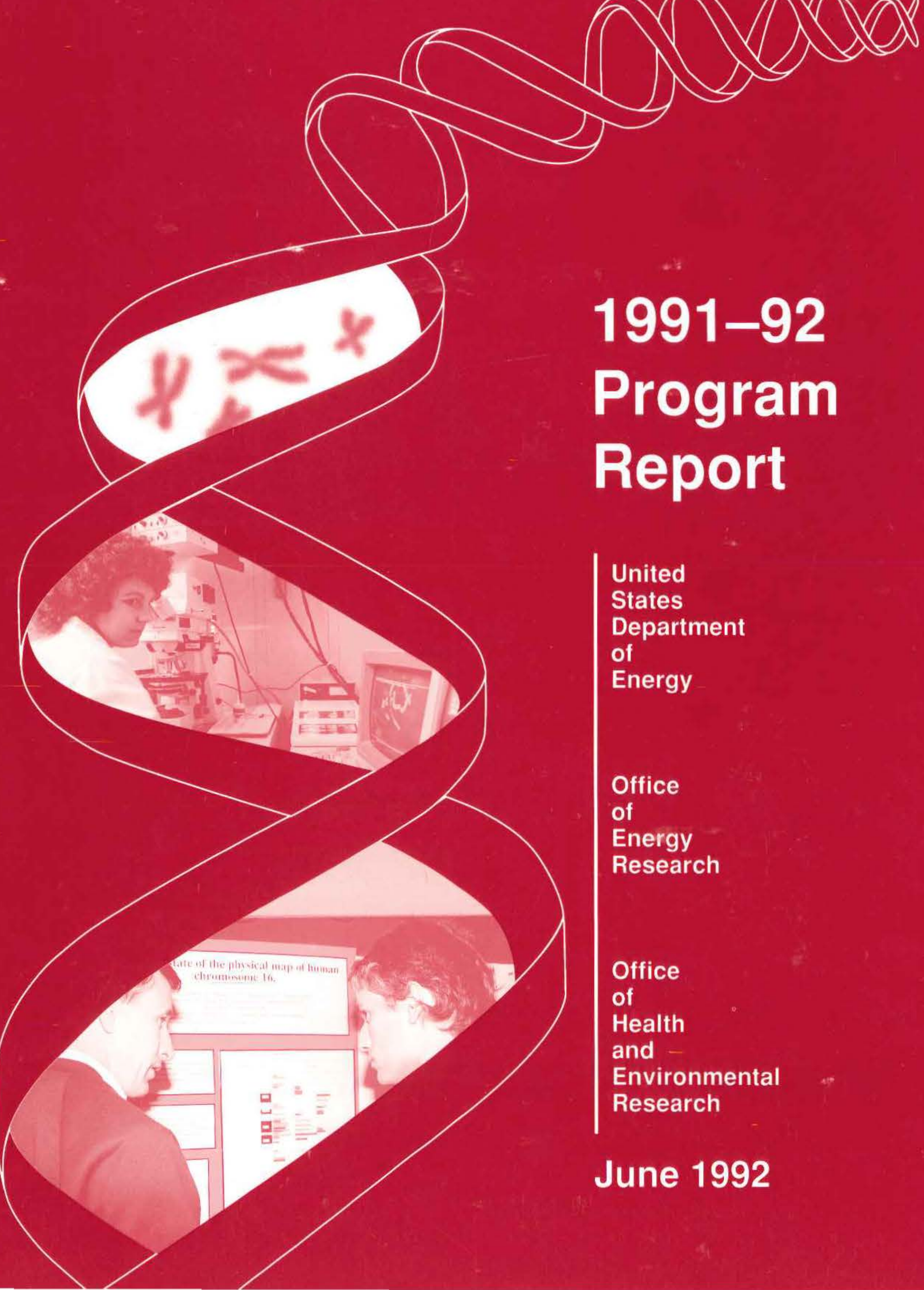
DOE/ER-0544P

Human Genome



1991–92 Program Report

United
States
Department
of
Energy

Office
of
Energy
Research

Office
of
Health
and
Environmental
Research

June 1992

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: 301/903-5051
Internet: "drell@mailgw.er.doe.gov"

Human Genome Management
Information System
Oak Ridge National Laboratory
P.O. Box 2008
Oak Ridge, TN 37831-6050
615/576-6669, Fax: 615/574-9888
BITNET: "bkq@ornlstc"
Internet: "bkq@ornl.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.

DOE/ER-0544P

Human Genome

1991–92 Program Report

Date Published: June 1992

U.S. Department of Energy
Office of Energy Research
Office of Health and Environmental Research
Washington, D.C. 20585

Major Events in the Development of the DOE Human Genome Program

Meeting held in Santa Fe, New Mexico, to explore Human Genome Initiative feasibility

DOE Human Genome Initiative announced

Pilot projects pursued at national laboratories: physical mapping, informatics, and development of critical resources and technologies for genomic analysis

HERAC recommendations made for the Human Genome Initiative

First interagency workshop held on genomic resources and informatics

LBL and LANL designated as human genome centers

First management and program plans published

First peer-review panel for human genome research proposals held

NRC recommendations made for a national human genome project

First primary R&D awards made

First Small Business Innovative Research awards made in genome research

DOE-NIH Memorandum of Understanding for human genome research signed

DOE Human Genome Coordinating Committee formed

Human Genome Management Information System established

First issue of *Human Genome Quarterly* published by HGMIS

DOE-NIH Five-Year Plan developed

First DOE Contractor-Grantee Workshop held at Santa Fe, New Mexico

First biannual meeting of NIH-DOE Joint Subcommittee on the Human Genome held

First Joint Mapping Working Group meeting held

DOE-NIH Five-Year Plan for the Human Genome Project presented to Congress

First DOE ELSI awards made

First Joint Working Group on Ethical, Legal, and Social Issues meeting held

DOE *Human Genome 1989-90 Program Report* published

First Joint Informatics Task Force meeting held

First joint DOE-NIH issue of *Human Genome News* published

First Joint Sequencing Working Group meeting held

LLNL designated as third human genome center

DOE-NIH annual planning and evaluation retreat held

DOE Human Genome Distinguished Postdoctoral Fellowships initiated

DOE plans announced to generate STS mapping sites at expressed loci

Joint NIH-DOE Five-Year Plan officially implemented

Second Contractor-Grantee Workshop held at Santa Fe, New Mexico

First DOE Cooperative Research and Development Agreement signed

ADA revisions recommended by ELSI Working Group

Special peer-review panel held for STS production from cDNAs

First meeting of the Joint Working Group on the Mouse held

First DOE Human Genome Distinguished Postdoctoral Fellowships awarded

DOE-NIH support of Genome Data Base begun

cDNA/STS technology awards made

NIH-DOE annual planning and evaluation retreat held

First peer-review panel held for ELSI

Moscow SBH Workshop sponsored by DOE, HUGO, and The Wellcome Trust

ACRONYM LIST

ADA	Americans with Disabilities Act
DOE	U.S. Department of Energy
ELSI	Ethical, Legal, and Social Issues
HERAC	Health and Environmental Research Advisory Committee
HGMIS	Human Genome Management Information System
HUGO	Human Genome Organization (International)
LANL	Los Alamos National Laboratory
LBL	Lawrence Berkeley Laboratory
LLNL	Lawrence Livermore National Laboratory
NIH	National Institutes of Health
NRC	National Research Council
R&D	Research & Development
USDA	U.S. Department of Agriculture
SBH	Sequencing by hybridization
STS	Sequence tagged site

86 87 88 89 90 91 92

Foreword

Acquiring complete knowledge of the organization, structure, and function of the human genome—the master blueprint of each of us—is the broad aim of the Human Genome Project. It is a new kind of program in biology, both in its size and focus on a limited set of goals and in its dependence on the development and use of technology. The coordinated U.S. Human Genome Project was officially initiated by the Department of Energy (DOE) Office of Health and Environmental Research and the National Institutes of Health (NIH) National Center for Human Genome Research (NCHGR) in FY 1991 with the publication in April 1990 of *Understanding Our Genetic Inheritance; The U.S. Human Genome Project: The First Five Years 1991–1995*. The DOE effort, which began very modestly almost 4 years before, is now over 5 years old. Taking stock of what has been done and what remains to be done is particularly appropriate at this time.

That the ambitious scientific goal of the Human Genome Project can now be imagined is the result of the revolution occurring in biology during the last 20 years. Modern biological science has achieved a profound but still quite incomplete level of understanding of how the diversity of all living things is determined. This insight, along with scientific and technical advances in other fields, has brought unprecedented power both in being able to analyze and manipulate genetic structures and to use and store large quantities of genetic information. DOE is uniquely positioned to bring together expertise in physics, chemistry, engineering, and computer science to help solve fundamental biological problems and to exploit exciting opportunities presented by the Human Genome Project. Genome research will also contribute to the department's role in providing the scientific foundation for understanding the health effects of radiation and of chemical insults to the genome.

The DOE program stresses mapping, the development of sequencing technologies and instrumentation, and informatics. Informatics refers to computational approaches in acquiring, storing, distributing, analyzing, and manipulating vast amounts of mapping and sequence data that will result from the project. Another important program component studies the ethical, legal, and social issues arising from use of the generated data, particularly in the privacy and confidentiality of genetic information. Cutting across all DOE biological and environmental research programs are several science education activities.

The Human Genome Project is a closely cooperative activity between NIH and DOE. NCHGR is an important and essential participant. Internationally, the formation of the Human Genome Organization and the establishment of national genome projects by an increasing number of countries indicate the fascination and promise of this effort on the collective imaginations of many nations. In addition to the inherent excitement about increased knowledge of human life, the project offers the promise of many new opportunities for benefiting humanity through the development of new diagnostics, pharmaceuticals, and therapies for a multitude of human diseases; a wide range of improvements will flow from other biotechnology advances. Further expected benefits include improved risk assessment for individuals and populations exposed to agents that impact genetic material, as well as possible applications of the data to environmental and remediation issues.

To be successful, the program must continue to focus on clear objectives for mapping and sequencing and to incorporate the flow of technological developments into the efforts of all working laboratories. Strategies must be planned carefully and in a comprehensive

Foreword

fashion as the next phase begins, in which mapping and sequencing results proliferate and technologies mature. Planning must be project-wide and include interagency planning at ever-earlier stages.

This report describes the status of the DOE Human Genome Program and its accomplishments to date. Research highlights are noted from the program as a whole and from the three principal DOE human genome centers at Lawrence Berkeley Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory. These national laboratory facilities of DOE have been especially successful because they are organized to focus efforts, foster interdisciplinary projects, and use advanced technologies, some developed for other purposes, toward program goals. Essential work is also reported from 41 different research universities. Remarkable progress has been made in advanced instrumentation and informatics.

A further indication of the increasing development of the DOE program is the simple statistic that the 1989–90 report had 157 pages and included 57 abstracts of work involving 211 scientists. The current program report contains over 240 pages and includes more than 150 abstracts of work involving over 400 investigators, essentially a doubling of DOE program size.

The Human Genome Project ultimately will create scientific resources for the next wave of advances in biology and medicine. As the project is completed, accomplishments will dwarf those that have occurred in the biological sciences since the advent of recombinant DNA technologies. By the same token, the ethical and social consequences of the uses of this new knowledge must be considered as the knowledge is acquired; if this knowledge is responsibly obtained and applied, the next decade of biological research will be history's most fruitful and rewarding by any measure.

David J. Galas, Associate Director
Office of Health and Environmental Research
Office of Energy Research
U.S. Department of Energy

Preface

This is the third report summarizing the Department of Energy (DOE) Human Genome Program, its content, progress, and accomplishments. Since the program's conception in 1986 and initiation in 1987 by the DOE Office of Health and Environmental Research (OHER), its broad objectives have rapidly gained both national and international support. The program has made important strides in the development and application of technologies and tools that are required for the cost-effective characterization of the molecular nature of the human genome.

This country's Human Genome Project is jointly administered by OHER and the National Center for Human Genome Research of the National Institutes of Health. A successful effort to characterize the molecular nature of human inheritance will require continuing international cooperation involving scientists from many countries. A number of other nations have begun substantial efforts to map and sequence the human genome and those of key model organisms. Although intellectual property issues threaten some aspects of international cooperation, increasing exchange of information has led to more involvement of the international community in discovery, acceleration of the pace of the research, and increased cost-effectiveness. International communication is facilitated by regular meetings to update the maps of individual chromosomes and by contributions to databases such as the Genome Data Base and nucleic acid sequence databanks. Through such databases a worldwide data aggregation and distribution system is being developed to exchange information regarding the genome.

Aided by funding from the Human Genome Project, serious study is under way on ethical, legal, and social issues that are becoming more urgent because of the rapid growth in knowledge of human genetics. It is important to develop and disseminate deeper and more widespread understanding of these dynamic issues and of the choices available for families, the law, and society. An educated public is required to make intelligent choices in this area. The national genome project is now the largest provider of funds for study of such issues.

A key to the long-term success of the program is the initial phase of intensive resource and technology development that requires input and involvement from many scientific and engineering disciplines. Exciting contributions have already been made to biomedical knowledge and biotechnology, and such advances are certain to continue at an ever-increasing rate. Announcements of discovery of important disease genes have become commonplace. Within 10 years nearly all the perhaps 100,000 genes that make up the human genome are likely to be found. Within 15 years the program is expected to culminate in a reference DNA sequence of the entire genome.

Never has such a mass of data flowed into biology and medicine. An understanding of how genetic variations account for much of the richness and adventure of human diversity will be greatly increased. More practically, there can be little doubt of tremendous payoffs in terms of diagnoses and, ultimately, specific therapies for many human diseases.

Moreover, new technologies and rapidly developing analytical tools to characterize the human genome will have widespread impact beyond human health. They will find application in revealing the genetic inheritance of many organisms of potential scientific and commercial interest and will provide an important stimulus to broaden and deepen the impact of modern biology in areas such as energy, environmental protection and waste treatment, agriculture, and the materials sciences.

Of particular importance is the facile access to proteins that rapidly follows discovery of their genes. As a result of genome projects, we will soon be in a position to begin the systematic large-scale characterization of proteins and their structure. The interplay of molecular biology, structural studies, high-performance computing, and advanced molecular graphics will certainly lead to an understanding of macromolecular structure-function relationships. The scientific and economic implications of such a predictive understanding cannot be overestimated. It is the key to full realization of the potential of modern biology.

Intense X-ray light and neutrons produced by unique, large, and expensive machines (synchrotrons and reactors) at DOE laboratories are important national resources for the determination of biological structure and, hence, for the national effort in biotechnology. A central goal of OHER is to provide access to these machines by making facilities and technical support available to structural-biology users, a need that has been projected to increase tenfold in the next several years.

Finally, as Robert Sinsheimer elegantly pointed out in *The FASEB Journal* (November 1991), the Human Genome Project is an epic venture of discovery that will in time clarify many endlessly and fruitlessly debated mysteries of human nature. With this project we are launched upon a new stage of the age-old quest to illuminate the record of the human past—the prehistory of our species as recorded in the genetic script or blueprint for our being. When complete, the project will have provided us with an unprecedented resource—the complete text of our genetic endowment. It will be seen as a turning point in human history.

David A. Smith

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

Acknowledgements

The DOE Office of Health and Environmental Research gratefully acknowledges the contributions made by genome research grantees and contractors in submitting abstracts, photographs, captions, and narratives. The Human Genome Management Information System at Oak Ridge National Laboratory (managed by Martin Marietta Energy Systems, Inc., for the U.S. Department of Energy under contract DE-AC05-84OR21400) collected and organized the information, prepared the manuscript, and implemented the design and production of this publication.

Contents

Human Genome
1991–92 Program Report

Introduction

History of the Human Genome Project 1

DOE-NIH Coordination 4

Scientific Five-Year Goals of the U.S. Human Genome Project 5

Highlights of Research Progress

Mapping 8

Informatics 15

Sequencing 18

Activities Addressing Ethical, Legal, and Social Issues Related to Human Genome Project Data 22

Technology Transfer and Industrial Collaboration 24

Human Genome Center Research Narratives

Lawrence Berkeley Laboratory 30

Lawrence Livermore National Laboratory 36

Los Alamos National Laboratory 46

Program Management Infrastructure

DOE OHER Mission 54

Program Management Task Group 57

Field Coordination 58

- Human Genome Coordinating Committee 58
- Human Genome Management Information System 60
- Human Genome Distinguished Postdoctoral Fellowships 61

Resource Allocation 62

Interagency Coordination 64

- Joint DOE-NIH Activities 64
  - Joint Mapping Working Group 65
  - Joint Informatics Task Force 66
  - Joint Sequencing Working Group 66
  - Joint Working Group on Ethical, Legal, and Social Issues 66
  - Joint Working Group on the Mouse 66

Contents

Other U.S. Genome Research 66

U.S. Department of Agriculture 66

National Science Foundation 66

Howard Hughes Medical Institute 67

International Coordination

HUGO: Worldwide Genome Research Coordination 71

UNESCO: Promoting the Interests of Developing Countries 72

Abstracts of DOE-Funded Research

Project Categories and Principal Investigators 77

Resource Development 82

Physical and Genetic Mapping 99

Mapping Instrumentation 111

Sequencing Technologies 119

Informatics 139

Ethical, Legal, and Social Issues 156

Infrastructure 163

Small Business Innovation Research Phase I 165

Small Business Innovation Research Phase II 168

Completed Projects 171

Appendices

A. Primer on Molecular Genetics 191

B. Conferences, Meetings, and Workshops Sponsored by DOE 219

C. Members of the DOE Health and Environmental Research Advisory Committee 223

D. Members of the DOE-NIH Joint Working Groups 225

E. Glossary 229

Index to Principal and Coinvestigators Listed in Abstracts 243

Acronym List (Inside back cover)

Introduction

The U.S. Human Genome Project is the national coordinated 15-year effort to characterize all the human genetic material—the genome—by improving existing human genetic maps, constructing physical maps of entire chromosomes, and ultimately determining the complete sequence of the deoxyribonucleic acid (DNA) subunits in the human genome. Parallel studies are being carried out on selected model organisms to facilitate the interpretation of human gene function. The ultimate goal of the U.S. project is to discover all of the more than 100,000 human genes and render them accessible for further biological study.

For more information on the science of genomics, see Appendix A, "Primer on Molecular Genetics," p. 191. Terms are defined in the Glossary, p. 229. An acronym list is on the inside back cover.

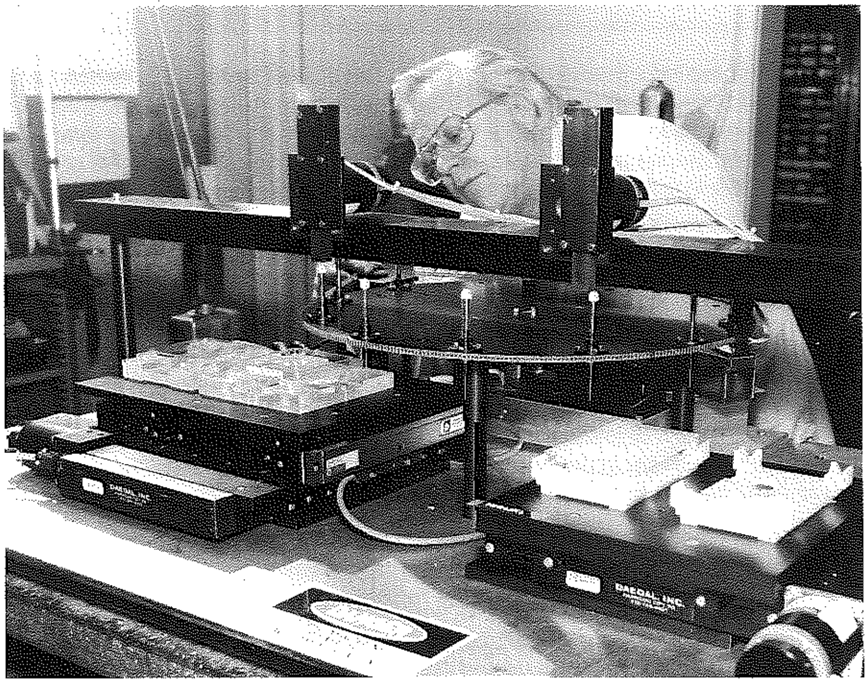
Current technology could probably be used to attain the objectives of the Human Genome Project, but the cost and time required would be unacceptable. For this reason, a major feature of the first 10 years of the project is to optimize existing methods and develop new technology to increase efficiency in DNA mapping and sequencing by 1 or 2 orders of magnitude. The genome will eventually be sequenced using continually evolving technologies and revolutionary methods not in existence today.

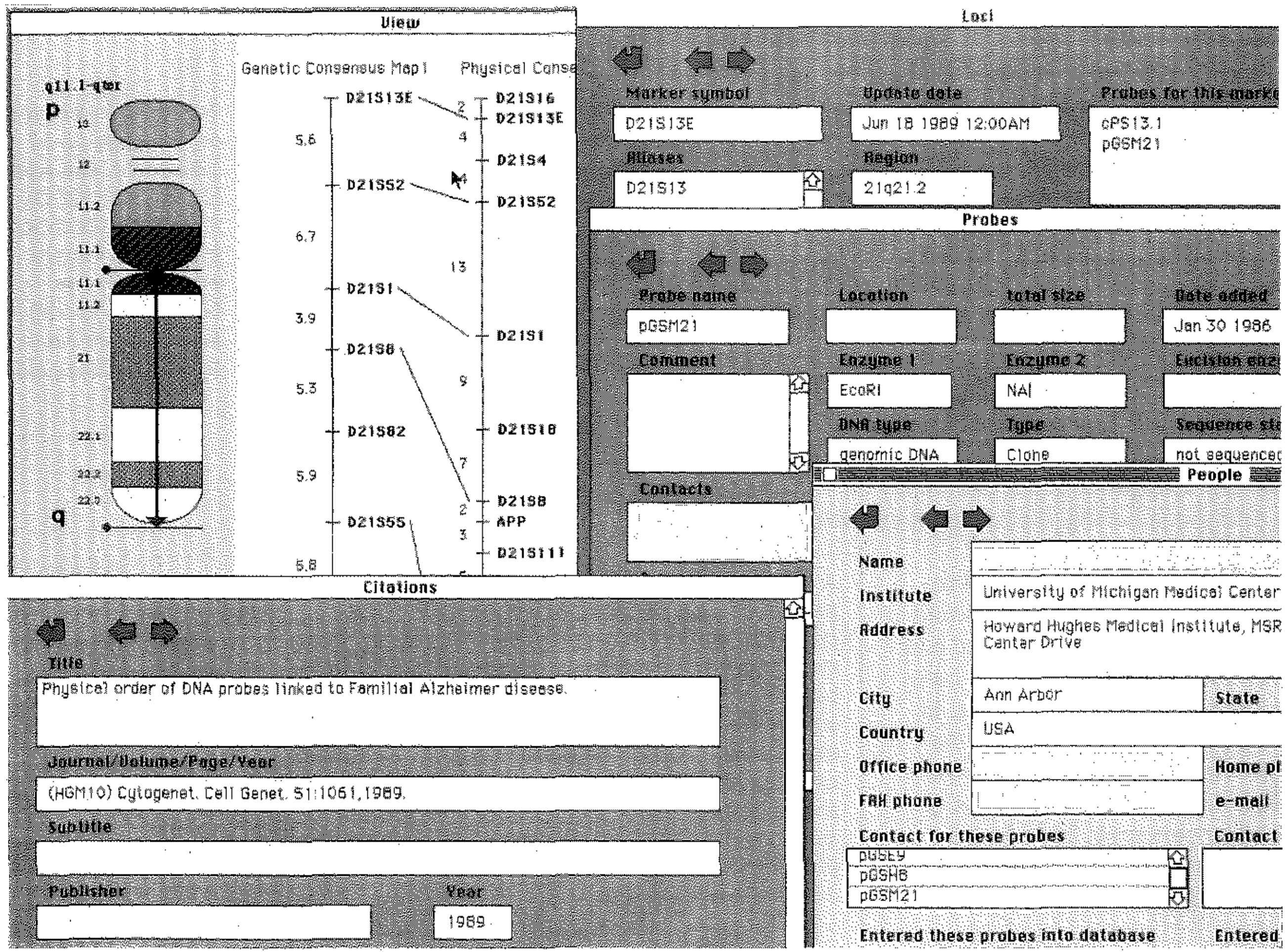
Information obtained as part of the Human Genome Project will dramatically change almost all biological and medical research and dwarf the catalog of current genetic knowledge. In addition, both the methods and the data developed as part of the project are likely to benefit investigations of many other genomes, including a large number of commercially important plants and animals.

History of the DOE Human Genome Program

A brief history of the U.S. Department of Energy (DOE) Human Genome Program will be useful in a discussion of the objectives of the DOE program as well as those of the collaborative U.S. Human Genome Project. The Office of Health and Environmental Research (OHER) of DOE and its predecessor agencies—the Atomic Energy Commission and the Energy Research and Development Administration—have long sponsored research into genetics, both in microbial systems and in mammals, including basic studies on genome structure, replication, damage, and repair and the consequences of genetic mutations.

A High-Speed Colony Picker. *Images of yeast or bacterial colonies grown in petri dishes are analyzed to determine the precise locations of the colonies containing target DNA sequences. A computer-controlled x-y table then moves the dish so that a particular colony is picked by one of a series of tungsten needles fixed to a carousel rotating in a plane parallel to the table. Meanwhile, a similar table containing a 96-well microtiter plate positions itself so the needle can deposit the material in a well of the plate. Designed and built at LBL, this device can pick about 1500 colonies per hour. In its first major test, the machine picked and arrayed a 10,000-clone library from Washington University in St. Louis. The device can be serviced by a robot. (Photograph provided by Joseph Jaklevic and Edward Theil, LBL. For more information, refer to Jaklevic in the Index to Principal and Coinvestigators.)*





Loci

Marker symbol: D21S13E
Update date: Jun 18 1989 12:00AM
Probes for this marke: cPS13.1 pGSM21
Aliases: D21S13
Region: 21q21.2

Probes

Probe name: pGSM21
Location:
total size:
Date added: Jan 30 1986
Comment:
Enzyme 1: EcoRI
Enzyme 2: NAI
Excision enz:
DNA type: genomic DNA
Type: Clone
Sequence st: not sequenced
Contacts:

People

Name:
Institute: University of Michigan Medical Center
Address: Howard Hughes Medical Institute, MSR
Center Drive
City: Ann Arbor
State:
Country: USA
Office phone:
Home pl
FAX phone:
e-mail
Contact for these probes: pGSE9 pGSH8 pGSM21
Contact
Entered these probes into database
Entered

Citations

Title: Physical order of DNA probes linked to Familial Alzheimer disease.
Journal/Volume/Page/Year: (HGM10) Cytogenet. Cell Genet. 51:1061,1989.
Subtitle:
Publisher:
Year: 1989

Chromosome Information System (CIS). *Developed by the LBL Human Genome Center computing group, CIS allows molecular biologists to manipulate various kinds of genomic information without knowing the underlying database structure and implementation. CIS has three system functions: a graphical user interface, a database, and an intermediate layer of software that implements the translation from the biological perspective of the user interface to the relational database model. CIS incorporates mapping data about chromosomes, maps (genetic and physical), markers (e.g., loci and probes), sequences (primers and sequence tagged sites), and relevant reference information. Using this program, biologists can search and navigate through maps derived from current experimental results, collaborating laboratories, and public databases; edit and compare maps through direct manipulation; and interact with other programs by either accepting or providing data.*

Queries for retrieving maps can be formulated by selecting a region on the chromosome shown on the screen or by choosing from a list of available data. The graphical presentation provides access to all data pertaining to a particular region on the chromosome for the novice user and also allows visual comparison of different kinds of maps. Experienced users can choose from a list of entries to gain rapid access to known information. Information shown on the map can be followed to reveal other relevant information (e.g., owners or contact persons, bibliographic citations, and pointers to other databases) or to trace data to experiments that originally placed the object on the map.

The above display illustrates the hypertext capabilities of CIS: the selected map is shown on the top left panel of the figure; clicking on the label D21S13E opens a window revealing data around that locus. Clicking in the probe, people, and citation fields of the locus window retrieves the related information in multiple windows.

(Photograph provided by the LBL Human Genome Center. For more information, refer to S. Lewis, J. McCarthy, and M. Zorn in the Index to Principal and Coinvestigators.)

In 1984, OHER and the International Commission on Protection Against Environmental Mutagens and Carcinogens cosponsored a conference in Alta, Utah, which highlighted the growing roles of recombinant DNA technologies. Substantial portions of the meeting's proceedings were incorporated into the Congressional Office of Technology Assessment report, *Technologies for Detecting Heritable Mutations in Humans*, in which the value of a reference sequence of the human genome was recognized.

Acquisition of such a reference sequence was, however, far beyond the capabilities of biomedical research resources and infrastructure existing at that time. Although the small genomes of several microbes had been mapped or partially sequenced, the detailed mapping and eventual sequencing of 24 distinct human chromosomes (22 autosomes and the sex chromosomes X and Y) that together comprise an estimated 3 billion subunits was a task some thousandsfold larger.

DOE OHER was already engaged in several multidisciplinary projects contributing to the nation's biomedical capabilities, including the GenBank® DNA sequence repository, which was initiated and sustained by DOE computer and data-management expertise. Several major user facilities supporting microstructure research were developed and are maintained by DOE (see box, p. 55). Unique chromosome-processing resources and capabilities were in place at Los Alamos National Laboratory and Lawrence Livermore National Laboratory. Among these were the fluorescence-activated cell sorter (FACS) systems to purify human chromosomes within the National Laboratory Gene Library Project for the production of libraries of DNA clones. The availability of these monochromosomal libraries opened an important path—a practical means of subdividing the huge total genome into 24 much more manageable components.

With these capabilities, OHER began in 1986 to consider the feasibility of a dedicated human genome program. Leading scientists were invited to the March 1986 international conference at Santa Fe, New Mexico, to assess the desirability and feasibility of implementing such a project. With virtual unanimity, participants agreed that ordering and eventually sequencing DNA clones representing the human genome were desirable and feasible goals. With the receipt of this enthusiastic response, OHER initiated several pilot projects. Program guidance was further sought from the DOE Health Effects Research Advisory Committee (HERAC, see Appendix C for a list of current members).

The HERAC Recommendation. The April 1987 HERAC report recommended that DOE and the nation commit to a large, multidisciplinary, scientific, and technological undertaking to map and sequence the human genome. DOE was particularly well suited to focus on resource and technology development, the report noted; HERAC further recommended a leadership role for DOE because of its demonstrated expertise in managing complex and long-term multidisciplinary projects involving both the development of new technologies

Investigator performing computer analysis of chromosome 19 map data at the Human Genome Center, LLNL.

and the coordination of efforts in industries, universities, and its own laboratories. Evolution of the nation's Human Genome Project further benefited from a 1988 study by the National Research Council (NRC) entitled *Mapping and Sequencing the Human Genome*, which recommended that the United States support this research effort and presented an outline for a multiphase plan.

DOE-NIH Coordination

The National Institutes of Health (NIH) was a necessary participant in the large-scale effort to map and sequence the human genome because of its long history of support for biomedical research and its vast community of scientists. This was confirmed by the NRC report, which recommended a major role for NIH. In 1987, under the leadership of Director James Wyngaarden, NIH established the Office of Genome Research in the Director's Office. In 1989 this office became the National Center for Human Genome Research (NCHGR), directed by James D. Watson. After Watson's resignation in April 1992, Michael Gottesman was appointed NCHGR Acting Director.

In addition to extramural support for research projects in physical mapping and the development of index linkage markers and technology, NIH also provides support for genetic mapping based on family studies and, following NRC recommendations, for studies on several relevant model organisms. DOE-supported genome research is focused almost exclusively on the human genome through support of large-scale physical mapping, resource and instrumentation technology development, and improvements in computational and database capabilities and research infrastructure. A significant portion of the DOE Human Genome Program is allocated to the DOE national laboratores.

In several important areas, DOE and NIH cooperate to support critical resources such as the Genome Data Base (GDB) at Johns Hopkins University. Cofunded since 1991 as the central international repository of human chromosome mapping data, GDB is expected to receive supporting funds from other nations. DOE and NIH also cooperate to support joint workshops; a number of ethical, legal, and social issues projects; and the *Human Genome News* newsletter.

Joint task groups under the DOE-NIH Joint Subcommittee on the Human Genome meet periodically to define program needs and develop recommendations for their parent DOE and NIH committees. OHER and NCHGR cosponsor workshops and meetings of the task groups on mapping; sequencing; informatics; the use of the mouse as a mammalian model; and—in a departure from most scientific programs—ethical, legal, and social issues related to data produced in the project.

Many other highlights of the DOE OHER program follow in the succeeding sections of this report, including reports from the human genome centers; further details of program infrastructure, management, and coordination; resource allocation; and abstracts of individual research projects.

Scientific Five-Year Goals of the U.S. Human Genome Project from the NIH-DOE Five Year Plan* [Implemented October 1, 1990 (FY 1991)]

1. Mapping and Sequencing the Human Genome

Genetic Mapping

Complete a fully connected human genetic map with markers spaced an average of 2 to 5 cM apart. Identify each marker by a sequence tagged site (STS).

Physical Mapping

Assemble STS maps of all human chromosomes with the goal of having markers spaced at approximately 100,000-bp intervals.

Generate overlapping sets of cloned DNA or closely spaced unambiguously ordered markers with continuity over lengths of 2 Mb for large parts of the human genome.

DNA Sequencing

Improve current and develop new methods for DNA sequencing that will allow large-scale sequencing of DNA at a cost of $0.50 per base pair.

Determine the sequence of an aggregate of 10 Mb of human DNA in large continuous stretches in the course of technology development and validation.

2. Model Organisms

Prepare a mouse genome genetic map based on DNA markers. Start physical mapping on one or two chromosomes.

Sequence an aggregate of about 20 Mb of DNA from a variety of model organisms, focusing on stretches that are 1 Mb long, in the course of developing and validating new and improved DNA sequencing technology.

3. Informatics—Data Collection and Analysis

Develop effective software and database designs to support large-scale mapping and sequencing projects.

Create database tools that provide easy access to up-to-date physical mapping, genetic mapping, chromosome mapping, and sequencing information and allow ready comparison of the data in these several data sets.

Develop algorithms and analytical tools that can be used in the interpretation of genomic information.

4. Ethical, Legal, and Social Considerations

Develop programs directed toward understanding the ethical, legal, and social implications of Human Genome Project data. Identify and define the major issues and develop initial policy options to address them.

5. Research Training

Support research training of pre- and postdoctoral fellows starting in FY 1990. Increase the number of trainees supported until a steady state of about 600 per year is reached by the fifth year.

Examine the need for other types of research training in the next year (FY 1991).

6. Technology Development

Support automated instrumentation and innovative and high-risk technological developments as well as improvements in current technology to meet the needs of the genome project as a whole.

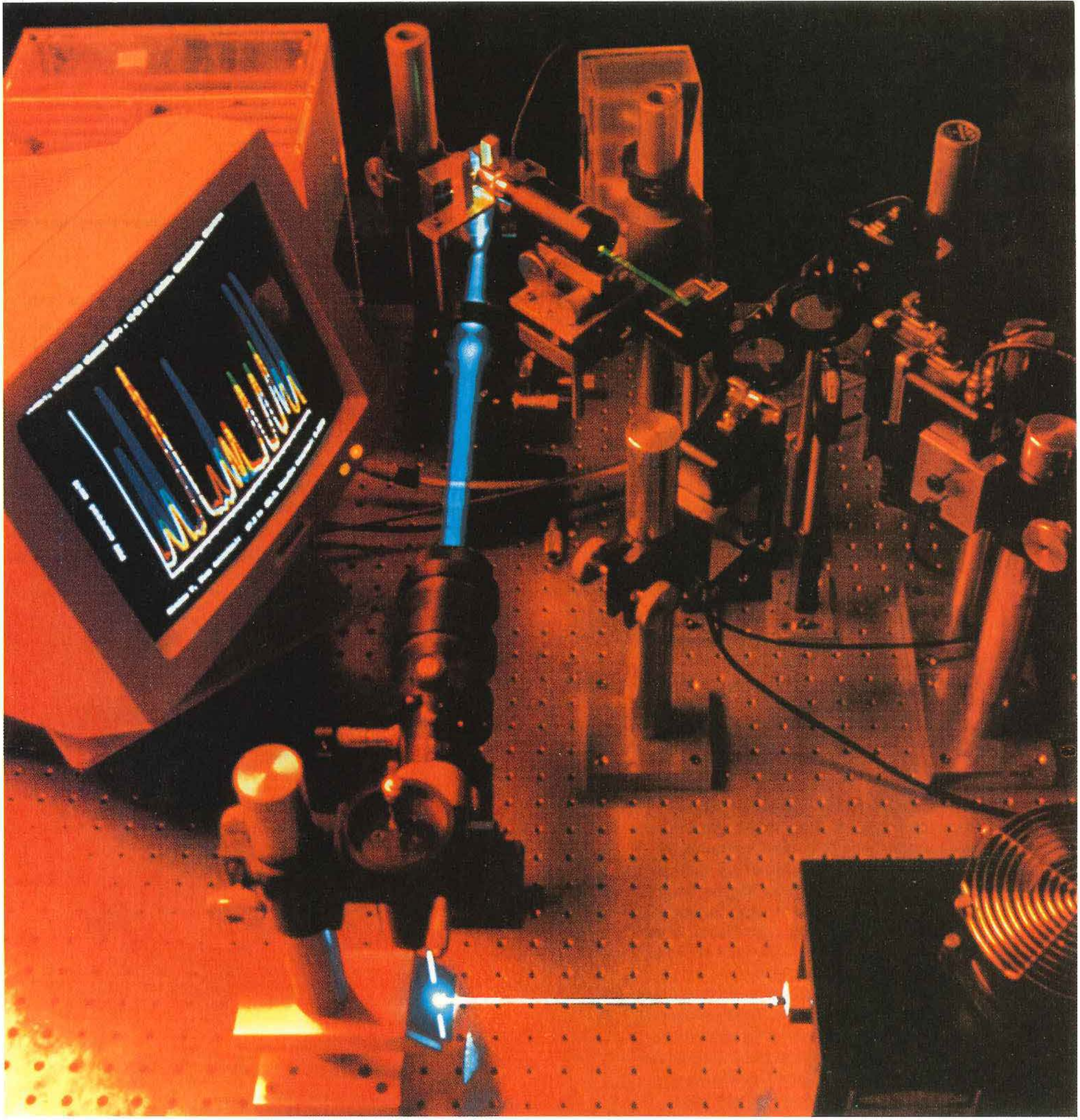
7. Technology Transfer

Enhance the already close working relationships with industry.

Encourage and facilitate the transfer of technologies and of medically important information to the medical community.

**Understanding Our Genetic Inheritance; The U.S. Human Genome Project: The First Five Years FY 1991–1995,* DOE/ER-0452P, U.S. Department of Health and Human Services and U.S. Department of Energy, April 1990.

High-Speed DNA Sequencing. *The sequence of bases in a DNA fragment is determined during capillary electrophoresis using a laser-based procedure. A low-power argon ion laser (right foreground) emits a blue-white excitation beam, which is turned at a right angle and focused onto DNA fragments in a capillary gel (seen as a faint thread at the center top of picture) during a simulated sequencing experiment. The focused beam causes each dye-tagged DNA base to fluoresce in a yellow-green emission beam at a particular identifiable wavelength. The emitted fluorescence is collected and analyzed by a detector (right center of picture). Accumulated sequence data can be seen on the computer monitor at left. Efficient heat dissipation in the ultrathin capillary gels allows the use of high voltages to separate DNA fragments in a very short time. (Photograph by Bruce Fritz provided by Lloyd Smith, University of Wisconsin, Madison. For more information, refer to Smith in the Index to Principal and Coinvestigators.)*



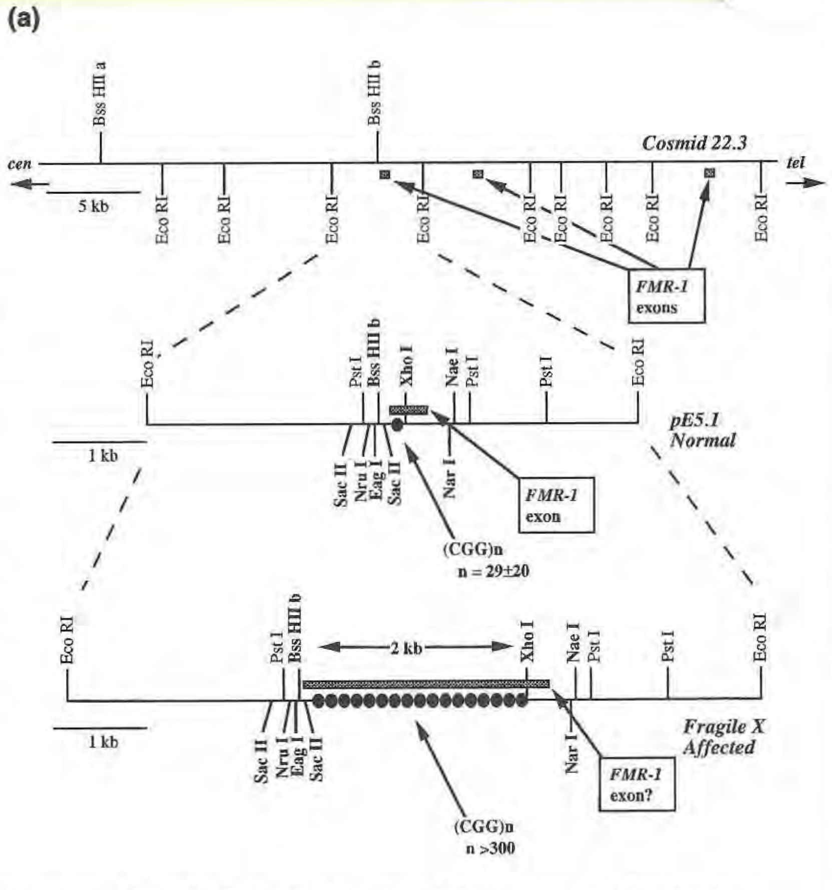
Highlights of Research Progress

Mapping

A major goal for DOE and NIH, as stated in the Five Year Plan (p. 5) for the Human Genome Project officially implemented in FY 1991, is to develop refined physical maps of chromosomes. Increasingly detailed maps will provide biomedical scientists with rapid access to important areas on chromosomes through their specific markers and ordered sets of DNA clones.

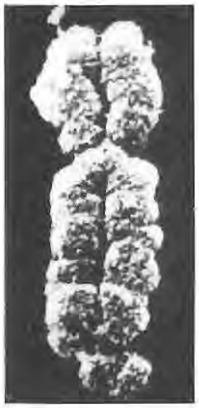
Physical Map Construction

DOE sponsors both extensive physical mapping studies and supportive resource and technology development. Physical mapping of chromosomes 5, 11, 16, 17, 19, 21, 22, and X has been or is being supported directly. Increasingly detailed maps facilitate access to important chromosomal loci through their constituent markers and ordered DNA clones.



***(a) Map of the Fragile X Site in Cosmid 22.3, Subclone pE5.1, and the Presumed Structure in an Affected Fragile X Chromosome.** Restriction maps and other relevant features are given for cosmid and plasmid subclones from yeast artificial chromosome 209G4 [A. J. M. H. Verkerk et al.,* Cell ***65**, 905–14 (1991)] with rare-cutting restriction sites in bold type. The hypothesized structure of the fragile X mutation is shown as multiples of the CGG repeat region represented by black circles, and exons of the FMR-1 gene are located at the black bars. Note that CGG repeats are within an exon of FMR-1. FMR-1 expression is reduced or absent in fragile X patients [M. Pieretti et al.,* Cell ***66**, 817–22 (1991)]. The most common allele in the human population contains 29 CGG repeats; the range observed in normals varies from 6 to 46 repeats. These expand to permutation alleles (52 to >200 repeats) in unaffected individuals in fragile X pedigrees and go up to much larger numbers (600 to >2000) of repeats in affected fragile X individuals [Y.-H. Fu et al.,* Cell ***67**, 1047–58 (1991)].*

(b)



***(b) Induced Fragile X Chromosome Seen in Scanning Electron Microscopy.** The fragile site appears as a gap in the lower portion of the left chromatid, with the fragment partially detached. [Photograph first published in C. J. Harrison et al., "The Fragile X: A Scanning Electron Microscope Study,"* J. Med. Genet. ***20**, 280–85 (1983).]*

(Figure and photograph provided by David Nelson, Baylor College of Medicine. For more information, refer to Nelson and C. T. Caskey in the Index to Principal and Coinvestigators.)

The earliest concerted mapping efforts began on chromosome 16 at the Los Alamos National Laboratory (LANL) Center for Human Genome Studies and on chromosome 19 at the Lawrence Livermore National Laboratory (LLNL) Human Genome Center. These efforts have achieved excellent progress (see detailed narratives, pp. 46 and 36, respectively) through the development of effective multidisciplinary teams and efficient methods for generating clone "fingerprints." The fingerprints provide data for recognizing clone pairs that overlap, facilitating the construction of increasingly larger sets of overlapping clones, called contigs. Approximately 90% of chromosomes 16 and 19 is now represented by fingerprinted clones, and multiclone contigs span at least 80% of their length. Initial contig assembly methodologies are complemented by strategies designed to finish the physical maps and align them with genetic maps. This progress, together with the many contributions from other research groups (presented in the Abstracts section of this report), shows that resources and technologies required to achieve the mapping goals stated in the Five Year Plan are rapidly being realized.

Page numbers for research abstracts of investigators noted in parentheses can be located in the "Index to Principal and Coinvestigators Listed in Abstracts," p. 243.

National Laboratory Gene Library Project (NLGLP)

Among the resources most crucial to mapping progress are the libraries of clones representing each of the human chromosomes. Their availability reduces the total genome mapping effort to 24 smaller, more-manageable mapping projects. This chromosome-specific clone library production from physically purified chromosomes depends on the unique LANL and LLNL chromosome-sorting facilities maintained through the DOE NLGLP. These library resources are either distributed from the laboratories or through the American *Type Culture* Collection. As of December 1991 over 620 chromosome-specific libraries were distributed as resources for entire chromosome mapping efforts and for more-selective gene hunts. Current library production is focused on the needs of the major chromosome mapping projects (L. Deaven, LANL; P. de Jong, LLNL).

Recombinant Clone Types

Other biological resources are also being developed to further chromosome mapping progress. These resources include several useful genetic elements or recombinant DNAs and their cellular hosts. The largest elements are the intact, single human chromosomes maintained in somatic cell hybrids, such as single human chromosome/hamster-host cell hybrids. They are valuable for sorting out the human chromosomes for construction of single-chromosome libraries. Insert sizes of recombinants range from millions to a few hundred bases. Recombinant cosmid clones with 40- to 50-kb human DNA inserts predominated in the early contig-building efforts and continue to be a basic resource (refer to Abstracts: Resource Development, p. 82).

Monochromosomal Yeast Artificial Chromosomes (YACs)

YACs with inserts of 200 kb and larger, whose initial development was pioneered with NIH support, are now widely used in physical mapping projects. The recently developed capability to produce YACs from flow-sorted chromosomes is making available monochromosomal YAC libraries to speed mapping projects (M. McCormick, L. Deaven, and R. Moyzis, LANL). These libraries are made up of YACs containing human DNA inserts. This contrasts with libraries made from somatic cell hybrids, which are made up of YACs that contain mostly nonhuman DNA inserts.

Highlights of Research Progress: Mapping

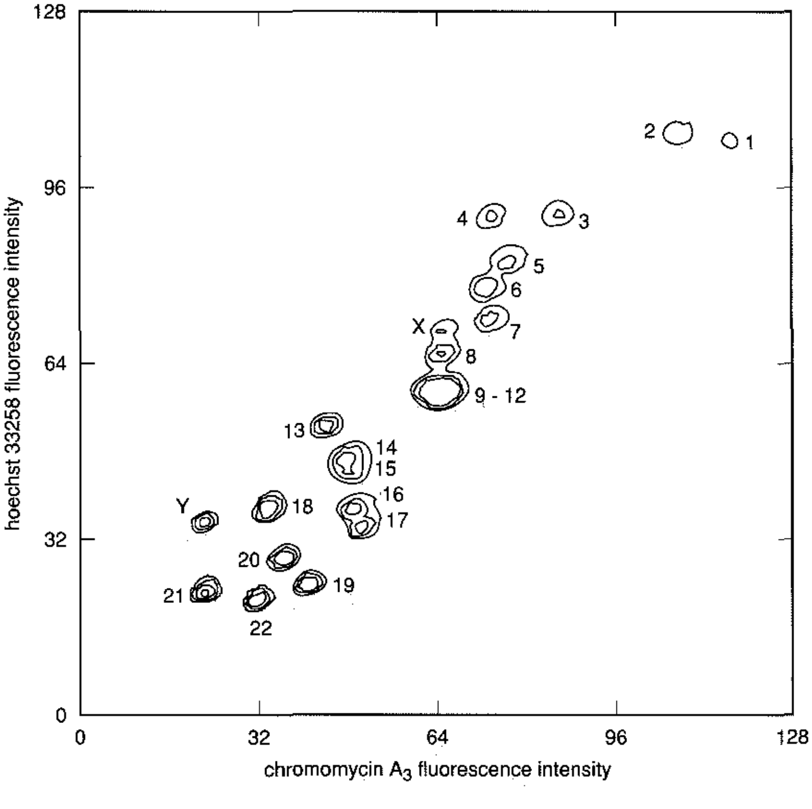
Clone Library Array and Analysis

When user laboratories maintain clone libraries in the same arrayed-format addressing system, the information obtained from these libraries is maximized because the accumulated data from different laboratories can be readily combined. The tedious task of arraying thousands of DNA clones has been greatly alleviated through the development and implementation of automated or robotic processing systems (T. Beugelsdijk and P. Medvick, LANL; J. Jaklevic, Lawrence Berkeley Laboratory (LBL); and A. Olsen, LLNL). These systems are being increasingly utilized in clone analyses and in comparisons needed for overlap detection.

Multiplexed Clone Overlap Detection

Overlap detection of sequence homologies by DNA hybridization is speeded by multiplexing strategies in which the processing of pools of clones or their derivative probes replaces the more tedious analysis of individual clones. Multiplexing was first implemented by the chromosome 11 mapping group (G. Evans, Salk Institute for Biological Studies). Several second-generation multiplexing schemes are now being implemented to speed overlap detection both within libraries and between members of different types of libraries (J.-F. Cheng, LBL; P. de Jong, LLNL).

Analysis of Human Chromosomes Using Flow Cytometry. *Flow cytometry distinguishes chromosomes on the basis of their DNA content. The human chromosome flow histogram shown in the figure displays vertically the number of fluorescence events vs the fluorescence intensities of the DNA stains chromomycin A3 and Hoechst 33258. Chromosomes 9, 10, 11, and 12 are so similar in DNA content that they form a single overlapping peak and must be sorted from each other by using rodent-human hybrid cells containing a single (or a few) human chromosomes. Flow-sorting technology was pioneered at LANL and at LLNL. Chromosome-specific libraries constructed under the National Laboratory Gene Library Project at LANL and LLNL are available from the American* Type Culture *Collection in Rockville, Maryland. The availability of such rich sources of probes from specific chromosomes has facilitated human gene mapping and genetic disease diagnosis. [Figure first published in L. S. Cram et al., "Polyamine Buffer for Bivariate Human Flow Cytogenetic Analysis and Sorting,"* Methods Cell Biol. ***33****, 377–83 (1990), copyright© 1990 by Academic Press, Inc. Figure provided by John Fawcett, Life Sciences Division, LANL. For more information, refer to J. H. Jett in the Index to Principal and Coinvestigators.]*



Messenger RNA/cDNAs Used To Generate Sequence Tagged Sites (STSs)

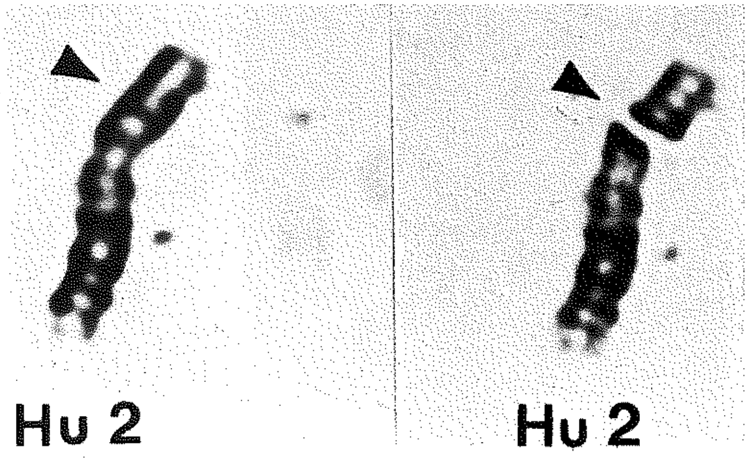
STS marking of DNA clones provides a common language for uniting the results obtained with different types of recombinant DNAs and varied approaches to map generation. An STS is a short, unique DNA sequence (generally 100 to 300 bp) that distinguishes a chromosomal locus. The STS segment can be selectively amplified within the entire genome by the polymerase chain reaction to provide an identifying tag for any DNA clone containing the site. DOE is emphasizing the use of STSs for expressed genes, as represented by their derivative cDNAs. Mapping these STSs onto contigs and to their chromosomal loci is thus rapidly placing genes on the developing chromosome maps (refer to Abstracts: Resource Development, p. 82).

Microdissection Libraries

Chromosome microdissection can facilitate region-specific mapping efforts, such as the localized ordering of clones on the much longer chromosomes, by identifying sets of clones derived from the specific region. Region-specific probes can also serve in the identification of locally expressed genes by selectively displaying their counterparts within complex cDNA libraries (F.-T. Kao, Eleanor Roosevelt Institute).

Libraries of Hybrid Somatic Cells with Partial Human Chromosomes

Aberrant chromosomes arising from rearrangement processes can be moved into host rodent cells, providing for the maintenance of a human subchromosomal segment. A large hybrid set has been assembled for chromosome 16 (G. Sutherland, Adelaide Children's Hospital, South Australia). These partial chromosomes together define over 100 chromosomal segment "bins" to which clones, contigs, and other DNA markers can be assigned by DNA hybridization tests. This resource system is greatly speeding the completion of the chromosome 16 map.



Microdissection of a DNA Fragment from the Short Arm of Human Chromosome 2. *Fragments are physically cut from a chromosomal region and amplified by the polymerase chain reaction (PCR) method. The PCR products are cloned into plasmids to construct a library representing the dissected region. The microclones from the library can be used to isolate corresponding cosmid and yeast artificial chromosome clones with large inserts to establish contigs and for high-resolution physical mapping. The microclones containing unique sequence inserts can also be used to screen cDNA libraries to isolate expressed gene sequences from defined regions of the genome. Left: before dissection; Right: after dissection. (Photograph provided by Fa-Ten Kao, Eleanor Roosevelt Institute. For more information, refer to Kao in the Index to Principal and Coinvestigators.)*

Highlights of Research Progress: Mapping

Fluorescence In Situ Hybridization (FISH)

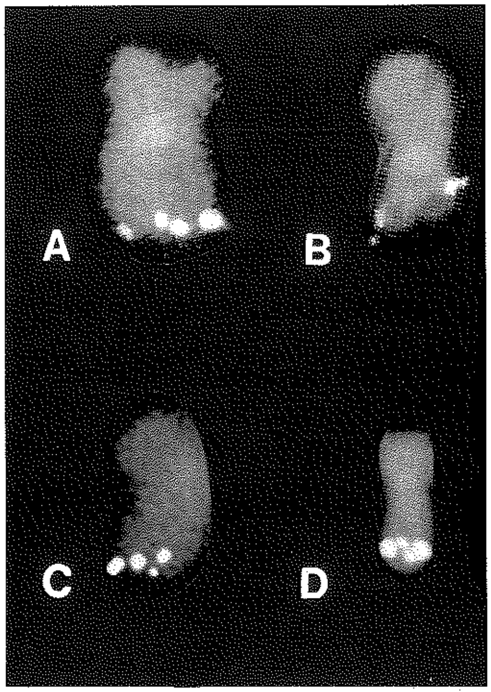
The previous mapping of DNA clones by FISH onto metaphase chromosomes has now been extended to the much less condensed interphase and pronuclear DNAs. Mapping onto less-condensed chromosomes increases spatial resolution and the capacity to order closely spaced markers. As a component of evolving mapping strategies, FISH is serving to locate and orient cosmid contigs on intact chromosomes and measure distances between the cosmids as well as to mapped cDNAs. (J. Gray, University of California; J. Korenberg, Cedars-Sinai Medical Center; B. Trask, LLNL).

Fragile X Locus Cloned

The fragile X locus has been cloned and its mode of action is being characterized (C. T. Caskey and D. L. Nelson, Baylor College of Medicine; and collaborators). Fragile X syndrome may be the most common form of inherited mental retardation. About 1 in 1500 males and 1 in 2500 females are affected by the syndrome, which is caused by a high mutation frequency at the fragile X locus.

Myotonic Dystrophy Locus Cloned

The gene responsible for myotonic dystrophy, an autosomal dominant disease, has been identified and cloned. The structural defect is characterized by a tandemly repeated segment of DNA within or close to the coding region on 19q13.3. The extent of the amplified region appears to be associated with the severity of the disease (C. T. Caskey, Baylor College of Medicine; P. de Jong and A. Carrano, LLNL; and collaborators).



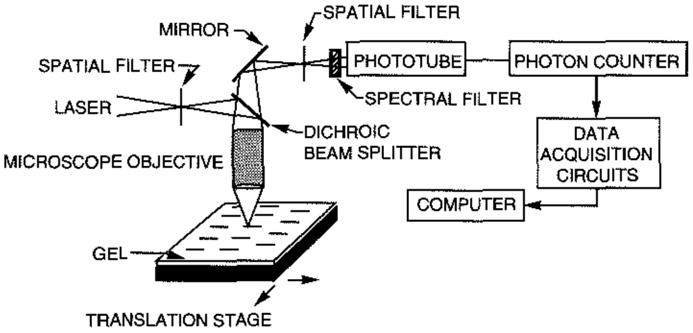
(a)

(b)

Localizing Cosmid Clones by Fluorescence In Situ Hybridization.
(a) *Four different pairs of cosmid clones (light spots on chromosomes labeled A through D) are mapped to the long arm of chromosome 11. This is a method for determining the relative orientation of landmark clones (i.e., the order of the clones) before isolating the corresponding yeast artificial chromosome clones to fill in the gaps between mapped cosmids. Two cosmids are labeled and hybridized together, and the distances separating them on the DNA is estimated. This figure also shows the variation in DNA marker location relative to the entire chromosome structure.*

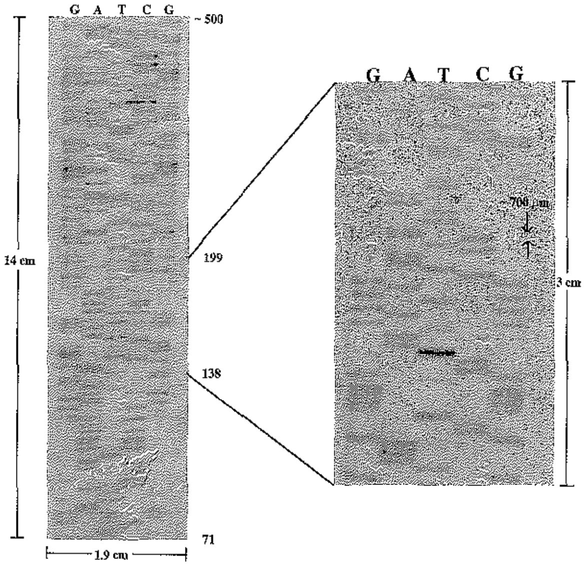
(b) *An investigator is shown using a confocal laser scanning microscope and digital imaging system to analyze the position of cosmid clones on chromosome 11. (Photographs provided by Glen Evans, Salk Institute for Biologi-cal Studies. For more information, refer to Evans in the Index to Principal and Coinvestigators.)*

Laser-Excited Confocal Fluorescence Gel Scanner. *This apparatus detects DNA separated on sequencing, agarose mapping, and pulsed-field gels with a high degree of sensitivity. Other advantages include its off-line detection apparatus (not tied to the electrophoresis system) and its immediate display of a quantitative image for data analysis.*



(a)

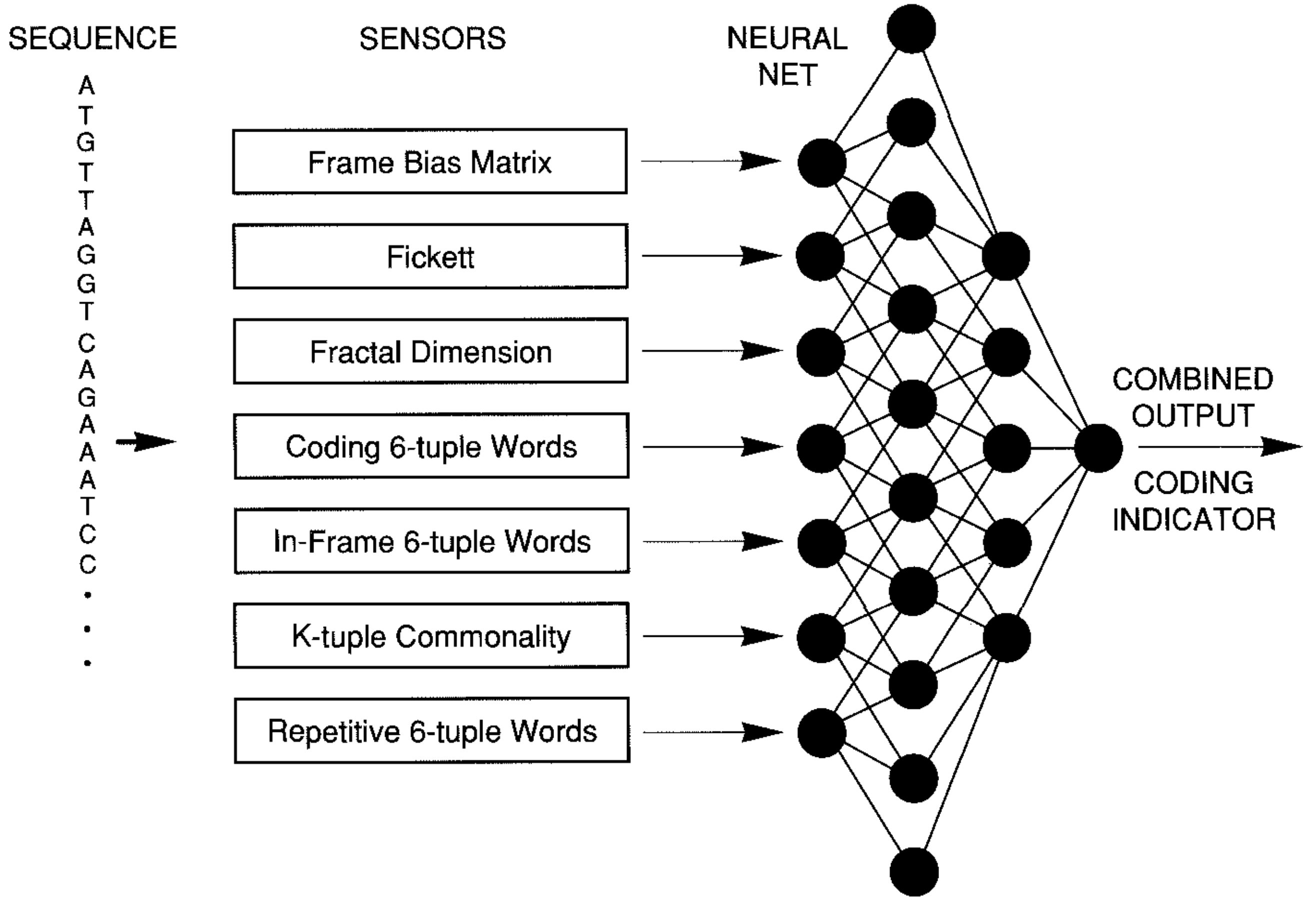
(a) *The 488-nm exciting laser light is reflected by a long-pass dichroic beamsplitter to the long-working-length microscope objective. The fluorescence is collected by the objective, passed back through the beamsplitter, and reflected by a mirror to a focus at the detection spatial filter. Spectral filters are used to eliminate residual Rayleigh and Raman scattering. The output of the phototube is passed through the data acquisition circuits and stored as an image on a computer. A computer-controlled XY-translation stage is used to sweep the gel past the laser beam.*



(b)

(b) *This image of a DNA sequencing gel was recorded by using the laser-excited confocal fluorescence gel scanner. M13mp18 DNA was sequenced using the Sanger method incorporating a fluorescently labeled primer. The left panel presents a 1.9 cm by 14 cm portion of the gel image spanning the M13 sequence from 71 to 500 bases beyond the primer. The right panel presents an expanded view of the M13 sequence from 138 to 199 bases. Labeled DNA fragments (30 femtomoles) were loaded per 3-mm lane on a 400-µm-thick Hydrolink sequencing gel. Bands as small as 1 mm by 200 µm were detected; the limiting sensitivity is about 10 attomoles of fluorescently labeled fragments per band. This work demonstrates the feasibility of detecting bands on miniaturized sequencing gels.*

[Figures first published in M. A. Quesada et al., "High Sensitivity DNA Detection with a Laser-Excited Confocal Fluorescence Gel Scanner," BioTechniques ***10****, 616–25 (1991). Figures provided by Richard Mathies, University of California. For more information, refer to Mathies and A. N. Glazer in the Index to Principal and Coinvestigators.]*



Coding Recognition Module (CRM) and GRAIL System. *Extracting relevant biological information from human DNA will require automated computer-based interpretation of vast amounts of uncharacterized sequence data. Efforts at ORNL have focused on the development of technologies that use artificial intelligence and parallel computation to localize and characterize genes and other biologically important features in DNA sequence data. The foundation of this project is a unique set of tools developed for recognizing the variable sequence patterns characteristic of gene sequence features. Each of these tools consists of two parts: (1) a group of algorithms, or sensors, to measure sequence attributes related to the feature of interest and (2) a neural network that learns to recognize the feature by examining output from the sensor algorithms.*

The CRM shown in the figure incorporates a group of seven sensor algorithms, each designed to indicate the probability that a given sequence position is within a coding region. After a suitable training procedure, the neural network learns to interpret the sensor outputs and, when provided with sensor data from a test sequence, to make highly accurate decisions about the location of coding DNA. The coding-recognition capabilities of CRM were combined with a rule-based interpreter and user interface to produce ***GRAIL*** *(**G**ene **R**ecognition and **A**nalysis **I**nternet **L**ink), which allows users to submit DNA sequences by e-mail and electronically receive for each potential exon (1) an analysis of potential exon positions with strand assignment and preferred reading-frame determination, (2) a qualitative evaluation, and (3) a protein database comparative search. [Figure first published in E. C. Uberbacher and R. J. Mural, "Locating Protein-Coding Regions in Human DNA Sequences by a Multiple Sensor-Neural Network Approach,"* Proc. Natl. Acad. Sci. USA ***88****, 11261–65 (December 1991). Figure provided by E. Uberbacher, ORNL. For more information, refer to Uberbacher, Mural, and R. Mann in the Index to Principal and Coinvestigators.]*

Informatics

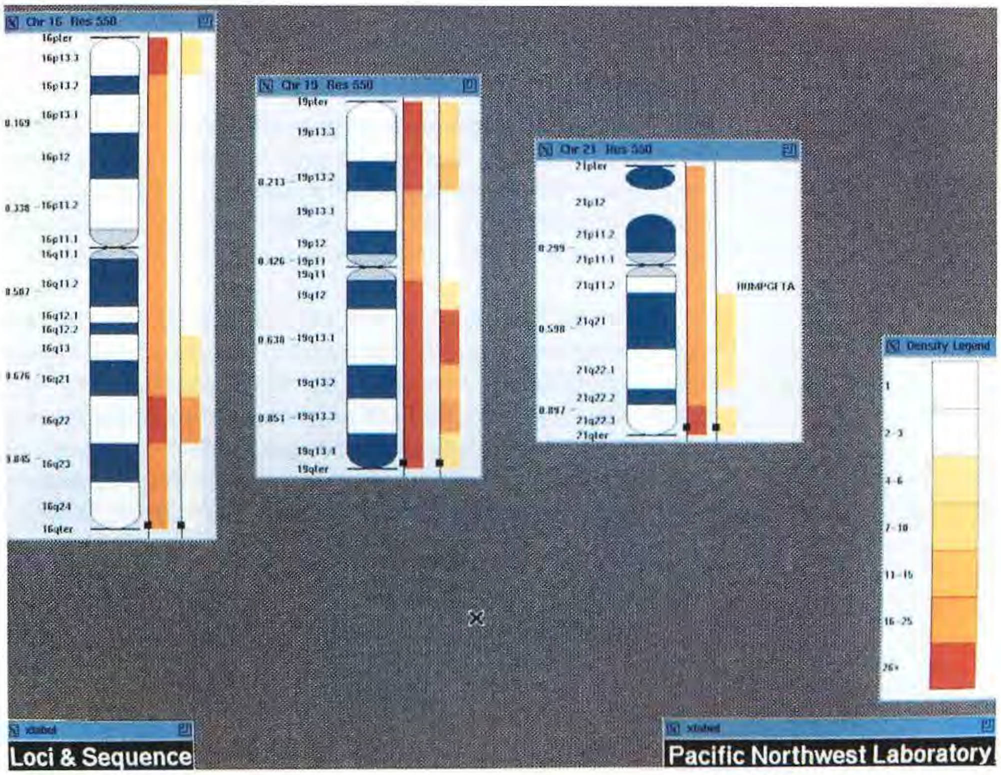
Multiple informatics capabilities will be crucial to the successful application of data derived from the genome project. Informatics expertise, software, and hardware are being developed in the following areas: chromosome map assembly, databases, DNA sequence analysis, and laboratory automation.

Map Assembly

Algorithms for automatically assembling physical maps from cloned fingerprint data have been further improved (E. Branscomb, LLNL; M. Cinkosky, V. Faber, J. Fickett, and D. Torney, LANL).

Software permitting fast parallel computations on multiple computers was developed to speed computation-intensive mapping analyses (E. Branscomb, LLNL).

A computer communication and interrogation system is being assembled to minimize redundancy during the production of STS chromosomal markers from cDNAs. Participating laboratories will rapidly query distant databases to determine the novelty of a candidate mRNA/cDNA before further pursuing the STS-generation process.



Color-Density Mapping of Objects Taken from Genome Data Base and GenBank® Using GnomeView. *GnomeView provides a graphic view of mapped and sequenced regions and their density. Two successive queries were sent out to chromosomes 16, 19, and 21, the first for mapped loci and the second for sequenced regions. Results of queries are shown to the right of the blue and white ideogram for each chromosome. Mapping details in both figures include a list of all loci found; short descriptions are available in a separate window (not shown). (Photograph provided by Richard Douthart, PNL. For more information, refer to Douthart in the Index to Principal and Coinvestigators.)*

Highlights of Research Progress: Informatics

Databases

Graphical interfaces for mapping databases were constructed to display several different types of aligned chromosomal data and provide expandable views [R. Douthart, Pacific Northwest Laboratory (PNL); J. Fickett, LANL; S. Lewis, Lawrence Berkeley Laboratory (LBL); R. Overbeek, Argonne National Laboratory (ANL)].

The electronic Laboratory Notebook database and similar databases are being continuously expanded to include new data types as mapping strategies evolve (J. Fickett, LANL).

The internationally available Genome Data Base (GDB), housed at Johns Hopkins University and cofunded since September 1991 by DOE and NIH, is the primary reference database for human chromosome mapping data produced in the United States and abroad. The organizational structure of GDB is shown on the opposite page (P. Pearson, GDB).

In a collaboration between LLNL and GDB, computer system interfaces have been devised for automatically transferring large amounts of data from mapping centers to GDB for integration into and updating of chromosome maps.

Enhancements of the GenBank® DNA sequence database located at LANL continue. Primarily supported by NIH with contributions from DOE, GenBank exchanges data daily with European and Japanese databases. GenBank has expanded its electronic data-publishing facilities and has reached agreements with a number of journals to facilitate electronic publication of large volumes of DNA sequence data (J. Cassatt, NIH).

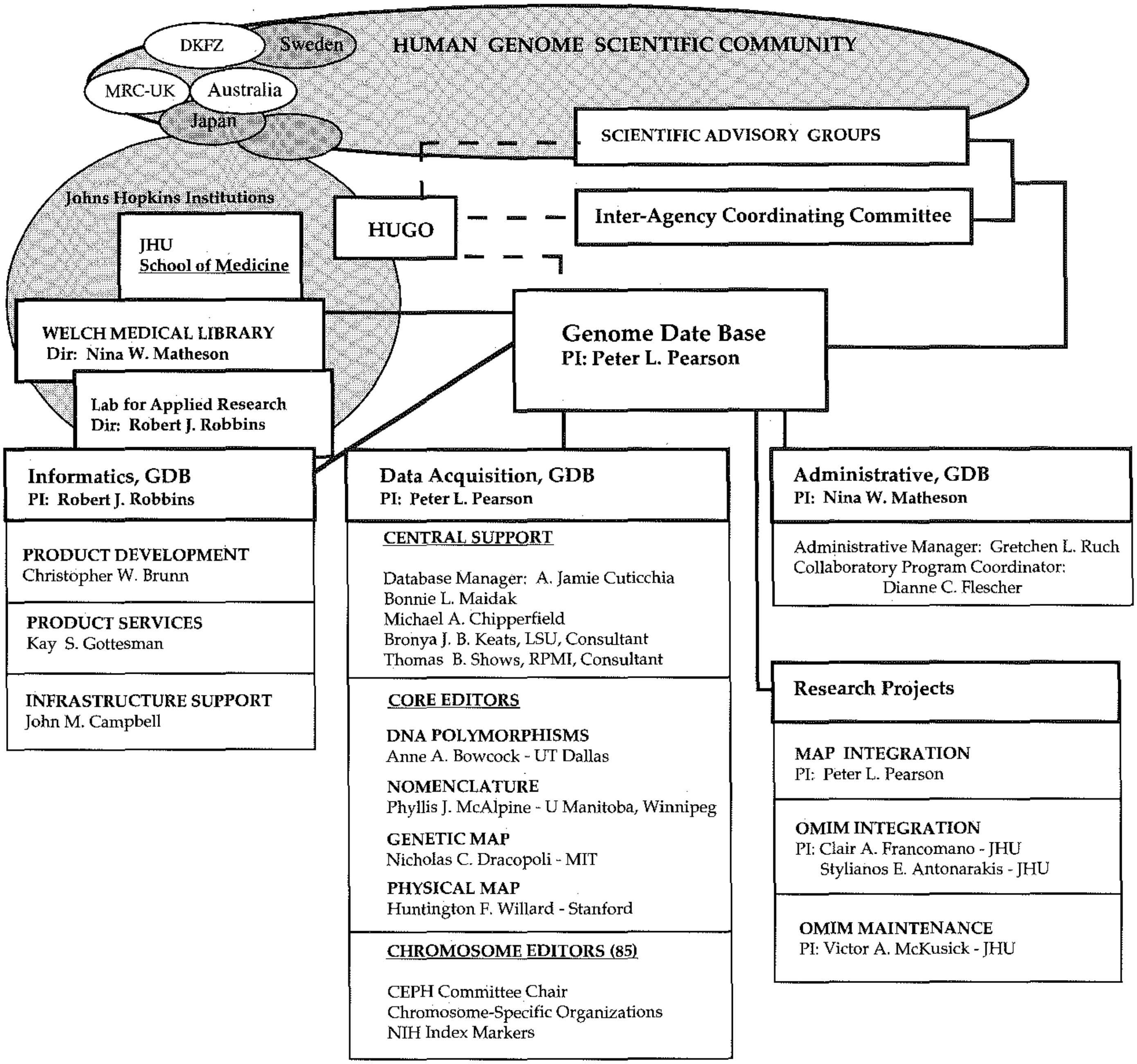
Sequence Analysis

gm, developed at New Mexico State University, is the first DNA sequence analysis algorithm capable of recognizing and ordering the set of protein-coding regions (exons) from among the noncoding regions (introns) comprising a gene, rather than predicting isolated protein-coding sequences. **gm** has been distributed to laboratories worldwide (C. Fields, now at NIH, and C. Soderlund, now at LANL).

Gene Recognition and Analysis Internet Link (GRAIL), a novel neural network–based algorithm for identifying exons within DNA sequences, is online at Oak RIdge National Laboratory (ORNL) to serve the biological community by automatically analyzing sequences. From a number of examples, this artificial intelligence system learns several distinct sequence characteristics through which exons can be recognized. GRAIL automatically accepts input sequences sent to ORNL over Internet and returns the output analysis to the sender (R. Mural and E. Uberbacher, ORNL).

Laboratory Automation

Advances continue in the linking of laboratory instruments directly to data-acquisition computers and analysis software at the LANL, LLNL, and LBL human genome centers.



GDB Organizational Chart

[Figure provided by Nina W. Matheson, GDB]

Sequencing

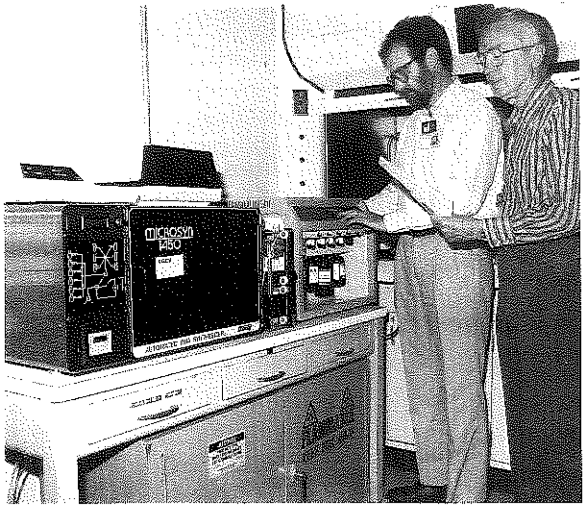
Highlights of Research Progress

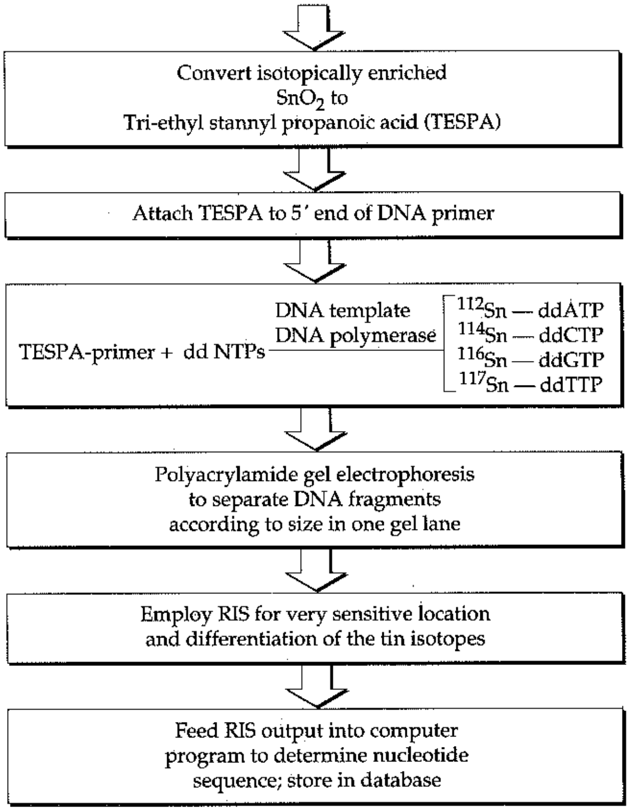
The DOE Human Genome Program has supported both evolutionary (incremental, gel-based) improvements to classical sequencing methods and several revolutionary (completely novel, gel-less) technologies. Steady advances have occurred in the evolutionary area with the implementation of automated sample preparation, multiplex sequencing, and strategies that minimize the need for prior subcloning.

Gel Sequencing Approaches

Multiplex sequencing systems have matured enough for transfer to the commercial sector (G. Church, Harvard Medical School; R. Gesteland, University of Utah).

The readout of multiplexed gels and blots using stable isotopes as nucleic acid labels has the potential to increase sequencing speeds by at least a factor of 10 because resonance ionization mass spectroscopy is capable of differentiating many isotopes (H. Arlinghaus, Atom Sciences, Inc.; K. B. Jacobson, ORNL).




Convert isotopically enriched SnO_2 to Tri-ethyl stannyl propanoic acid (TESPA)
Attach TESPA to 5′ end of DNA primer
TESPA-primer + dd NTPs
DNA template
DNA polymerase
^{112}Sn — ddATP
^{114}Sn — ddCTP
^{116}Sn — ddGTP
^{117}Sn — ddTTP
Polyacrylamide gel electrophoresis to separate DNA fragments according to size in one gel lane
Employ RIS for very sensitive location and differentiation of the tin isotopes
Feed RIS output into computer program to determine nucleotide sequence; store in database

DNA Synthesizer. *Using a synthesizer that attaches up to about 50 mononucleotides in any order, researchers synthesize oligonucleotides for DNA sequencing by hybridization and sequencing with stable isotopes. These promising new sequencing approaches eliminate the use of radioisotopes and their attendant problems such as personnel exposure to radiation and the prohibitive cost of radioactive waste disposal. In stable isotope DNA sequencing, as shown in the diagram, the synthesizer attaches a hexylamine that reacts with the ester of triethyl stannyl propanoic acid (TESPA) to provide tin-labeled DNA. Subsequent detection of fragments by resonance ionization spectroscopy (RIS) promises to speed sequencing by at least a factor of 10 due to its ability to differentiate many isotopes. [Reprinted with permission from K. B. Jacobson and H. F. Arlinghaus, "Development of Resonance Ionization Spectroscopy for DNA Sequencing and Genome Mapping,"* Anal. Chem. ***64****, 315–28 (1992), copyright 1992, American Chemical Society. Photograph provided by K. Bruce Jacobson, ORNL. For more information, refer to Jacobson and R. S. Foote in the Index to Principal and Coinvestigators.]*

Chemiluminescent label systems are now substituting for the less-desirable radioactive labels in many applications (I. Bronstein, Tropix, Inc.).

Systems have been developed to retain chromosome continuity information by bypassing the customary subcloning step in the sequencing of recombinant DNAs (D. Berg, Washington University; C. Berg and L. Strausbaugh, University of Connecticut; J. Dunn and F. Studier, Brookhaven National Laboratory; R. Gesteland and R. Weiss, University of Utah).

Fractionation speeds on capillary and very thin slab gels are 10-fold faster than on traditional thick gels (N. Dovichi, University of Alberta, Canada; B. Karger, Northeastern University; L. Smith, University of Wisconsin).

The fluorescence/luminescence detection of fractionated nucleic acids has been significantly improved to allow detection of the smaller amounts of DNA loaded on capillary and thin slab gels (N. Dovichi, University of Alberta; R. Mathies, University of California; E. Yeung, Ames Laboratory).

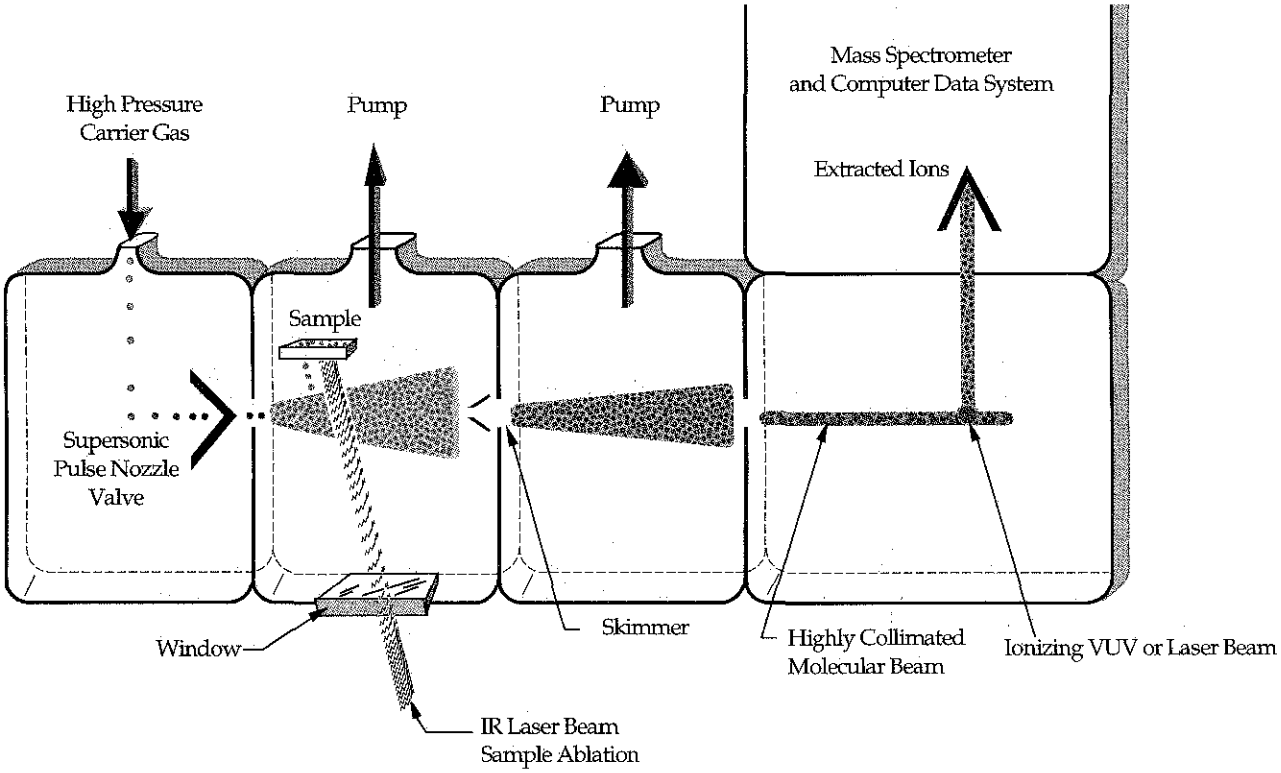
Over 300 kb have been sequenced from human and mouse T-cell receptors, providing fundamental new insights into the molecular biology of the immune response (L. Hood and T. Hunkapiller, California Institute of Technology).

Gel-less Sequencing Technologies

The technology for interrogating or sequencing clones by hybridization with short oligomers has passed a second proof-of-concept test. Three unknown DNA fragments were fully and accurately sequenced (R. Crkvenjakov and R. Drmanac, ANL).

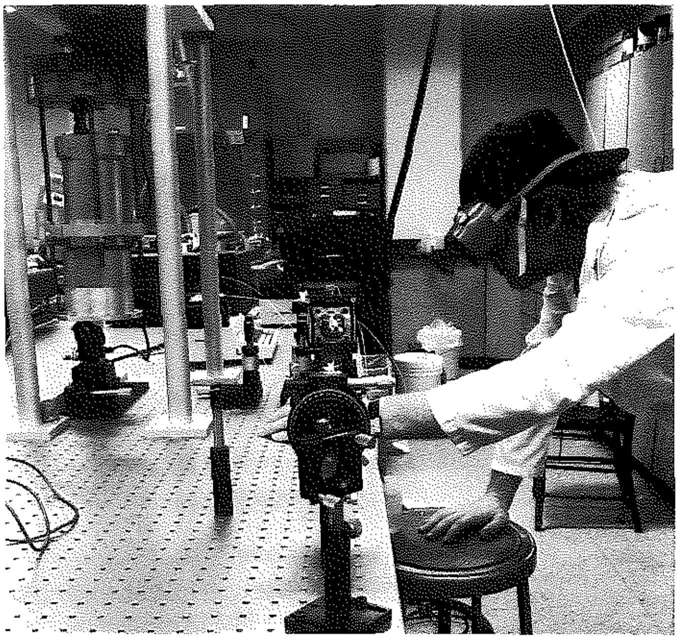
In research and development for single-molecule sequencing by processive nucleotide release, the capacity to detect single nucleotides by laser-induced fluorescence has been demonstrated (R. Keller and J. Jett, LANL).

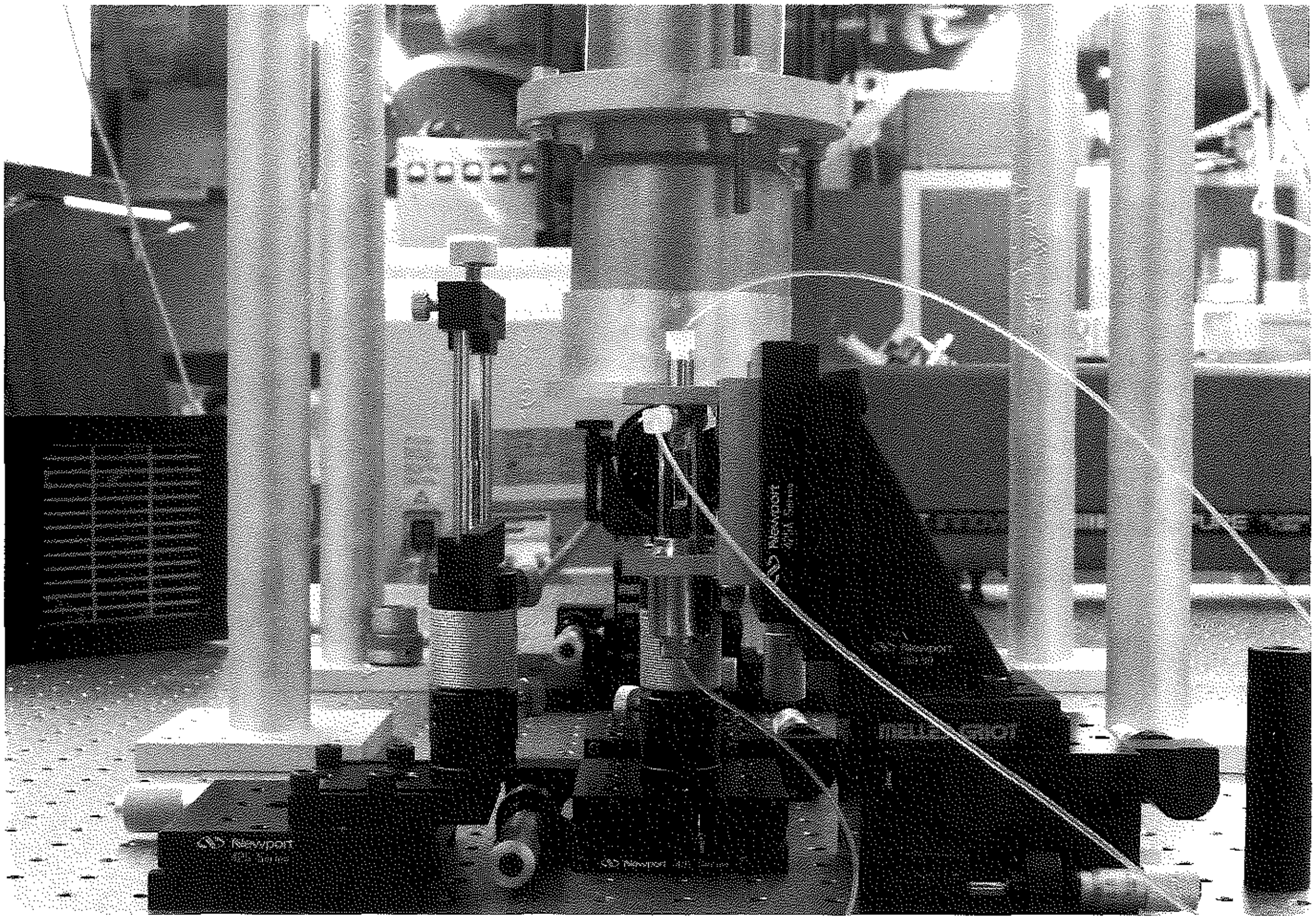
Progress is being made in developing methods to sequence DNA using lasers coupled to a mass spectrometer. The great advantage of these approaches is that the mass spectrum can be acquired in milliseconds (C. Chen, ORNL; J. Jaklevic, W. Benner, and J. Katz, LBL; L. Smith and B. Chait, University of Wisconsin; R. Smith, PNL; P. Williams and N. Woodbury, Arizona State University).



Sequencing by Mass Spectrometry. *Several gel-less sequencing technologies are now under development. For the particular approach shown in this figure, the DNA sample is ablated from a metal surface with an infrared (IR) laser and collimated before being ionized to near-threshold photoionization by a vacuum ultraviolet (vuv) laser. Choosing the minimum photon energy necessary for exceeding the ionization threshold leaves almost no excess energy to cause ion fragmentation. Because only parent ions are made, data analysis by reflectron time-of-flight mass spectrometry is simplified. The goals of this approach are to eliminate the need for both gel electrophoresis and radioactive tagging and to greatly increase the DNA sequencing rate. (Figure provided by C. H. Winston Chen, ORNL; accepted in 1992 for publication by* Rapid Commun. Mass Spectrom. *For more information, refer to Chen, M. G. Payne, and K. B. Jacobson in the Index to Principal and Coinvestigators.)*

DNA Sequence Detection by Imaging with a Charge-Coupled Device (CCD) Camera. *A CCD camera is operated in a unique readout mode known as time delay integration (TDI), which facilitates observation of fluorescent species in a flow-cytometer cell. DNA sequence analysis is accomplished by tagging DNA bases with fluorescent labels, cleaving the tagged nucleotides in a flowing stream, and detecting the fluorescent emission with a CCD camera. The camera can be used for spatially discriminating the fluorescent signal from the bulk background emission. In TDI mode, pixel imaging of the CCD array is clocked at the same rate as image translation; hence, the entire fluorescence signal can be integrated into a single charge packet while the background emission is distributed over the entire CCD array. The close-up photograph of the CCD camera (below) shows the excitation and imaging optics required to implement this technique. (Photographs provided by M. Bonner Denton, University of Arizona. For more information, refer to Denton and R. Keller in the Index to Principal and Coinvestigators.)*





Activities Addressing Ethical, Legal, and Social Issues Related to Human Genome Project Data

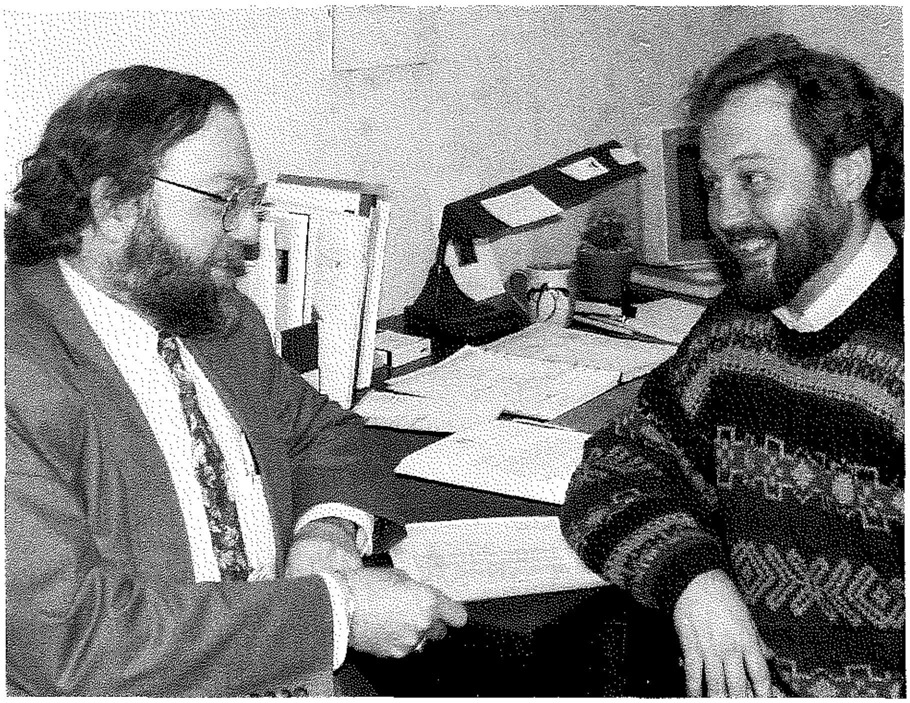
Highlights of Research Progress

In FY 1991, DOE activities on ethical, legal, and social issues (ELSI) included two conferences, three education projects, and three research projects. The first conference, Justice and the Human Genome, held in November 1991 at the University of Illinois College of Medicine, considered discrimination that could result from the use of genetic information about ethnic and other groups. The second conference, held in March 1992 at the Texas Medical Center Institute of Religion, focused on Genetics, Religion, and Ethics.

The three education projects on the science and the societal implications of data produced in the Human Genome Project, listed with their preparers, include (1) a module to be developed and distributed to all U.S. high school biology teachers (Biological Sciences Curriculum Study); (2) an educational television series, "Medicine at the Crossroads," which will address the role of genetics in understanding and treating disease (WNET, New York, cofunded with NIH and the National Science Foundation); and (3) a program of hands-on workshops for public officials and other nonscientists (Cold Spring Harbor Laboratory).

The three ongoing research projects, listed with the institutions developing them, are (1) a study of ethical issues arising from the rapid proliferation of genetic tests that can predict future disease in otherwise healthy individuals [National Academy of Sciences (NAS) Institute of Medicine, cofunded with NIH]; (2) a legal study of confidentiality protection for genetic data (Shriver Center); and (3) a study to consider problems in funding young investigators in biological and biomedical sciences (NAS).

Daniel Drell (DOE Human Genome Program) and Eric Juengst [NIH National Center for Human Genome Research (NCHGR)] discuss their agencies' respective programs on the ethical, legal, and social issues that may arise from data produced by human genome research. [Photo provided by Leslie Fink, NIH NCHGR.]



In its first 2 years, the DOE Human Genome Program funded a variety of ELSI activities, noted above.`To avoid being spread too thinly, the ELSI component of the DOE Program now focuses on confidentiality and privacy concerns raised by increased genetic data about individuals. This sensitive, personal information, which may predict disorders before symptoms occur or treatments are available, can affect a person's self-image, employability, status in the eyes of others, and ability to obtain health insurance. Since genetic knowledge can also lead to better understanding of disease causation and to more-accurate assessments of environmental affronts, a balance must be achieved between the health of the public and the privacy interests of the individual.

The DOE Human Genome Program is funding six new projects covering ELSI activities in research and education. One of the three projects investigating genetic discrimination will compare two states (Florida and Georgia), contrasting their genetic testing, screening, and counseling programs and the impact on different ethnic and socioeconomic communities. Another will examine the impact of two genetic conditions (cystic fibrosis and sickle cell disease) on African-Americans and Caucasians. A third will identify particular social institutions that may engage in discrimination and will consider whether the discrimination, if present, is the result of ignorance or systematic policy. A fourth project will explore in detail (a) the effect of genetic knowledge on the right of privacy and (b) the uses of genetic information in public health planning. A fifth project will develop a program of educational workshops for secondary and high school science teachers, focused on both the science and the ethical, legal, and social issues arising from data generated by human genome research. A sixth project will involve a second educational television series, "The Secret of Life" (WGBH, Boston), which will address the current revolution in molecular biology and genetics.

Other activities include conferences on Genes and Human Behavior: A New Era? (October 1991); Computers, Freedom, and Privacy (March 1992); and Science, Technology, and Ethical Responsibility (scheduled for June 1992).

While very challenging issues are raised by genome research, solutions are not simple; defensible rights often exist on both sides of any issue. Further research is needed, as well as activities to promote public awareness and assist in policy development. Also, with the increasing use of computers to assemble, store, and organize data (including genetic data) into large databases, the issues of security and access control become more acute. To begin reorienting and better defining the scope of ELSI activities in the DOE program, the DOE-NIH Joint ELSI Working Group has established a collaborative effort on privacy to identify an ELSI research agenda and develop a more detailed approach to some of these concerns.

Technology Transfer and Industrial Collaboration

Highlights of Research Progress

Technology transfer, considered one of the three most important facets of the DOE mission (along with meeting the nation's defense and energy needs), is enhancing U.S. investment in research and technological competitiveness. By creating new products, markets, and jobs, the rapid deployment of technology from the research laboratory to the marketplace can play an important role in vitalizing the U.S. economy. A vast potential exists for commercial development of genome resources and technology; applications to clinical medicine have already begun.

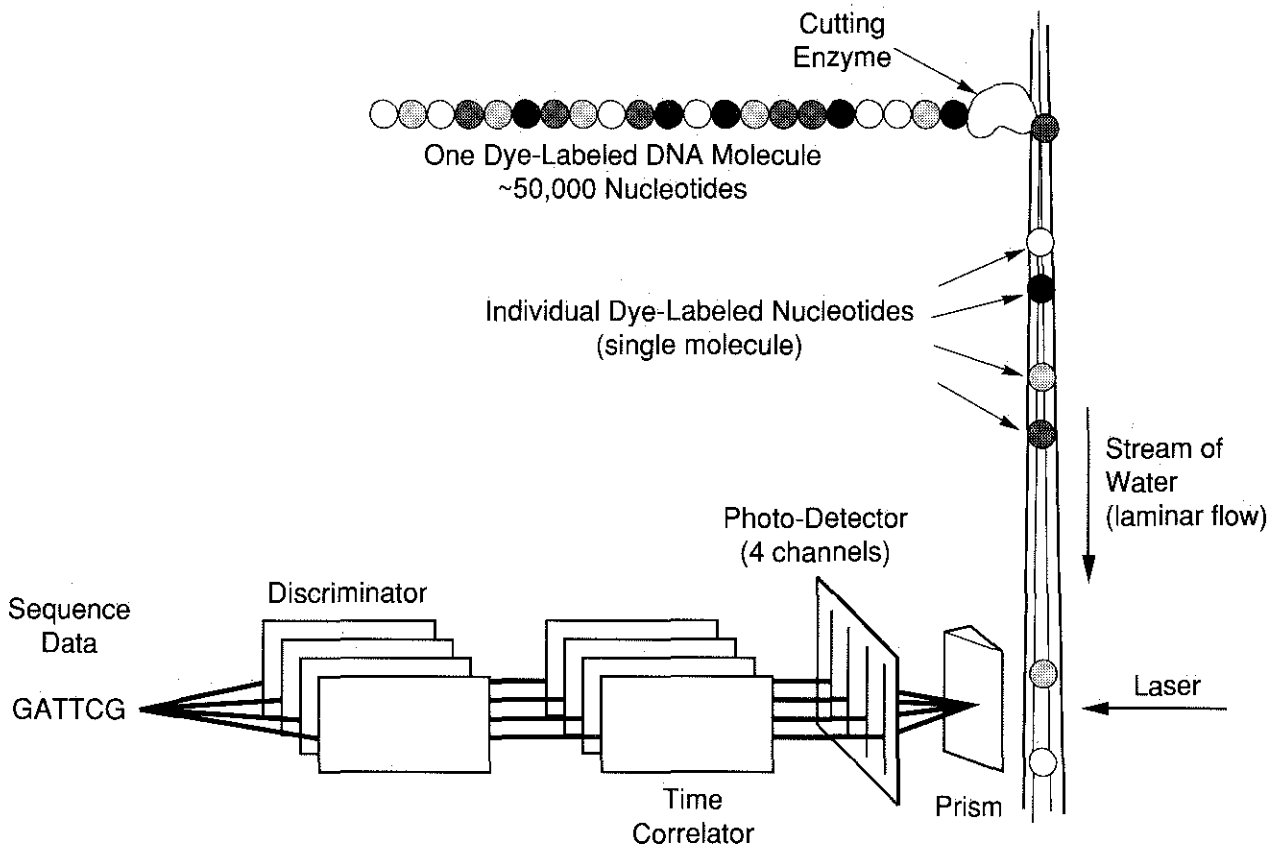
All participants in the Human Genome Program are encouraged to engage in active collaborations with the private sector and transfer their resources and technologies for commercial development.

Each national laboratory has a technology transfer office. The LLNL, LBL, and LANL human genome centers provide a variety of opportunities for collaborations on joint projects or for obtaining direct access to technology. They are also exploring additional ways to increase cooperation with the private sector; a number of interactive projects are now under way, and additional interactions are in the preliminary stages. In some instances, private industries are marketing technologies developed at DOE-sponsored research laboratories and are providing research funds or other resources to the centers; other collaborative programs involve joint development of technologies and their applications to achieve project goals.

One mechanism being used by the DOE national laboratories is the Cooperative Research and Development Agreement (CRADA). The first CRADA in the genome project, established by DOE in the spring of 1991, was between Life Technologies, Inc. (LTI) and the LANL Center for Human Genome Studies for technologies developed in the single-molecule sequencing project. In this project an LTI-modified DNA polymerase will be used to label a single DNA strand with four different fluorescent, base-specific tags. After an exonuclease cuts the labeled nucleic acid base pairs from the DNA, the labeled bases will be induced to fluoresce as they pass sequentially through a focused laser beam. The bases can be identified and counted by a sensitive photodetector (see figure on p. 25 for more information). If successful, the technology will allow sequencing of 50,000-bp DNA fragments at 1000 bp/s. LTI will have the first opportunity to license products resulting from the joint effort and would pay royalties to LANL under such a license.

Potential commercial advancements in the Human Genome Program have also been recognized outside the genome community. *Research and Development* magazine selected an achievement by Edward Yeung and other Ames Laboratory scientists as one of the 100 most significant developments of 1991. This R&D 100 Award was given for the development of a user-friendly instrument that detects with extremely high sensitivity the fluorescent molecule concentration (based on laser-excited fluorescence), an improvement that may lead to routine high-speed DNA sequencing by capillary gel electrophoresis. A U.S. patent for portions of this technology has been issued, and several commercial manufacturers are considering the possibilities of marketing the instrument.

A technology pioneered by LLNL to identify chromosomal abnormalities (e.g., aneuploidy, translocations, and deletions) has been licensed to Imagenetics, Inc., a medical diagnostics firm that will manufacture the technology and provide funding for future research and



Rapid DNA Sequencing Based on Single Molecule Detection. *A DNA rapid sequencer detects single chromophores by laser-induced fluorescence in flowing sample streams. After the bases in a single DNA fragment are fluorescently labeled, the fragment is attached to a support and moved into a flowing sample stream. Individual bases are detected by laser-induced fluorescence as they are cleaved from the DNA molecule by a cutting enzyme. As a molecule passes through a focused laser beam, it is repeatedly cycled from the ground electronic state to an excited electronic state with the emission of a photon on each cycle. Under optical saturation conditions, the number of photons emitted (the burst size) can approach the transit time of the molecule across the laser beam divided by the fluorescent lifetime. The burst size can be thousands or hundreds of thousands of photons, ultimately limited by the photostability of the molecule. Because the photon bursts are correlated in time, they can be distinguished from the background fluorescence. The emission spectra of the four fluorescent tags allow identification of the four bases.*

The DNA rapid sequencer was first developed by LANL, which has licensed use of the technology to Life Technologies, Inc., in the first Cooperative Research and Development Agreement (CRADA—a technology transfer mechanism) established by the DOE Human Genome Program. An advantage of this approach is the ability to sequence large DNA fragments, thereby significantly reducing the amount of subcloning and the number of overlapping sequences required to assemble megabase segments of sequence information. Another advantage is the elimination of radioactive materials used in other sequencing methods. [Reprinted by permission of the publisher from "Rapid DNA Sequencing Based Upon Single Molecule Detection," by Lloyd M. Davis et al., Genet. Anal.: Tech. Appl. ***8****(1), 1–7, Copyright 1991 by Elsevier Science Publishing Co., Inc. Figure provided by Richard Keller, LANL. For more information, refer to Keller in the Index to Principal and Coinvestigators.]*

Highlights of Research Progress: Technology Transfer and Industrial Collaboration

development. This technology involves the use of specially developed fluorescent dyes called Whole Chromosome Paints™ to detect diseases such as cancers and leukemia. Whole Chromosome Paints are being marketed by LTI.

Some other technology transfers from DOE-sponsored genome research, both at the national laboratories and extramurally, are highlighted below. In progress or awaiting finalization are many more developments and agreements, some of which cannot be disclosed at this time because of their proprietary nature.

Resources. Collaborative agreements have aided in the further development of several new technologies used in genome research, as well as in their commercial applications. New methods are being evaluated for use in isolating mRNA, chromosomes, and restriction fragments; in amplifying hybridization signals; and in extending DNA molecules. In addition, bacterial host strains have been developed that give greater stability to cosmid constructs containing human DNAs. Improvements are being made in DNA detection methods by the development of new probes, stains, and fluorescent dyes.

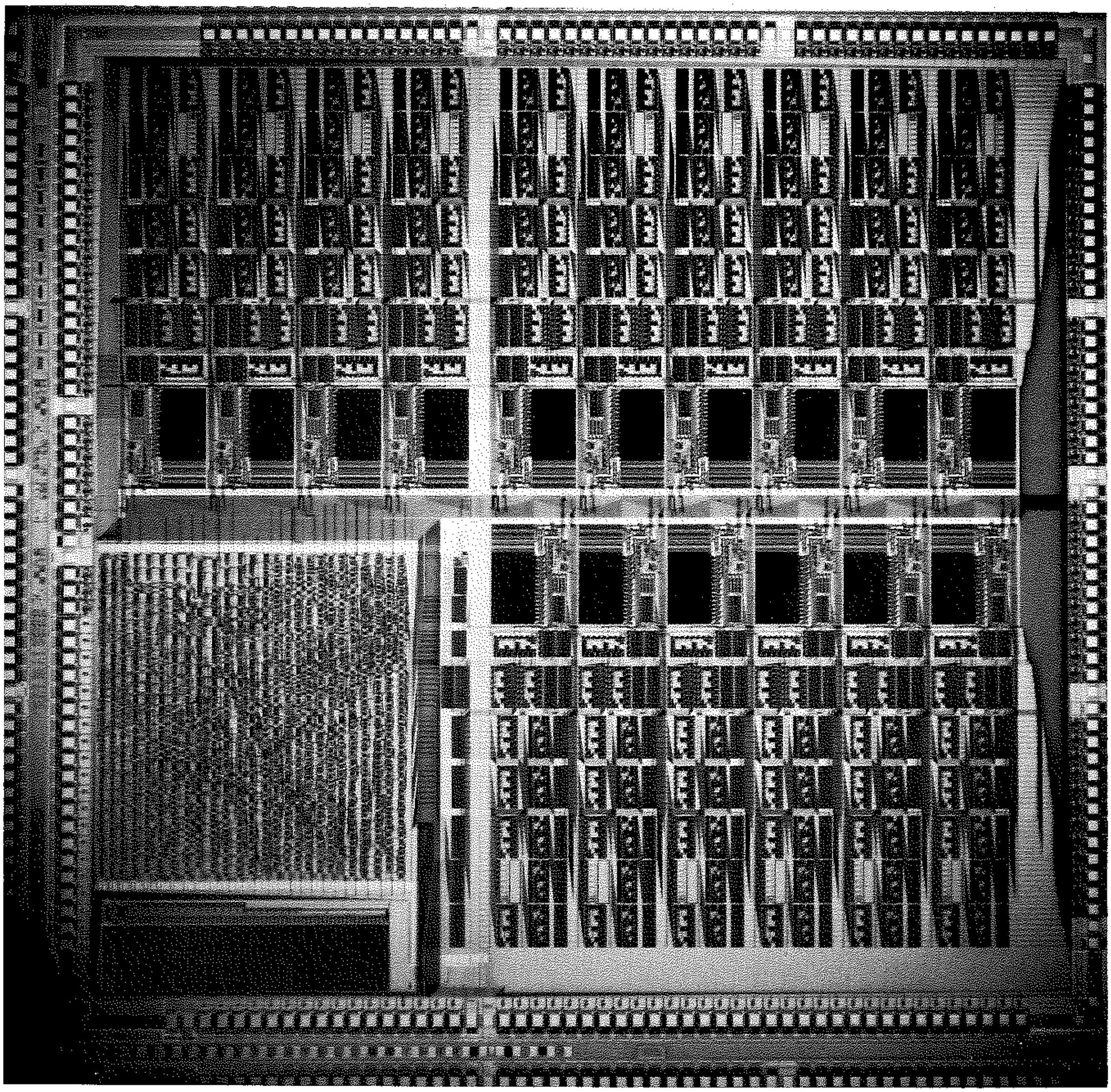
As a result of the recent cloning of the fragile X gene, several companies are negotiating for licenses to develop assays for diagnosing fragile X syndrome, probably the most frequently inherited form of mental retardation.

Hardware. Automation and enhancement of data collection and analysis has been the goal of many collaborations with the commercial sector. Equipment is being designed to automate (1) the production of high-density arrays on agarose or filters and (2) clone fingerprinting by gel electrophoresis (as well as the data collection and analysis software). Advanced applications for robotic systems are also being developed.

The resolution of DNA fragments is being enhanced by improvements in pulsed-field gel electrophoresis. Resonance ionization spectroscopy is being modified to enable rapid detection of stable isotope labels on DNA following gel electrophoresis. A commercial gel scanner is being developed for reading DNA gels.

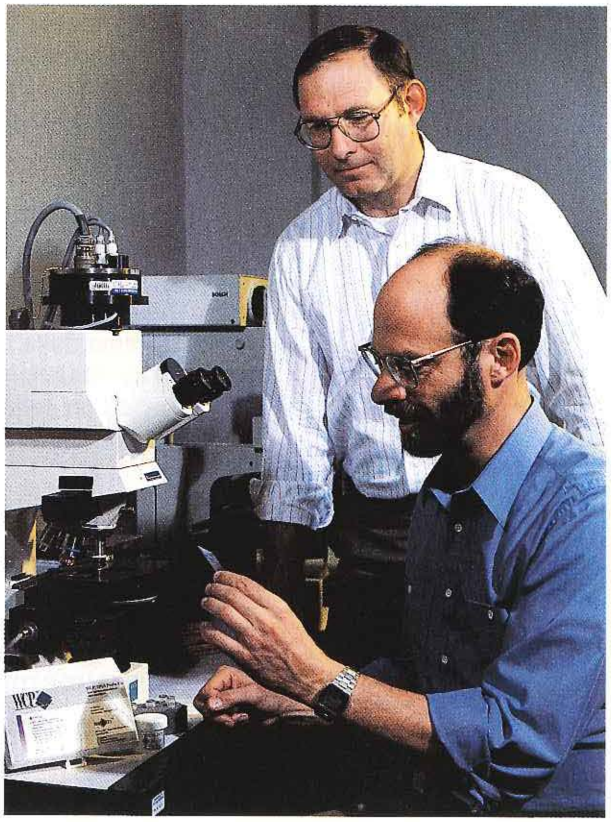
Software. To aid physical map construction, programs are being designed for efficient clone analysis. Several other image-analysis programs are being developed, including data-capture software for images from video screens in combination with a DNA molecule imaging system.

Sequencing. Multiplex sequencing technologies are being used to sequence pathogenic microbes.



Biological Information Signal Processor (BISP) Chip Used to Analyze DNA Sequence Data. *The BISP chip measures somewhat larger than 1cm^2 and has 400,000 transistors and 208 IO (Input Output) pins. Each chip has 16 processing elements (visible in the photograph) and operates at a maximum rate of 12.5 MHz. The chip is a silicon implementation of the Smith/Waterman dynamic programming algorithm for finding similarities between related but different strings of data. (Photograph provided by Tim Hunkapiller, California Institute of Technology. For more information, refer to Hunkapiller and L. Hood in Index to Principal and Coinvestigators.)*

New Technology Identifies Chromosome Abnormalities. *Developed at LLNL, whole chromosome painting permits significantly faster and more accurate identification of human chromosomal abnormalities frequently present in cancer, such as the exchange of material between chromosome numbers 7 (red stain) and 12 (aqua stain) shown in the large photograph on p. 29. A normal 7 can be seen at the far right. To its left, another copy of 7 has an aqua tip on its top, indicating the presence of material from 12. Similarly, the lower aqua-colored chromosome indicates a normal 12, while the upper aqua one has a red piece from 7. The display of two distinct colors on a single chromosome indicates a chromosomal exchange. The remaining chromosomes in this photograph retained a blue dye that is not specific for any particular chromosome.* ***(Inset)*** *Biomedical scientists and coinventors of the new technology prepare to look over a sample. This technology has been licensed by LLNL to Imagenetics, Inc., a medical diagnostics firm based in Naperville, Illinois. Whole Chromosome Paints™ will be sold to research laboratories worldwide by Life Technologies, Inc. (Gaithersburg, Maryland). (Photographs provided by Joe Gray, University of California, San Francisco, and LLNL.)*


BOSCH
WCP

Human Genome Center Research Narratives

Lawrence Berkeley Laboratory

Since its inception in 1987, the Lawrence Berkeley Laboratory (LBL) Human Genome Center has focused on developing the necessary research and analytical technology to speed genome mapping and decrease the cost of sequencing. Over the last year, LBL has strengthened its ties with the University of California, Berkeley, particularly in the biological sciences. This collaboration fosters interdisciplinary activities in biology, instrumentation, and informatics.

Biology

The biology component at LBL is concentrating on developing and improving mapping and sequencing strategies for human chromosome 21. To achieve these goals, investigators in each biology project draw on the expertise of the center's instrumentation and computing groups.

Two major biology projects are under way, and a third is in development. Physical mapping at LBL is focused on a 10-Mb region of human chromosome 21, and over 90 unique chromosome 21–specific yeast artificial chromosomes (YACs) have been located by fluorescence in situ hybridization (FISH). A new method has been developed that permits rapid isolation of chromosome-specific YACs, using probes isolated from flow-sorted chromosome libraries from Lawrence Livermore National Laboratory. In addition, cDNAs specific to a given YAC are being isolated by an automatable procedure based on magnetic beads.

The second major biology effort involves testing new approaches to physical mapping and genomic sequencing. These projects exploit current methods, such as FISH and appropriate pooling strategies, for efficient isolation of overlapping clones. In addition, new work has begun on subcloning and ordering libraries of clones for mapping and on the use of gamma delta transposons as the primer site for sequencing studies. Increased efficiency in constructing physical maps results from a clone-limited strategy for generating maps based on sequence tagged sites (STSs). This nonrandom selection strategy reduces the number of STS assays required and produces contigs that cover a larger fraction of the genome.

The third biology project is aimed at developing automated methods for generating genetic maps. A simple filter assay will be used to detect heterozygosity at mapped loci in yeast, mice, and human DNA samples.

Instrumentation

The instrumentation program within the LBL Human Genome Center has two major areas of effort: (1) biology and instrumentation development and support and (2) new instrumentation development based on emerging technologies. Supporting activities include the design and fabrication of gel boxes, automation of protocols on existing robotic frameworks, and the installation and networking of a variety of image-acquisition systems. In addition, advanced robotic [high-speed colony picking, robotic-based polymerase chain reaction, and DNA synthesis] and laboratory systems integration is under development.

Efforts to produce new, adaptable technologies for the genome program include optimizing large-molecule detection systems; designing versatile optical fluorescence systems for multiplex labeling; and developing microfabricated arrays for application to large-scale clone libraries, sequencing by hybridization, and other procedures. The use of computer-controlled robotic systems provides a mechanism for automatically capturing the vast amount of data generated by laboratory operations. This requires a close coordination between hardware and software development in laboratory system design that goes far beyond automation of a few discrete protocols.

Informatics

A major part of the computing and instrumentation effort is driven by biology projects. The center's computing group focuses on specific applications in four major areas: raw data acquisition and analysis, information tracking and management, data interpretation and comparison analysis, and development of software tools. Visual data for mapping (including in situ pictures, autoradiograms, ethidium gels, and chemiluminescent staining) are handled by BioPix, a set of programs that assemble and integrate data from image capture to analysis. A similar system is being developed for sequence data. The Chromosome Information System (CIS) allows biologists to search, edit, and compare various maps, markers, and related reference information and to interact with other programs to exchange data. The laboratory data analysis system uses existing software packages and provides system management and support throughout the center. New, in-house analysis packages are being devised for sequence alignment and assembly. Software development tools permit rapid design and modification of database management systems, thus facilitating increased productivity, vendor independence, and conceptual clarity.

Achievements

- Over 90 independent YACs averaging 100 kb were regionally assigned to human chromosome 21 by FISH. These YACs include genetic markers to help integrate maps.

- Two hundred unique probes were isolated for chromosome 21 and are being used to identify YACs from genomic libraries.

- A rapid cDNA clone-screening method uses immobilized YAC clones to screen cDNA libraries, which are then localized on specific chromosomes. An alternative screening method uses individual YACs or cosmids attached to magnetic beads to isolate specific cDNAs, a method that can be readily automated to speed identification of coding sequences for physical mapping.

- Marker-selected libraries, highly enriched for clones containing $(CA)_n$ repeats, were constructed from primary genomic libraries. These enriched libraries increase the efficiency of screening almost 50-fold.

- A probe-mapping procedure determines the distance between the probe and the chromosome or YAC end. This method, which uses X rays to break large DNA pieces randomly, can be used to map cDNAs and to estimate the length of entire genes.

- A double-ended, clone-limited strategy for physical mapping of chromosomes was devised. This strategy maps chromosomes on the order of 100 Mb and should result in larger contigs with a minimum of assays.

- CIS, developed by the genome center computing group, was used to produce consensus maps at workshops on human chromosomes 3 and 21 and is being expanded for use with a number of plant species in the Plant Genome Program of the U.S. Department of Agriculture.

- High-level database design tools have been developed to permit molecular biologists to define data objects in a way that captures biological concepts. The software automatically generates low-level commands for a commercial database management system, facilitating the evolutionary development of modular system components. These tools are also being used by researchers to design the Superconducting Super Collider database and the Integrated Genome Database.

- A variety of mechanical, electrical, and chemical means have been used to manipulate DNA molecules; these methods include stretching molecules physically by externally applied electrical fields and guiding the molecules through grooves in a glass surface; digesting and separating single molecules; and picking up, transporting, and releasing DNA with scanning tunneling microscope (STM) tips.

- Investigation of the feasibility of using STM for visualizing the individual bases of single-stranded DNA has shown that while purines and pyrimidines can be distinguished from each other, two bases in the same class cannot be differentiated by this method.

- A fast, filter-based assay was developed to identify single base-pair polymorphisms, eliminating the need for gel assays.

- Higher throughput was achieved through the construction of a dedicated high-speed colony-picking workstation. The pick rate is 10 to 20 times faster than the initial picking system and both faster and more accurate than a highly qualified human. The new picker arrayed an entire library of over 10,000 clones in 1 day.

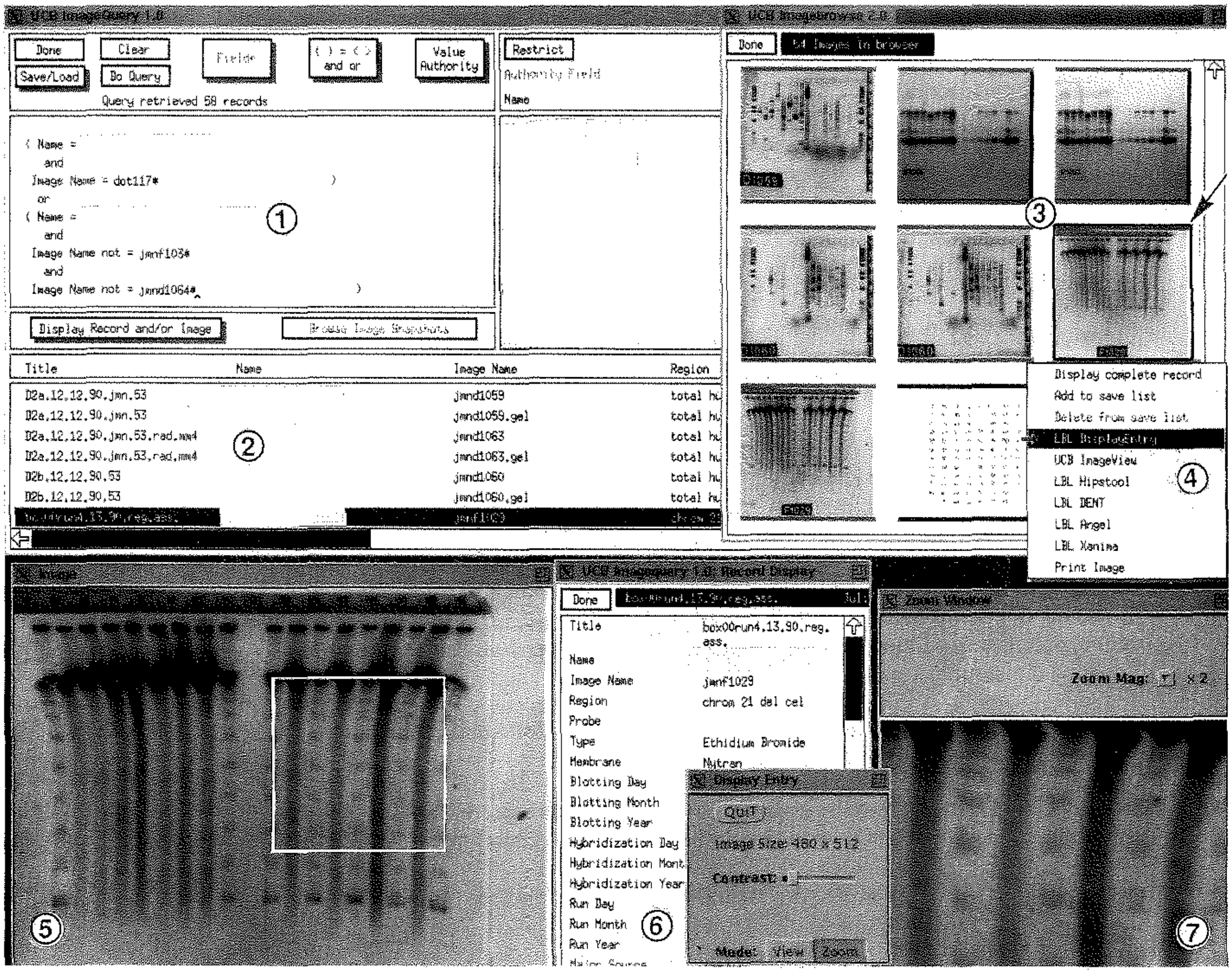
- Robots have been modified for use with a number of chemistry protocols, including cosmid and YAC library replication, various pooling schemes, and high-density filter array production. Using the robot to replicate libraries has made copies available to researchers in the private sector and in other national laboratories.

Future Plans

- Construction of a 10-Mb contig of human chromosome 21 based on overlapping YACs. The sequence will be determined by the most efficient strategy available.
- Sequencing of a P1 clone. Subclone assembly will use a nonrandom strategy, and primer sequences will originate in the transposon gamma delta.
- Construction of chromosome genetic maps of human chromosomes 16 and 19 in collaboration with other DOE genome centers. A simple gel-based heterozygosity assay is being developed to support this research.
- Development of a computational biology program within the computing group to design and implement new algorithms for sequence assembly. Preliminary data will come from collaborations with other genome centers.
- Design and implementation of a software tool suite for managing information and for optimizing the unique strategy of particular research groups. As large-scale sequencing projects develop, new acquisition and analysis software will be integrated into CIS.
- Implementation of QUEST, a database tool that will provide a single entry point to the conceptual data model. QUEST will then implement automatically any changes in the user interface, the database query procedures, and the database schema definition.
- Optimization of improved detectors and the associated mass spectrometry system for large biological molecules.
- Automation of handling and analysis of dot-blot hybridization experiments and the implementation of a high-speed colony-picking apparatus.

For more information on the LBL Human Genome Center, contact Jasper Rine, Director, or Sylvia Spengler, Deputy Director, at 510/486-4943.

(a)



Query retrieved 58 records

```
( Name =
   and
 Image Name = dot117*                    )
  or
( Name =
   and
 Image Name not = jmnf103*
   and
 Image Name not = jmnd1064*                )
```

Name

Title	Name	Image Name	Region
D2a.12.12.90.jmn.53		jmnd1059	total hu
D2a.12.12.90.jmn.53		jmnd1059.gel	total hu
D2a.12.12.90.jmn.53.rad.mm4		jmnd1063	total hu
D2a.12.12.90.jmn.53.rad.mm4		jmnd1063.gel	total hu
D2b.12.12.90.53		jmnd1060	total hu
D2b.12.12.90.53		jmnd1060.gel	total hu
box00run4.13.90.reg.ass.		jmnf1029	chrom 21

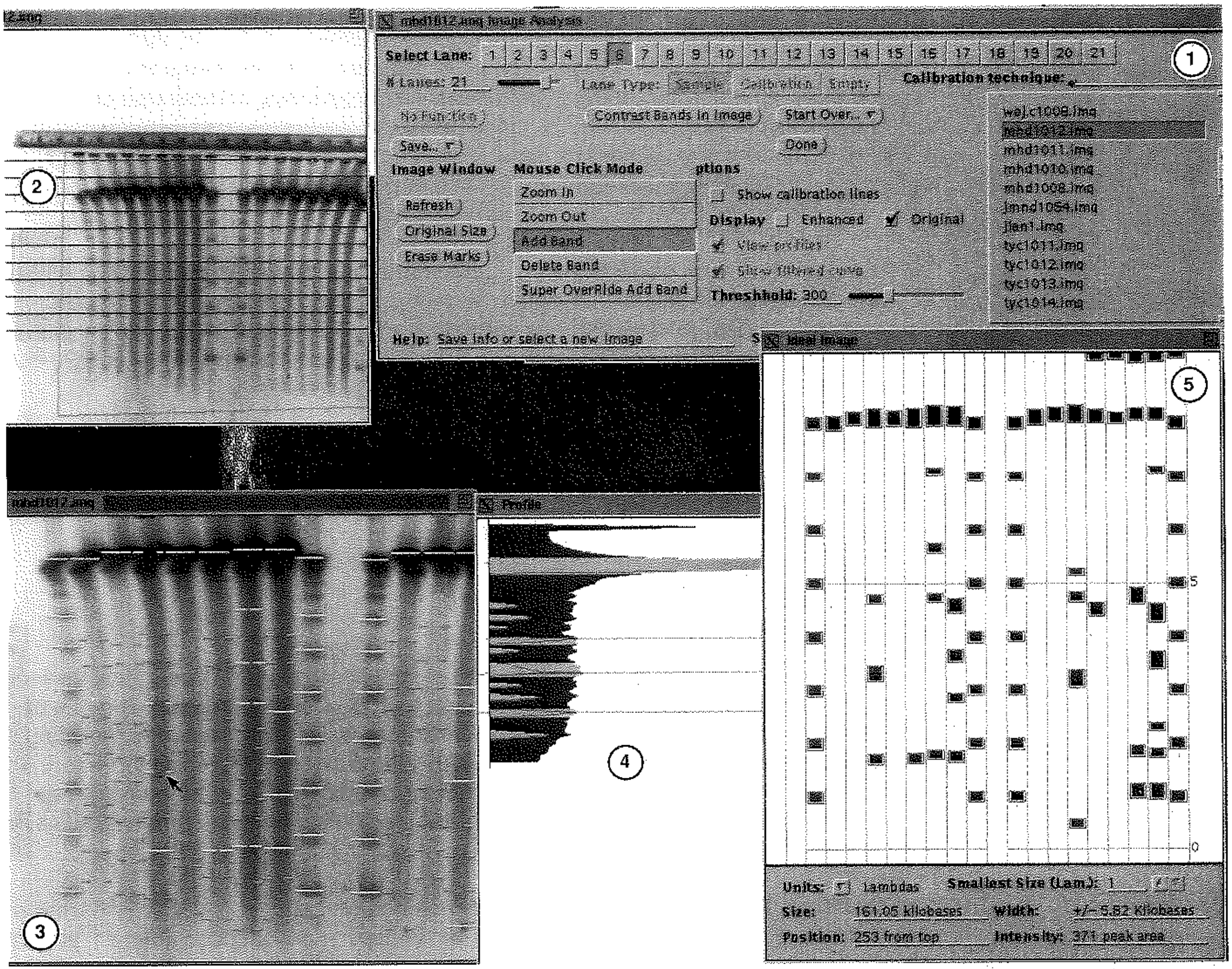
Title	box00run4.13.90.reg.ass.
Name	
Image Name	jmnf1029
Region	chrom 21 del cel
Probe	
Type	Ethidium Bromide
Membrane	Nytran
Blotting Day	
Blotting Month	
Blotting Year	
Hybridization Day	
Hybridization Mont	
Hybridization Year	
Run Day	
Run Month	
Run Year	

LBL Human Genome Center Computing Group. *This group actively supports and develops software for molecular biology and researches new techniques. ImageQuery [obtained and adapted from the University of California, Berkeley (UCB)] and the Chromosome Information System (see figure on p. 2) illustrate tools developed or adapted at LBL to aid in raw data acquisition and analysis and experimental data management.*

ImageQuery is a modular, distributed image database and analysis system for production chromosome mapping. ImageQuery integrates commercial and research software for image capture and annotation. Users select by queries or by browsing through image icons and associated text, and ImageQuery returns a list with postage stamp–sized icons of the images. Users can then select from this list to retrieve the full description or execute a program for a detailed image analysis.

(a) *ImageQuery database interface lets users select images of interest from a computer archive for display and analysis. This workstation screen shows several "windows" from a typical ImageQuery session. (1) A user specifies criteria to select a subset of images. (2) A one-line summary description is displayed for each retrieved image. Individual images can be selected for further manipulation by choosing summary lines (highlighted in black) or (3) by clicking on individual image icons. (4) Using the pop-up menu shown under a selected image (see arrow), (5) the user displays the full image, (6) its descriptive components, and (7) a zoom view of the image outlined in window 5.*

(b)



Help: Save info or select a new image

Lanes: 21

Threshhold: 300

Units: lambdas

Smallest Size (Lam.): 1

Size: 161.05 kilobases

Width: +/- 5.82 Kilobases

Position: 253 from top

Intensity: 371 peak area

Other programs for processing, analysis, and display are invoked directly from ImageQuery. Imaging devices, archival mass storage, user workstations, and software used in conjunction with this system all run in LBL's network-based computing environment, which includes the communications protocol TCP/IP, the Network File System, and the ethernet local area network.

(b) *Electrophoresis gel image analysis software is invoked from ImageQuery. This example shows (1) an experimental AnGel program control panel; (2) results from running automatic lane detection for a selected portion of the image; (3) results from automatic band detection overlaid on original image; (4) an intensity profile for a selected band; and (5) the resulting idealized image, which can be stored in the database for use in automated map construction programs. (Photographs on pp. 34–35 provided by LBL Human Genome Center. For more information, refer to S. Lewis and F. Olken in the Index to Principal and Coinvestigators.)*

Lawrence Livermore National Laboratory

Human Genome Center Research Narratives

The Human Genome Center at Lawrence Livermore National Laboratory (LLNL) is a multidisciplinary team effort that brings together chemists, biologists, molecular biologists, physicists, mathematicians, computer scientists, and engineers in an interactive research environment. Many of these individuals have previously collaborated on research projects in molecular biology, cytogenetics, mutagenesis, and instrumentation, as well as in the National Laboratory Gene Library Project (NLGLP). These projects have contributed substantially to the identification and characterization of human DNA repair genes, specifically the three on chromosome 19 that are a focus of interest at LLNL.

The short- and long-term goals of the LLNL effort are to (1) develop biological and physical resources useful for genome research, (2) model and evaluate DNA mapping and sequencing strategies, (3) couple these resources and strategies in an optimal way to construct ordered clone maps and DNA sequences of human chromosomes, and (4) use the map and sequence information to study genome organization and variation. To achieve these goals, the Human Genome Center is organized into three broad research and support areas, each consisting of multiple projects led by a principal investigator. Extensive interaction occurs within and among all projects that have as their common goal the construction of ordered clone maps of the human genome. The program structure of the center includes a core facility and projects that focus on physical mapping and enabling technologies.

Research and Support Areas

Coordination and collaboration take place with other research groups throughout the world that are involved in the genome initiative or other mutual scientific interests. The role of LLNL in the Human Genome Project is seen as encompassing several areas, including technology development, map construction, map interpretation, and integration with ongoing and new programs in structural biology and mutagenesis. The following three components are highly interactive; individual staff members often have responsibilities in more than one component.

Core facilities. The administrative group is concerned with budget oversight, external and internal meeting coordination, preparation of center reports, training coordination, property and space management, safety oversight, and secretarial support. The scientific core provides general support to the physical mapping effort, including cell culture and DNA extraction; library, probe, and clone management; oligonucleotide synthesis; fluorescence-based restriction mapping; and DNA sequencing. The core also facilitates material distribution to collaborators in the external community.

Mapping activities. Five projects represent the coordinated effort to obtain an overlapping set of clones for human chromosome 19 and to further characterize genomic organization:

- **Assembly, closure, and characterization of a chromosome 19 contig map.** The goal of this project is to construct an overlapping set of cosmid clones using a variety of techniques. An automated fluorescence-based restriction-fragment fingerprinting strategy is used to establish a foundation map of cosmid contigs. The contig closure effort will focus on using yeast artificial chromosomes (YACs) and cosmids with two hybridization-based techniques; one is based on fragments generated from *Alu* sequence primers or sequence tagged sites (STSs) by the polymerase chain reaction (PCR) and the second on RNA transcripts generated from the ends of cloned inserts.

- **Interdigitation of the physical and genetic maps of human chromosome 19.** The goals of this effort are to locate known genetic markers on the expanding contig map, to coordinate the isolation of chromosome 19–specific STSs, and to localize them on the cosmid map.

- **DNA sequence mapping by fluorescence in situ hybridization (FISH).** This project exploits the power of FISH on metaphase chromosomes, interphase cells, and pronuclear DNA. FISH will be used to determine the location of genes of interest and the relative order and orientation of the cosmid contigs.

- **cDNA mapping.** The goal of this project is to isolate, sequence, and map cDNAs–expressed in a variety of human tissues—that will become the STSs on which future studies of genetic organization and gene function will be based.

- **New mapping strategies.** New methods useful for library construction, contig closure, and overlap detection will be developed and validated. Focus is on improving *Alu*-PCR–based technology and pooling schemes to achieve closure of the chromosome 19 map with cosmids and YACs.

Enabling technologies. The following groups provide computational, resource, and instrumentation support for research activities:

- **Computational support for the Human Genome Center.** This group is responsible for mathematical modeling of mapping and sequencing strategies and the development and application of data analysis algorithms and software. They are also responsible for the construction and maintenance of interactive relational databases that enable internal and external data access, including development of graphical visualization tools.

- **NLGLP.** This project, a joint effort with Los Alamos National Laboratory, draws upon LLNL experience in flow instrumentation and chromosome sorting to construct human chromosome–specific libraries in lambda and cosmid vectors for use in physical mapping and other studies.

- **Instrumentation for cytogenetics and gene mapping.** This group is responsible for developing instrumentation to facilitate flow systems analysis and chromosome sorting and to support FISH.

Accomplishments

The LLNL Human Genome Center has made excellent progress in the construction of an ordered set of cosmids for chromosome 19, the development and application of new biochemical and mathematical approaches for constructing ordered clone maps, the automation of fingerprinting chemistries, and high-resolution imaging of DNA. Major accomplishments are highlighted below.

Human Genome Center Research Narratives: LLNL

- Considerable progress has been made toward the closure of the chromosome 19 physical map. More than 10,000 cosmids have been analyzed by an automated fluorescence-based fingerprinting approach and assembled into over 870 contigs that span about 80% of the chromosome. FISH has been used to locate over 400 cosmids and 117 contigs on the cytological map, and more than 70 known genetic markers have been located on cosmid contigs. Closure of the gaps between contigs is under way using YACs and cosmids.

- Cosmid contigs analyzed in the carcinoembryonic antigen (CEA) gene family region of chromosome 19 were found to be tightly linked over relatively short stretches of DNA. This gene family of about 22 members appears to span a contiguous region of about 1 Mb. With probes made from the ends of these contigs, hybridization techniques were applied to join contigs established by fingerprinting into larger contigs. In addition, almost 2 Mb surrounding the myotonic dystrophy locus were linked with cosmids and YACs.

- More than 20 clones containing DNA sequences corresponding to a number of important genes and regions that map to chromosome 19 were isolated from two separate YAC libraries. Among these clones were the region encoding the LDL receptor and ApoE gene, two important components of the regulation of cholesterol and triglyceride metabolism in humans. Similarly, a region was isolated that encodes a family of serine proteases called Kallikreins, whose role is the specific proteolytic activation of peptide hormones and growth factors. Clones of these regions are being used for the structural analysis and mapping of these genes.

- A structural defect found in the cloned gene linked to the autosomal dominant disease myotonic dystrophy has been identified through an international collaboration. This chromosome 19 defect, which is characterized by a tandemly repeated segment of DNA within or close to the coding region on q13.3, is similar to that seen in the fragile X syndrome. The extent of the amplified region appears to be associated with the severity of the disease.

- The gene for DNA ligase 1 was mapped to the long arm of chromosome 19. A defect in this gene may be associated with increased cancer risk. This is the fourth gene involved in DNA metabolism that has been mapped to this region of chromosome 19.

- Significant progress was accomplished in defining the organization of the cytochrome P450 genes mapping to chromosome 19. Multiple members of each of the three subfamilies were identified. The cosmids containing these genes will be useful resources for studies of the function and physiological importance of the genes.

- Three levels of resolution of FISH have been developed and applied to localize and orient cosmids. Localizing cosmids to metaphase chromosomes provides a resolution of about 1 to 3 Mb. Localization to somatic interphase cells gives a resolution of from 50 kb to 1 Mb and hybridization to sperm pronuclei a 20-kb to 1-Mb resolution. With FISH, a linear relationship was demonstrated between physical distance and genomic distance of 20 kb up to at least 800 kb in pronuclei derived from human spermatozoa. With a single probe, the presence of multiple copies of the closely related genes of the CEA family has been detected in human sperm pronuclei. Single and multicolor hybridizations are routinely performed.

- A reproducible method of mapping YACs by FISH has been developed. This procedure involves isolating YACs with pulsed-field gels, digesting with the restriction enzyme *Mbo* I, ligating to oligonucleotide linker adapters, and amplifying with PCR. The products are then mapped onto human metaphase chromosomes by standard FISH methods.

- The technique of *Alu*-PCR has been further exploited. To isolate region-specific DNA probes from human-rodent hybrid cell lines, previously developed PCR procedures were expanded. Human sequences are preferentially amplified using PCR primers specific for repeats of the human *Alu* repeat family. Several new primers have been developed that amplify human DNA sequences very efficiently, further facilitating probe isolation from human genome regions present in the available hybrids. Many different human sequences amplify from the hybrids; individual probe sequences are obtained by subsequent cloning in plasmid vectors in *Escherichia coli*. To expedite this, ligation-independent cloning has been developed to increase efficiency of cloning and eliminate the common background of clones that do not contain recombinant DNA molecules. In addition, an efficient procedure has been developed to clone the PCR products common to two cell lines. This method—coincidence cloning—permits a further enrichment for sequences derived from defined regions of the genome.

- Clone-pooling schemes have been developed to facilitate screening of both cosmid and YAC libraries. Each clone is present in a number of different pools, reducing the number of DNA samples that must be deposited on a high-density filter for hybridization-based screening and the number of tubes needed for PCR-based screening. Since each clone is defined by a unique combination of pools, the screening of pools by probe hybridization permits identification of the recombinants shared by a number of pools. This approach was used very successfully to screen a 10,000-clone cosmid library. The idea also was used to consolidate a 60,000-clone YAC library into about 1800 sample pools. Results demonstrated that hybridization-positive YAC pools can, indeed, be distinguished from hybridization-negative YAC pools, thus allowing the efficient identification of YAC clones.

- Human YACs were isolated from a library constructed using a monochromosomal 19 hybrid cell line. The YACs vary in size between 120 and 350 kb. One of the analyzed YACs carries sequences from the telomere region of chromosome 19, and another maps to the centromere region of chromosome 19 by FISH.

- A second-generation suite of robust, reliable computer programs was completed for signal preparation and analysis of chromosome 19 restriction fragment fingerprints. These programs implement methods for random noise suppression, background subtraction, and color decorrelation. A new program (TIMEWARP) was also completed to map peak locations in a gel to a common coordinate system by dynamic programming and shape-preserving spline interpolation.

- The Sybase database has been enhanced to contain all the laboratory notebook and experimental data important to physical map construction. This includes clone repository information, restriction fragment fingerprinting, and data on probe hybridization and FISH. The database is coupled to the graphical browser so the end user can retrieve many of the experimental results in graphical form.

- The graphical database browser was enhanced to run Human Genome Project data remotely over Internet. The browser's ability to link to multiple databases at external collaborator sites has been demonstrated.

- In a collaborative effort, automatic transnetwork methods for transferring physical mapping results to the central Genome Data Base (GDB) at Johns Hopkins were built, tested, and implemented by GDB and LLNL. This work was in support of DOE concerns that all laboratories should effect mechanisms to ensure that data are made available to the appropriate public databases after a suitable time period. Prototype methods were implemented, tested, and publicly demonstrated for logically linking our database with the major sequence and mapping databases (GenBank® and GDB). Direct transnetwork queries that logically integrate these data sets are now feasible.

- As part of NLGLP, high-speed flow sorting was used to purify individual human chromosomes for cloning. Large-insert phage and cosmid libraries have been made for chromosomes 9, 12, 18, 19, 21, 22, and Y. Several libraries have been distributed to users and evaluation sites. In addition, the high-speed sorter has been rebuilt with new fluidics to optimize sterility and with new electronics to increase the purity of the sorted material.

- Construction of a new high-speed chromosome sorter was completed. This instrument has new digital acquisition electronics, a new fluidic system, and a more stable sample stream. The instrument analyzes chromosomes at the rate of up to 20,000/s and can reliably produce 250 to 1000 ng of sorted chromosome DNA equivalents per day.

- Using scanning tunneling microscopy (STM), individual images of the bases adenine and thymine were obtained at atomic resolution, indicating that a scanning-probe microscopy technique can discriminate between purines and pyrimidines.

- Several technologies have been transferred to industry. They include software for analysis and graphical display of physical map data, sequence information for the commercialization of *Alu*-PCR primers, and vectors for the construction of cosmid libraries. In addition, collaborative research programs with industry have continued in the areas of fluorescence-based restriction fragment analysis, development of pulsed-field gel systems, development and testing of automated and high-throughput plasmid/cosmid DNA extraction, and development and testing of a robot for high-density colony replication on filters.

Future Plans

The LLNL genome center's first priority is to complete, to the extent possible, an ordered clone map of chromosome 19; this physical map will likely be a composite linear array of cosmid, lambda, and YAC clones. It will be correlated with the genetic map to assist the scientific community in localizing and isolating all genes from chromosome 19. State-of-the-art technology will be used to sequence selected high-interest regions of the chromosome. Once the technology has been validated for map construction of a large portion of chromosome 19, efforts will be directed to chromosome 2.

When Human Genome Project emphasis shifts from mapping to sequencing, exploration will turn to rapid automated DNA sequencing methods that can use large fragments such as cosmids or YACs as templates. STM and X-ray imaging technologies under development at LLNL are expected to contribute to advancements in sequencing.

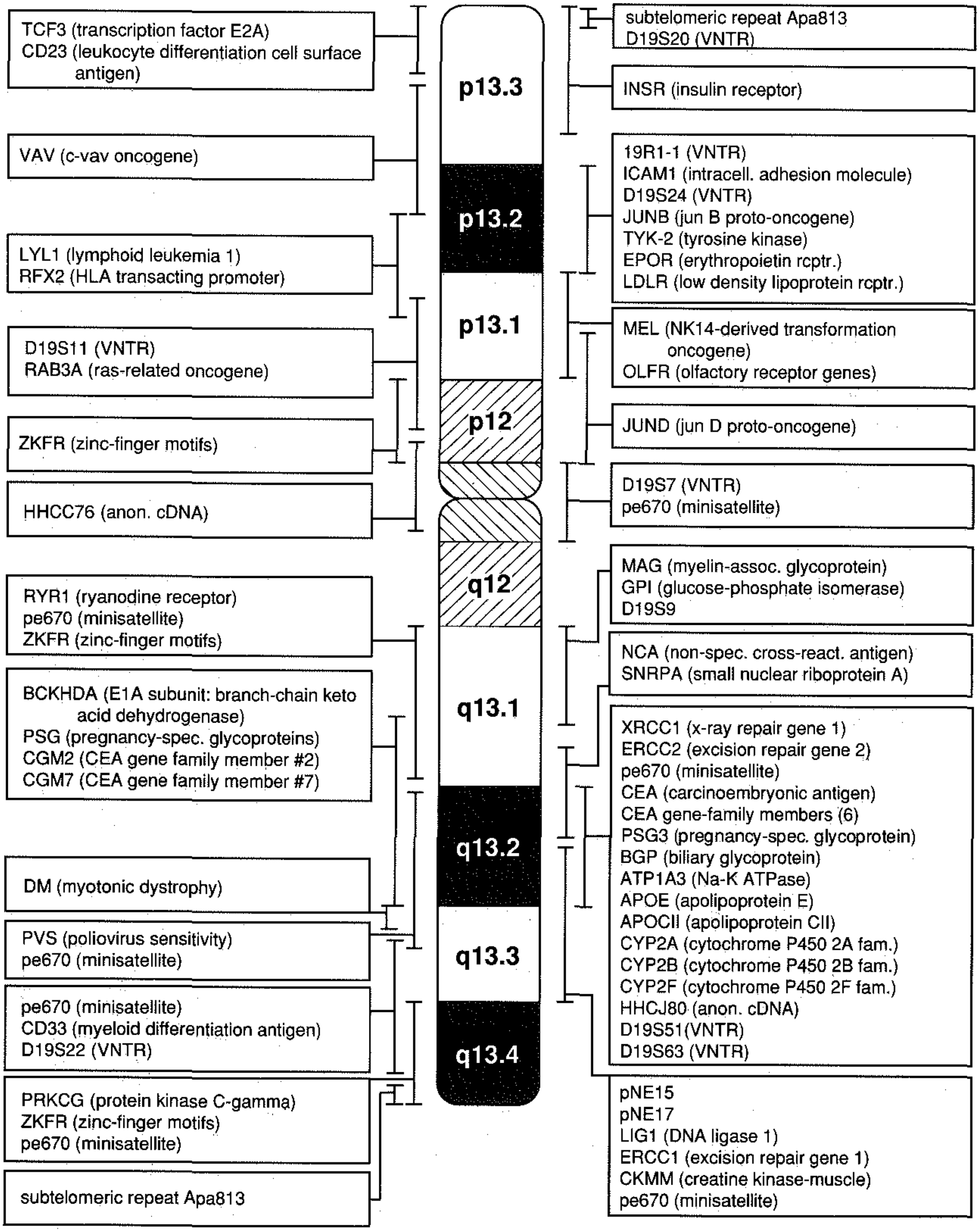
Automation is an essential element of physical mapping. New processes and instruments will be explored to reduce the need for human intervention in highly repetitive tasks. A number of instruments for clone manipulation and biochemical processes will be considered for automation.

An effort to map and sequence the cDNAs expressed in a variety of human tissues has recently been initiated. These cDNAs will be used to generate STSs and will serve as the foundation for future studies of gene organization and gene function.

Assisting the scientific community in completing ordered clone maps is critical and will remain a high priority. LLNL intends to serve as a resource laboratory for clones and for map information on chromosomes of interest. Ultimately, map and sequence information will be used to study the global architecture of the chromosome and also to evaluate human somatic and genetic variation, both spontaneous and induced.

For more information on the LLNL Human Genome Center, contact Anthony Carrano, Director, at 510/422-5698 or Leilani Corell, Administrator, at 510/423-3841.

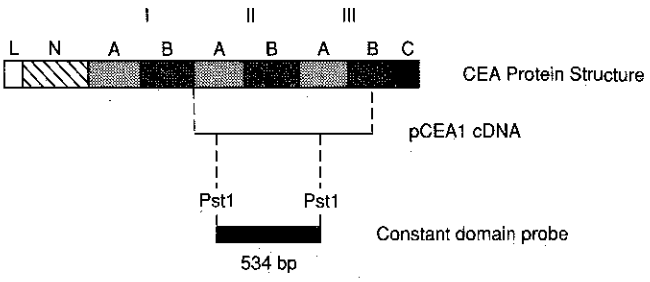
Known Genes and Genetic Markers with which Chromosome 19 Cosmids Constructed at LLNL Are Associated. *The vertical lines represent the general areas to which the genes and markers have been localized; the gene for myotonic dystrophy (DM) was recently mapped to a specific area on chromosome band q13.3. In addition to genes, several types of genetic markers have been localized to chromosome 19; these include repetitive DNA [repeated DNA segments such as minisatellite DNA and variable number of tandem repeated DNA (VNTRs) that may or may not contain genes] and DNA structural patterns (e.g., a zinc-finger motif involved in binding DNA to specific proteins needed for DNA transcription). Determination of the chromosomal positions for markers provides important reference points along the chromosome, facilitating the orientation and localization of genes and other markers. [Figure (by Barbara Trask; submitted 1992) provided by Linda Ashworth, LLNL Human Genome Center. For more information, refer to Trask and Ashworth in the Index to Principal and Coinvestigators.]*


TCF3 (transcription factor E2A)
CD23 (leukocyte differentiation cell surface antigen)
VAV (c-vav oncogene)
LYL1 (lymphoid leukemia 1)
RFX2 (HLA transacting promoter)
D19S11 (VNTR)
RAB3A (ras-related oncogene)
ZKFR (zinc-finger motifs)
HHCC76 (anon. cDNA)
RYR1 (ryanodine receptor)
pe670 (minisatellite)
ZKFR (zinc-finger motifs)
BCKHDA (E1A subunit: branch-chain keto acid dehydrogenase)
PSG (pregnancy-spec. glycoproteins)
CGM2 (CEA gene family member #2)
CGM7 (CEA gene family member #7)
DM (myotonic dystrophy)
PVS (poliovirus sensitivity)
pe670 (minisatellite)
pe670 (minisatellite)
CD33 (myeloid differentiation antigen)
D19S22 (VNTR)
PRKCG (protein kinase C-gamma)
ZKFR (zinc-finger motifs)
pe670 (minisatellite)
subtelomeric repeat Apa813
p13.3
p13.2
p13.1
p12
q12
q13.1
q13.2
q13.3
q13.4
subtelomeric repeat Apa813
D19S20 (VNTR)
INSR (insulin receptor)
19R1-1 (VNTR)
ICAM1 (intracell. adhesion molecule)
D19S24 (VNTR)
JUNB (jun B proto-oncogene)
TYK-2 (tyrosine kinase)
EPOR (erythropoietin rcptr.)
LDLR (low density lipoprotein rcptr.)
MEL (NK14-derived transformation oncogene)
OLFR (olfactory receptor genes)
JUND (jun D proto-oncogene)
D19S7 (VNTR)
pe670 (minisatellite)
MAG (myelin-assoc. glycoprotein)
GPI (glucose-phosphate isomerase)
D19S9
NCA (non-spec. cross-react. antigen)
SNRPA (small nuclear riboprotein A)
XRCC1 (x-ray repair gene 1)
ERCC2 (excision repair gene 2)
pe670 (minisatellite)
CEA (carcinoembryonic antigen)
CEA gene-family members (6)
PSG3 (pregnancy-spec. glycoprotein)
BGP (biliary glycoprotein)
ATP1A3 (Na-K ATPase)
APOE (apolipoprotein E)
APOCII (apolipoprotein CII)
CYP2A (cytochrome P450 2A fam.)
CYP2B (cytochrome P450 2B fam.)
CYP2F (cytochrome P450 2F fam.)
HHCJ80 (anon. cDNA)
D19S51(VNTR)
D19S63 (VNTR)
pNE15
pNE17
LIG1 (DNA ligase 1)
ERCC1 (excision repair gene 1)
CKMM (creatine kinase-muscle)
pe670 (minisatellite)

Assembly and Analysis of the Human Chromosome 19 Physical Map in the Carcinoembryonic Gene Family Region. *Integrating the genetic linkage and physical maps is important for maximizing the usefulness of physical maps for future applications, such as the isolation of genes associated with genetic diseases.*

This group of figures depicts the identification and verification of cosmids associated with genetically mapped markers for carcinoembryonic antigen (CEA), a human tumor marker widely used in clinical oncology that is also present during fetal development and in some normal adult tissues.

(a)



CEA is a member of a large family of closely related glycoproteins whose genes show a high degree of sequence similarity and map to the q13.1–13.2 region of chromosome 19. The CEA gene family consists of the CEA subfamily and the pregnancy-specific glycoprotein (PSG) subfamily. The CEA subfamily contains the CEA gene itself, the nonspecific cross-reacting antigen (NCA), biliary glycoprotein (BGP), and at least three additional CEA gene family members (CGM). The PSG subfamily encodes a large number of very closely related proteins found in large amounts in the placenta during pregnancy and in tumors of trophoblastic (early embryogenesis) origin.

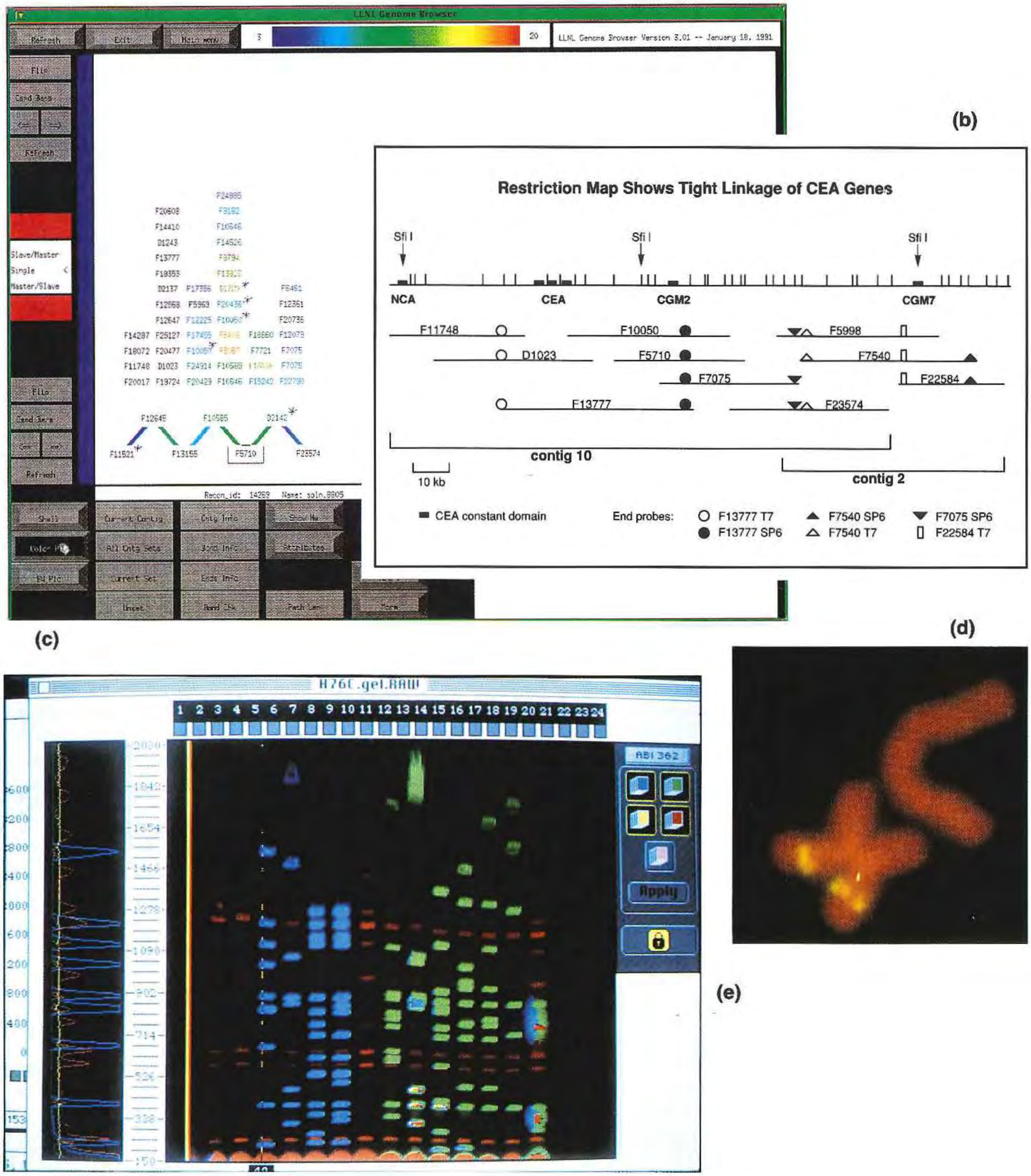
(a) Screening a chromosome 19 library for CEA family genes. *All members of the CEA family contain varying numbers of constant domains, represented by A and B. (CEA constant domains are homologous to the immunoglobulin constant domain.) A probe for the constant-domain repeat of the CEA coding sequences was used to identify CEA-positive cosmids in a human chromosome 19 library generated from flow-sorted chromosomes. Hybridization with this probe identified 238 positive cosmid clones in a six to eightfold coverage of the chromosome, suggesting the existence of 30 to 40 CEA-related genes. Cosmid DNA from the positive clones was analyzed using a high-resolution, fluorescence-based restriction fingerprinting technique, and overlapping clones were assembled into contigs. This strategy allowed assembly of 197 CEA-positive cosmids, together with 103 additional cosmids, into 9 significant contigs consisting of 18 to 55 cosmids per contig.*

(b) Restriction map verifies overlaps, shows tight linkage of CEA genes. *An* Eco*R I restriction map of cosmids in contigs 10 and 2 verified the overlaps established by fingerprinting. Probes made from cosmid clones at the ends of contigs were then used to screen the library for clones that would extend the contigs. Hybridization with end probes further validated the overlap and indicated that these two contigs should be merged, reducing the total number of contigs to eight. Screening with the CEA constant-domain probe for specific members of the CEA and PSG subgroups indicated the location of four closely spaced CEA family genes (NCA, CEA, CGM2, CGM7) in this contig of about 175 kb.*

(c) Screen image of contig 10. *This figure shows the image produced by "Browser," a database visual-access tool developed at LLNL. The likelihood of overlap between adjacent cosmids on the tiling path (i.e., overlapping members) is indicated by the colored bars, with colors ranging from blue (indicating low overlap) to red (indicating high overlap). Redundant cosmids are listed above the tiling-path member with which they most closely overlap. In this display, the clone F5710 [shown here boxed and also in* ***(b)****] has been selected, and overlapping clones are displayed. Asterisks indicate clones that were hybridized in situ.*

(d) A human metaphase spread that was hybridized with a cosmid from the CEA gene family region. *The chromosomes are stained with propidium iodide (red), and the site of cosmid hybridization is marked by yellow-green fluorescence as the result of a fluorescein tag on the hybridization probe. Two copies of chromosome 19 overlap each other in the metaphase spread, and the sites of hybridization are adjacent. The hybridization band is broad since many of the members of the CEA family share a common sequence domain.*

(e) Automated fluorescence-based restriction fragment analysis. *Using the same technology established for fluorescence-based cosmid fingerprinting, fluorochrome-labeled linkers are attached to the ends of restriction fragments and analyzed on a horizontal agarose-based fluorescence scanning system (GENESCANNER™, from Applied Biosystems, Inc.). Internal size standards (red dye) are present in each lane, and the cosmid DNAs can be labeled with one of three additional fluorochromes. This method allows rapid accumulation of restriction digest information and has been applied to the analysis and verification of the CEA and other contigs. (Photographs and figures on pp. 44–45 provided by Leilani Corell, LLNL Human Genome Center. For more information, refer to E. Branscomb, A. Carrano, P. de Jong, G. Lennon, H. Mohrenweiser, A. Olsen, and B. Trask in the Index to Principal and Coinvestigators.)*


LLNL Genome Browser
Refresh
Exit
Main menu
3
20
LLNL Genome Browser Version 3.01 -- January 18, 1991
File
Card Base
Refresh
Slave/Master
Single
Master/Slave
F24885
F20608 F3182
F14410 F10646
D1243 F14526
F13777 F8794
F18353 F13813
D2137 F17356 D1725 F6461
F12568 F5953 F20436 F12351
F12647 F12225 F10060 F20736
F14287 F25127 F17455 F18560 F12073
F18072 F20477 F10057 F7721 F2075
F11748 D1023 F24914 F10585 F7075
F20017 F19724 F20429 F10646 F15242 F22790
F12645
F10585
D2142
F11521
F13155
F5710
F23574
Recon_id: 14269 Name: soln.8805
Shell
Color Pic
BW Pic
Current Contig
All Cntg Sets
Current Set
Unset
Cntg Info
Bond Info
Ends Info
Bond Chk
Show Me
Attributes
Path Len
(b)
Restriction Map Shows Tight Linkage of CEA Genes
Sfi I
Sfi I
Sfi I
NCA
CEA
CGM2
CGM7
F11748
F10050
F5998
D1023
F5710
F7540
F7075
F22584
F13777
F23574
contig 10
10 kb
contig 2
▬ CEA constant domain
End probes:
○ F13777 T7
● F13777 SP6
▲ F7540 SP6
△ F7540 T7
▼ F7075 SP6
▯ F22584 T7
(c)
H76C.gel.RAW
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
2030
1842
1654
1466
1278
1090
902
714
526
338
150
ABI 362
Apply
(d)
(e)

Los Alamos National Laboratory

Human Genome Center Research Narratives

The Center for Human Genome Studies at Los Alamos National Laboratory (LANL) provides direction, coordination, and technical oversight for the LANL portion of the DOE Human Genome Program. The center draws scientific talent from six technical divisions at LANL. Molecular biologists, chemists, physicists, mathematicians, computer scientists, and engineers are contributing to progress in physical mapping, technology development, and informatics. Although a specific goal is the assembly of a complete physical map for human chromosome 16, much of the work is broadly supportive of the worldwide Human Genome Project. Collaborative research and development programs have also been initiated with private-sector and other institutions involved in human genome research. The major technical subdivisions of the center are physical mapping, technology development, and informatics. Activities are also under way at the center to explore ethical, legal, and social issues arising from genome research data and to transfer technology developed within the center's projects.

Physical Mapping

Physical mapping includes the development of conceptual advances in mapping strategy and the construction of a physical map of chromosome 16. The physical map will be composed of phage, cosmid, and YAC contigs ordered by repetitive sequence fingerprinting. These ordered contigs will be integrated with the genetic linkage map, the cytogenetic map, and known gene sequences on chromosome 16. The final map, along with its eventual translation into a sequence tagged site (STS) map, will provide the means for rapid access to any region of the chromosome for further analysis. In addition, the ordered clone sets will be available for eventual sequencing.

Technology Development

Technology development efforts include the application of robotics to the handling and storage of DNA fragments, the development and application of methods for the construction of DNA libraries from flow-sorted chromosomes, and the development of new methods for rapid, inexpensive, large-scale sequencing. All these projects are or will be supportive of the physical mapping of chromosome 16, and they also contribute to the larger genome program. For example, the construction and distribution of various kinds of libraries from sorted chromosomes is playing a significant role at many of the genome research centers.

Informatics

Informatics efforts involving the collection and analysis of genome-related data will play an increasingly important role in the genome project. LANL has a long history of expertise in this research area and will continue to lead in providing these essential resources.

Ethical, Legal, and Social Issues (ELSI) Activities

The center also sponsors active participation in ELSI studies related to data produced by human genome research and is compiling a comprehensive literature bibliography in collaboration with Georgetown University. LANL scientists participated in a series of discussions on ELSI issues sponsored by the University of California Humanities Research Institute.

Technology Transfer

LANL will continue to put a high priority on collaborations with private industry to use the skills and resources of the private sector and to ensure effective technology transfer to the U.S. commercial sector. The first Cooperative Research and Development Agreement (CRADA) involving human genome research activity was signed in 1991 by LANL and Life Technologies, Inc. (LTI).

Recent Progress and Future Directions

Construction of a physical map of chromosome 16. The chromosome-mapping strategy at LANL involves the rapid generation of cosmid contigs representing around 60% of the target chromosome, followed by directed gap closure with yeast artificial chromosomes (YACs). The first phase of this goal, the rapid generation of nucleation contigs on chromosome 16, has been completed [Stallings et al., *Proc. Natl. Acad. Sci. USA* **87**: 6218–22 (1990)]. An approach for identifying overlapping cosmid clones by exploiting the high density of repetitive sequences in human DNA was used to generate 553 contigs following the fingerprinting of over 4500 individual cosmid clones. These contigs represent more than 80% of the euchromatic arms of chromosome 16 and were constructed with about one-fourth as many cosmid fingerprints as random strategies requiring 50% minimum overlap detection.

Nucleating at specific regions allows (a) the rapid generation of large (>100 kb) contigs in the early stages of contig mapping and (b) the production of a contig map with useful landmarks [i.e., $(GT)_n$ repeats] for rapid integration of the genetic and physical maps. All 4500 fingerprinted cosmids in contigs and singlets have been rearrayed on high-density filters. Such filters already provide investigators with access to more than 90% of chromosome 16, with a 60% probability that any region is already present in a contig. These high-density chromosome-specific cosmid filter arrays have also proved useful for YAC fingerprinting with repetitive sequence polymerase chain reaction (PCR) techniques. In collaboration with the laboratories of David Ward (Yale University) and David Callen (Adelaide Children's Hospital, Australia), 130 of these arrayed cosmids have been regionally localized via in situ hybridization or somatic cell hybrid panels. The average gap (containing only singlets), approximately 65 kb in length, can be easily closed with YACs. A single walk from each end of current contigs should, statistically, reduce the number of contigs to approximately 50, one of the 5-year goals of the Human Genome Project (i.e., 1- to 2-Mb contigs; >95% coverage). To facilitate closure, LANL investigators are constructing from monochromosomal hybrids and flow-sorted material both a total genomic YAC library (from cell line GM130, using the vectors pJS97 and pJS98; currently onefold representation) and chromosome 16 YAC clones. One hundred STS markers are being generated to key contigs. Extensive analyses of the DNA sequences obtained from contig ends are in progress using multiple approaches to identify potential coding regions. These approaches include nucleotide and translated amino acid sequence homology searches against GenBank, using BLAST and FASTA, and the new adaptive network program, GRAIL, developed and made available by the Oak Ridge National Laboratory. Current progress with YAC closure indicates that the complete physical map of chromosome 16 will be achieved in the next few years.

Low-abundance repetitive DNA sequences identified on chromosome 16. Chromosome 16–specific, low-abundance repetitive DNA sequences (designated CH16LARs) have been identified during construction of the cosmid contig map of this chromosome. CH16LARs were initially identified by in situ hybridization of cosmid and YAC clones to normal human chromosomes (in collaboration with David Ward). The cosmid clones all came from contig 55. The hybridization signals were unusually intense and occurred on four regions of human chromosome 16: bands p13, p12, p11, and q22. Contig 55 contains more clones than any other contig (78 clones or 2% of all clones fingerprinted thus far). Ordering clones within contig 55 is not possible because the presence of these low-abundance repetitive DNA sequences has generated false overlaps. The regions containing CH16LARs may cover as much as 5% of the euchromatic arms of chromosome 16 (~5 Mb of DNA). One CH16LAR sequence (CH16LAR1) was cloned and sequenced, and a minisatellite type of repetitive sequence was identified. The region containing CH16LARs is of biological interest since the pericentric inversion breakpoints commonly found in myelomonocytic leukemia fall within these regions [Mitelman, *Hereditas* **104**:113 (1986)]. Alternative strategies for mapping and ordering clones from this region are being implemented.

Construction and distribution of DNA libraries from flow-sorted chromosomes: National Laboratory Gene Library Project (NLGLP). NLGLP is a cooperative project between LANL and Lawrence Livermore National Laboratory. Investigators at LANL have cloned a set of complete digest libraries into the *Eco*R I insertion site of Charon 21A; they are available from the American *Type Culture* Collection, Rockville, Maryland. Sets of partial digest libraries in the cosmid vector s*Cos*1 and in the phage vector Charon 40 are being constructed for human chromosomes 4, 5, 6, 8, 10, 11, 13, 14, 15, 16, 17, 20, and X. Individual human chromosomes are first sorted from rodent-human hybrid cell lines until about 1 µg of DNA has been accumulated. The sorted chromosomes are then examined for purity by in situ hybridization, and the DNA is extracted and partially digested with the restriction enzyme *Sau* 3AI, dephosphorylated, and cloned into vectors. Partial digest libraries have been constructed for chromosomes 4, 5, 6, 8, 11, 13, 16, 17, and X. Purity estimates from sorted chromosomes, flow-karyotype analysis, and plaque or colony hybridization indicate that most of these libraries are 90 to 95% pure. Additional cosmid library constructions and arrays of libraries having five- to tenfold genomic coverage into microtiter plates are in progress. Libraries have been constructed in M13 or bluescript vectors to generate STS markers for selecting chromosome-specific inserts from a genomic YAC library. LANL has also cloned sorted DNA into YAC vectors and expects to construct a series of YAC libraries representing individual chromosomes (see below).

A YAC library for human chromosome 21. YACs have been constructed using DNA isolated from aliquots of flow-sorted human chromosome 21. Chromosomes were prepared from the somatic cell hybrid WAV-17, which contains chromosome 21 as the only human chromosome. DNA isolated from sorted chromosomes was restricted with either *Cla* I or *Eag* I or both *Not* I and *Nhe* I, ligated to YAC vectors pJS97 and pHS98, and transformed into *Saccharomyces cerevisiae* strain YPH 250. The transformation efficiency of YACs ranged from 600 to 2500 cfu/µg of sorted DNA. About 1200 human YACs with an average size of 200 kb have been identified. The locations of 20 random YACs on chromosome 21 were confirmed by hybridization to somatic cell hybrid mapping panels. Three YACs that hybridize to D21S55 have been identified and are being used to initiate construction of a physical map of the Down's syndrome region of chromosome 21. Sixty YAC clones from the chromosome 21 library were localized on chromosome 21 by in situ hybridization. The results indicate that the library contains inserts that are well distributed along the length of the chromosome and that the frequency of chimeric inserts is low (below 3%). A collaboration between the genome centers at LANL and Lawrence Berkeley Laboratory (LBL) will use the library for comprehensive physical mapping of chromosome 21. The ability to construct chromosome-specific YAC libraries from sorted chromosomes will facilitate isolation of disease genes and construction of long-range physical maps of complex genomes. LBL is working on chromosome 21 in cooperation with LANL.

Chromosome-specific STS libraries. Specific STSs have been systematically generated using flow-sorted chromosomes. DNA from about 200,000 chromosomes was digested with either one or two restriction enzymes (usually *Bam*H I and *Hin*d III) and cloned directly into bacteriophage M13mp18. One-pass sequencing was conducted, either manually or with a Dupont Genesis 2000 automated sequencer. DNA sequences were analyzed for the presence of sequence similarity to common human repetitive sequences, and appropriate PCR oligomers were synthesized. An acceptable STS-PCR assay yielded the appropriately sized product from both the hybrid cell line DNA containing only the human chromosome of interest and the pools of 384 anonymous YAC clones, spiked with 5 ng/µl total human DNA. To date, over 340 kb of anonymous DNA sequence from human chromosomes 5 and 7 have been analyzed. Two hundred STS markers for chromosome 7 have been generated in collaboration with Maynard Olson's laboratory at Washington University [Green et al., *Genomics* (in press)], and the first 100 STS markers for chromosome 5 are currently being generated in collaboration with John Wasmuth's laboratory at the University of California, Irvine; 50 STSs for chromosome 5 have been regionally localized. The overall efficiency of PCR reactions yielding appropriate products, with the anonymous genomic sequences from flow-sorted chromosomes, has been approximately 75%. GRAIL analyses indicate that approximately 15% of both the chromosome 16 STSs and the randomly selected STSs for chromosomes 5 and 7 contain putative coding regions.

Informatics. The Laboratory Notebook database, designed to manage all information necessary for map assembly, has been expanded to include sequences, STS mapping information, and grid hybridization data, as well as clone fingerprints and completed maps. The forms-based interface is being expanded to provide easy access to the new tables. Graphical interfaces and innovative algorithms to aid map assembly have been prototyped and are being refined. Integrated, multilevel maps are increasing in importance. A strong emphasis for the coming year will be to implement the Software for Integrated Genome Map Assembly (SIGMA) system, which was designed to aid in display, assembly, evaluation, and editing of integrated maps.

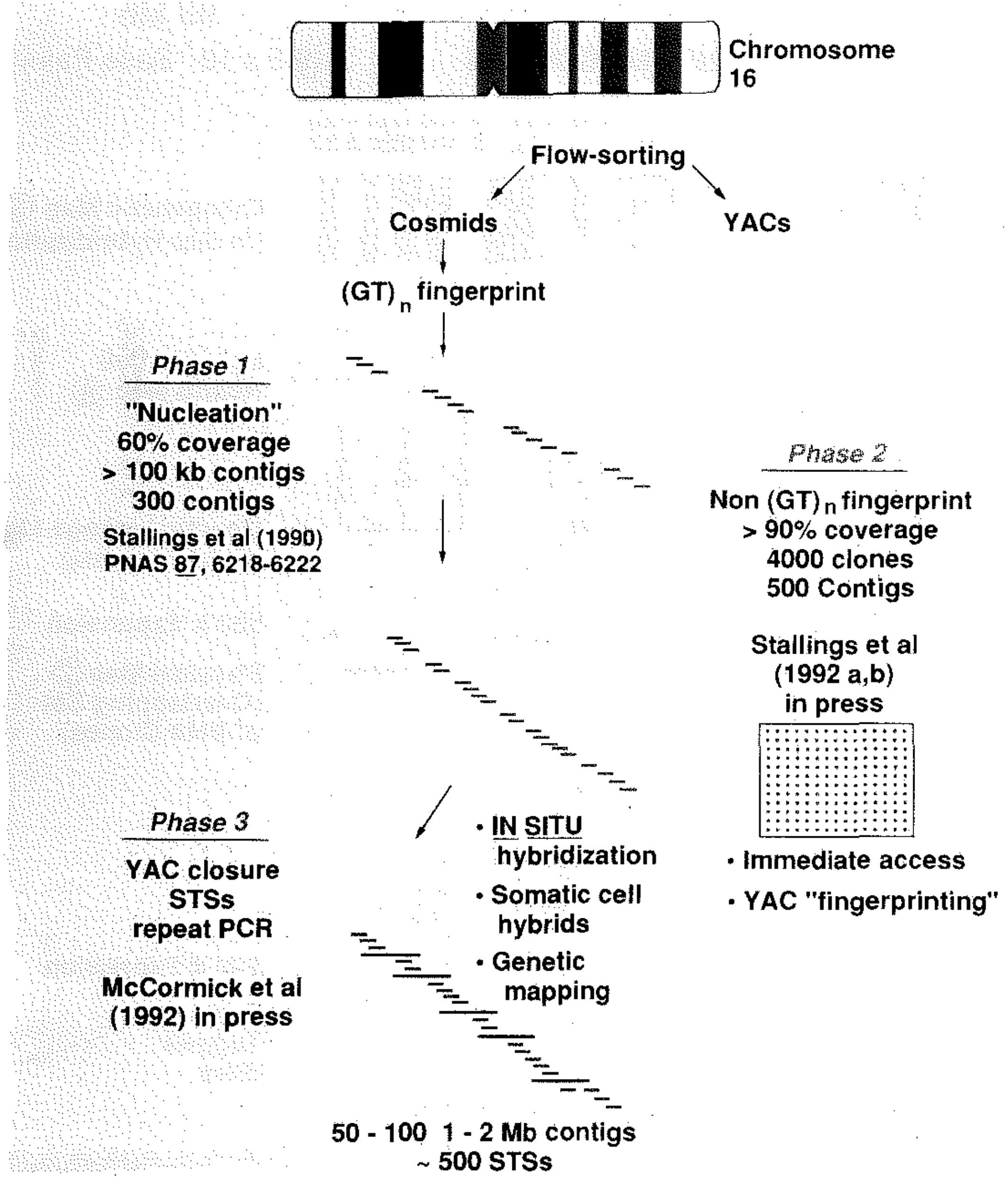
DNA sequencing based upon single-molecule detection in flow cytometry. This project addresses the problem of rapidly sequencing bases in large fragments of DNA. A DNA fragment of about 40 kb will be labeled with base-identifying tags and suspended in the flow stream of a flow cytometer capable of single-molecule detection. The tagged bases will be sequentially cleaved from the single fragment and identified as the liberated tag passes through the laser beam. A sequencing rate of 100 to 1000 bases/s on DNA strands of around 40 kb is projected [*Genet. Anal.* **8:** 1 (1991)]. Accomplishments of this project are as follows:

Human Genome Center Research Narratives: LANL

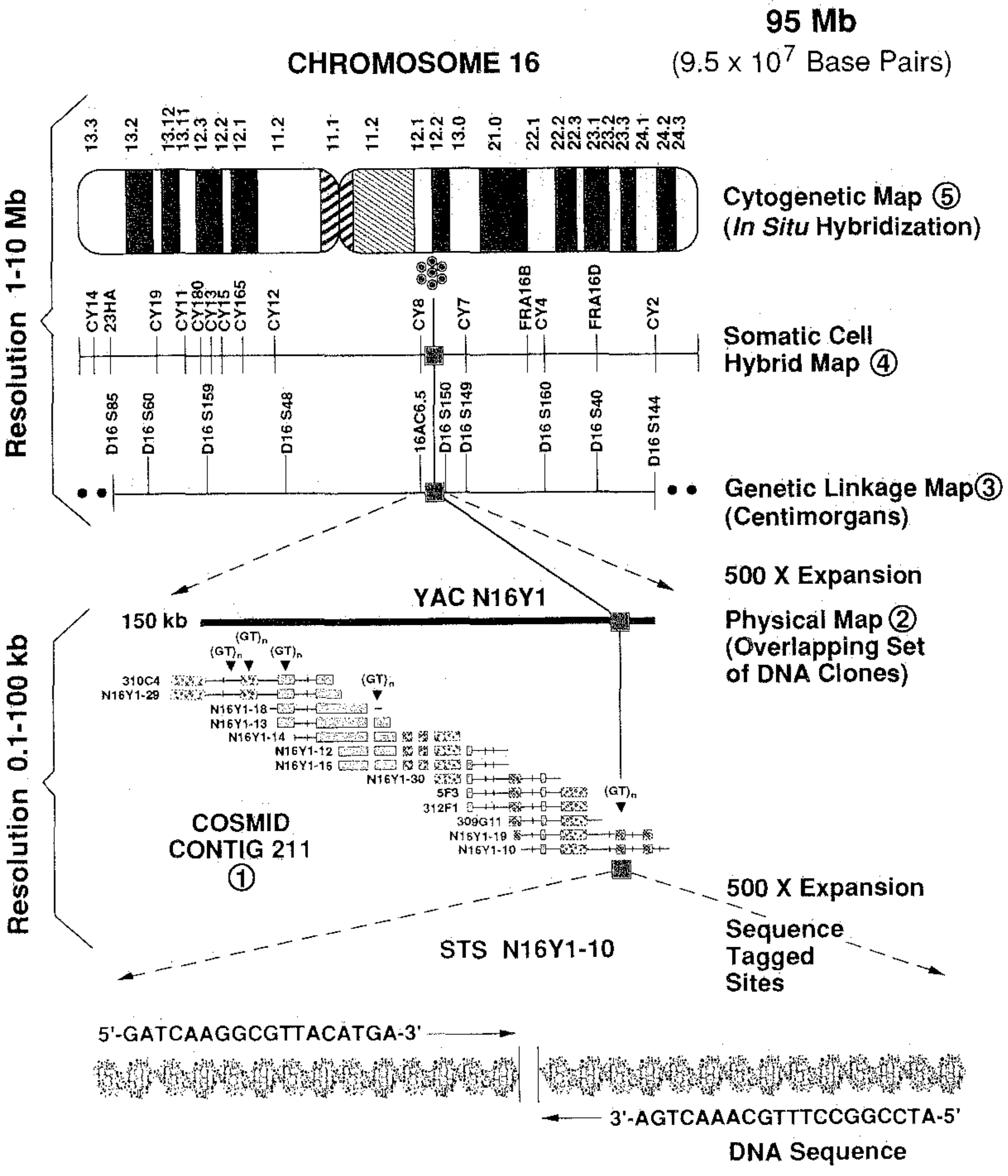
- Signed CRADA with LTI for joint research on DNA sequencing. LTI will offer expertise in nucleic acid chemistry and enzymology, and LANL will specialize in detection technology and DNA handling. LTI will commercialize the technique [for more information, refer to the figure on p. 25 and to *Human Genome News*, **3**(1): 5 (May 1991)].
- Detected several different kinds of single fluorescing molecules with ~85% efficiency and low error rates [*Chem. Phys. Lett.* **174:** 553 (1990)].
- Observed photon bursts simultaneously from rhodamine-6G and Texas Red, using both a doubled Nd/YAG and a synchronously pumped dye laser for excitation and dual-wavelength detection.
- Synthesized DNA fragments up to 500 nucleotides long that contain one fluorescent nucleotide and three normal nucleotides. DNA synthesis was observed with rhodamine-dCTP, rhodamine-dATP, rhodamine-dUTP, fluorescein-dATP, and fluorescein-dUTP. This work was a collaboration with LTI.
- Digested the fluoresceinated DNAs described above by six different exonucleases: native T4 polymerase, native T7 polymerase, Klenow fragment of *Escherichia coli* pol I, exo III, *E. coli* pol III holoenzyme, and snake venom phosphodiesterase. LTI also participated in these investigations.

Robotic workcell for DNA filter array construction. A gantry robot–based workcell has been assembled to array small spots of DNA in an interleaved format. Grid densities on these membrane filters can be varied from 576 to 9216 spots per 22 cm^2. The robot picks a microtiter plate from a dispenser, scans a barcode label, removes the plate cover, and inserts a 96-pin gridding tool into the plate wells. The tool is then positioned at the appropriate place on the membrane, and the solutions on the pins are transferred as spots. The gridding tool is washed and sterilized, the lid replaced on the microtiter plate, and the plate placed into a receiving stacker. The entire sequence is repeated with new plates until the desired array has been constructed.

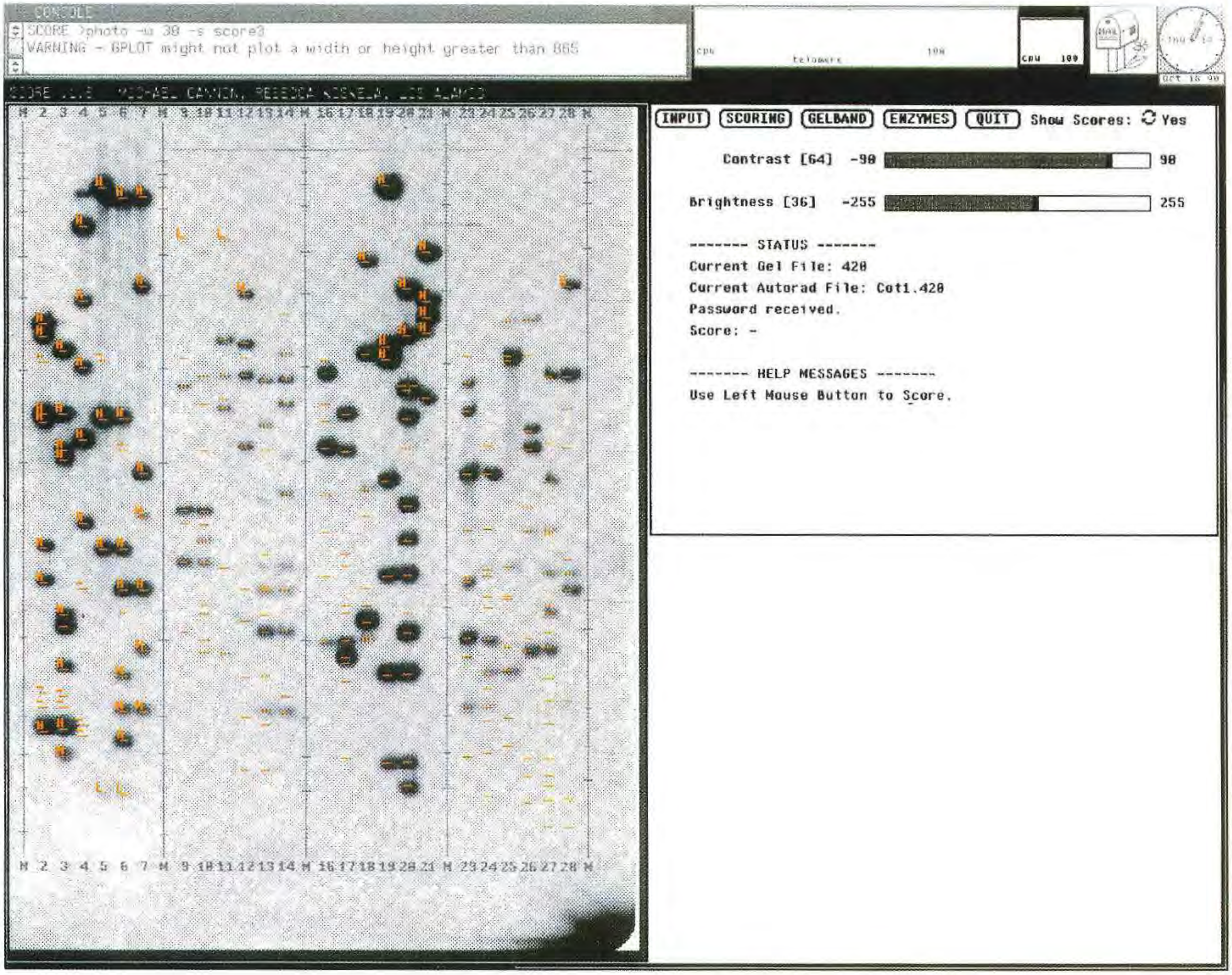
For more information on the LANL Center for Human Genome Studies, contact Robert K. Moyzis, Director, or Larry Deaven, Deputy Director, at 505/667-3912.



Strategy for Mapping Chromosome 16 at LANL. (a) *In phase 1 of the mapping strategy, a cosmid library was constructed from flow-sorted chromosome 16 DNA. Individual clones from this library were fingerprinted using restriction fragment digests (*Eco*R I,* Hind *III, and* Eco*R I/*Hind *III double digest) followed by electrophoretic separation of fragments, capillary transfer to membrane filters, and hybridization with selected classes of ^{32}P-labeled repetitive sequences (see Stallings et al., "Physical Mapping of Human Chromosomes by Repetitive Sequence Fingerprinting,"* Proc. Natl. Acad. Sci. USA ***87****, 6218–22, 1990). In phase 2 more than 4000 clones were fingerprinted, resulting in over 500 contigs representing about 80% of the euchromatic arms of chromosome 16. These contigs and the remaining singlet clones have been arrayed in microtiter plates, providing access to more than 90% of the chromosome. Through the use of in situ hybridization and somatic cell hybrid panels, 105 contigs have been regionally localized with an average gap of about 65 kb between contigs. Phase 3 focuses on filling in the gaps by generating sequence tagged site markers from contig ends and selecting appropriate yeast artificial chromosome (YAC) clones from genomic and chromosome-specific YAC libraries.*



(b)*This diagram illustrates how a cloned DNA fragment can be related to the cytogenetic map and how sequence tagged site (STS) markers (a short segment of sequenced DNA) can aid in integrating the genetic and physical maps. The location of an STS (STS N16Y1-10) is shown from the bottom up in (1) an ordered set of cosmid clones (cosmid contig 211); (2) a 150-kb YAC insert (YAC N16Y1); (3) a genetic linkage map of chromosome 16 between two genetic markers (16AC6.5 and D16S150); (4) a somatic cell hybrid map between two markers (CY8 and CY7); and (5) a cytogenetic map between 16q12.1 and 16q12.2, as determined by in situ hybridization. (Figures on pp. 51 and 52 provided by Monica Fink, LANL. For more information, refer to R. Moyzis, L. Deaven, M. McCormick, and R. Stallings in the Index to Principal and Coinvestigators.)*


CONSOLE
cpu
telomere
100
cpu 100
Oct 16 90
SCORE V1.0 MICHAEL CANNON, REBECCA KOSKELA, LOS ALAMOS
M 2 3 4 5 6 7 M 9 10 11 12 13 14 M 16 17 18 19 20 21 M 23 24 25 26 27 28 M
INPUT
SCORING
GELBAND
ENZYMES
QUIT

```
SCORE >photo -w 38 -s score3
WARNING - GPLOT might not plot a width or height greater than 865
```

```
Show Scores: Yes

     Contrast [64]  -98                    98

Brightness [36]   -255                     255

------- STATUS -------
Current Gel File: 420
Current Autorad File: Cot1.420
Password received.
Score: -

------- HELP MESSAGES -------
Use Left Mouse Button to Score.
```

SCORE, a LANL Program for Computer-Assisted Scoring of Southern Blots. *The physical mapping technique used by LANL requires the scoring of a large number of Southern blots. In the Southern blotting procedure, size-fractionated DNA is transferred by capillary action from an electrophoresis gel to a membrane. Following transfer, the membrane-bound DNA can be hybridized to radioactively labeled probe DNA. Membrane-bound sequences that anneal to the probe DNA, and thus contain similar sequences, can then be visualized by autoradiography.*

SCORE retains the use of expert human judgment while taking over much of the scoring task drudgery. The primary functions of the program are to keep track of band and lane locations for the user and to store the resulting data directly into a database. This is done by making an aligned overlay of the fluorescent gel image (orange dashes on the screen) and the autoradiogram blot image (gray spots). Use of SCORE has resulted in greatly increased efficiency and accuracy in gathering data for physical mapping. [Photograph first published in T. M. Cannon et al., "A Program for Computer-Assisted Scoring of Southern Blots," BioTechniques ***10****, 764–67 (1991). Photograph provided by James Fickett, LANL. For more information, refer to Fickett and D. Torney in the Index to Principal and Coinvestigators.]*

Program Management Infrastructure

DOE OHER Mission

Genetics and radiation biology have been a long-term concern of the DOE Office of Health and Environmental Research (OHER) and DOE forerunners—the Atomic Energy Commission (AEC) and the Energy Research and Development Administration (ERDA). In the United States, the first federal support for genetics research was through AEC. In the early days of nuclear energy development, the focus was on radiation effects and later broadened under ERDA and DOE to include the health implications of all energy technologies and their by-products (see "Enabling Legislation" in box below). Today, an extensive program of OHER-sponsored research on genomic structure, maintenance, damage, and repair continues at the national laboratories and universities. Some major components of OHER genetics research are (1) molecular cloning and characterization of DNA repair genes, (2) improvement of methodologies and resources for quantitating and characterizing mutations, and (3) the focused resource and technology development needed to map and sequence the human genome—the Human Genome Program.

Human exposure to environmental factors and the body's response to such factors are a major concern. Unavoidable genome-damaging agents in the environment include natural radiation sources, such as the components of sunlight, cosmic rays from space, and radon from the earth. Both inorganic and organic chemicals, some natural to the environment

Enabling Legislation

The Atomic Energy Act of 1946 (P.L. 79-585) provided the initial charter for a comprehensive program of research and development related to the utilization of fissionable and radioactive materials for medical, biological, and health purposes.

The Atomic Energy Act of 1954 (P.L. 83-703) further authorized AEC "to conduct research on the biologic effects of ionizing radiation."

The Energy Reorganization Act of 1974 (P.L. 93-438) provided that responsibilities of ERDA shall include "engaging in and supporting environmental, biomedical, physical and safety research related to the development of energy resources and utilization technologies."

The Federal Nonnuclear Energy Research and Development Act of 1974 (P.L. 93-577) authorized ERDA to conduct a comprehensive nonnuclear energy research, development, and demonstration program to include the environmental and social consequences of the various technologies.

The DOE Organization Act of 1977 (P.L. 95-91) instructed the department "to assure incorporation of national environmental protection goals in the formulation and implementation of energy programs; and to advance the goal of restoring, protecting, and enhancing environmental quality, and assuring public health and safety," and to conduct "a comprehensive program of research and development on the environmental effects of energy technology and programs."

and others generated by human commerce and energy-related processes, put people at risk. Normal biological functions also contribute to the risk of genetic damage when the body's own cells produce potentially damaging molecules in the course of metabolic processes such as defensive actions against microbes, detoxification of harmful environmental substances, and cell proliferation. Even DNA is not completely stable chemically; its normal methyl-cytosine constituent has a low but measurable rate of spontaneous mutagenic change.

Systems that reverse many types of DNA damage have evolved to include a wide range of repair mechanisms within cells of all species. Humans show great diversity in this capacity, with repair-gene deficiencies showing up as sensitivity to DNA damage from low-level radiation and in diseases such as cancer. Some human genes that contribute to DNA repair processes have been characterized, and others await detection and molecular cloning. A goal of the OHER program is to improve the capabilities for diagnosing individual susceptibility to genome damage.

The genome program is providing fundamental information about the linear structure of chromosomes and genes, but understanding gene function requires other types of knowledge. Elucidating the three-dimensional (3-D) structure of proteins is crucial in explicating their functions. To advance these studies, several unique facilities for 3-D microstructure research, developed and maintained at DOE laboratories (see box on DOE facilities), are increasingly in demand by molecular biologists.

To carry out its national research and development obligations, OHER conducts the following activities:

- Sponsors research and development projects at universities, in the private sector, and at DOE national laboratories;

Major DOE Facilities and Resources Relevant to Molecular Biology Research

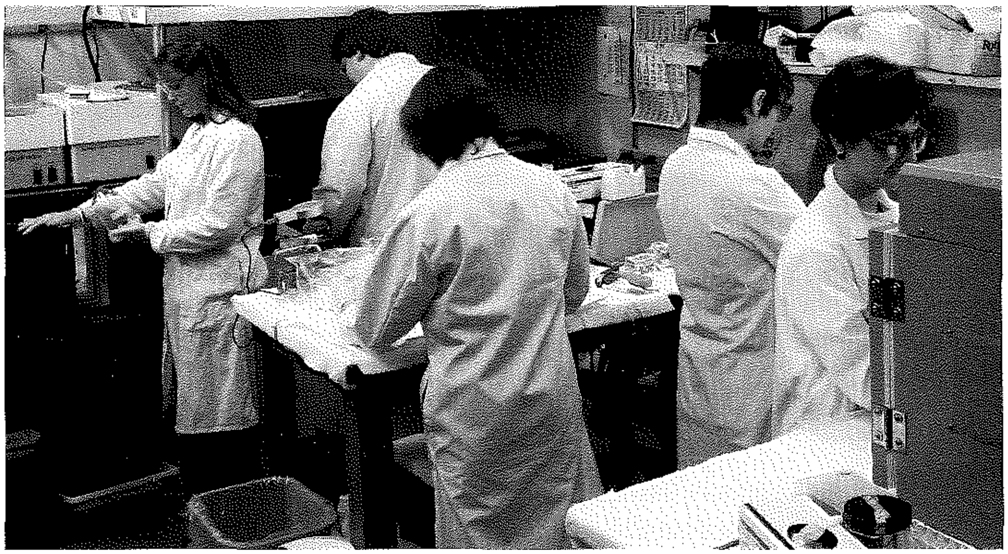
Center for X-Ray Optics	**LBL**
GenBank® Data Sequence Repository	**LANL**
High Flux Beam Reactor	**BNL**
Los Alamos Neutron Scattering Center	**LANL**
National Flow Cytometry Resource	**LANL**
National Laboratory Gene Library Project	**LANL, LLNL**
Protein Structure Data Bank	**BNL**
National Synchrotron Light Source	**BNL**
Scanning Transmission Electron Microscope Resource	**BNL**
Stanford Synchrotron Radiation Laboratory	**Stanford**
GRAIL, Online Sequence Interpretation Service	**ORNL**

Program Management Infrastructure: DOE OHER Mission

- Uses the unique capabilities of multidisciplinary DOE national laboratories for the nation's benefit;
- With advice from the scientific community and other sectors of government, considers novel, beneficial initiatives; and
- Provides expertise on various governmental working groups.

David J. Galas has directed OHER, an office of the DOE Office of Energy Research, since April 1990. He also serves under the White House Office of Science and Technology Policy as Cochair of the Committee on Life Sciences and Health and as Chairman of its Subcommittee on Biotechnology Research. John C. Wooley became OHER Deputy Associate Director in June 1992.

The Human Genome Program, conceived as an Initiative within OHER, is administered primarily through the Health Effects and Life Science Research Division, directed by David A. Smith. The Medical Applications and Biophysical Research Division, directed by Robert W. Wood, monitors the instrumentation sector of the Human Genome Program and, more broadly, sponsors research and development of resources and instrumentation having biomedical and biotechnological applications.



Researchers preparing samples for chromosome 19 mapping studies at the LLNL Human Genome Center.

Program Management Task Group

The Human Genome Program Management Task Group (see box for list of members) reports to the OHER Director and works to coordinate the following within OHER:

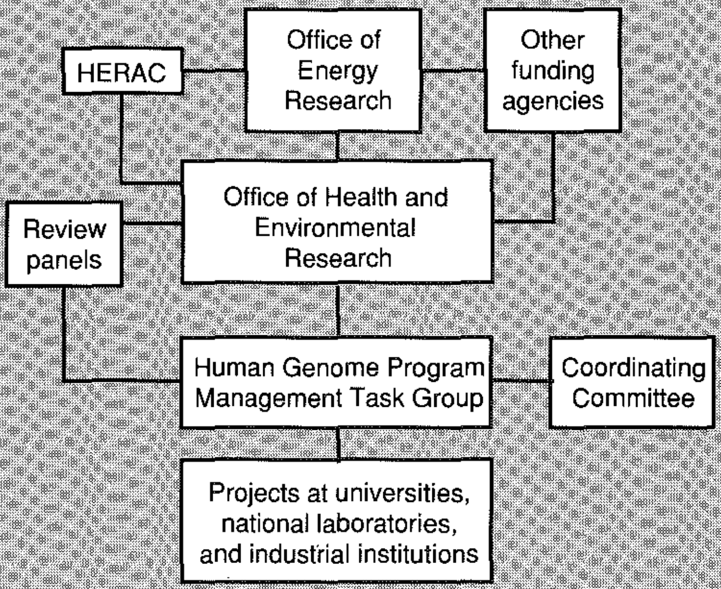
- peer review of research proposals, using both prospective and retrospective evaluations and
- administration of awards, collaboration with all concerned agencies and organizations, organization of periodic workshops, and responses to the needs of the developing program.

DOE Human Genome Program Management Task Group in 1992

David A. Smith, Chair	Molecular biologist
Ann M. Barber	Computational biologist
Benjamin J. Barnhart	Geneticist
Daniel W. Drell	Biologist
Gerald Goldstein	Physical scientist
Murray Schulman	Radiation biologist
Jay Snoddy*	Molecular biologist
Marvin Stodolsky	Molecular biologist
John C. Wooley	Biophysicist

*On detail from Argonne National Laboratory.

DOE Human Genome Program Management and Coordination


HERAC
Office of Energy Research
Other funding agencies
Review panels
Office of Health and Environmental Research
Human Genome Program Management Task Group
Coordinating Committee
Projects at universities, national laboratories, and industrial institutions

Field Coordination

Program Management Infrastructure

Human Genome Coordinating Committee (HGCC)

Another component of the OHER management structure, HGCC was formed in October 1988 to represent DOE genome program researchers along with observers from other government and private agencies (see box for list of HGCC members). Members of the Human Genome Program Management Task Group are ex-officio members of HGCC, and they participate in the regularly scheduled HGCC meetings. HGCC responsibilities include the following:

- assisting OHER with overall coordination of DOE-funded genome research;
- facilitating the development and dissemination of novel genome technologies;

Human Genome Coordinating Committee Members in 1992

Elbert W. Branscomb, Computational Biologist, Human Genome Center, Lawrence Livermore National Laboratory

Charles R. Cantor, Principal Scientist, DOE Human Genome Program, Lawrence Berkeley Laboratory

Anthony V. Carrano, Director, Human Genome Center and Leader, Biomedical Sciences Division, Lawrence Livermore National Laboratory

C. Thomas Caskey, Director, Institute for Molecular Genetics, Baylor College of Medicine

David J. Galas, Office of Health and Environmental Research, DOE

Raymond F. Gesteland, Professor and Cochair, Department of Human Genetics, University of Utah; Investigator, Howard Hughes Medical Institute Laboratory for Genetic Studies at the Eccles Institute, University of Utah

Leroy E. Hood, Director, Center for Integrated Protein and Nucleic Acid Chemistry and Biological Computation; Director, Cancer Center, California Institute of Technology

Robert K. Moyzis, Director, Center for Human Genome Studies, Los Alamos National Laboratory

Jasper Rine, Director, Human Genome Center, Lawrence Berkeley Laboratory

Robert J. Robbins, Director, Welch Medical Library for Applied Research in Academic Information, Johns Hopkins University

David A. Smith, Office of Health and Environmental Research, DOE

Lloyd M. Smith, Assistant Professor, Analytical Division, Department of Chemistry, University of Wisconsin, Madison

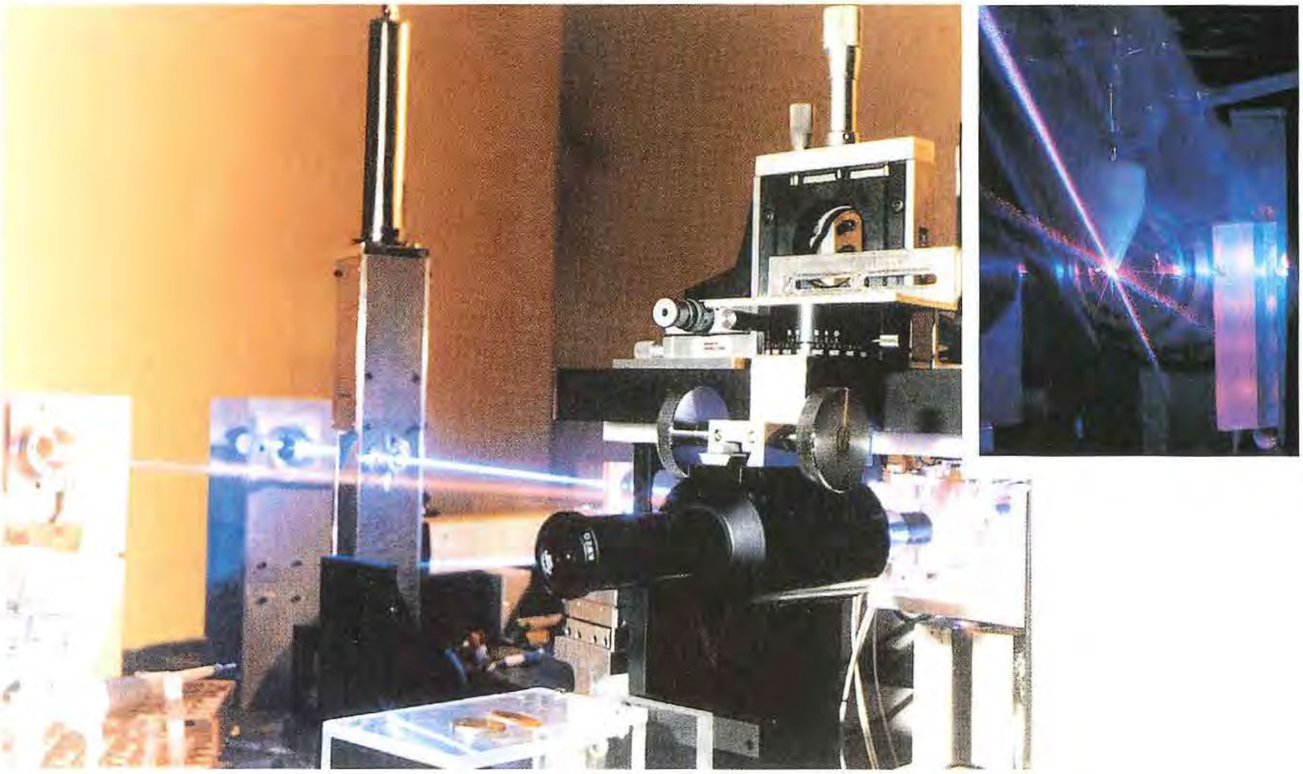
John C. Wooley, Office of Health and Environmental Research, DOE

HGCC Executive Officer: **Sylvia J. Spengler**, Deputy Director Human Genome Center, Lawrence Berkeley Laboratory

- ensuring proper management and sharing of data and samples;
- participating with other national and international efforts; and
- recommending establishment of *ad hoc* task groups to analyze specific areas, such as ethical, legal, and social issues; informatics requirements; mapping and sequencing technologies; use of the mouse as a model organism; cost of resource distribution; and use of chromosome flow-sorting facilities.

A Principal Scientist is a member of HGCC, reports to the Human Genome Program Task Group regarding the responsibility of keeping the program at the leading edge of genome research, and conveys recommendations on broad scientific policies to HGCC. Currently serving as a Principal Scientist is Charles R. Cantor, Lawrence Berkeley Laboratory.

Separating Chromosomes by Flow Cytometry. *The flow sorter pictured below analyzes and separates chromosomes isolated from cells synchronized in the cell division stage when they are condensed and stable. The chromosomes are stained with fluorescent dyes that bind specifically to DNA. As the chromosomes flow one by one past a laser beam* ***(Inset)****, they are distinguished by measuring the intensity of their laser-activated fluorescence, which is proportional to their DNA content. Individual chromosomes of any desired type can then be directed to collection tubes. Pioneered at LANL and further developed at LANL and LLNL, flow sorting generates the chromosome collections needed for the construction of corresponding clone libraries, a project carried out at LANL and LLNL through the National Laboratory Gene Library Project. The availability of human chromosome–specific libraries (through LANL, LLNL, and the American* Type Culture *Collection in Rockville, Maryland) reduces the total genome mapping effort to 24 smaller, more manageable mapping projects. (Photographs provided by Ger van den Engh, LLNL. For more information, refer to van den Engh and P. de Jong in the Index to Principal and Coinvestigators.)*



Program Management Infrastructure: Field Coordination

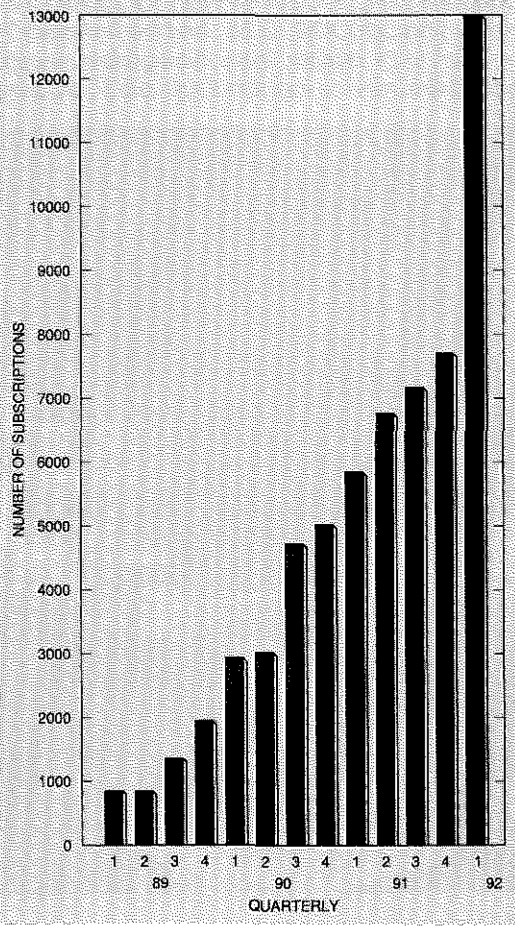
Human Genome Management Information System (HGMIS)

As an aid to the DOE Human Genome Program Task Group, communication and information services are provided by HGMIS at Oak Ridge National Laboratory. In this role HGMIS facilitates international communication among management and research personnel and informs other interested persons about genome research. HGMIS publications, such as the bimonthly newsletter *Human Genome News* and technical and program reports, are available to anyone interested in the genome project. *Human Genome News* is jointly supported by OHER and the NIH National Center for Human Genome Research (NCHGR).

Subscribers to the newsletter number over 13,000 and 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; the press; and other interested individuals. In the first quarter of 1992, over 5000 Genome Data Base users were added to the mailing list. Subscribers outside the United States include more than 3000 individuals and institutions in 48 countries.

GROWTH OF NEWSLETTER MAILING LIST

FROM APRIL 1989 TO APRIL 1992



(HGMIS Contact:
615/576-6669,
Fax: 615/574-9888.)

Human Genome Distinguished Postdoctoral Fellowships

In 1990 OHER established the Human Genome Distinguished Postdoctoral Research Program to support research on projects related to the DOE Human Genome Program. The postdoctoral program developed from a 1988 recommendation of the DOE Energy Research Advisory Board to "increase support through expansion of the targeted (science and engineering) graduate and postgraduate research fellowship programs with emphasis given to energy-related areas of greatest projected human resource shortages." Recipients of the first fellowships, awarded in FY 1991, are listed below.

Fellowship appointments are tenable at DOE and university laboratories having substantial DOE-sponsored research projects supportive of the Human Genome Program. Fellows will participate in advanced genetics-related research, interact with outstanding professionals, and become familiar with major issues while making personal contributions to the program's goal of mapping and sequencing the human genome. This interaction, involving the exchange of ideas, skills, and technologies, will benefit the fellow, the host laboratory, and the DOE program.

These fellowships complement the Alexander Hollaender Distinguished Postdoctoral Fellowships initiated by OHER. The Hollaender Fellowships, established in memory of the 1983 recipient of the prestigious DOE Enrico Fermi Award, provide support in all areas of OHER-sponsored research. Both postdoctoral programs are administered by Oak Ridge Associated Universities, which is a university consortium and DOE contractor.

1991 DOE Human Genome Distinguished Postdoctoral Fellows*

Xiaohua Huang (Stanford University, Biophysical Chemistry)
Host: *University of California, Berkeley*

Ben Koop (Wayne State University, Molecular Biology and Genetics)
Host: *California Institute of Technology*

Carol Soderlund (New Mexico State University, Computer Science)
Host: *Los Alamos National Laboratory*

Harold Swerdlow (University of Utah, Bioengineering)
Host: *University of Utah*

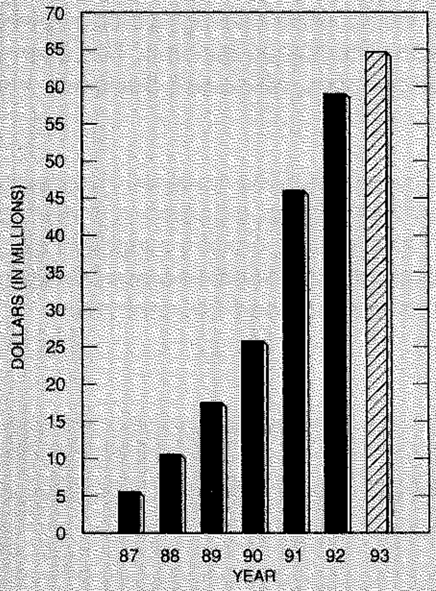
*Contact: Linda Holmes: 615/576-3192, Fax: 615/576-0202.

Resource Allocation

Program Management Infrastructure

Reports by the Health and Environmental Research Advisory Committee (HERAC) and the National Research Council (NRC) recommended that national funding for the Human Genome Project increase to a sustaining yearly level of $200 million. DOE program expenditures were $5.5 million in FY 1987, $10.7 million in FY 1988, $17.5 million in FY 1989, $25.9 million in FY 1990, $46 million in FY 1991, and $59 million in FY 1992. The proposed presidential budget for the DOE Human Genome Program in FY 1993 is $64.7 million (graph). DOE-sponsored research is conducted in a variety of institutions (upper table). The lower table categorizes research expenditures for FY 1992.

Expenditures and FY 1993 Presidential Budget for the DOE Human Genome Program



Types of Institutions Conducting DOE-Sponsored Genome Research

8	National laboratories
3	Other federal organizations
41	Academic institutions
10	Private-sector institutions
12	Nonacademic, commercial organizations

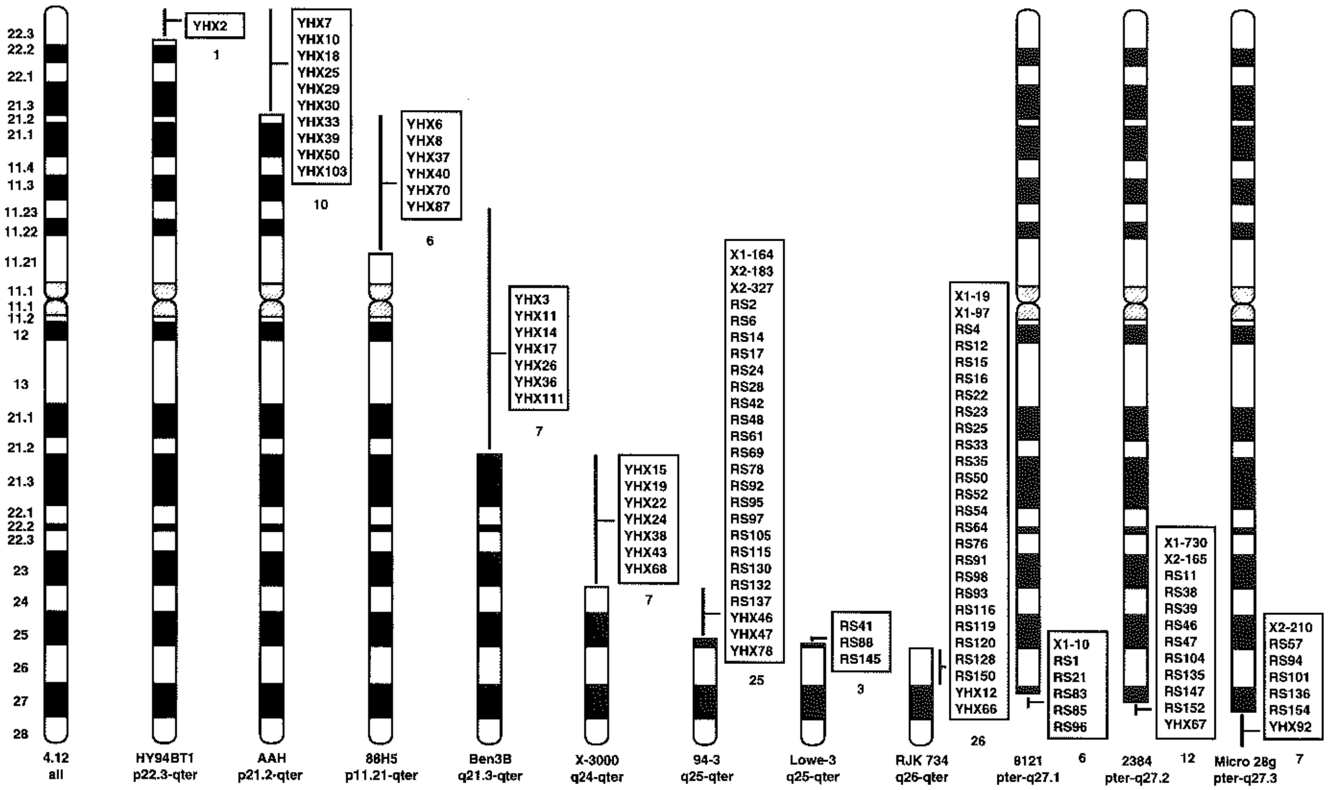
Human Genome Program Funds Distribution in FY 1992 (in $K)
(Commitments as of May 1, 1992)

	Project Type					
Organization Type	**Mapping & Sequencing**	**Instrumentation Development**	**Informatics**	**ELSI**	**Totals**	**Percent of 56800[1]**
DOE laboratories	23671	7559	5122	236	36588	64.4
Academic sites	5462	3341	4528	736	14067	24.8
Institutions (nonprofit)	2173	0	602	847	3622	6.4
NIH laboratories	680	0	0	0	680	1.2
Companies and SBIR[2]	1550	0	314	392	2256	3.9
All organizations	33536	10900	10566	2211	57213	
[Percent of 56800]	[59.0]	[19.2]	[18.6]	[3.9]	[100.7][3]	

[1] Total allocation of $59 million less capital equipment funds of $2.2 million.

[2] Small Business Innovation Research grants.

[3] Excess occurs because funding for genome SBIR projects is received from the DOE-wide SBIR program, to which OHER contributes.



Assignment of 110 Yeast Artificial Chromosome Clones Derived from the Human X Chromosome to Regions Between Translocation and Deletion Breakpoints in an X Chromosome Mapping Panel. *Specifically localized were 72 clones from a library generated from the Xq24-qter region of the chromosome and 38 clones from a library generated from the entire chromosome. Below each ideogram are the names of the hybrid cell lines used for each breakpoint and the regions retained. Several clones were identified in medically interesting regions; one set of three overlapping clones crossed a chromosomal translocation implicated in Lowe syndrome. The clones represented here provide material for developing long-range contigs and physical maps of these intervals and for isolating and characterizing several disease loci; the human X chromosome has been implicated in 200 to 300 single-gene disorders. [Figure first published in David L. Nelson et al., "*Alu*-Primed Polymerase Chain Reaction for Regional Assignment of 110 Yeast Artificial Chromosome Clones from the Human X Chromosome: Identification of Clones Associated with a Disease Locus,"* Proc. Natl. Acad. Sci. USA ***88****, 6157–61 (1991). Figure provided by D. L. Nelson, Baylor College of Medicine. For more information, refer to Nelson and C. T. Caskey in Index to Principal and Coinvestigators.]*

Interagency Coordination

Program Management Infrastructure

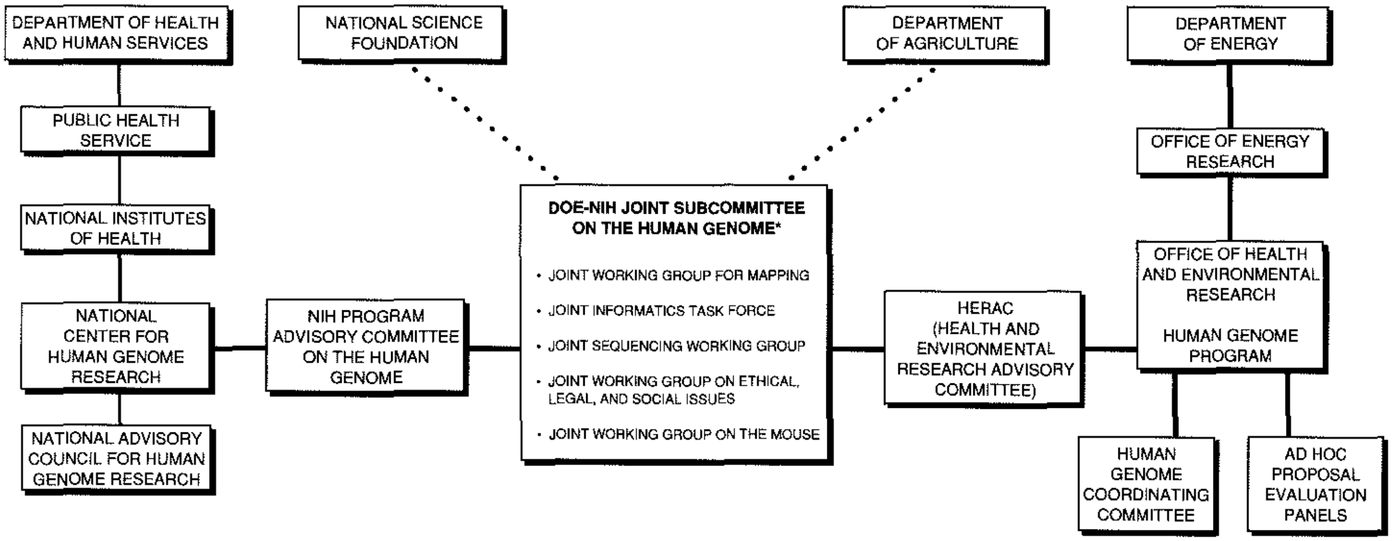
Joint DOE-NIH Activities

The NIH Human Genome Program, led by NIH NCHGR, has emphasized the study of disease genes in the construction of complete genetic and physical maps of the genomes of humans and selected model organisms. NIH is also developing new technologies and information systems to manage mapping and sequencing data.

In the fall of 1988 DOE and NIH began coordinating their human genome research programs under the Memorandum of Understanding, an outgrowth of the HERAC and NRC reports, "to foster interagency cooperation that will enhance the human genome research capabilities of both agencies." More information on NCHGR-sponsored projects and infrastructure may be obtained by contacting the NCHGR Office of Communications at 301/402-0911.

Joint DOE-NIH Subcommittee on the Human Genome in 1992

Cochairs:	
Paul Berg (PACHG)	Stanford University School of Medicine
Sheldon Wolff (HERAC)	University of California, San Francisco
Charles R. Cantor	Lawrence Berkeley Laboratory (HGCC)
Anthony V. Carrano	Lawrence Livermore National Laboratory (HGCC)
Joseph L. Goldstein	University of Texas Southwestern Medical Center
Leroy E. Hood	California Institute of Technology
Leonard S. Lerman	Massachusetts Institute of Technology (HERAC)
Victor A. McKusick	Johns Hopkins Hospital
Robert K. Moyzis	Los Alamos National Laboratory (HGCC)
Maynard V. Olson	Washington University School of Medicine (PACHG)
MaryLou Pardue	Massachusetts Institute of Technology (HERAC)
Mark L. Pearson	E. I. du Pont de Nemours & Company (PACHG)
Diane C. Smith	Xerox Corporation (PACHG)
Robert T. Tjian	University of California, Berkeley
Nancy S. Wexler	Columbia University (PACHG)
John C. Wooley	Office of Health and Environmental Research, DOE
Ex Officio Members:	
David J. Galas	Office of Health and Environmental Research, DOE
Mark S. Guyer	National Center for Human Genome Research, NIH
Elke Jordan	National Center for Human Genome Research, NIH
David A. Smith	Office of Health and Environmental Research, DOE
Michael Gottesman	National Center for Human Genome Research, NIH


DEPARTMENT OF HEALTH AND HUMAN SERVICES
PUBLIC HEALTH SERVICE
NATIONAL INSTITUTES OF HEALTH
NATIONAL CENTER FOR HUMAN GENOME RESEARCH
NATIONAL ADVISORY COUNCIL FOR HUMAN GENOME RESEARCH
NIH PROGRAM ADVISORY COMMITTEE ON THE HUMAN GENOME
NATIONAL SCIENCE FOUNDATION
DEPARTMENT OF AGRICULTURE
DOE-NIH JOINT SUBCOMMITTEE ON THE HUMAN GENOME*
• JOINT WORKING GROUP FOR MAPPING
• JOINT INFORMATICS TASK FORCE
• JOINT SEQUENCING WORKING GROUP
• JOINT WORKING GROUP ON ETHICAL, LEGAL, AND SOCIAL ISSUES
• JOINT WORKING GROUP ON THE MOUSE
HERAC (HEALTH AND ENVIRONMENTAL RESEARCH ADVISORY COMMITTEE)
DEPARTMENT OF ENERGY
OFFICE OF ENERGY RESEARCH
OFFICE OF HEALTH AND ENVIRONMENTAL RESEARCH
HUMAN GENOME PROGRAM
HUMAN GENOME COORDINATING COMMITTEE
AD HOC PROPOSAL EVALUATION PANELS

*Various organizations and federal agencies interact through the NIH-DOE Joint Subcommittee meetings.

A national plan, primarily authored by NIH and DOE, for a coordinated multiyear research project was presented to Congress in early 1990. *Understanding Our Genetic Inheritance, The U.S. Human Genome Project: The First Five Years (1991–1995)* detailed a comprehensive spending plan and optimal strategies for mapping and sequencing the human genome. Referred to as the Five Year Plan, it calls for open biannual meetings of the DOE-NIH Joint Subcommittee on the Human Genome. The joint subcommittee invites reports from experts, including those on national and international genome efforts; medical genetics; and related ethical, legal, and social issues as they pertain to data produced in the project. The subcommittee is made up of members from the NIH Program Advisory Committee on the Human Genome (PACHG) and from the DOE HERAC or the HGCC members appointed by HERAC. The subcommittee reports to its parent committees—PACHG and HERAC.

Many workshops and meetings have since been cosponsored by the two agencies (see Appendix B). In addition, the Joint Subcommittee on the Human Genome has established five joint working groups that meet regularly to address specific areas of genome research and make recommendations to the joint subcommittee. The objectives of these five joint working groups, listed below, include establishing research priorities; identifying research, training, and technical needs; and coordinating U.S. research activities with those of other countries. Members of the working groups represent various disciplines. (Membership lists of the working groups are included in Appendix D.)

Joint Mapping Working Group. The mapping working group encourages development and use of methodologies to integrate genetic linkage and physical maps, meet project mapping goals, and identify informatics needs associated with map generation and completion.

Joint Informatics Task Force (JITF). An *ad hoc* committee, JITF prepared a comprehensive report on genome information needs and data analysis tools. The report was presented to the DOE-NIH Joint Subcommittee on the Human Genome in January 1992.

Joint Sequencing Working Group. The sequencing working group investigates and makes recommendations on research and technology development priorities to enable the sequencing of 3 billion nucleotides of human DNA within 15 years.

Joint Working Group on Ethical, Legal, and Social Issues (ELSI). ELSI identifies and addresses the social concerns that may arise as genome technology is developed and genetic data becomes available; stimulates bioethics research; promotes education of professional and lay groups; and collaborates with international groups such as the Human Genome Organization (HUGO), United Nations Educational, Scientific, and Cultural Organization (UNESCO), and the European Community (see next section).

Joint Working Group on the Mouse. The mouse working group was established to develop a strategy for efficiently using the mouse to accomplish mapping project goals as outlined in the Five Year Plan. This strategy will take advantage of the extensive genetic map data amassed on the mouse. Because of numerous similarities between mouse and human genomes, these studies are considered essential to understanding human biology and to interpreting more complex data obtained in studies of humans.

Other U.S. Genome Research

U.S. Department of Agriculture (USDA). USDA has implemented a Plant Genome Research Program to foster and coordinate research on single and multigenic traits related to agricultural, forestry, and environmental concerns. The goal of this 5-year program is to improve plant varieties by locating important genes and markers on chromosomes, determining gene structure, and transferring genes to improve the performance of economically important crops such as corn, wheat, soybeans, and pine. Use of these "molecular breeding" techniques will increase U.S. competitiveness in the world marketplace.

National Science Foundation (NSF). NSF coordinates an interagency research effort to map and sequence the small genome of *Arabidopsis thaliana*, a simple weed that provides an ideal model for studying plant biochemistry, genetics, and physiology. Knowledge of the function of every *Arabidopsis* gene will be applicable to the understanding and manipulation of higher plants and to genome research in general. These studies are also supported by DOE, NIH, and USDA as part of their own genome initiatives, and the four agencies coordinate their *Arabidopsis* activities. NSF also has instrumentation, computational, and informatics programs that support genomics research, in addition to individual awards in genetics and molecular biology.

Howard Hughes Medical Institute (HHMI). HHMI, a private medical research organization, contributes to the genome effort through its support of biomedical research primarily at university molecular biology and genetics laboratories. In addition, HHMI has cosponsored several genomics conferences and, between 1985 and September 1991, supported the collection and dissemination of genome mapping data through a network of databases.

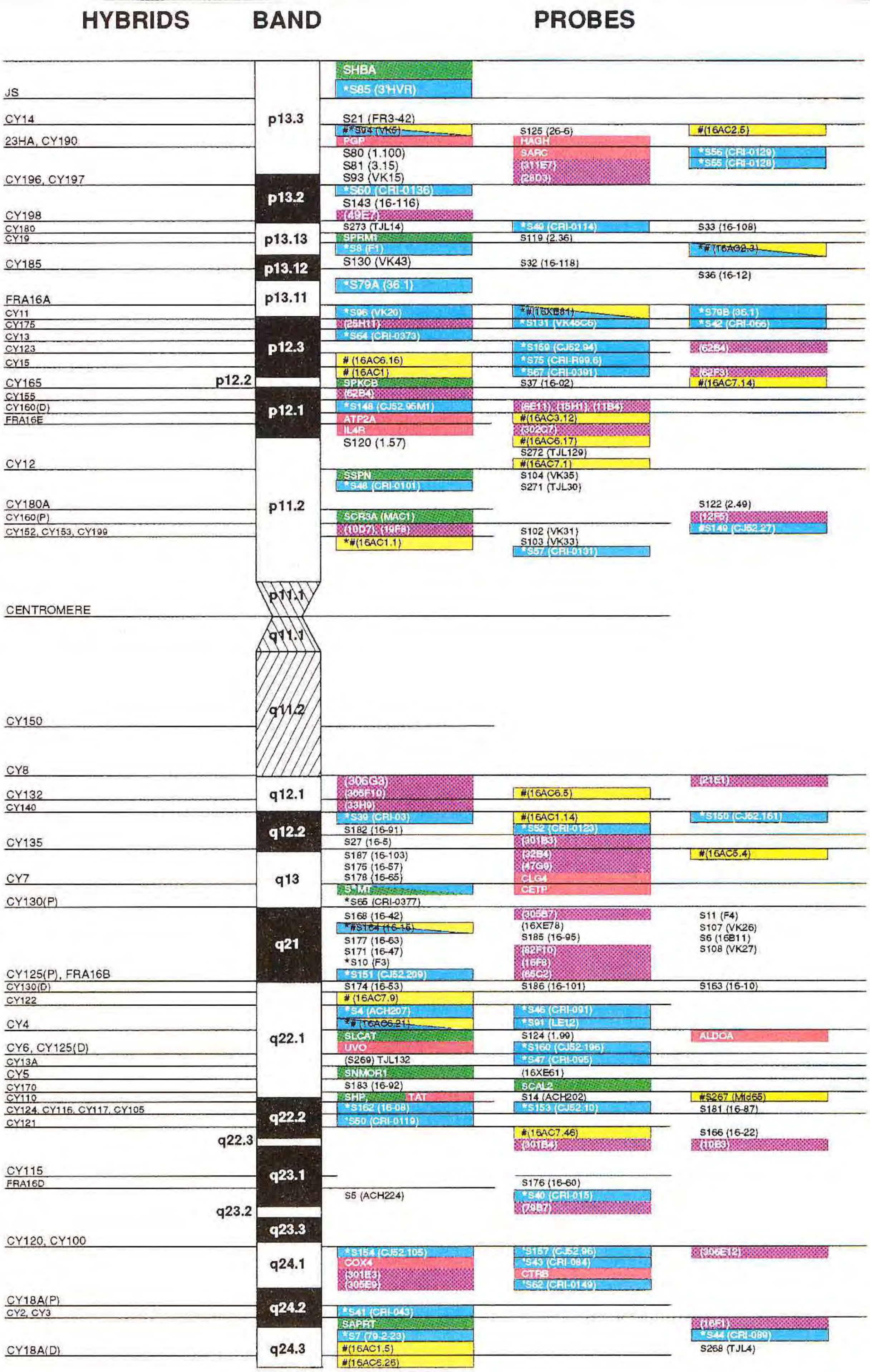
Researchers reviewing chromosome 19 map data on a slide viewer at LLNL Human Genome Center.

Physical Map of Human Chromosome 16 Depicting Mapped Genes and Probes. *Listed to the left of the ideogram are mouse-human somatic cell hybrids (CY numbers) containing breakpoints of human chromosome 16, as indicated by the horizontal line. Hybrids generally contain the region of chromosome 16 from the breakpoint to the tip of the long arm. Hybrids CY160, CY130, and CY125 are interstitial deletions.*

Listed to the right are genes and probes that have been localized by Southern blot analysis or polymerase chain reaction methods. The colors are coded as follows:

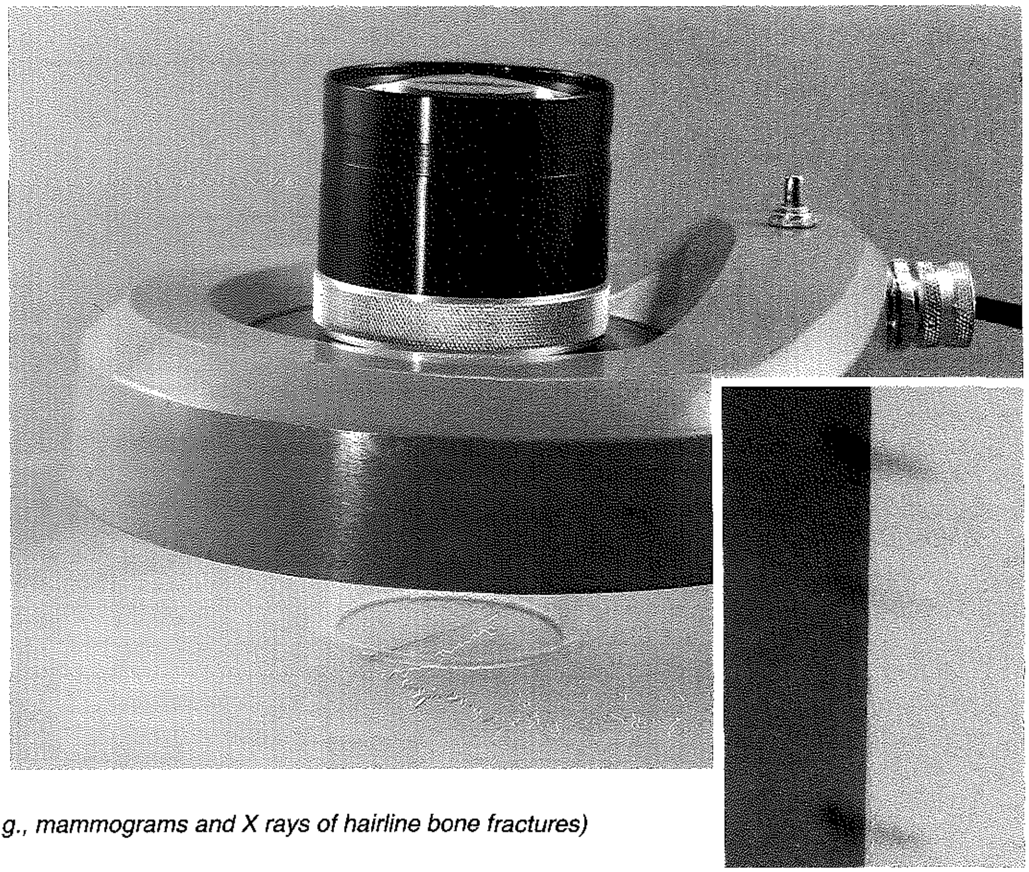
- *pink* *mapped genes,*
- *green* *mapped genes converted to sequence tagged sites (STSs),*
- *yellow* *STSs that are* $(AC)_n$ *microsatellite repeats,*
- *blue* *markers typed in Centre d'Etude du Polymorphisme Humain families, and*
- *purple* *cosmids that are members of the LANL contigs.*

(Figure provided by David Callen, Adelaide Children's Hospital. For more information, refer to Callen and G. R. Sutherland in Genomics *(1992, in press) and in the Index to Principal and Coinvestigators.)*


HYBRIDS
BAND
PROBES
SHBA
*S85 (3'HVR)
JS
CY14
p13.3
S21 (FR3-42)
#*S94 (VK5)
S125 (26-6)
#(16AC2.5)
23HA, CY190
PGP
HAGH
S80 (1.100)
SARC
*S56 (CRI-0129)
S81 (3.15)
(311E7)
*S55 (CRI-0128)
S93 (VK15)
(28D3)
CY196, CY197
p13.2
*S60 (CRI-0136)
S143 (16-116)
CY198
(49E7)
CY180
S273 (TJL14)
*S40 (CRI-0114)
S33 (16-108)
CY19
p13.13
SPRM1
S119 (2.36)
*S8 (F1)
*# (16AC2.3)
CY185
p13.12
S130 (VK43)
S32 (16-118)
S36 (16-12)
*S79A (36.1)
FRA16A
p13.11
CY11
*S96 (VK20)
*#(16XE81)
*S79B (36.1)
CY175
(25H11)
*S131 (VK45C6)
*S42 (CRI-066)
CY13
*S64 (CRI-0373)
CY123
p12.3
*S159 (CJ52.94)
(62B4)
CY15
# (16AC6.16)
*S75 (CRI-R99.6)
# (16AC1)
*S67 (CRI-0391)
(62F3)
CY165
p12.2
SPKCB
S37 (16-02)
#(16AC7.14)
CY155
(62B4)
CY160(D)
p12.1
*S148 (CJ52.95M1)
(6E11), (15H1), (11B4)
FRA16E
ATP2A
#(16AC3.12)
IL4R
(302C7)
S120 (1.57)
#(16AC6.17)
S272 (TJL129)
CY12
#(16AC7.1)
SSPN
S104 (VK35)
*S46 (CRI-0101)
S271 (TJL30)
CY180A
p11.2
S122 (2.49)
CY160(P)
SCR3A (MAC1)
(12F5)
CY152, CY153, CY199
(10D7), (19F8)
S102 (VK31)
#S149 (CJ52.27)
*#(16AC1.1)
S103 (VK33)
*S57 (CRI-0131)
p11.1
CENTROMERE
q11.1
q11.2
CY150
CY8
(306G3)
(21E1)
CY132
q12.1
(305F10)
#(16AC6.5)
CY140
(33H9)
*S39 (CRI-03)
#(16AC1.14)
*S150 (CJ52.161)
q12.2
S182 (16-91)
*S52 (CRI-0123)
CY135
S27 (16-5)
(301B3)
S187 (16-103)
(32B4)
#(16AC5.4)
S175 (16-57)
(47G6)
CY7
q13
S178 (16-65)
CLG4
S-MT
CETP
CY130(P)
*S65 (CRI-0377)
S168 (16-42)
(305B7)
S11 (F4)
*#S164 (16-15)
(16XE78)
S107 (VK26)
q21
S177 (16-63)
S185 (16-95)
S6 (16B11)
S171 (16-47)
(82F10)
S108 (VK27)
*S10 (F3)
(15F8)
CY125(P), FRA16B
*S151 (CJ52.209)
(65C2)
CY130(D)
S174 (16-53)
S186 (16-101)
S163 (16-10)
CY122
# (16AC7.9)
CY4
*S4 (ACH207)
*S46 (CRI-091)
*# (16AC6.21)
*S91 (LE12)
q22.1
SLCAT
S124 (1.99)
ALDOA
CY6, CY125(D)
UVO
*S160 (CJ52.196)
CY13A
(S269) TJL132
*S47 (CRI-095)
CY5
SNMOR1
(16XE61)
CY170
S183 (16-92)
SCAL2
CY110
SHP, TAT
S14 (ACH202)
#S267 (Mfd65)
CY124, CY116, CY117, CY105
q22.2
*S162 (16-08)
*S153 (CJ52.10)
S181 (16-87)
CY121
*S50 (CRI-0119)
q22.3
#(16AC7.46)
S166 (16-22)
(301B4)
(10B3)
CY115
q23.1
S176 (16-60)
FRA16D
S5 (ACH224)
*S40 (CRI-015)
(79B7)
q23.2
q23.3
CY120, CY100
*S154 (CJ52.105)
*S157 (CJ52.96)
(306E12)
q24.1
COX4
*S43 (CRI-084)
(301E3)
CTRB
(305E9)
*S62 (CRI-0149)
CY18A(P)
q24.2
CY2, CY3
*S41 (CRI-043)
APRT
(16F1)
*S7 (79-2-23)
*S44 (CRI-089)
CY18A(D)
q24.3
#(16AC1.5)
S268 (TJL4)
#(16AC6.26)

International Coordination

Genomic research is being carried out in countries throughout the world. The two international organizations described on the next two pages are working to coordinate and facilitate national efforts. HUGO includes a number of DOE and NIH genome investigators and administrators. HUGO and UNESCO have been informed of dedicated genome programs in the following nations and international agencies: Commonwealth of Independent States (formerly U.S.S.R.), Denmark, European Community, France, Germany, Hungary, Italy, Japan, Netherlands, United Kingdom, and United States.



Un-Dimmer. *The Un-Dimmer is a simple contrast-enhancing viewer used to read X-ray, autoradiographic, and other films. By using a reflector under the film instead of a light box, this apparatus improves the contrast of faint objects that appear on a light background. The viewing light comes from above and passes through the film twice instead of once, as in conventional transmission light boxes. This double passage of light multiplies the apparent optical density of an object. The Un-Dimmer is useful in viewing DNA sequencing and hybridization films. Broader applications may include medical radiology (e.g., mammograms and X rays of hairline bone fractures) and industrial radiography.*

Band contrast on a sequencing gel is nearly doubled with the use of the Un-Dimmer.

(Photographs provided by Jack B. Davidson, ORNL. For more information, refer to Davidson in the Index to Principal and Coinvestigators.)

HUGO: Worldwide Genome Research Coordination

HUGO, formed by scientists to coordinate worldwide genome mapping and sequencing, now has regional offices in the United States (Bethesda, Maryland) and Europe (London) and a satellite office in Moscow. A Pacific office is under development in Osaka, Japan.

HUGO offices were funded initially by several charitable organizations. In 1990 HHMI awarded HUGO a 4-year, $1 million grant to support the HUGO Americas office; in that same year The Wellcome Trust provided a 3-year grant, with the first year's funds amounting to over $400,000, to assist with activities in the European office. The Imperial Cancer Research Fund (U.K.) provides support for the HUGO president's office, and the Osaka office has received private support as well. To support future activities, HUGO directors intend to raise funds from various countries that have active genome research programs.

HUGO members are elected; there are over 400 members from 32 countries. The international officers in 1992: Sir Walter Bodmer (United Kingdom), President; Charles R. Cantor (United States), Vice-President; Andrei Mirzabekov (Russia), Vice-President; Kenichi Matsubara (Japan), Vice-President; Bronwen Loder (United Kingdom), Secretary; and Robert Sparkes (United States), Treasurer. Each office operates with its own trustees.

The objectives of HUGO include

- fostering collaboration to avoid unnecessary competition or duplication of effort and to coordinate human genome research with model organism studies;
- coordinating exchanges of relevant data and materials;
- educating researchers and the public on the scientific, ethical, social, legal, and commercial implications of the research; and
- acting as a clearinghouse for genome-related information, such as relevant conferences, worldwide genome programs and researchers, and database and material availability. A training program may be initiated to encourage the spread of new and promising technologies.

HUGO has established expert international *ad hoc* advisory committees on mapping workshops and databases, informatics, ethics, mouse mapping, and intellectual property and ownership.

Single-chromosome workshops are crucial to the success of the Human Genome Project. Working with the funding agencies, HUGO is playing a central role in the coordinated development of such meetings and has assisted in planning workshops for chromosomes 2, 3, 13, 16, 19, and X in 1992. HUGO expects to work with the scientific community to select workshop chairs and to assist in fundraising and organizing and running these and future meetings. Chromosome workshops and other meetings are listed in Appendix B.

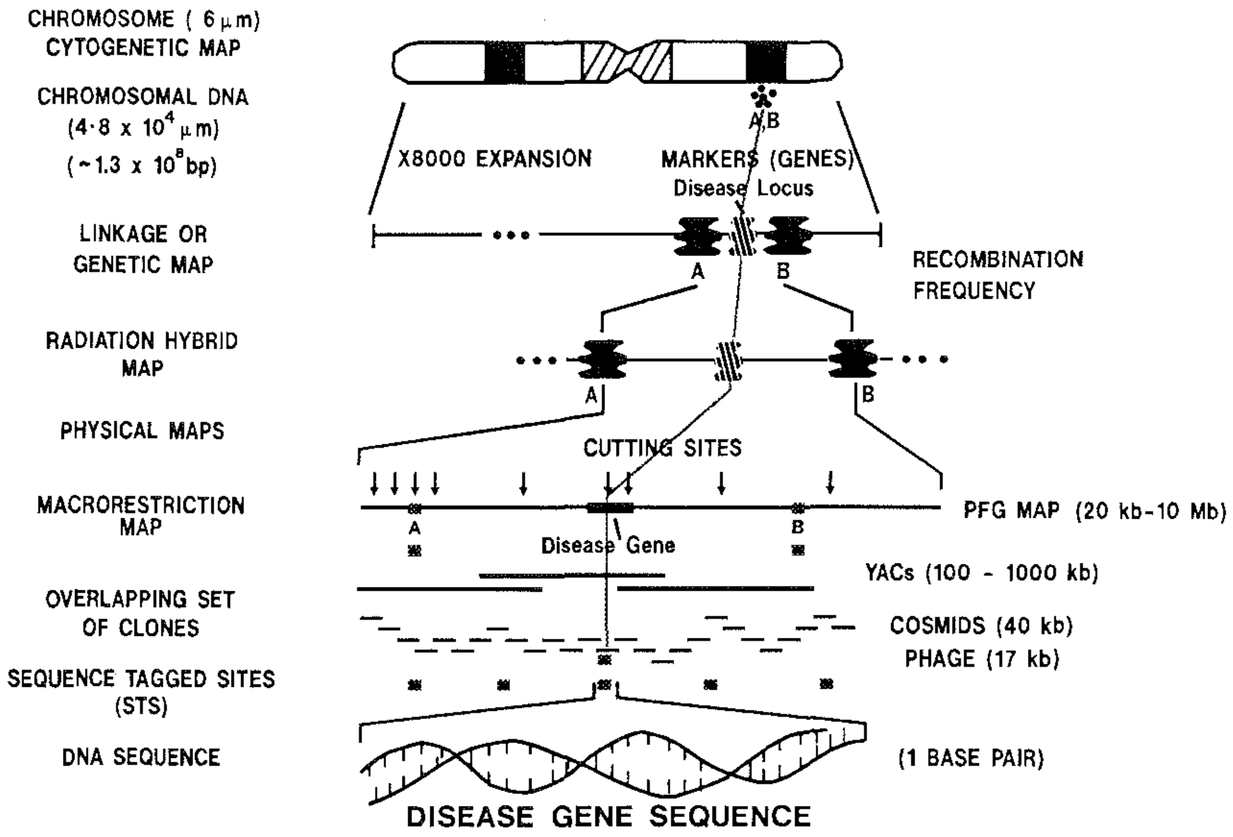
UNESCO: Promoting the Interests of Developing Countries

International Coordination

A UNESCO Human Genome Program was approved for 1990–91 at the 25th session of the UNESCO General Conference. Attendees concluded that full knowledge of the human genome is vitally important and that UNESCO could be influential in stimulating governments and agencies to support coordinated programs. UNESCO expects to play a key role in promoting the interests of developing countries. The Scientific Coordinating Committee (SCC), composed of 13 scientists, plans and implements the program, which was budgeted at $350,000 for the first year; SCC members include representatives selected from geographic regions and from international genome organizations such as HUGO. Members of SCC and of the UNESCO Secretariat agreed that UNESCO will concentrate its activities on access to and use of data obtained from human genome mapping and sequencing research, as well as on related ethical and social issues.

UNESCO emphasizes the use of training programs as one of the best means of obtaining cooperation and diminishing the gap between developed and developing countries. The Third World Academy of Sciences (TWAS) joined UNESCO in sponsoring a training program that provided 19 fellowships in 1991 to awardees from Algeria, Argentina, Cameroon, Chile, China, Costa Rica, Cyprus, Czechoslovakia, Egypt, Guinea, India, Indonesia, Myanmar, the Republic of Korea, Peru, Spain, Ukraine, Russia, and Yugoslavia. The 1- to 3-month fellowships enable scientists from developing countries to carry out research in well-established scientific centers and to learn new research techniques. UNESCO and TWAS are also jointly compiling a directory to identify third-world genome researchers and their needs.

To avoid overlap with other genome projects, UNESCO focuses on communication among countries about major trends and regional efforts, one of which, the Latin American Human Genome Program, was established during a UNESCO-supported symposium in Chile in 1990. The first annual UNESCO South-North Human Genome Conference was held in 1992 in Caxambu, Brazil, to increase interaction between scientists from developed countries and those of the third world. The second conference is planned for Thailand in 1993, and the third will probably take place in China in 1994.



Multiple Levels of Human Chromosome Mapping. *The line running vertically through the diagram represents the tracking of markers A and B through progressively more precise levels of mapping. In this way, investigators can follow a candidate disease gene from the coarsest to the finest map resolution, which is the DNA sequence. The cytogenetic map provides the lowest level of resolution, measuring the distance between chromosomal features (i.e., bands or breakpoints) visible under the light microscope. Chromosome banding can resolve features to about 5 Mb. The linkage or genetic map measures the recombination frequency between two linked markers, which can be genes or polymorphisms (A and B in this diagram.) Radiation hybrid maps are produced by breaking chromosomes with radiation and then identifying the fragment carrying the marker (the breakpoint); the resolution of these maps is comparable to that of linkage maps. At the next resolution level, macrorestriction fragments of 1 to 2 Mb are separated and the markers localized and mapped. Finer mapping resolution is provided by ordered libraries of yeast artificial chromosomes (YACs), which have insert sizes from 100 to 1000 kb. Ordered libraries of cosmids have smaller insert sizes, usually about 40 kb, and produce higher-resolution maps. The DNA base sequence is the highest-resolution map, with sequence tagged sites (STSs) used as unique reference points. (Figure provided by C. E. Hildebrand, LANL.)*

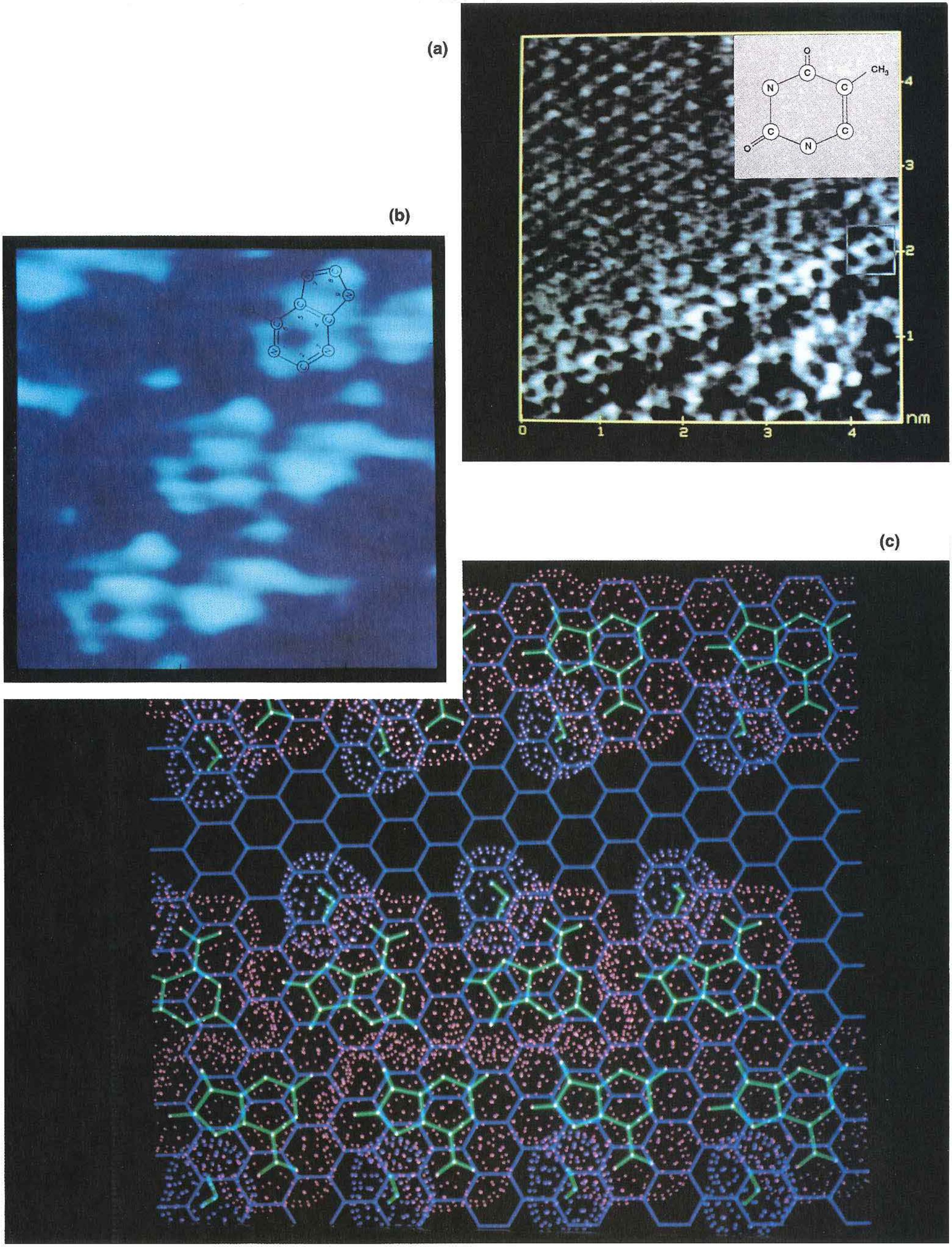
Scanning Tunneling Microscope (STM) Images of the Bases Adenine and Thymine on Graphite. *Organized lattices of adenine and thymine molecules were produced by applying aqueous solutions of the bases to heated graphite.*

(a) *A thymine solution formed rows of molecules containing only the single-ring structure that is characteristic of pyrimidines. (Box overlay identifies one molecule. Molecular model of thymine is inset for comparison with the photo image.)*

(b) *High resolution of adenine molecules revealed bimolecular rows with each molecule displaying the double-ring purine structure; note molecular adenine overlay. (Adenine molecules on the right side of each row were distorted and did not display a complete ring structure.)*

(c) *Computer model of the adenine lattice is at an energy minimum consistent with STM data. The lattice is stabilized primarily by tight van der Waals packing of hydrated adenines to each other and to the underlying graphite substrate. Colored dots represent van der Waals radii (the points at which atoms begin to repel each other) for nitrogen, carbon, hydrogen, and oxygen atoms. The blue honeycomb grid represents the hexagonal arrangement of the carbon atoms in the graphite substrate. The model was generated using the molecular dynamics program ENCAD. Applications of STM to DNA analysis and sequencing may directly impact the progress of the human genome effort by eventually providing a new, electronic method for sequencing DNA at 3 orders of magnitude faster than existing methods.*

[Photographs first published in M. J. Allen et al., "Scanning Tunneling Microscope Images of Adenine and Thymine at Atomic Resolution," Scanning Microsc., *in press (1991). Photographs provided by Biomedical Sciences Division, LLNL. For more information, refer to R. L. Balhorn and W. Siekhaus in the Index to Principal and Coinvestigators.]*


(a)
CH₃
O
C
N
C
C
C
N
O
nm
0
1
2
3
4
1
2
3
4
(b)
(c)

Abstracts of DOE-Funded Research

The abstracts in this section (beginning on p. 82) were contributed by DOE Human Genome Program grantees and contractors. Names of the principal investigators are in bold print. Telephone and telefax numbers and electronic mail address are for the first person named unless the investigator to be contacted is identified with a superscript †. An index of project categories and principal investigators begins on the facing page, and an index of all investigators named in the abstracts is given at the end of this report.

Project Categories and Principal Investigators

Principal investigators of the research projects described by the abstracts in this section are listed here under their respective subject categories.

RESOURCE DEVELOPMENT

Raghbir S. Athwal ... 82
E. Morton Bradbury and Joe M. Gatewood ... 82
Jan-Fang Cheng ... 83
Larry L. Deaven ... 83
Larry L. Deaven ... 84
Pieter J. de Jong ... 84
Joe Gray ... 85
Cynthia L. Jackson ... 86
Fa-Ten Kao ... 86
Julie R. Korenberg ... 87
Michael J. Lane, Peter Hahn, and John Hozier ... 88
Gregory G. Lennon and Anthony V. Carrano ... 88
Christopher H. Martin and Michael J. Palazzolo ... 89
MaryKay McCormick, Larry Deaven, and Robert Moyzis ... 89
Donald T. Moir ... 90
Robert K. Moyzis ... 91
David L. Nelson ... 91
William C. Nierman and Donna R. Maglott ... 92
Mihael H. Polymeropoulos ... 92
Venigalla B. Rao ... 93
Eugene Rinchik and Richard Woychik ... 93
Eugene Rinchik and Richard Woychik ... 94
Michael J. Siciliano ... 94
James M. Sikela ... 95
Marcelo Bento Soares ... 96
Betsy M. Sutherland ... 96
J. Craig Venter ... 97
Jean-Michel H. Vos ... 97
Geoffrey Wahl ... 98
Philip A. Youderian ... 98

PHYSICAL AND GENETIC MAPPING

C. Thomas Caskey and David L. Nelson ... 99
Jan-Fang Cheng ... 100
Pieter J. de Jong and Anthony V. Carrano ... 100
Glen A. Evans ... 101
Jeffrey C. Gingrich ... 102
Christopher H. Martin and Michael J. Palazzolo ... 103
Harvey W. Mohrenweiser ... 103
Anne S. Olsen ... 104
Melvin I. Simon ... 105

Abstracts: Project Categories and Principal Investigators

Cassandra L. Smith ... 105
R. L. Stallings and R. K. Moyzis ... 106
Lisa J. Stubbs and Eugene Rinchik ... 107
Grant R. Sutherland ... 108
Barbara Jo Trask and Anthony V. Carrano ... 109

MAPPING INSTRUMENTATION

Tony J. Beugelsdijk ... 111
Charles R. Cantor ... 111
J. Calvin Giddings ... 112
James F. Hainfeld ... 112
Joseph Jaklevic ... 113
James H. Jett, John C. Martin, and Mark E. Wilder ... 114
Barry L. Karger ... 114
William F. Kolbe, Joseph E. Katz, and Joseph M. Jaklevic ... 115
Leonard Lerman ... 115
Patricia A. Medvick ... 116
Donald C. Uber ... 117
Edward S. Yeung ... 117

SEQUENCING TECHNOLOGIES

Thomas M. Brennan ... 119
Gilbert M. Brown ... 119
C. H. Winston Chen, Marvin G. Payne, and K. Bruce Jacobson ... 120
George M. Church ... 121
Radomir Crkvenjakov and Radoje Drmanac ... 121
Jack B. Davidson ... 122
M. Bonner Denton and Richard Keller ... 123
Norman J. Dovichi ... 124
Thomas L. Ferrell, Robert J. Warmack, and David P. Allison ... 125
Robert S. Foote ... 125
Raymond F. Gesteland ... 126
Leroy E. Hood ... 126
K. Bruce Jacobson ... 127
Joseph M. Jaklevic and W. Henry Benner ... 128
Joseph M. Jaklevic and W. F. Kolbe ... 128
James H. Jett, Richard A. Keller, John C. Martin, and E. Brooks Shera ... 129
Raoul Kopelman, John Langmore, and Bradford Orr ... 130
Christopher H. Martin and Michael J. Palazzolo ... 131
Richard A. Mathies and Alexander N. Glazer ... 132
David Mead and Lloyd M. Smith ... 132
Michael C. Pirrung ... 133
Charles C. Richardson ... 133
Arthur D. Riggs ... 135

Lloyd M. Smith 135
Lloyd M. Smith and Brian Chait 136
Richard D. Smith 136
Linda D. Strausbaugh and Claire M. Berg 137
F. William Studier and John J. Dunn 138
Peter Williams and Neal Woodbury 138

INFORMATICS

Elbert Branscomb 139
James Cassatt 140
Richard J. Douthart 140
Vince Faber and David Torney 141
James W. Fickett 141
Christopher A. Fields and Carol A. Soderlund 142
Tim Hunkapiller 143
Jerzy Jurka 144
Eugene Lawler 144
Charles B. Lawrence and Eugene W. Myers 144
Suzanna Lewis and Frank Olken 145
Suzanna Lewis, Manfred Zorn, and John McCarthy 145
Victor Markowitz 146
Victor Markowitz, Tim Hunkapiller, and Suzanna Lewis 147
Thomas G. Marr and Andrew Reiner 148
Frank Olken 148
Peter L. Pearson 148
David B. Searls 150
Jude W. Shavlik and Michiel O. Noordewier 151
David Torney 152
Edward Uberbacher and Richard Mural 153
Edward Uberbacher, Richard Mural, and Reinhold Mann 154
Manfred D. Zorn 155

ETHICAL, LEGAL, AND SOCIAL ISSUES

Paula Apsel and Graham Chedd 156
Ruth E. Bulger 156
C. Thomas Caskey and J. Robert Nelson 157
Debra L. Collins, R. Neil Schimke, and Linda Segebrecht 157
Troy Duster 158
Frank Grad 159
Joseph D. McInerney 159
Marvin Natowicz 160
George Page and Stefan Moore 160
Phillip Reilly 161
Ralph W. Trottier and Lee A. Crandall 161
Jan Witkowski 161

Abstracts: Project Categories and Principal Investigators

INFRASTRUCTURE

Charles R. Cantor 163
Linda Holmes and Alfred Wohlpart 163
Amanda Lumley and James Wright 163
Betty K. Mansfield and John S. Wassom 164
Sylvia Spengler 164
Michael S. Yesley 164

Small Business Innovation Research Phase I (6 1/2 MONTHS)

Heinrich F. Arlinghaus 165
Frederic R. Furuya 166
Gerald D. Hurst 166
Robert G. Lowery 166

New SBIR Phase I (Effective July 27, 1992)

W. Holt Anderson 167
Irena Bronstein 167
Jay Flatley 167
John R. Hartman 167
Douglass J. McAllister 167
Gordon Ringold 167
Darlene B. Roszak-MacDonell 167

Small Business Innovation Research Phase II (2 YEARS)

Norman G. Anderson 168
Heinrich F. Arlinghaus 168
John R. Hartman 169
Christopher Martin and Irena Bronstein 170

COMPLETED PROJECTS

Projects in this category are either completed or no longer receiving support through the DOE Human Genome Program.

Resource Development

William C. Nierman and Donna R. Maglott 171
Sherman M. Weissman and David C. Ward 171

Mapping and Mapping Instrumentation

Stylianos E. Antonarakis and Philip A. Hieter 172
David F. Barker 172
Gerald Entine 173
Glen A. Evans 174
Gunter A. Hofmann 174

Leonard S. Lerman 175
Ronald A. McKean 175
George M. Storti 176

Sequencing Technologies

Heinrich F. Arlinghaus 177
Rodney L. Balhorn and Wigbert Siekhaus 177
Douglas E. Berg 178
Richard P. Haugland 179
Charles C. Richardson 180
James Trebes 181

Informatics

Peter Alexander 182
Ross Overbeek 182
Stanley Schwartz 183
Jeffrey M. Stiegman 183
Jeffrey M. Stiegman and Angela Corona 184
Charles D. Storman 185
John S. West 185

Ethical, Legal, and Social Issues

Peggy Fischer 186
Mark Lappé and Timothy Murphy 186
Lance Hoffman; Betsy Fader; Jonathan Beckwith 186

Monochromosomal Hybrids for the Analysis of the Human Genome

Raghbir S. Athwal
University of Medicine and Dentistry of New Jersey, New Jersey Medical School,
Newark, NJ 07103-2714
201/456-5215, Fax 201/456-3644

The objective of this project is to produce a panel of rodent-human hybrid cell lines, each containing a single different human chromosome (monochromosomal hybrids) marked with an *Ecogpt* gene. The presence of the *Ecogpt* gene affords stable retention of the human chromosomes by allowing their selection in a medium containing mycophenolic acid and xanthine (MX medium). These hybrid panels will provide an invaluable resource for genome fractionation and physical mapping of human chromosomes. In addition, these hybrids will be useful in the genetic analysis of complex phenotypes, such as growth regulation in tumor cells and repair of radiation-induced DNA damage.

The experimental approach for producing these hybrids involves "tagging" of the chromosomes in normal diploid human cells by retroviral vector-mediated transfer of the *Ecogpt* gene. Since a transferred gene can integrate at random in the recipient genome, independent gene transfer clones may each carry the selectable marker integrated into a different site and perhaps a different chromosome. We have already produced 153 independent gene transfer clones, now being analyzed for the identity of *Ecogpt*-marked chromosomes. A marked chromosome in human cells is identified by cloning the human DNA, flanking the site of vector integration by PCR amplification, and then mapping the cloned DNA fragment to a specific chromosome. We have already identified cell lines for 5 chromosomes through this approach and expect to generate a bank of 24 human cell lines each carrying *Ecogpt* integrated into a different chromosome.

Normal human cells have a limited life span. To perpetuate the marked chromosomes, these cells must be fused with immortalized human tumor cells. To facilitate the transfer of single chromosomes to recipient cells by microcell fusion, human hybrid cells are fused with mouse A9 cells to generate mouse-human segregating hybrids. Mouse-human hybrid cells are used as microcell donors to transfer the marked human chromosome to mouse or Chinese hamster cells. Microcell hybrids isolated by selection in MX medium are analyzed to confirm the identity and integrity of the transferred chromosome. The ultimate objective of this project is to generate two sets of monochromosomal hybrids, one in mouse and the other in Chinese hamster cells. We have already produced monochromosomal hybrids for chromosomes 1, 2, 5, 6, 7, 9, 13, 15, 17, and 21.

An Investigation of Gene Organization Within the Human Genome Utilizing cDNA Sequencing

E. Morton Bradbury and **Joe M. Gatewood**
Center for Human Genome Studies, Life Sciences Division, Los Alamos National Laboratory,
Los Alamos, NM 87545
505/667-2690, FTS 843-2690, Fax 505/665-3024, FTS Fax 843-3024

About 85% of the DNA in mature human spermatozoa is extremely condensed and transcriptionally inactive. The remaining DNA represents a sequence-specific subset of the genome associated with a unique subset of minor histone variants. We hypothesize that this minor DNA component, which contains the ribosomal genes, also contains the genes responsible for directing postfertilization paternal-specific gene expression. To test this hypothesis we are examining the distribution of tissue-specific cDNA sequences within human sperm chromatin. Both hydatidiform mole and ovarian teratoma RNA will be converted to cDNA and sequenced. Sequences from both genomic compartments of sperm will be mapped to individual chromosomes using a procedure based on sequence tagged sites (STSs). Our goal is to assign approximately 2500 cDNA STSs to the human genome map over a 5-year period. These assignments will determine the chromosomal distribution of genes in the condensed and active chromatin compartments and provide the framework for integrating the physical, genetic, and functional human genome maps.

Isolation of Chromosome-Specific cDNA Clones

Jan-Fang Cheng and Victor Boyartchuk
Human Genome Center, Cell and Molecular Biology Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-6549, FTS 451-6549, Fax 510/486-6816, FTS Fax 451-6816, Internet: "jfcheng@lbl.gov"

A new method of isolating cDNA clones from genomic DNA is being investigated. This method involves purification of large genomic DNA (e.g., yeast artificial chromosome or cosmid DNA), immobilization of the purified DNA to magnetic beads, and hybridization of cDNA to the immobilized "target" DNA on beads. The initial test using a cosmid containing glycinamide ribonucleotide transformylase (GART) as the target DNA showed that the majority of the cDNA recovered from a HeLa cDNA library are actually GART cDNA. We are now in the process of optimizing a number of variables in this cDNA fishing strategy. For example, we need to determine the concentrations of target DNA and beads and the number of cosmid clones that can be combined in one hybridization reaction. The advantage of this method is that the hybridization reactions can be carried out in microtiter plates with a small volume of buffer so that a large number of genomic clones can be used as target DNA at the same time.

Chromosome Structure and Function

Larry L. Deaven, Evelyn Campbell, and Mary Campbell
Center for Human Genome Studies, Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
505/667-3114, FTS 843-3114, Fax 505/665-3024, FTS Fax 855-3024, Internet: "moyzis@flovax.lanl.gov"

This project is designed to provide new information on the relationship between chromosome damage and function. A number of physical and chemical agents associated with energy production are known to alter chromosome structure or induce changes in the chromosome number of normal mammalian cells. Our studies are designed to document these changes and to elucidate cellular responses to their initiation and perpetuation.

Our current approach involves the improvement of flow cytogenetic analysis. Pure fractions of human chromosomes are essential for constructing chromosome-specific DNA libraries and may prove to be a critical element in assembling a physical map of the human genome. Chromosome libraries and a human genome physical map will provide a new generation of tools for the detection of chromosomal changes and the correlation of these changes with disease. The chromosome-specific libraries that we construct are distributed to a large number of laboratories for use in genetic and physical mapping. Libraries with high levels of purity save considerable amounts of time and expense in the overall Human Genome Project. Therefore, we are attempting to optimize the production of pure chromosome fractions by improving methods for chromosome isolation, handling isolated chromosomes, and detecting purity levels in sorted chromosomes.

We have assembled a set of rodent-human hybrid cell lines that are useful for sorting each human chromosome. We continue to attempt to find additional lines that have better flow resolution and increased chromosomal stability during long-term culture. The in situ hybridization techniques developed for analysis of cell lines and determination of sorted-chromosome purity are also being used for mapping YACs to metaphase chromosomes. Spots of chromosomes sorted directly on nitrocellulose filters have been used to map cDNA probes of unknown chromosomal origin.

Abstracts: Resource Development

Human Recombinant DNA Library

Larry L. Deaven, Jon L. Longmire, MaryKay McCormick, Deborah L. Grady, and Robert K. Moyzis
Center for Human Genome Studies, Life Sciences Division, Los Alamos National Laboratory,
Los Alamos, NM 87545
505/667-3114, FTS 843-3114, Fax 505/665-3024, FTS Fax 855-3024,
Internet: "moyzis@flovax.lanl.gov"

The goal of the National Laboratory Gene Library Project is the production of chromosome-specific human gene libraries and their distribution to the scientific community for (1) studies of the molecular biology of genes and chromosomes, (2) study and diagnosis of genetic disease, and (3) physical mapping of chromosomes. This is a cooperative project using the flow-sorting and molecular-cloning expertise at Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL).

The specific aim of the project's first phase was the production of complete digest libraries from each of the human chromosomal types purified by flow sorting; the average insert size expected was about 4 kb. The bacteriophage lambda vector used was Charon 21A, which has both *Eco*R I and *Hind* III insertion sites accommodating human DNA fragments of 0 to 9.1 kb. Each laboratory produced a complete set of chromosome-specific libraries—LANL with *Eco*R I and LLNL with *Hind* III. The small-insert libraries are deposited in a repository at the American *Type Culture* Collection, Rockville, Maryland; more than 3200 aliquots representing over 600 chromosome-specific libraries have been distributed to over 750 laboratories worldwide.

The second phase of the project is now under way. It involves construction of partial-digest libraries with larger inserts in more advanced, recently developed lambda vectors (9 to 23 kb) and in cosmid vectors (33 to 46 kb). These large-insert libraries have characteristics that are better suited to basic studies of gene structure and function, organization of genes on chromosomes, and ordering of cloned sequences. The second-phase strategy is to divide responsibility for the genome between the two laboratories, with LLNL cloning chromosomes 1, 2, 3, 7, 9, 12, 18, 19, 21, 22, and Y, and LANL cloning chromosomes 4, 5, 6, 8, 10, 11, 13, 14, 15, 16, 17, 20, and X. In this way, each chromosomal type will be cloned into both lambda and cosmid vectors. Cosmid libraries have been constructed and arrayed in microtiter plates at LANL for chromosomes 5, 6, 8, 11, 13, 16, and 17 in Charon 40 and s*Cos*1. Libraries have also been constructed in M13 or bluescript vectors for chromosomes 5, 7, and 14. They provide a rich source of sequence tagged site markers for the insert selection for specific chromosomes in genomic yeast artificial chromosome (YAC) libraries. We have also constructed a library for chromosome 21 in the YAC vectors pJS97 and pHS98.

Gene Libraries for Each Human Chromosome: Construction and Distribution

Pieter J. de Jong, Barbara Trask, Ger J. van den Engh, and Anthony V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory,
Livermore, CA 94550
510/423-8145, FTS 543-8145, Fax 510/423-3608, FTS Fax 543-3608, Internet: "pieter@pcr.llnl.gov"

The goal of the National Laboratory Gene Library Project is to produce chromosome-specific human gene libraries and distribute them to the scientific community for studies of the molecular biology of genes and chromosomes, study and diagnosis of genetic diseases, and physical mapping of chromosomes. This cooperative project employs the flow-sorting and molecular-cloning expertise at Lawrence Livermore National Laboratory (LLNL) and Los Alamos National Laboratory (LANL).

The specific aim of the project's first phase was the production of complete digest libraries from each of the human chromosomal types purified by flow sorting. Each laboratory produced a complete set of small-insert, chromosome-specific libraries in the lambda vector Charon 21A, with LANL using *Eco*R I and LLNL using *Hind* III. These libraries were deposited at American *Type Culture*

Collection in Rockville, Maryland, which has distributed over 3200 aliquots representing over 600 chromosome-specific libraries to more than 750 laboratories worldwide. All LLNL libraries have been subcloned into the Bluescribe plasmid vector, which facilitates their use as DNA probes and permits the preparation of RNA end-probes.

The project's second phase, now under way, involves the construction of libraries with large inserts in lambda replacement vectors and in cosmid vectors. These large-insert libraries are better suited to studies of gene structure and function, organization of genes on chromosomes, and ordering of cloned sequences. The LLNL group is cloning chromosomal types 1, 2, 3, 7, 9, 12, 18, 19, 21, 22, and Y, and LANL researchers are cloning the complementary set of chromosomes (4, 5, 6, 8, 10, 13, 14, 15, 16, 17, 20, and Y). At LLNL we are using the lambda vector Charon 40 (10- to 25-kb inserts) and the cosmid vectors Lawrist 5 or Lawrist 16 (34- to 46-kb inserts).

We constructed large lambda libraries from flow-sorted chromosomes 19, 22, and Y and Lawrist libraries from chromosomes 9, 12, 19, 21, 22, and Y. In addition, we prepared a large cosmid library from chromosome 22 using an F-plasmid–derived vector (pFOS1), in collaboration with H. Shizuya and M. Simon from the California Institute of Technology, Pasadena.

Molecular Cytogenetics and Computer-Assisted Microscopy

Joe Gray, Dan Pinkel, Wen-Lin Kuo, Damir Sudar, and Don Peters
Department of Laboratory Medicine, Division of Molecular Cytometry, University of California, San Francisco, CA 94143-0808
415/476-3461, Fax 415/476-8218, Internet: "gray@lcaquips.ucsf.edu"

The goal of this project is to improve speed and accuracy in detecting and characterizing cells that carry genetic aberrations by developing fluorescence in situ hybridization (FISH) with physically mapped probes and computer-assisted fluorescence microscopy. To accomplish this goal, we will develop a set of nucleic acid markers distributed along each chromosome for use as probes for FISH. The probes will be selected to optimize computer-assisted analysis of metaphase spreads and interphase nuclei and will contain elements targeted to the rapidly expanding set of genetic loci implicated in human disease. The design and production of such a resource contributes to the human health objectives of the human genome effort.

Key features of the project are (1) Primary probes distributed at 2- to 4-Mb intervals along each chromosome for general localization of aberrations. Secondary reagents composed of more closely spaced or overlapping probes may be added as they are developed by the Human Genome Project. This will permit more precise definition of an aberration that has been roughly localized using the primary reagent. Initial work will be concentrated on the primary reagent since this reagent can be assembled now with relatively modest effort. (2) Differential probe labeling and multicolor analysis instrumentation to allow simultaneous analysis at many loci. (3) Probe distribution selected to facilitate automated analysis. (4) Probe target size optimized for metaphase and interphase analysis (i.e., ~100 to 400 kb in extent) so the hybridization efficiency is high and the interphase hybridization signal is tightly confined. (5) Probes to loci known to be involved in genetic disease included in the set (e.g., probes to 21q22.3, Rb1, erbB-2, c-myc, BCR, and ABL). (6) Probes selected, when possible, to contain polymorphic regions so they can be tied to the genetic map and used in loss-of-heterozygosity studies or in linkage analysis.

Physically mapped probes will be obtained from DOE and NIH mapping projects when possible. However, many chromosomes are not likely to be well populated by mapped probes in the near future. We propose to cover these regions by multiplex mapping, using probes from chromosome-specific cosmid or yeast artificial chromosome libraries. Each multiplex hybridization maps one clone from each of several libraries. These clones are grown, labeled, and hybridized as a unit. Thus most metaphase spreads will show a single hybridization domain on each of the target chromosomes. The probe locations will be mapped using semiautomated fractional-length analysis.

We have already automated much of this process including multicolor image acquisition, calculation of the chromosome axis, cosmid location, and fractional-length calculation. Addition of automated metaphase findings should yield a system capable of analyzing several multiplex hybridizations per hour, yielding a throughput of ~100 cosmids/d.

Computer-assisted microscopy will be developed to support two aspects of applying the finished probe sets to biological and clinical problems: (1) Automated analysis of metaphase chromosomes for aberrations. Proper selection of the hybridization probes should simplify automated analysis by allowing development of a banding pattern with maximum discrimination between chromosomes of similar size and shape. (2) Interphase cytogenetic analysis. Areas to be developed include hybridization domain enumeration to facilitate aneuploidy detection, quantification of gene amplification, and detection of gene loss; detection of rare, genetically aberrant cells (e.g., residual leukemic cells or metastatic cells in the peripheral blood or bone marrow; and correlated analysis of cellular phenotype (e.g., immunophenotype) and genotype (e.g., FISH signal).

An Improved Method for Producing Radiation Hybrids Applied to Human Chromosome 19

Cynthia L. Jackson and Hon Fong L. Mark
Rhode Island Hospital and Brown University, Providence, RI 02903
401/277-4370, Fax 401/277-8514

Our long-term goal is the development and utilization of radiation-hybrid cell lines containing overlapping regions of human chromosome 19 marked with a retroviral vector. The vector contains features designed to facilitate gene mapping, such as a new dominant selection system, sites for rare-cutter restriction enzymes, and a tRNA suppressor gene that allows easy cloning of the insertion site. A monochromosomal microcell hybrid containing a marked human chromosome 19 was isolated first. The retroviral vector insertion site in human chromosome 19 was cloned and localized to the 19p13-q12 region.

Radiation hybrids were produced at doses of 1000 to 8000 rads. Because of the selectable marker, these hybrids should have a stable fragment of chromosome 19 containing the chromosome region surrounding the inserted marker. These hybrids will be fully characterized by several methods, including in situ hybridization, polymerase chain reaction, and Southern analysis. Additional retroviral vectors will be used to obtain several hybrids containing chromosome 19 marked in different regions. The location of each marker will be identified; regions of interest will be isolated from rodent-human hybrids by irradiation and fusion. An overlapping set of radiation hybrids, each containing only a small fragment of the original chromosome 19 will be produced and their chromosome 19 regions will be characterized and analyzed for markers. These hybrids will aid in physically mapping chromosome 19 and in isolating specific sequences.

Chromosome Region-Specific Libraries for Human Genome Analysis

Fa-Ten Kao and Jing-Wei Yu*
Eleanor Roosevelt Institute and Department of Biochemistry, Biophysics, and Genetics, University of Colorado Health Sciences Center; Denver, CO 80262
303/333-4515, Fax 303/333-8423
*Eleanor Roosevelt Institute, Denver, CO 80262

Molecular analysis and fine structure mapping of the human genome require isolation of large numbers of DNA probes from defined chromosomal regions. A direct approach is to remove the region of interest by chromosome microdissection and to clone the dissected DNA sequences by polymerase chain reaction (PCR)-assisted microcloning. We plan to develop these techniques and apply them to the construction of region-specific libraries to facilitate physical mapping of the human genome.

Specific objectives are (1) to develop PCR microcloning using universally amplified DNA sequences from microdissected chromosome fragments by the linker-adaptor method; (2) to develop PCR microcloning by the terminal deoxynucleotidyl transferase method; and (3) to use this technology to construct region-specific libraries for human chromosomes 2 and 5 and make the libraries available to the human genome centers at Lawrence Livermore National Laboratory and Los Alamos National Laboratory for physical mapping. Our longer-term goal is the application of this technology to other human chromosomes or important chromosomal regions to construct region-specific libraries, enabling a more efficient analysis of the human genome.

Partial Bibliography

F.-T. Kao and J.-W. Yu, "Chromosome Microdissection and Cloning in Human Genome and Genetic Disease Analysis," *Proc. Natl. Acad. Sci. USA* **88**, 1844–48 (1991).

Human cDNA Mapping Using Fluorescence In Situ Hybridization

Julie R. Korenberg
Department of Pediatrics, Medical Genetics, Cedars-Sinai Medical Center, University of California, Los Angeles, CA 90048
310/855-6451, Fax 310/967-0112

The ultimate goal of this research is to create a cDNA map of the human genome. Mapping will be approached using the techniques of high-resolution fluorescence in situ hybridization (FISH).

The human genome is estimated to consist of some 100,000 genes, up to 30,000 of which may be expressed in the brain. One approach to the rapid identification of these genes[1] has resulted in the cloning and partial sequencing of cDNAs for 337 new genes, 48 similar to genes in other organisms and 30 known genes. The application of these expressed sequence tags to genome mapping and as genetic disease markers is limited by the lack of high-resolution information about their positions in the genome.

On the basis of our current technology for high-resolution mapping of cDNA clones, we will initiate a pilot project to test the use of FISH in the rapid mapping of cDNAs to single human chromosome bands.

To discover the inherent limitations to be overcome in a large-scale cDNA mapping project, it is necessary to examine a diversity of representative sources of cDNAs; identify desirable and undesirable cDNA characteristics; and determine the resolution, accuracy, sensitivity, and throughput of current methodology.

[1] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and J. C. Venter, "Complementary DNA Sequences: Expressed Sequence Tags and the Human Genome Project," *Science* **252**, 1651–56 (1991).

Abstracts: Resource Development

Construction of a Human Genome Library Composed of Multimegabase Acentric Chromosome Fragments

Michael J. Lane, Peter Hahn,* and **John Hozier****
Departments of Medicine and Microbiology and *Department of Radiology, State University of New York-Health Science Center at Syracuse, Syracuse, NY 13210
315/464-5446, Fax 315/464-8255
**Department of Medical Genetics, Florida Institute of Technology, Melbourne, FL 32901

The objective of this project is the creation of a human genome library composed of 5-Mb elements carried in mouse EMT-6 cells. In addition, new methods will be developed to detect overlaps between the elements. This project will involve infection or transfection of a human cell line with a selectable marker, selection of infected human cells, and lethal irradiation of a polyclonal population of the cells. Irradiation will be followed by polyethylene glycol-mediated fusion to a recipient mouse cell line with the capacity to rescue introduced DNA as circular double-minute chromosomes.

This type of library will significantly reduce (1) the time necessary to isolate genes that have been genetically mapped to less than 5 cM apart and (2) the number of independent elements required to construct a physical map of the human genome.

The cDNA Genome: Strategies and Results with Particular Reference to Human Chromosome 19

Gregory G. Lennon, Harvey Mohrenweiser, Ger van den Engh, and **Anthony V. Carrano**
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94551
510/422-5711, FTS 532-4711, Fax 510/423-3608, FTS Fax 543-3608, Internet: "greg@mendel.llnl.gov"

The goals of this project are to improve methods for identifying and analyzing transcribed regions of the genome and applying these methods to isolating and mapping chromosome-specific cDNAa. Reference cDNA libraries are being created and arrayed robotically at high density (25 to 60 clones per cm^2) on hybridization membranes. These filters are being used to identify cDNAs of interest and to study patterns of expression. Specific cDNAa from these and other libraries [Sherman Weissman (Yale University)] are being sequenced for STS production and for comparison to sequence databases. Additional methods of characterizing sequence, such as sequencing by hybridization, are being developed to further the identification of interesting cDNAs and associated diagnostic polymorphisms. For mapping purposes, work is progressing on developing rapid methods of determining the chromosomal localization of unsequenced cDNAs.

The set of genomic cosmid clones comprising the current tiling path of chromosome 19 is also being arrayed and filters produced. This set is being used in the cDNA effort in three ways: first, the location of HTF islands is being determined and then the HTF-positive cosmids used to screen for cDNAs; second, the tiling-path set forms a basis for sequencing and selecting chromosome-specific transcribed sequences using both direct selection and exon-trapping techniques; and third, 5′·and 3′ cDNA ends are being used separately as hybridization probes on tiling-path filters to link previously unlinked contigs together. This research, which is currently focused on human chromosome 19, integrates the genetic and physical maps as it aids the search for new genes and genetic markers.

Identification and Characterization of Expressed Chromosomal Sequences

Christopher H. Martin,[†] Carol A. Mayeda, and **Michael J. Palazzolo**[†]
Human Genome Center, Cell and Molecular Biology Division, Lawrence Berkeley Laboratory,
Berkeley, CA 94720
[†]510/486-5909, FTS 451-5909, Fax 510/486-6816, FTS Fax 451-6816, Internet: "chrism@lbl.gov" or "michaelp@lbl.gov"

Our major effort in the past has been to isolate and characterize a large fraction of cDNA molecules that correspond to expressed genomic sequences. When we began these experiments, a number of technical problems had to be solved to accomplish these goals, including the ability to (1) clone large numbers of non-cross-hybridizing cDNAs, (2) isolate full-length cDNAs, (3) rapidly sequence these cDNA molecules, (4) generate reagents that could be used to determine the positional and temporal restraints on the expression of the corresponding genes, and (5) map the cDNAs to the corresponding position in the genome.

Over the past few years we have developed procedures that overcome these technical obstacles, and we are using these methods to begin a large-scale molecular characterization of the expressed sequences in the nervous system. We have already cloned almost 2500 different cDNA molecules. Future experiments would focus on elucidating the DNA sequence of these molecules and identifying their genomic position.

Partial Bibliography

M. J. Palazzolo, D. R. Hyde, K. Vijay Raghavan, K. Mecklenburg, S. Benzer, and E. Meyerowitz, "Use of a New Strategy to Isolate and Characterize 436 Drosophila cDNA Clones Corresponding to RNAs Detected in Adult Heads but not in Early Embryos," *Neuron* **3**, 527–39 (1989).

Application of Flow-Sorted Chromosomes to the Construction of Human Chromosome–Specific Yeast Artificial Chromosomes

MaryKay McCormick, Linda Meincke, Mary Campbell, Evelyn Campbell, John Fawcett, **Larry Deaven**,* and **Robert Moyzis***
Life Sciences Division and *Center for Human Genome Studies, Los Alamos National Laboratory,
Los Alamos, NM 87545
505/665-4438, FTS 855-4438, Fax 505/665-3024, FTS Fax 855-3024, Internet: "mkm@flovax.lanl.gov"

Flow-sorted human chromosomes have been used as a source of DNA for construction of chromosome-specific libraries in cosmid and lambda vectors. We have investigated the use of flow-sorted chromosomes as a source of DNA for construction of chromosome-specific yeast artificial chromosomes (YACs) that accommodate inserts 10 times larger than those cloned in cosmid or lambda vectors. WAV-17, a somatic cell hybrid containing chromosome 21 as the only human chromosome, has been used to sort chromosome 21 at purities of 85 to 95%. The size of the DNA isolated from the chromosomes was analyzed by pulsed-field gel electrophoresis and then used to construct YACs. About 5 x 10^6 chromosomes 21 (~500 ng) were sorted on an Epics V flow cytometer (Coulter Corporation) into a 5-mL centrifuge tube coated with 400 μL of 1.0% low-melting agarose (LMA) and concentrated by centrifugation at 1500 g for 30 min. The chromosome-containing LMA was recovered from the tube and washed in buffer containing proteinase K to remove chromosomal proteins before DNA restriction. The chromosomal DNA was digested to completion using restriction endonucleases that have infrequent recognition sites (*Cla* I, *Eag* I, of *Not* I and *Nhe* I together). DNA isolated from the sorted chromosomes was cloned entirely in the presence of the LMA, and average size clones of 200 kb have been obtained at efficiencies of 600 to 2500 cfu/μg of DNA sorted. This size and efficiency is sufficient to construct human chromosome–specific libraries in YACs, which will greatly facilitate the construction of physical maps of human chromosomes. Libraries have been constructed for chromosomes 16 and 21.

Chimera-Free, High-Copy-Number YAC Libraries and Efficient Methods of Analysis

Donald T. Moir
Collaborative Research, Inc., Waltham, MA 02154
617/275-0004, Fax 617/891-5062

The overall goal of this research is to define procedures for generating chimera-free yeast artificial chromosome (YAC) libraries in new vectors that permit facile screening and chromosomal walking. To accomplish this goal, our specific aims are to (1) define YAC cloning methods that minimize the formation of molecules with "chimeras" or multiple inserts; (2) develop new vectors and methods that facilitate library screening and manipulation of insert DNA; (3) incorporate these advances and build a large-insert YAC library of human DNA equivalent to five genomes; and (4) prepare high-density screening blots from the amplified YAC clones in this library for distribution to other laboratories.

The results of this research will reveal the role of in vitro co-ligation and in vivo recombination mechanisms in generating chimeric YACs and will provide methods for minimizing chimera production. Two differently marked YACs or YAC-ligation reactions will be mixed and used to transform wild-type and recombination-deficient host strains carrying auxotrophies for the markers on the YAC arms. Analysis of the marked arms of the resulting YACs by auxotrophic phenotype and Southern blots will indicate the frequency of YACs with recombinant arms due to recombinational mechanisms and the effect of recombination-deficient hosts on chimera formation. Co-ligation frequencies will be assessed and reduced by using an excess of oligonucleotide linker/adaptors to saturate the human DNA insert ends before ligation to vector arms.

The YAC vector will be modified to provide higher copy numbers of YACs than now possible; permit specific tagging and capturing of YAC vector-insert ends, whole inserts, or whole YACs; and permit specific radiation-sensitive gene (RAD) independent introduction of new DNA segments into existing YACs. Selectable markers whose copy number can be selected in varying degrees and whose products are not needed for DNA replication will be used to generate clones with over 20 YAC copies per cell. Triple-helix formation by short polypyrimidine third strands will be exploited to permit specific binding of oligonucleotide tags to each end of the YAC cloning site. These tags will be used to capture YACs and inserts for probe generation and sequencing. In addition, a synthetic yeast site-specific recombinase (FLP) recombination target (FRT) sequence will be added to the vector to permit convenient retrofitting of existing YACs.

A YAC library equivalent to five genomes will be made in the new vector by methods experimentally shown to eliminate chimera generation. High-density screening blots from amplified microcolonies will be prepared by means of a Biomek 1000 automated laboratory workstation. The entire library will be arrayed on only 20 membranes, each the size of a microtiter dish, for screening by hybridization with labeled polymerase chain reaction products or vector-insert end fragments. Each membrane will be able to be reprobed ten or more times. Multiple copies of these screening blots will be prepared and shipped to numerous researchers to test the concept of making a YAC library readily accessible to the entire research community. Multiple copies of the library itself will be stored on 20 high-density membranes soaked in 20% glycerol and maintained at -70°C. These copies will be available for shipment to research centers or clone repositories such as ATCC that will serve as centers for distributing positive clones to investigators using the screening blots.

In summary, the approaches described here will permit construction of YAC libraries, essentially chimera-free and with amplifiable YACs that can be tagged and captured for convenient analysis or readily retrofitted to incorporate further improvements in vector technology.

Genome Organization and Function

Robert K. Moyzis, Julie Meyne, and Robert L. Ratliff
Center for Human Genome Studies, Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
505/667-3912, FTS 843-3912, Fax 505/665-3024, FTS Fax 855-3024, Internet: "moyzis@flovax.lanl.gov"

The mechanisms for organizing the mammalian genome and the genetic and nonmutational alterations accompanying abnormal phenotypic change are important aspects in defining the effects of energy-related technologies. Determining the genetic variability in these mechanisms provides a rational basis for establishing thresholds for toxic substances, for making valid cross-species extrapolations, and eventually for identifying individuals at risk.

The ultimate objective of this project is to determine the molecular mechanisms by which higher organisms organize and express their genetic information. Applications of these basic investigations will include the development of novel approaches for detecting human genetic diseases and for measuring the effects of low-level ionizing radiation and carcinogen exposure. A combination of biochemical, biophysical, and recombinant DNA techniques are being used to identify, isolate, and determine the roles of DNA sequences involved in long-range genomic order.

Current efforts are focused on determining the organization and function of human repetitive DNA sequences. Major achievements in the last year included (1) the demonstration that the human telomere sequence $(TTAGGG)_n$, identified and isolated by our laboratory, is the vertebrate telomere and (2) the cloning of 100- to 250-kb human telomere fragments as yeast artificial chromosomes. Future studies will be directed toward the definition and isolation of other "functional" repetitive DNA regions such as centromeres.

Isolation of cDNAs from the Human X Chromosome and Derivation of Related STSs

David L. Nelson
Institute for Molecular Genetics, Baylor College of Medicine, Houston, TX 77030
713/798-6552, Fax 713/798-6521, Internet: "nelson@mbir.bcm.tmc.edu"

The overall aim of this proposal is the isolation of expressed sequences from the human X chromosome, their conversion to sequence tagged sites (STSs), and association with large-insert (YAC) clones. These goals will be accomplished through several strategies, and the collection of these data will provide STSs for integration into the large-scale physical map of the chromosome currently under construction. The collection of cDNAs and their sequences from the X chromosome constitutes an important goal of genome analysis and will provide a rich source of information. The methods proposed to isolate and characterize genes from this region are general in nature and will be applicable to other regions of the human genome.

The methods proposed are the following:

(1) Sublocalization of cDNA-based STSs with X-chromosome localization, provided by Mihael Polymeropoulos and Craig Venter at NIH, and identification of associated YAC clones.

(2) Isolation of human-specific cDNA clones of heterogeneous nuclear RNA (hnRNA) through the use of somatic cell hybrids retaining the X chromosome as their only human material. This procedure will use oligonucleotides specific to the *Alu*-repeat element as a primer for cDNA synthesis or polymerase chain reaction amplification.

(3) Identification and isolation of expressed regions from YAC clones by hybrid selection of cDNAs and introduction of YAC DNA sequences into cultured mouse L cells, followed by analyzing the resulting RNA for the presence of human-specific expression or through exon-trapping methods.

Sequences will be characterized by regional assignment, cross-species hybridization to identify conservation, and DNA sequence-homology analysis. This catalog of genes from the X chromosome will provide significant insight into the organization of the genome, evolutionary relationships of X chromosomes among mammalian species, and the etiology of many of the hundreds of X-linked human genetic diseases.

Multiplex Mapping of Human cDNAs

William C. Nierman, **Donna R. Maglott**, and Scott Durkin
American *Type Culture* Collection, Rockville, MD 20852-1776
301/231-5559, Fax 301/770-1848

Our objective is to determine the utility of multiplexing polymerase chain reactions (PCRs) for mapping cDNA sequences to chromosomes and subchromosomal regions. PCR primers will be designed using sequence data generated by J. Craig Venter's laboratory at the NIH National Institute of Neurological Disorders and Stroke, where cDNA clones from human brain libraries have been partially sequenced.[1] We will then (1) test the primers for multiplexed specific amplification from human genomic DNA; (2) combine useful primer pairs for multiplexed PCR amplification from a somatic-cell hybrid cell mapping panel; (3) determine the presence or absence of the specific amplification products from each cell line DNA by electrophoretic analysis using the Applied Biosystems, Inc. (ABI) 373A automated sequencer; and (4) analyze the pattern of amplification results from the hybrid panel to identify the chromosomal origin of the cDNA sequence. A high throughput can be attained with a minimum allocation of resources by multiplexing the amplification reactions and analyzing the reaction products on the ABI sequencer.

In addition to determining chromosomal assignments for expressed sequence tags, this project will also provide primer sequence data for subsequent subchromosomal localizations, and generate a broad data set from which to evaluate strategies for identifying functional primer sequences from cDNA sequence data.

[1] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, O. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and J. C. Venter, "Automated Partial Sequencing of cDNAs: Expressed Sequence Tags and the Human Genome," *Science* **252**, 1651–56 (1991).

Chromosomal Localization of Brain cDNAs

Mihael H. Polymeropoulos, Hong Xiao, and Carl R. Merrill
Neuroscience Center at St. Elizabeth's Hospital, National Institute of Mental Health, Washington, DC 20032
202/373-6077, Fax 202/373-6087

Of the estimated 100,000 human genes, 30,000 are expressed in the human brain; some 20,000 of these are believed to be brain-specific. About 2000 genes have been mapped so far, reflecting a small proportion of the total number of expressed genes. The advent of automated sequencing has enabled the partial sequencing of several hundred expressed brain genes (cDNAs). The long-term goal of this project is to establish the chromosomal location for 1600 of these cDNAs and to develop strategies that will increase the throughput for cDNA mapping.

In our approach, cDNA chromosomal locations will be established by developing a sequence tagged site (STS) for each sequenced brain cDNA; the chromosomal location for each STS will be determined by analysis for segregation of the amplified products in two human-rodent somatic-cell hybrid panels.

This approach can be integrated fully in ongoing mapping efforts by providing easily identifiable landmarks based on DNA sequence. Furthermore, the mapping of brain cDNAs will facilitate the identification of genes involved in inherited central nervous system disorders by providing candidate genes in areas where linkage has already been established or is being pursued.

A Bacteriophage T4 In Vitro DNA Packaging System To Clone Long DNA Molecules

Venigalla B. Rao, Vishakha Thakhar, and Lindsay W. Black*
Department of Biology, The Catholic University of America, Washington, DC 20064
202/319-5271, Fax 202/319-5721
*Department of Biological Chemistry, University of Maryland Medical School, Baltimore, MD 21201

Bacteriophage T4 packages about 170 kb of its DNA that includes 2% terminal redundancy in a strictly headful manner. We recently purified the various packaging components of T4 and developed an in vitro DNA packaging system. This system is being used to clone about 150 kb of foreign DNA and to construct genomic libraries. We have cloned the 95-kb *tryp* and the 100-kb *nad Not* I fragments of *Escherichia coli* into the P1-1ox vectors. (The P1-ox vectors were developed by Nat Sternberg at DuPont.) We have also generated genomic libraries of *E. coli* DNA from partial *Bam*H I digests. The clones obtained were 40 to 135 kb. We are now addressing various aspects of this system for construction of 150-kb human genomic libraries, including isolation of large quantities of 150-kb intact DNA fragments and improvement of cloning efficiency.

Development of an Embryonic Stem (ES) Cell–Based System for the In Vitro Generation of Germline Deletion Complexes Throughout the Mouse Genome

Eugene Rinchik and **Richard Woychik**
Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-8077
615/574-0953, FTS 624-0953, Fax 615/574-1283, FTS Fax 624-1283

The physical and genetic information accumulated through analyses of the large deletion complexes in mouse chromosomes 2, 4, 7, and 9 (involving the *a, b, c, p, d,* and *se* loci) has underscored the value of these and other types of chromosome rearrangements as tools for defining and merging fine-structure physical and functional maps. The phenotypes observed when deletions are made homozygous provide a measure for constructing first-stage functional maps of megabase regions. Importantly, the syndromes manifested in such situations may provide the only true models for certain types of syndromic genetic disease often associated with large-scale cytogenetic abnormalities in humans. Deletion complexes are also useful in high-efficiency "saturation-mutagenesis" experiments designed to generate fine-structure "point"-mutation maps of megabase stretches of chromosomes that can subsequently be correlated with detailed molecular/physical maps. However, available panels of deletion mutations represent the accumulated products of 30 years of germ-cell radiation-mutagenesis experiments and in aggregate span only 2 to 3% of the genome surrounding the loci of the morphological specific-locus test. The potential relatively high cost of generating these types of mutations on a genome-wide scale solely by classical radiation- and chemical-mutagenesis techniques has thus far limited their availability as tools for whole-genome analysis.

This project explores the possibility of developing simple mutagenesis procedures in embryonic stem (ES) cells that could be used for efficient in vitro generation of deletion complexes throughout the mouse genome. ES cells in culture that has been stably transfected with DNA constructs containing both positive and negative selectable markers will be exposed to X rays at levels that will produce deletions in the DNA of the host chromosomes. Radiation-induced deletions that cover the integration site containing the transfected construct will remove the negative selectable marker, allowing that clone to grow in an appropriate selective medium. Since each independent deletion event is likely to be a different-sized deletion, multiple clones from a given insertion will constitute a deletion complex. Thus, cells from a number of different ES-cell clones can then be used for an in vitro physical-mapping reagent. Importantly, they can also be used to generate chimeras for introducing members of a given in vitro deletion complex into the mouse germline to create breeding stocks that carry heterozygous deletions of any given genomic region.

For more information on this subject, refer to the abstracts by L. Stubbs and E. Rinchik, E. Rinchik and R. Woychik, and E. Uberbacher and R. Mural.

Development of a Large-Scale Targeted Mutagenesis Program for Determining Organismal Function of Specific Human Genes

Mike Mucenski, Bill Lee, **Eugene Rinchik,**[†] and **Richard Woychik**
Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-8077
[†]615/574-0953, FTS 624-0953, Fax 615/574-1283, FTS Fax 624-1283

Cloning and mapping mouse genomic sequences containing regions homologous to human cDNAs are important first steps in understanding the function of genes in the human genome. This information by itself, however, is of limited value for establishing the function of any given gene in the context of the whole organism. Targeted mutagenesis of specific genes by homologous recombination in embryonic stem (ES) cells is now possible in the mouse. This capability is likely to be one of the most powerful tools for beginning to understand the function of any genes identified as cDNA clones in the course of physically mapping the human genome. At present, however, it is unclear whether the existing technology for generating targeted mutations can be implemented on a genome-wide scale. Therefore, this project will determine the feasibility of producing targeted mutations at ORNL on a scale as large as our production of transgenic mice with the pronuclear microinjection procedure. All parameters related to generating null mutations by homologous recombination in ES cells will be analyzed and streamlined for future genome-wide application of this technology. If the feasibility of producing targeted mutations on an expanded scale can be demonstrated, "knockout" (null) mutations in mouse genes corresponding to selected human cDNAs can then be produced in an efficient and cost-effective manner.

For more information on this subject, refer to the abstracts by L. Stubbs and E. Rinchik, E. Rinchik and R. Woychik, and E. Uberbacher and R. Mural.

Sequence Tagged Sites for Human Chromosome 19 cDNAs

Michael J. Siciliano and Anthony V. Carrano[*]
Department of Molecular Genetics, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030
713/792-2910, Fax 713/794-4394
[*]Lawrence Livermore National Laboratory, Livermore, CA 94550

In this project we will implement a procedure for cloning human chromosome–specific expressed-gene sequences in order to develop 900 sequence tagged sites (STSs) for human chromosome 19 expressed genes; successful strategies and technologies should be applicable to other chromosomes. We plan to use heterogeneous (unspliced) RNA transcripts (hnRNA) to construct an hn-cDNA library from somatic cell hybrids (human-Chinese hamster) monochromosomal for human chromosome 19. The primers used for cDNA synthesis from the hnRNA will recognize 5′ splice junction sites that will enrich for the isolation of expressed exons. We will isolate hn-cDNA library clones containing human inserts from the library and verify them for human origin and regional assignment on Southern blots containing appropriate somatic cell hybrids.

Lawrence Livermore National Laboratory (LLNL) will sequence up to 400 bp of these human inserts. Sequence analysis programs for each insert will determine whether or not a similar fragment has been sequenced previously, estimate the probable exon segment of the unique sequence region, and evaluate the presence in the fragment of *Alu*-linked poly-A tandem repeats associated with high levels of polymorphism. To make the segment an STS, we will synthesize polymerase chain reaction (PCR) primers for the probable exon segment of the unique sequence region of each fragment. PCR primers will also verify the STS position on chromosome 19 by concordant segregation analysis on hybrid clone panels by PCR and will be available to LLNL for STS assignment to the chromosome physical map being developed there. The same PCR primers will be used to determine the STS expression in RNA messages from a series of tissue types.

Partial Bibliography

P. Liu, R. Legerski, and M. J. Siciliano, "Isolation of Human Transcribed Sequences from Human-Rodent Somatic Cell Hybrids," *Science* **246**, 813–15 (1989).

cDNA/STS Map of the Human Genome: Methods Development and Applications Using Brain cDNAs

James M. Sikela, Akbar S. Khan, Arto K. Orpana, Andrea S. Wilcox, Janet A. Hopkins, and Tamara J. Stevens
Department of Pharmacology, University of Colorado Health Sciences Center, Denver, CO 80262
303/270-8637, Fax 303/270-7097, Internet: "sikela_j%maui@vaxf.colorado.edu"

The goal of this research project is to contribute to the generation of the detailed cDNA/sequence tagged site (STS) expression map of the human genome by converting human brain cDNAs to STSs. Ideally, such a map would be composed of a collection of unique cDNAs, each in the form of an STS with a known chromosomal location and therefore easily accessible to interested investigators. In addition, each cDNA would be stored in a microtiter plate well and have its own address. Toward this end, this project will involve the selection of large numbers of individual cDNAs from directionally cloned human brain libraries that have been prescreened to enrich for unique cDNAs. Partial DNA sequence from each cDNA will be obtained by single-pass, high-throughput automated DNA sequencing using a single sequencing primer for all clones. To identify potential functional relatedness to other genes and to minimize duplication of effort, sequence information for each cDNA will be compared to sequences in internal and external DNA and protein databases. For STS generation, polymerase chain reaction (PCR) primers designed from 3′ untranslated regions will be used to assign cDNAs to specific chromosomes and chromosomal regions by PCR analysis of somatic cell and, when available, radiation hybrid DNAs. Improvements in the speed and accuracy of these strategies will be investigated.

Partial Bibliography

A. S. Khan, A. S. Wilcox, J. A. Hopkins, and J. M. Sikela, "Efficient Double Stranded Sequencing of cDNA Clones Containing Long Poly(A) Tails Using Anchored Poly(dT) Primers," *Nucleic Acids Res.* **19**, 1715 (1991).

A. S. Wilcox, A. S. Khan, J. A. Hopkins, and J. M. Sikela, "Use of 3′ Untranslated Sequences of Human cDNAs for Rapid Chromosome Assignment and Conversion to STSs: Implications for an Expression Map of the Genome," *Nucleic Acids Res.* **19**, 1837–43 (1991).

Abstracts: Resource Development

Chromosome-Specific cDNAs and Sequence Tagged Sites

Marcelo Bento Soares, Pierre Jelenc,* Stephen Brown, Maria de Fatima Bonaldo, and Agiris Efstratiadis*
Department of Psychiatry and *Department of Genetics and Development, Columbia University, New York, NY 10032
212/960-2313, Fax 212/795-5886

The goals of this project are to construct and normalize cDNA libraries from human tissues, develop methods for assigning cDNAs to particular chromosomes and chromosomal regions, and use the assigned cDNAs to generate sequence tagged sites (STSs). Our strategy involves the generation of cDNA libraries by using novel vectors (lafmids) that we have constructed and tested. After cDNA is cloned in lafmids, single-stranded versions of the libraries can be used for library normalization by a kinetic approach. Hybridization between normalized and chromosome-specific libraries can then be used to assign cDNAs to chromosomal regions. Both the assigned cDNAs and the corresponding exon-containing genomic clones will be used as sources of STSs.

Synthetic Endonucleases

Betsy M. Sutherland and Gary A. Epling*
Biology Department, Brookhaven National Laboratory, Upton NY 11973
516/282-3293 or -3380, FTS 666-3293, Fax 516/282-3407, FTS Fax 666-3407
*Chemistry Department, University of Connecticut, Storrs, CT 06268

Synthetic endonucleases can be constructed to cleave at regions of functional importance. Constructing efficient synthetic endonucleases requires knowledge of optimum strategies for labeling, determining binding properties of labeled molecules, and studying cleavage properties of cleaving moieties, alone and when conjugated to the protein that confers binding specificity. We have developed new quantitative nonradioactive assays that allow determination of the extent and specificity of binding of T7 RNA polymerase labeled with Rose-Bengal (RB/RNAP). These assays also yield information about levels and specificity of photocleavage by this synthetic endonuclease. We found that standard conditions for binding T7 RNA polymerase to promoter-containing DNA required large excesses of RB/RNAP per promoter and allowed polymerase binding to DNA without promoters. We therefore devised new binding conditions that resulted in both higher levels of binding and greater specificity for promotor-containing DNA molecules.

We studied the effect of different levels of Rose-Bengal polymerase labeling on binding and cleavage. Generally, low levels of Rose-Bengal labeling (≤~5 RB/RNAP) allow the labeled polymerase to bind at approximately the same levels as the unlabeled polymerase. At higher levels of labeling (~10 to 15 RB/RNAP), binding RB/RNAP to DNA containing a T7 promoter is retained, but nonspecific binding increases. We have thus chosen intermediate levels of polymerase labeling (~5 to 10 RB/RNAP) for cleavage of promoter-containing DNA. Under optimal conditions, the RB/RNAP cleaves DNA containing a T7 promoter but does not cleave a similar DNA lacking T7 promoter sequences.

Several strategies have been developed to optimize cleavage properties of the cleaving moiety. These include clarification of the chemical mechanism involved in the photocleavage, as well as modification of the cleaving moiety chemical structure to enhance the desired photoreactivity. Mechanistic studies suggest that photo-induced electron transfer, rather than singlet oxygen sensitization, is the key step to photo-nicking. New compounds that may improve the electron-transfer efficiency of RB are under development as potential cleaving moieties.

Partial Bibliography

B.·M. Sutherland, P. Y. Bennett, K. Conlon, G. A. Epling, and J. C. Sutherland, "Quantitation of Supercoiled DNA Cleavage in Non-Radioactive DNA: Application to Ionizing Radiation and Synthetic Endonuclease Cleavage," *Anal. Biochem.* (submitted 1991).

Expressed Sequence Tags (ESTs) from Human Brain cDNAs for Genome Mapping

J. Craig Venter, Mark Adams, Mark Dubnick, Chris Fields, Jenny Kelley, Anthony Kerlavage, Ruben Moreno, and James Nagle
Section of Receptor Biochemistry and Molecular Biology, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892
301/496-8800, FTS 496-8800, Fax 301/480-8588, FTS Fax 496-8588, Internet: "jcventer@loglady.ninds.nih.gov"

Large-scale cDNA sequencing and mapping offer many unique opportunities for understanding the structure of the human genome and gene expression. The human brain is estimated to contain 20,000 to 30,000 unique mRNA species, which represent one-third to one-half of all human genes. Automated DNA sequencing technology has made possible the initial characterization of large numbers of cDNA clones.

This project focuses on construction and analysis of directionally cloned cDNA libraries and automated DNA sequencing of 3′ untranslated sequence of brain clones for mapping. Efforts will be made to screen the brain libraries before sequencing to eliminate redundant clones by a combination of approaches, including normalization, subtraction, and differential hybridization.

As the project progresses, we will continue to identify unique clones and hope to maintain at least 4000 new ESTs per year. Newly developed pools of clones for subtraction and hybridization screening will be made available to other researchers in the field. In addition, sequences will be given to any researcher who desires to perform physical or genetic mapping of ESTs.

In preliminary studies, sequence data were collected from over 600 human brain cDNA clones. These sequences have proven valuable in several ways: many showed significant similarity to genes from other organisms, providing a clue to the function of human genes; over 40 have been localized to chromosomes by using polymerase chain reaction screening of somatic-cell hybrid DNA panels, and several have been used to generate STSs for mapping human chromosomes. These sequences will be used eventually to confirm the coding region in genomic sequence and to further our understanding of the expressed gene complement of the brain.

Partial Bibliography

M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Morril, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. K. McCombie, and J. C. Venter, "Complementary DNA Sequencing: Expressed Sequence Tags and the Human Genome Project," *Science* **252**, 1651–56 (1991).

Development of Human Virus-Based Genomic Library of 150- to 200-kb Inserts

Jean-Michel H. Vos, Tian-Qiang Sun, and Sharon Michael
Department of Biochemistry and Biophysics and Lineberger Comprehensive Cancer Research Center, University of North Carolina, Chapel Hill, NC 27599-7295
919/966-6888, Fax 919/966-3015, Internet: "vos@med.unc.edu"

Libraries of 150- to 200-kb inserts of human genomic DNA would be a very useful resource in global mapping efforts. The objective of this project is to develop and demonstrate methods for rapidly cloning and propagating such large-size DNA inserts entirely in human cells and for using these fragments for gene isolation, gene mapping, and mutation detection. The proposed cloning strategy is based on the development of virus-mediated gene-transfer technology.

Our specific objectives are to generate a human genomic library that (1) covers the range of insert sizes suitable for mapping by sequence tagged sites (i.e., 100-kb DNA fragments); (2) is stably maintained in human cells for preservation of human genomic ordering and imprinting; and (3) is virus based for easy production and efficient shuttling between human cells. Using 150- to 200-kb average insert sizes cloned on a mini-Epstein-Barr virus (EBV) plasmid in human lymphoblastoid cells, we plan to pursue three specific goals:

1. Establishment of episomal human genomic libraries on mini-EBV vectors in human helper lymphoblastoid cell lines,
2. Production of human genomic libraries packaged as infectious EBV virions, and
3. Preparation of EBV-based infectious stocks of chromosome-specific human genomic libraries.

In establishing the library, the project will develop a valuable collection of reagents useful in mapping, isolating, sequencing, and assaying genes and other functional genomic elements.

Isolation of Specific Human Telomeric Clones by Homologous Recombination and YAC Rescue

Geoffrey Wahl and Linnea Brody
Gene Expression Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037
619/453-4100 Ext. 587, Fax 619/455-1349

A new strategy will be used for direct isolation of telomeric and subtelomeric regions from human and other mammalian chromosomes. The procedure involves homologous recombination to target a vector to telomere proximal regions. The vector is designed to enable the regions flanking both sides of the insertion site to be rescued directly as yeast artificial chromosome (YAC) clones. The vector is constructed with selectable markers to enable amplification of the region flanking the insertion site, and it contains specialized restriction sites that are cut infrequently in the mammalian genome. Both these features should facilitate the rescue of YACs up to and possibly exceeding 1 Mb and aid in developing long-range restriction maps of the targeted regions.

Initial proposed targets are the telomeric regions from chromosomes 19, 16, 5, and 2. Additional telomeric regions will be isolated as the cosmids or other clones derived from telomere proximal regions become available to provide the sequences required to prepare the targeting substrates. We anticipate that development of the telomere rescue strategy will greatly facilitate the goal of obtaining a complete map of the human and other genomes.

New Hosts and Vectors for Genome Cloning

Philip A. Youderian and Phillip Greener
California Institute for Biological Research, La Jolla, CA 92037
619/535-5471, Fax 619/535-5472

We are developing a novel *Escherichia coli* host/vector system for cloning large (250- to 450-kb) inserts of foreign DNAs to facilitate rapid mapping of complex genomes. This involves (1) constructing *E. coli* hosts with different combinations of mutations that prevent the rearrangement of DNA inserts with features of highly repetitive human DNAs, including direct and inverted repeats, Z-DNA sequences, (CA)n, and (A)n; (2) constructing novel macron vectors carrying large foreign DNA inserts as invertible segments of single-copy, amplifiable *E. coli* plasmids; and (3) further testing these systems by attempting to make an ordered-clone map of human chromosome arm 21q (about 40,000 kb).

Physical and Genetic Mapping

Mapping and Ordered Cloning of Human X Chromosome

C. Thomas Caskey and **David L. Nelson**[†]
Institute for Medical Genetics, Baylor College of Medicine, Houston, TX 77030
[†]713/798-4787 or -6552, Fax 713/798-7383 or -6521, Internet: "nelson@mbir.bcm.tmc.edu"

The overall goal of this project remains the isolation of the human X chromosome in overlapping cloned DNA segments. This will involve using the yeast artificial chromosome (YAC) vector system to produce large-insert clones, assigning them to specific chromosomal regions and characterizing contiguous clonal arrays.

YAC DNA libraries will be constructed (1) from human-hamster somatic cell hybrids that retain the entire X chromosome or portions of it as their only human material and (2) from flow-sorted human X chromosomes. Additional YAC clones will be collected from other preexisting libraries, such as the total human library constructed at Washington University in St. Louis and recently transferred to Baylor.

Specific YAC DNA fragments will be amplified from total yeast DNA using *Alu*-polymerase chain reaction (PCR), which allows sequence amplifications between two adjacent *Alu* repeats. Hybridization with these *Alu*-PCR probes will allow regional assignment of each clone to somatic cell hybrid mapping panels that divide the X chromosome into about 40 to 50 subregions (each ~3 Mbp long).

Initial attempts to identify long stretches of adjacent clones will be performed among clones previously assigned to the same chromosomal subregion and will later be expanded to the entire chromosome. We plan to obtain end probes from each clone by *Alu*-PCR and use them to identify overlaps. Confirmation, overlap extent, and quality control of the clones used will be achieved by a variety of methods including *Alu* and long interspersed repeated sequence hybridization to digested yeast DNA and by *Alu*-PCR fingerprinting.

New methods are proposed for direct identification of clones from specific regions of the chromosome; for rapid production of sequence tagged sites from each clone; and for isolation of polymorphic genetic markers amenable to assay by PCR, which can be used to place the physically mapped regions onto the genetic map.

Partial Bibliography

S. A. Ledbetter and D. L. Nelson, "Genome Amplification Using Primers Directed to Interspersed Repetitive Sequences (IRS-PCR)," pp. 107–19 in *PCR: A Practical Approach*, ed. M. McLaren, IRL Press, London, 1990.

S. A. Ledbetter, D. L. Nelson, S. T. Warren, and D. H. Ledbetter, "Rapid Isolation of DNA Probes within Specific Chromosome Regions by Interspersed Repetitive Sequence (IRS) PCR," *Genomics* **6**, 475–81 (1990).

D. L. Nelson, "Applications of PCR Methods in Genome Mapping," *Curr. Opin. Genet. Dev.* **1**, 62–68 (1991).

D. L. Nelson, "Current Methods for YAC Clone Characterization," *Genet. Anal.: Tech. Appl.* **7**, 100–106 (1990).

D. L. Nelson, "Interspersed Repetitive Sequence Polymerase Chain Reaction (IRS PCR) for Generation of Human DNA Fragments from Complex Sources," *Methods,* in press (1991).

D. L. Nelson, A. Ballabio, M. F. Victoria, M. Pieretti, R. D. Bies, R. A. Gibbs, J. A. Maley, A. C. Chinault, T. D. Webster, and C. T. Caskey, "*Alu* PCR for Regional Assignment of 110 Yeast Artificial Chromosome Clones from the Human X Chromosome: Identification of Clones Associated with a Disease Locus," *Proc. Natl. Acad. Sci. USA* **88**, 6157–61 (1990).

D. L. Nelson, S. A. Ledbetter, L. Corbo, M. F. Victoria, R. Ramirez-Solis, T. D. Webster, D. H. Ledbetter, and C. T. Caskey, "*Alu* Polymerase Chain Reaction: A Method for Rapid Isolation of Human-Specific Sequences from Complex DNA Sources," *Proc. Natl. Acad. Sci. USA* **86**, 6686–90 (1989).

Massive Isolation and Contig Building of Chromosome-Specific YAC Clones

Jan-Fang Cheng and Julia Nikolic
Human Genome Center, Cell and Molecular Biology Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-6549, FTS 451-6549, Fax 510/486-6816, FTS Fax 451-6816, Internet: "jfcheng@lbl.gov"

I have developed a screening method that allows identification of multiple clones from a genomic yeast artificial chromosome (YAC) library using multiple probes simultaneously. This method is based on pooling of YAC clones, pooling of single-copy probes, and dot-blot hybridization. A genomic YAC library with 5-fold human genome coverage is being pooled into 48 microtiter plates, each well containing 12 YAC clones. DNA is being purified from these pools and dotted onto nylon membranes. By using a dot-blot apparatus, the whole genomic library can be spotted onto 12 filters, each 8 by 11 cm. To limit the necessity of secondary screening, we pooled YACs in two dimensions (i.e., vertically and horizontally). We are also constructing a bank of chromosome 21–specific probes from a flow-sorted cosmid library (a gift from Lawrence Livermore National Laboratory). These probes have been prescreened so that none of them carries human repeats. We have collected over 100 single-copy, chromosome 21 probes, and this number is expanding rapidly. In the hybridization of the first 60 probes against a YAC library, we have isolated over 200 YAC clones. These clones are currently being localized to chromosome 21 by in situ hybridization.

We are systematically using 20 probes to isolate over 80 YAC clones in one hybridization reaction. These isolated clones are then arrayed in microtiter plates for further contig building by hybridizations. Our goal is to generate a YAC contig and physical map for human chromosome 21.

New Strategies for Closure of the Chromosome 19 Contig Map

Pieter J. de Jong, Chris Amemiya, Charalampos Aslanidis, Jane Tang, Kathy Yokobata, and **Anthony V. Carrano**
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
510/423-8145, FTS 543-8145, Fax 510/423-3608, FTS Fax 543-3608, Internet: "pieter@pcr.llnl.gov"

Physical mapping of the human genome is a fast-moving research area. Virtually all cloning and mapping technologies now being applied at the Lawrence Livermore National Laboratory Human Genome Center are evolving continuously into more reliable, efficient, and generally applicable approaches. The main focus of this project is to design, optimize, and test novel procedures to facilitate region-by-region closure of the chromosome 19 map.

Inter-*Alu*–polymerase chain reaction (PCR) is used to amplify probe sequences or to generate distinctive PCR fragments for fingerprinting. *Alu*-PCR products can be reproduced from hybrid cells, pulsed-field gel DNA fragments, and different types of large-insert recombinant clones [e.g., cosmids or yeast artificial chromosomes (YACs)]. These varied background products can be used as a common language to assign the diverse types of recombinant clones to human genomic regions defined by hybrid cells. The products can also be used for rapid cross correlation of cosmid and YAC clones, for example, in using YAC clones to bridge gaps in the cosmid contig map. Multidimensional pooling of cosmid and YAC libraries facilitates the mapping of cosmids to YACs and vice versa. Such pools have been generated from the arrayed cosmid or YAC libraries by different pooling schemes, such as first-dimension pooling per microtiter dish, second-dimension per well position, and other pooling schemes per offset well position. *Alu*-PCR probes from the cosmids are hybridized to filter-bound *Alu*-PCR products generated from the YAC pools. The array positions of the corresponding YACs are then defined by unique combinations of hybridization-positive pools. Similarly, *Alu*-PCR probes from the YACs are hybridized to *Alu*-PCR products from the cosmid pools to define corresponding cosmid clones.

Finally, *Alu*-PCR, either *Alu*-fingerprinting or hybridization, is being used as an efficient tool to check the validity of existing contigs. As part of this integrated map-closure strategy, a large number of accessories to the *Alu*-PCR approach have been developed or are in an advanced state of development. They include coincidence, subtraction, and ligation-independent cloning; multidimensional pooling of arrayed clones; *Alu*-anchor PCR; and *Alu*-based chromosome walking strategies.

Physical and Transcription Mapping of Human Chromosome 11

Glen A. Evans, David McElligott, Steven Clark, Suzanne Clancy, Licia Selleri, Michael Smith, Merl Hoekstra, and Gary Hermanson
Molecular Genetics Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037
619/453-4100 ext. 279 or 376, Fax 619/558-9513, Internet: “gevans@salk-sd2.sdsc.edu” or “gevans@molly.sdsc.edu”

This laboratory has undertaken the construction of a physical map of human chromosome 11 by combined cosmid and YAC contig building. The goal of this project is to cover more than 90% of human chromosome 11 with ordered contigs larger than 2 Mb, localized to a precision of 2 Mb. The maps generated through this effort will consist of a set of overlapping DNA fragments in YACs or cosmids assembled into contigs spanning a set of reference markers and will contain only the locations of genes previously cloned by other methods. In parallel, a genetic map of polymorphic markers spaced at roughly 2-Mb intervals is being constructed in collaboration with others.

It is technically possible to make these physical and genetic maps much more robust and useful by superimposing information on the localization of transcribed sequences and obtaining the complete DNA sequence of a large number of these cDNA transcripts. The additional information to be added to the emerging chromosome 11 map includes (1) the location of transcription units on the physical map with respect to cosmids and YAC clones and contigs, (2) the complete DNA sequence of the corresponding cDNAs, (3) some general information on the tissue specificity of gene expression, and (4) an estimate of the quantity of the mRNA transcript. We aim to carry out cDNA cloning and sequencing studies along with the physical mapping effort.

The specific goals of this project are to

(1) Construct several high-quality arrayed cDNA libraries, most notably from brain, liver, and placenta. These libraries will be archived in 864-well microtiter plates produced for us by General Atomic Corporation. The libraries will be arrayed at high density on filter membranes and characterized by hybridization for the presence of repetitive sequences.

(2) Use these libraries to develop methods for cloning cDNAs corresponding to transcripts located within physical maps of portions of chromosome 11. Strategies will include direct hybridization of isolated cosmid and YAC clones to cDNA arrays, hybridization of cDNAs to YAC or cosmid DNA and rescue by polymerase chain reaction, the transient expression of genes following microinjection into rodent cells in culture, and rescue of cDNA sequences by homologous recombination in yeast.

(3) Develop methods for cloning chromosome 11–specific cDNAs in regions not covered by physical maps by hybrization of DNA from somatic cell hybrids to arrayed cDNA libraries.

(4) Characterize isolated cDNAs corresponding to mapped regions by grid hybrization. This method will determine abundancy and reduce duplication of effort, hybridization to standardized Southern and Northern blots, and DNA sequencing.

(5) Localize chromosome 11–specific cDNAs not included in a physical map by high-resolution fluorescence in situ hybridization and generate a reference set of sequence tagged sites from the cDNA sequences.

(6) Compare cDNA sequences determined in this project with sequence databases to detect similarity to known protein structural motifs.

Cloning and Characterization of Human Chromosome 21 YACs

Jeffrey C. Gingrich and Steven Lowry
Human Genome Center, Cell and Molecular Biology Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-6549, FTS 451-6549, Fax 510/486-6816, FTS Fax 451-6816

An initial goal of the Human Genome Project is to generate ordered clone libraries for large-scale sequencing efforts. At Lawrence Berkeley Laboratory, we are developing and refining methods for creating an ordered library of human chromosome 21 fragments in yeast artificial chromosomes (YACs). In addition to being the smallest human chromosome (~50 Mb), chromosome 21 contains a number of interesting genetic loci, including genes believed to be involved in Alzheimer's disease and Down's syndrome.

Initial YAC cloning efforts were from a hybrid cell line (mouse plus human chromosome 21). Chromosome 21–specific YACs were identified against the mouse background and characterized by a number of methods, in particular methods relying on the use of inter-*Alu* polymerase chain reaction (PCR) products from YACs.

The YACs are being regionally assigned by two methods. First, inter-*Alu* PCR products from the YACs are used as Southern hybridization probes against DNA from cell lines containing various parts of chromosome 21. Positive and negative hybridization results narrow down the chromosomal region from which the YAC is derived. Second, we are using biotinylated inter-*Alu* PCR products (in collaboration with Joe Gray at the University of California, San Francisco) or biotinylated total yeast DNA as probes for fluorescence in situ hybridization against chromosome spreads. Results from these experiments narrow down the YAC's location on the chromosome to within 2 to 4 Mb.

Regionally assigned YACs are presently being used as anchor points to create large contigs of chromosome 21 YACs by a novel PCR screening procedure using multidimensional pools of YACs. These multidimensional pools of YACs from a total human library are also being screened to isolate YACs that overlap other known chromosome 21 clones.

Finally, we are creating a fluorescence in situ hybridization (FISH) map of chromosome 21 by measuring the location of hybridization signal from previously characterized genetic and physical markers from chromosome 21. The FISH map is being used to locate YACs isolated from the hybrid cell line, as well as other anonymous YACs from chromosome 21 (for example, see the abstract by J.-F. Cheng).

A Clone-Limited STS Strategy for Physical Mapping

Christopher H. Martin,[†] Carol A. Mayeda, and **Michael J. Palazzolo**[†]
Human Genome Center, Cell and Molecular Biology Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
[†]510/486-5909, FTS 451-5909, Fax 510/486-6816, FTS Fax 451-6818, Internet: "michaelp@lbl.gov" or "chrism@lbl.gov"

The physical mapping of complex genomes is based on constructing a genomic library and then determining the overlaps between the mapping clone inserts to generate an ordered, cloned representation of nearly all sequences present in the target genome. We strongly favor a sequence tagged site (STS)–based strategy to map complex genomes.

One of the major difficulties in using an STS approach is the need to assay tens or hundreds of thousands of mapping clones with tens of thousands of STS markers. Many current STS selection strategies are based on relatively random selections of STSs; we have devised a nonrandom method. In our approach a five-hit mapping library is gridded to allow each clone to be maintained as a distinct entity. Both ends of each clone selected from this library are sequenced for use as STSs. Screening the library with such a pair of STSs will generate the first contig that should consist of about 9 to 11 clones, the selected clone, and about 4 to 5 STS-identified clones from each end.

The second set of paired STSs is derived from the selection of a second mapping clone chosen on the basis of not having been previously used as an STS source or identified as part of a contig by any other STS. This procedure is then repeated until every clone has been assigned to a contig.

Computer simulations suggest that our selection procedures, compared to a random approach, require fourfold fewer STS assays to build a map. Furthermore, the resulting map provides greater fractional genome coverage and is composed of larger contigs.

Partial Bibliography

M. J. Palazzolo, S. A. Sawyer, C. H. Martin, D. A. Smoller, and D. H. Hartt, "Optimized Strategies for Sequence-Tagged Site Selection in Genome Mapping," *Proc. Natl. Acad. Sci. USA* **88**, in press (1991).

Interdigitation of the Genetic and Physical/Cosmid Contig Maps of Human Chromosome 19

Harvey W. Mohrenweiser, Katherine M. Tynan, Elbert Branscomb, Pieter J. de Jong, Barbara J. Trask, and Anthony V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
510/423-0534, FTS 543-0534, Fax 510/422-2282, FTS Fax 532-2282, Internet: "harvey@cea.llnl.gov"

The most useful application of the cosmid overlap map will occur in the search for functionally significant loci, including disease genes. Chromosome 19 genetic markers will be the landmarks that define a set of ordered cosmids containing any locus of interest on the chromosome, thus creating a need for interconnection between the genetic and physical maps. The goal of this project is the interdigitation of the physical (contig) map and the genetic (expression) map of human chromosome 19; this involves (1) identifying cosmids containing genetic markers to serve as "anchors" for the physical map, (2) assigning cDNA clones to cosmids, and (3) generating sequence tagged sites (STSs) within mapped cosmids.

Anchor cosmids have been identified for 80 genes or markers, in collaboration with colleagues from about 40 different laboratories. These cosmids (about 6 per marker) have been assembled into contigs, along with overlapping cosmids from more than 9000 additional cosmids from this library. At least 3 of the 80 "tagged" contigs have been assigned to each chromosome band. Thirteen of these loci are the highly informative, genetically mapped markers comprising an initial reference or framework map for the chromosome. All analyzed tagged contigs map to the appropriate chromosome region by in situ hybridization. The anchor point effort will continue with about 50 available mapped probes remaining. (Positive cosmids have been identified for each of the 80 probes used in screening about 6-fold coverage of the library generated from flow-sorted chromosomes.)

Mapping cDNAs to cosmid contigs will identify specific cosmids for further analysis, including sequencing. Sixty-five previously mapped expressed genes have already been assigned to cosmids or contigs. They include 13 oncogenes or receptor genes used to map leukemia translocation breakpoints on 19p; 25 specific members of the immunoglobulin supergene family; and 4 unlinked genes involved in DNA metabolism. The carcinoembryonic antigen (CEA) gene family (having more than 20 members) has been assembled into 3 contigs spanning about 1.2 Mb. The CEA gene region is the focus of a concerted effort to establish a 1- to 2-Mb contig.

The genetic map will be completed by generating STSs and polymorphic markers from specific cosmids. As necessary for generating an STS, polymerase chain reaction primers for 10 of the 13 framework loci have been synthesized to generate a single appropriately sized amplification product from genomic DNA. Cosmids associated with the three additional appropriately spaced polymorphic markers have been identified thus far, providing the initial reference points for identifying other important genes.

Assembly, Closure, and Characterization of a Chromosome 19 Contig Map

Anne S. Olsen, Pieter J. de Jong, Chris Amemiya, Lori Johnson, Chira Chen, Linda Ashworth, Jesse Combs, Alex Copeland, and Anthony V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
510/423-4927, FTS 543-4927, Fax 510/423-3608, FTS Fax 543-3608, Internet: "olsen@ecor1.llnl.gov"

The goal of this project is to produce an overlapping cosmid map of human chromosome 19. This map presently consists of a foundation of cosmid contigs established by automated restriction fingerprint analysis. Over 8000 cosmid clones have been analyzed, with 4915 assembled into 721 contigs and 2391 cosmids on the minimum spanning path. We estimate that these contigs cover about 4.0 x 10^7 bp, or about 65% of the chromosome. Analysis of representative contigs by restriction mapping and overlap detection by hybridization have confirmed the validity of the contig assembly procedure. Many of the contigs have been anchored to the genetic map by assignment of mapped genes and loci. Many contigs have also been localized on the cytological map by in situ hybridization techniques at varying levels of resolution.

While we will continue to build and extend contigs with this random fingerprinting technique, our new emphasis will focus on closing the map by using directed walking techniques to extend and merge contigs. Hybridization techniques with RNA and DNA end probes, *Alu*-polymerase chain reaction (PCR), and *Alu*-vector PCR probes made from cosmids at the ends of contigs have been used to identify clones that extend the contigs. Screening with these probes has been made more efficient by use of high-density arrayed colony filters and multidimensional clone-pooling schemes.

We are now in the process of superimposing yeast artificial chromosomes (YACs) onto the cosmid map to aid in ordering contigs and to facilitate map closure. We will walk from cosmid contigs onto YACs both by PCR screening of YAC pools using primers derived from cosmid sequences and by screening YAC *Alu*-PCR pools using *Alu*-PCR products generated from cosmids. To identify cosmid clones corresponding to a given YAC clone, *Alu*-PCR products generated from the YAC will be hybridized to cosmid colony filters or to filters with *Alu*-PCR products of multidimensional cosmid pools. For YACs lacking *Alu* sequences, cosmid colony filters will be hybridized with probes made by labeling the YAC DNA. Regional localization of contigs by fluorescence in situ hybridization techniques or by hybridization with region-specific probes will enable us to focus on closure in a limited region.

The map will be completed at the resolution of cosmids where possible, but YACs and lambda clones will be added to the map where necessary to close gaps. Areas of particular interest identified in the process of contig construction are being analyzed more extensively. A detailed analysis of the carcinoembryonic antigen gene family region on 19q13.2 revealed the close linkage and genomic order of many of these genes and led to the identification of several new members of this family. This in-depth analysis of selected regions will further confirm the validity of contigs established by the fingerprinting and walking procedures.

Developing a Physical Map of Human Chromosome 22

Melvin I. Simon, Bruce Birren, and Hiroaki Shizuya
Biology Division, California Institute of Technology, Pasadena, CA 91125
818/356-3944, Fax 818/796-7066

The goal of this project is to use pulse alternating-current electrophoresis (PACE) and large-fragment cloning to derive a set of overlapping clones covering human chromosome 22. Much of the work involves the development of new or improved methods for cloning large DNA fragments and for handling, analyzing, and overlapping these clones. To create an overlapping clone map of human chromosome 22, a set of new bacterial artificial chromosome (BAC) vectors has been developed. They support the cloning of large DNA fragments, about 200 kb long. These BAC vectors are based on the *Escherichia coli* F-factor (fertility factor) and have been constructed to contain promoters for walking, a multiple cloning site, a *cos* site for cleavage reactions, rare-cutting sites surrounding the insert, and two selectable markers. A readily transformed, recombination and modification-deficient host strain has been developed as a repository for these clones.

A library of large fragments of chromosome 22 is being constructed in these vectors. An F-factor–based cosmid size library has been constructed in vectors called Fosmids. Source DNA for the libraries comes from hamster-human hybrid cells that contain either intact or deleted chromosome 22 and from flow-sorted chromosomes. The Fosmid clones were extremely stable and representative of much of chromosome 22. Some BAC clones larger than 150 kb were tested and found to be stable and to represent contiguous fragments of human DNA. Fingerprints of the clones from the different vector systems will be obtained by partially digesting the cloned DNA and labeling the *cos* sites with either radioactive or fluorescent tags. The *cos* sites will be cut with terminase and labeled by a hybridization-ligation reaction. Since each *cos* end has a different sequence, different oligos can be ligated to each site and a partial-digest map created from each end of the clone. The use of fluorescent tags attached by ligation allows the simultaneous use of different fluorochromes at each *cos* site; separate restriction analysis would be done for radiolabeled oligonucleotides. Detection of the restriction fragments would be performed on the PACE pulsed-field gel electrophoresis system. Computer algorithms will be used to construct the overlap map based upon the partial digest data.

Physical Structure of Human Chromosome 21

Cassandra L. Smith, Denan Wang, Kaoru Yoshida, Jesus Sainz, Carita Fockler, and Meire Bremer
Division of Chemical Biodynamics, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/643-6376, Fax 510/642-1188, Internet: "clsmith@ux5.lbl.gov"

A complete *Not* I macrorestriction map is being constructed for human chromosome 21. Macro restriction fragments are fractionated by pulsed-field gel electrophoresis, ordered, and regionally assigned using hybridization experiments. Genetically mapped sequences used as hybridization probes provide anchor points between the genetic and physical maps. De novo isolated linking clones identify contiguous fragments and are used to regionally assign *Not* I restriction fragments using a panel of hybrid cell lines. Probe hybridization to genomic DNA partially digested with *Not* I provides further information on adjacent fragments.

These approaches have assigned about 80% of the *Not* I restriction fragments. End-game strategies focus on chromosome 21 fragments contained in a hybrid cell line (WAV17). Hybridization of human-specific repetitive probes to *Not* I-digested WAV17 DNA revealed a minimum set of chromosome 21 restriction fragments. Single-copy sequences are assigned to the *Alu* fragments by hybridization. Inter-*Alu* polymerase chain reaction was used to generate single-copy probes from unassigned *Not* I fragments. These probes were used to regionally assign the *Not* I restriction fragments and to identify neighboring fragments. This links unassigned fragments to assigned fragments. The *Not* I map is being used to order chromosome 21 yeast artificial chromosomes by generating inter-*Alu* probes from *Not* I restriction fragments.

Partial Bibliography

J. F. Cheng and C. L. Smith, "YAC Cloning of Telomeres," *Genet. Anal.: Tech. Appl.* **7**, 119–25 (1990).

J. F. Cheng, C. L. Smith, and C. R. Cantor, "Structural and Transcriptional Analysis of a Human Subtelomeric Repeat," *Nucleic Acids Res.* **19**, 149–54 (1991).

H. Ichikawa, K. Simizu, A. Saito, D. Wang, J. Sainz, C. L. Smith, C. R. Cantor, H. Kobayashi, Y. Kaneko, H. Miyoshi, and M. Ohki, "Long-Distance Restriction Mapping of the Proximal Long Arm of Human Chromosome 21 with *Not* I Linking Clones," *Proc. Natl. Acad. Sci. USA*, in press (1991).

L. Pevny, S. Mita, E. A. Schon, J. Herbert, R. Mayeaux, M. T. Yu, and C. L. Smith, "Chromosome 21 Sequences Are Not Duplicated in Alzheimer's Disease: An Analysis By Pulsed Field Gel Electrophoresis," pp. 21–28, Vol. 126 in *Biotechnology and Human Genetic Predisposition to Disease*, UCLA Symposia on Molecular and Cellular Biology, New Series, ed. C. R. Cantor, T. Caskey, L. Hood, D. Kamely, and G. Omenn, Alan R. Liss, Inc., New York, 1990.

A. Saito, J. P. Abad, D. Wang, M. Ohki, C. R. Cantor, and C. L. Smith, "Construction and Characterization of a Human Chromosome 21 *Not* I Linking Library," *Genomics* **10**, 618–30 (1991).

Physical Mapping of Human Chromosome 16

R. L. Stallings, N. A. Doggett, C. E. Hildebrand, M. K. McCormick, L. L. Deaven,* D. F. Callen,** G. R. Sutherland,** K. Okumura,*** D. C. Ward,*** and **R. K. Moyzis***†
Life Sciences Division and *Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545
†505/667-2690, Fax 505/665-3024
**Department of Cytogenetics and Molecular Genetics, Adelaide Children's Hospital, North Adelaide, South Australia 5006, Australia
***Department of Human Genetics, Yale University School of Medicine, New Haven, CT 06510

Integration of contig physical maps with genetic linkage maps of human chromosomes will allow rapid access to cloned sequences spanning regionally localized genes, chromosomes breakpoints, and regions of allele loss. Toward this end, a cosmid contig map covering about 60% of human chromosome 16 has been constructed by repetitive sequence fingerprinting.[1] An additional 20 to 30% of this chromosome is represented as single fingerprinted cosmid clones.

Three basic approaches have been used to obtain regional localization of cosmid contigs: (1) fluorescence in situ hybridization, (2) somatic cell hybrid breakpoint mapping, and (3) hybridization of previously mapped genes and genetic markers to gridded arrays of cosmid clones on nylon membranes. A total of 128 contigs and singleton cosmid and YAC clones covering ~11.9 Mb (12.5% of chromosome 16) has been regionally localized on the chromosome.

A sequence tagged site (STS) map is also being constructed by obtaining sequence from the end clones of cosmid contigs. A total of 234 sequences (over 67 kb of sequence) has been obtained from M13'subclones of chromosome 16 cosmid clones; 39 STSs have been developed. Oligonucleotide primers from these STSs have been used in screening by polymerase chain reaction (PCR) both the Adelaide somatic cell hybrid breakpoint mapping panel and a chromosome 16–specific YAC library recently developed at Los Alamos. The YAC library, coupled with PCR and hybridization-based screening approaches, provides the tools necessary for obtaining 12-Mb contigs and progressing toward closure of the contig map.

1. R. Stallings, D. C. Torney, C. E. Hildebrand, J. Longmire, L. Deaven, J. Jett, N. Doggett, and R. Moyzis, "Physical Mapping of Human Chromosomes by Repetitive Sequence Fingerprinting," *Proc. Natl. Acad. Sci. USA* **87**, 6218–22 (1990).

Generating a Comparative Physical Map of Mouse Chromosome 7

Lisa J. Stubbs, **Eugene Rinchik**, and Estela Generoso
Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-8077
615/574-0848 or -0864, FTS 624-0848 or -0864, Fax 615/574-1283, FTS Fax 624-1283

Detailed comparisons of murine and human genetic maps have revealed that a number of linkage groups have been conserved, despite extensive chromosome reorganizations that have occurred throughout evolution. Whatever their evolutionary significance, the similarities and differences existing between murine and human genomes may provide valuable information regarding the structure and function of both. This project aims to expand the basis of those comparisons and begin the translation of that genetic basis into a detailed series of directly comparable physical maps.

Efforts are focused on the analysis of mouse chromosome 7 (Mmu 7), with special emphasis on regions with well-established human homology. The maps will be centered around conserved, mapped human cDNA clones to maximize potential for direct interspecies comparisons and to facilitate correlations with phenotype-based genetic maps of both species.

Although constructing extensive networks of overlapping restriction maps should already be possible in certain regions of the chromosome, significant gaps can also be expected. To fill those gaps, a large number of new Mmu 7 markers must be identified. For a source of new markers, we will concentrate on conserved human clones isolated from chromosomal segments with known Mmu 7 homology, including specific subregions of HSA 19q, 11q, 11p, 15q, and 16p. New markers will first be assigned to intervals by genetic methods and included in existing local maps or used to nucleate new regional networks.

The specific goal of this task is the complete physical linkage of markers, first throughout the well-marked segments of homology with HSA 19q13, 11p15.4–15.5, and 15q, and eventually throughout the length of Mmu 7.

This task is part of a multi-investigator program that combines genetic and physical mapping of mouse chromosomes with new strategies of germline mutagenesiss. The program has two major goals: (1) to contribute to the genetic and physical mapping of human chromosomes by phyically mapping the mouse genome within mouse-human homology regions; and (2) to complement other mouse-genetics centers by developing a system for relatively low-cost, major escalation of the mouse-mutation resource (targeted mutations and chromosomal rearrangements) that can be used by any present or future member of the mammalian-genome–analysis community.

For more information on this subject, refer to the abstracts by E. Rinchik and R. Woychik and by E. Uberbacher and R. Mural.

Abstracts: Physical and Genetic Mapping

Correlation of Physical and Genetic Maps of Human Chromosome 16

David F. Callen, Sinoula Apostolou, Elizabeth Baker, Liang Z. Chen, Helen Kozman, Sharon A. Lane, Julie Nancarrow, Hilary A. Phillips, Yang Shen, Andrew D. Thompson, Scott A. Whitmore, Norman A. Doggett,[*] Raymond L. Stallings,[*] C. Edgar Hildebrand,[*] John C. Mulley, Robert I. Richards, and **Grant R. Sutherland**[†]
Department of Cytogenetics and Molecular Genetics, Adelaide Children's Hospital, North Adelaide, South Australia 5006, Australia
[†](Int.) 61/8-204-7333 or -7284, Fax (Int.) 61/8-204-7384 or -7342
[*]Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545

A total of 49 mouse-human hybrids have been constructed to contain various portions of chromosome 16. Most human cell lines used as parents in these fusions contained balanced translocations identified by cytogenetics service laboratories. The breakpoints in these hybrids, together with the four major fragile sites on this chromosome, have been ordered by mapping genes and anonymous DNA probes. Our mapping efforts, now concentrating on a polymerase chain reaction approach, use multiplexed sets of sequence tagged site markers and A-C-repeat microsatellite sequences. At this stage chromosome 16 has been divided into 42 regions by mapping approximately 200 markers, including cosmid contigs developed at the Los Alamos National Laboratory.

Construction of a detailed genetic map of the complete chromosome has been facilitated by the physical mapping of a large series of probes typed in pedigrees from the Centre d'Etude Polymorphisme Humain (CEPH). Map detail will be refined by using highly informative A-C-repeat microsatellite sequences. Several sequences have been isolated and two have been typed on the CEPH panel.

This extensive hybrid cell panel, together with the detailed physical mapping now possible, is being used for cloning constitutional chromosome breakpoints, cancer breakpoints on chromosome 16, the Batten disease gene, and fragile sites. The rare folate-sensitive fragile site, FRA16A, has now been localized to a 400-kb *Sal* I fragment. In addition, cosmid contig maps spanning chromosome bands are being constructed. Extensive probings of high-density dot blots of cosmid clones with probes mapped in bands 16q22.1 and 16p12 are in progress.

Partial Bibliography

D. F. Callen, E. Baker, and S. Lane, "Re-evaluation of GM2346 from a Del (16) (q22) to t(4;16) (q35; q22.1)," *Clin. Genet.* **38**, 466–68 (1990).

D. F. Callen, E. Baker, H. J. Eyre, and S. A. Lane, "An Expanded Mouse-Human Hybrid Cell Panel for Mapping Human Chromosome 16," *Ann. Genet.* **33**, 190–95 (1990).

D. F. Callen, E. Baker, H. J. Eyre, J. E. Chernos, J. A. Bell, and G. R. Sutherland, "Reassessment of Two Apparent Deletions of Chromosome 16p to an Ins (11;16) and a T (1;16) by Chromosome Painting," *Ann. Genet.* **33**, 219–21 (1990).

D. F. Callen, E. Baker, S. Lane, J. Nancarrow, A. Thompson, S. A. Whitmore, K. Holman, R. I. Richards, D. H. MacLennan, R. Berger, D. Cherif, I. Jarvela, L. Peltonen, G. R. Sutherland, and R. M. Gardiner, "Regional Mapping of the Batten Disease Locus (CLN3) to Human Chromosome 16p12," *Am. J. Hum. Genet.*, in press (1991).

L. Z. Chen, P. C. Harris, S. Apostolou, E. Baker, K. Holman, S. A. Lane, J. K. Nancarrow, S. A. Whitmore, R. L. Stallings, C. E. Hildebrand, R. I. Richards, G. R. Sutherland, and D. F. Callen, "A Refined Physical Map of the Long Arm of Chromosome 16," *Genomics* **10**, 308–12 (1991).

M. J. Dixon, E. Haan, E. Baker, D. David, N. McKenzie, R. Williamson, J. Mulley, M. Farrall, and D. Callen, "Association of Treacher Collins Syndrome and Translocation 6p21.31/16p13.11: Exclusion of the Locus from These Candidate Regions," *Am. J. Hum. Genet.* **48**, 274–80 (1991).

V. J. Hyland, K. E. W. Fernandez, D. F. Callen, R. N. MacKinnon, E. Baker, K. Friend, and G. R. Sutherland, "Assignment of Anonymous DNA Probes to Specific Intervals of Human Chromosome 16 and X," *Hum. Genet.* **83**, 61–66 (1989).

V. J. Hyland, G. K. Suthers, K. Friend, R. MacKinnon, D. F. Callen, M. H. Breuning, T. Keith, V. A. Brown, P. Phipps, and G. R. Sutherland, "Probe VK5B, is Located in the Same Interval as the Autosomal Dominant Adult Polycystic Kidney Disease Locus, *PKD1*," *Hum. Genet.* **84**, 286–88 (1990).

H. M. Kozman, A. K. Gedeon, S. Whitmore, G. K. Suthers, D. F. Callen, G. R. Sutherland, and J. C. Mulley, "Addition of MT, D16S10, D16S4, and D16S91 to the Linkage Map Within 16q12.1–q22.1," *Genomics*, in press (1991).

J. C. Mulley, N. Barton, and D. F. Callen, "Localisation of Human *PGP* and *HAGH* Genes to 16p13.3," *Cytogenet. Cell Genet.* **53**, 175–76 (1990).

M. A. Pritchard, E. Baker, S. A. Whitmore, G. R. Sutherland, R. L. Idzerda, L. S. Park, D. Cosman, N. A. Jenkins, D. J. Gilbert, N. G. Copeland, and M. P. Beckmann, "The Interleukin-4 Receptor Gene (IL4R) Maps to 16p11.2–16p12.1 in Human and to the Distal Region of Mouse Chromosome 7," *Genomics* **10**, 801–806 (1991).

R. I. Richards, K. Holman, S. Lane, G. R. Sutherland, and D. F. Callen, "Human Chromosome 16 Physical Map: Mapping of Somatic Cell Hybrids Using Multiplex PCR Deletion Analysis of Sequence Tagged Sites," *Genomics*, in press (1991).

G. R. Sutherland, E. Baker, D. F. Callen, O. M. Garson, and A. K. West, "The Human Metallothionein in Gene Cluster is Not Disrupted in Myelomonocytic Leukemia," *Genomics* **6**, 144–48 (1990).

DNA Sequence Mapping by Fluorescence In Situ Hybridization

Barbara Jo Trask, Brigitte Brandriff, Katherine Tynan, Ger van den Engh, and **Anthony V. Carrano**
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
510/422-5706, FTS 532-5706, Fax 510/422-2282, FTS Fax 532-2282, Internet: "trask@flowcentral.llnl.gov"

Fluorescence in situ hybridization (FISH) facilitates the construction of a long-range map of chromosome 19 by (1) regionally localizing cosmids to chromosome bands; (2) confirming the accuracy of contigs by mapping cosmids representing their ends to metaphase and somatic interphase nuclei; (3) identifying contigs that are distributed along the chromosome at a density of about 3 Mbp (3 Mbp is the practical resolution of two-color hybridization to metaphase chromosomes); (4) ordering and orienting contigs relative to each other in two- or three-color experiments using metaphase, somatic interphase, or sperm pronuclear interphase targets; and (5) estimating the relative distance between contigs by mapping them to interphase and pronuclear targets. We will use probe order and map distance estimates to choose the best mapping approaches to bridge the gaps between contigs and complete the map of chromosome 19.

Cosmids are nonradioactively labeled with biotin or digoxigenin. Probe hybridization sites are labeled with fluorescent tags [e.g., Texas Red and FITC (green)] that can be viewed simultaneously. Sequences separated by 50 kbp and more can be ordered using a combination of targets of increasing decondensation (metaphase, somatic interphase, and pronuclei formed after fusion of eggs and sperm). We have validated these FISH mapping approaches by ordering cosmids from a 250-bp region surrounding the DHFR gene in CHO cells and a 2- to 3-Mbp region on Xq28. We have demonstrated a direct relationship between probe separation on the linear DNA molecule and the distance between hybridization sites in both interphase nuclei and pronuclei in the range from 50 kbp to several Mbp. Metaphase is the target of choice for ordering sequences separated by >2 Mbp.

We have regionally localized 318 cosmids to cytogenetic bands on chromosome 19; 87 of these are associated with genes or genetic markers, and 253 fall in 117 contigs that are well distributed along the chromosome. We use FISH to assay the validity of our contigs constructed by fingerprinting and overlap-detection algorithms. False contigs are identified as those whose members map to different bands or are >0.5 µm apart in interphase chromatin. Of the nearly 300 tiling path bonds assayed in a total of 72 contigs, we have identified 6 (2%) that contain a false join. This percentage is expected to decrease as more cosmids are entered into the contig solution. Contigs that map within the same band are ordered by measuring their proximity in interphase or pronuclear chromatin. Alternatively, cosmid order can be derived simply from the order of red and green fluorescent dots in nuclei that have been hybridized simultaneously with three or more probes. To date, FISH mapping has determined the order of about 21 contigs mapping around band 19q13.2, including members of the CEA family.

Technology Development for Large-Scale Physical Mapping

Tony J. Beugelsdijk, Patricia A. Medvick, Robert M. Hollen, and Randy S. Roberts
Los Alamos National Laboratory, Los Alamos, NM 87545
505/667-3169, FTS 843-3169, Fax 505/665-3911, FTS Fax 855-3911

Human genome mapping efforts have highlighted a huge area for potential automation of repetitive procedures. Our technology development efforts have centered on the construction of hybridization membranes for microtiter well plates, robotic control, and database development for initial storage of hybridization results.

A commercial system is available for constructing the hybridization array for a chromosome-specific library on a nylon filter. This product, however, requires continual human attendance, produces small-format grids, and has no associated clone-tracking capability. Our current automated gridding system, designed to accommodate high-density grids, has a 30-plate dispenser and restacker to permit unattended performance. To ensure information flow about the setup, the operator must supply initial startup data, such as required grid density and number of interleave patterns.

The hardware includes a Nutec gantry robot, a Zymark Company microtiter plate dispenser and restacker, a Keithley Instruments control system, a Symbol Technologies bar code scanner, a custom gridding tool, and an IBM personal computer. The software, originally written in C, has been converted to C++ in the object-oriented style of the Robot Independent Programming Language developed by Sandia National Laboratories to increase maintainability. An Rbase database provides location information to the robotic arm. We can now stack 30 trays at a time for unattended operation, gridding onto 1 or 2 membranes having 1 to 6 sectors each with an interleave density of 1, 4, 9, or 16 to 1.

Short-term improvement plans include changing the sample placement tool, adding two more microtiter plate dispensers, and developing a platelid holder. A reuseable tool with sterilization stations will provide a more consistent surface contact with the nylon membrane, contribute sterility, lower the operating cost, and reduce plastic waste associated with the disposable tool. Two additional microtiter plate dispensers will permit longer unattended membrane production and allow a full 6-sector with a 16-to-1 interleave density without refilling the dispensers. A platelid holder will permit the use of a more specialized toolholder on the robot arm and reduce robotic movement.

Instrumentation developments already under way include a colony- picking system for constructing primary microtiter plate arrays, a plate replicator for making working copies of the primary arrayed libraries, an imaging system to capture the hybridization experiment results, software to process and decode the images, and a colony-collocation system to place clones positive to a particular probe into adjacent wells in a clean microtiter plate. Information will be stored in an object-oriented database for perusal prior to entry into the laboratory notebook and the human genome database. The data can then be electronically accessed to facilitate physical map construction.

Overcoming Genome-Mapping Bottlenecks

Charles R. Cantor
Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-6900, FTS 451-6900, Fax 510/486-5282, FTS Fax 451-5282, Internet: "crcantor@lbl.gov"

We are using a combination of short- and long-term approaches to eliminate some bottlenecks in current methods of genome mapping. One major thrust is optimizing several potentially improved methods for chromosome fractionation. An ample supply of purified human chromosomes would be a boon to many mapping strategies, including the preparation of large-insert libraries such as yeast artificial chromosomes (YACs), the efficient assignment of many different types of clones to chromosomes or chromosome regions, and the efficient construction and completion of physical maps. A second major focus is developing efficient top-down strategies for assigning large-insert clones like YACs to chromosome regions and ordering these clones.

Among the major short-term projects are improved magnetic bead–based chromosome separations and improved pulsed-field gel separations of very large DNA pieces. We will also begin some studies to enhance the generation and fragmentation of very extended chromosomes or DNA molecules and eventually use them to develop efficient methods for assigning YACs to precise chromosome locations.

Field-Flow Fractionation of Chromosomes and DNA

J. Calvin Giddings
Department of Chemistry, University of Utah, Salt Lake City, UT 84112
801/581-6683, Fax 801/581-4353

The purpose of this work is to carry out studies of field-flow fractionation (FFF) and continuous SPLITT fractionation (CSF) to investigate their application to the separation and purification of chromosomes and DNA. Experimental and theoretical studies will use different versions of these instrumental systems to implement and enhance the performance of the separative process. The lead project is a collaborative program with Los Alamos National Laboratory, in which we will attempt to enhance the performance and throughput of chromosome-sorting methods by removing impurities (nuclei, chromosome clumps, DNA, and cell wall material) from chromosome preparations. Other goals are to separate chromosomes from each another with FFF and CSF, separate DNA fragments ranging in size from a hundred to a million bp, and improve instrumentation to increase throughput and effectiveness.

High-Resolution DNA Mapping by Scanning Transmission Electron Microscopy

James F. Hainfeld
Biology Department, Brookhaven National Laboratory, Upton, NY 11973
516/282-3372, FTS 666-3372, Fax 516/282-3407, FTS Fax 666-3407

The use of scanning transmission electron microscopy (STEM) to map and sequence DNA directly may complement the fast-growing technology of automatic sequencing. The advantages of a direct physical microscopy approach to sequencing include the ability to use very long fragments (10^5 to 10^6 bp) and to sequence them several orders of magnitude faster than would be possible with chemical methods. In addition, long DNA pieces do not present the problems associated with repetitive sequences. Although direct sequencing may be difficult to achieve with microscopy, high-resolution mapping is currently attainable. Again, potential use of pieces ~10^6 bp yields immediate advantages in the speed of discerning sequence organization.

Preliminary results were obtained using the following test system. A 622-bp sequence from pBR322 was excised with restriction enzymes and purified. A 128-bp fragment from the T7 virus was inserted at position 276 of the plasmid fragment, making a total of 720 bp. Denaturing and renaturing equal quantities of the 622-bp and 720-bp fragments resulted in a 50% formation of heteroduplexes, in which one 622-bp strand paired with a 720-bp strand and left the extra bases as a single-stranded loop. The next step involved synthesizing a 26-mer oligonucleotide complementary to a region of the single-stranded insert. This oligonucleotide was chemically modified by adding a sulfhydryl to the 5′ end and covalently attaching an undecagold cluster to it. The undecagold cluster contains 11 gold atoms and provides a small high-resolution marker visible in STEM; it has a gold core 0.8 nm in size. The oligonucleotides and heteroduplexes were then mixed under renaturing conditions and examined with STEM. We observed a kink at the 128-bp single-stranded insert position in the control heteroduplexes (with no gold clusters); also, total length and length to the insert were inconsistent with the proposed model. When the gold-oligonucleotide was hybridized, the gold cluster was visible as a tiny bright dot at the "V" vertex of the DNA. The gold cluster was about 10 Å (3 bp) from the base it labeled (3 bp); the accuracy of positioning a base from the end of DNA segments with STEM is ~2 bp. A total potential positional accuracy of 3 to 5 bp should prove useful in physical mapping.

Further work demonstrated RNA labeling. Yeast tRNA^{Phe} was altered to enzymatically introduce 2-thiocytidine (s^2C) at position 75. This was covalently coupled to an undecagold cluster. *Escherichia coli* tRNA^{Arg} with a naturally occurring s^2C at position 32 in the anticodon region was labeled similarly; this alteration required denaturation. The nucleic acid labels used are the smallest found to be stable in an electron microscope having a resolution of about 1 nm (3 bp). Their visualization by STEM provides the first example of high-resolution RNA labeling by electron microscopy.

Another project involved labeling a palindromic 12-mer having a 5′ sulfhydryl with the undecagold cluster. Upon renaturation one would expect to observe a short double-stranded piece of DNA separating a gold cluster at each end; this was borne out in our preliminary data, which showed pairs of clusters at the appropriate distance. This is the first time that a piece of DNA as small as a 12-mer could be discerned and identified directly with these labels in the electron microscope (EM).

These test cases demonstrate the potential of labeling specific bases for EM to obtain high-resolution structural information for mapping. Collaborators for this work included Kyra Carbone, Martha M. Simon, Philip Rappa, and Inan Feng (Brookhaven National Laboratory), Matias Sprinzl (Laboratorie für Biochemie, Bayreuth, Germany), and Miloslav Boublik (Roche Institute).

Partial Bibliography

J. F. Hainfeld, M. Sprinzl, V. Mandiyan, S. J. Tumminia, and M. Boublik, "Localization of a Specific Nucleotide in Yeast tRNA by Scanning Transmission Electron Microscopy Using an Undecagold Cluster," *J. Struct. Biol.* **107**, 1 (1991).

Automating the Analysis of Dot-Blot Hybridizations

Joseph Jaklevic, Tony Hansen, William Kolbe, Linda Sindelar, Edward Theil, and Donald Uber
Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory,
Berkeley, CA 94720
510/486-5647, FTS 451-5647, Fax 510/486-5857, FTS Fax 451-5857, BITNET: "jmj@lbl"

Screening and mapping cloned fragments by arraying them in dense grids on filters and hybridizing them with known probes is a widespread and powerful technique in genome analysis. Currently, manual and robotic arraying devices are being used at several laboratories to relieve the tedium of preparing filters. However, complete analysis of hybridization experiments also involves painstaking observation and tedious recordkeeping.

We are drawing on our previous experience in automatic colony picking to automate this process. After the probes are applied, the filters are imaged, usually using X-ray film or phosphor image plates, and the resulting images are digitized. Positive signals will be located automatically with image analysis software similar to that used for colony picking. Each positive hybridization location will then be mapped to the particular row and column of the original filter array. The grey-scale intensity will be integrated and normalized so that automatic comparisons can be made. The resulting values can be histogrammed to provide four levels of confidence: strongly positive, weakly positive, weakly negative, and strongly negative.

The availability of individual intensities in numerical form allows further quantitative analysis, which is not possible using visual inspection. Also, automatically capturing all such results simplifies inventory control and laboratory notebook records. For example, clones that hybridize to any combination of probes can be tracked to a specific well on a specific microtiter plate.

Positive signal coordinates must be carefully referenced to their proper positions on the filter array to avoid choosing an incorrect neighboring clone. To accomplish this, we have constructed several calibration images to which the actual working images will be mapped.

Advanced Flow Cytometry Technique Development

James H. Jett, **John C. Martin**, and **Mark E. Wilder**
Center for Human Genome Studies, Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545
505/667-3843, FTS 843-3843, Fax 505/665-3024, FTS Fax 855-3024, Internet: "jett@flovax.lanl.gov"

Achievement of Human Genome Project objectives will require the development of many new technologies in biology, instrumentation, and informatics. A Los Alamos National Laboratory (LANL) contribution has been flow cytometry, employed to isolate the chromosomes used as template material for the chromosome-specific libraries developed by the National Laboratory Gene Library Project (NLGLP).

Our current objective is to develop flow cytometry instrumentation for making measurements on cells, chromosomes, and molecules at rates up to 20,000/s. Specific program objectives include (1) increasing chromosome-sorting speeds to provide more material for NLGLP and the chromosome 16 physical mapping effort at LANL, (2) developing new ultrasensitive fluorescence-detection techniques with applications to high-speed DNA sequencing, and (3) supporting the LANL research community by providing new flow-measurement techniques and analytical tools to aid in understanding the complex multiparameter flow data.

Recent progress includes the prelimary implementation of DIDAC, a new digital data acquisition system for flow cytometry. A MicroVAX-based computer cluster is maintained for the LANL Life Sciences Division to aid the analysis of flow cytometric data and DNA sequence data.

Future efforts will focus on adding new capabilities to the high-speed sorter, completing a new data acquisition system to meet the needs of the next generation of flow cytometers, providing computer support to the chromosome 16 mapping effort, and furthering the development of sensitive fluorescent molecule detection for application to DNA sequencing.

DNA Separation by Pulsed-Field Capillary Electrophoresis

Barry L. Karger
Barnett Institute, Northeastern University, Boston, MA 02115
617/437-2867, Fax 617/437-2855

Our research objective is to optimize DNA separation techniques by combining the power of capillary gel electrophoretic procedures with the high-resolution capability of pulsed-field electrophoresis. We will use uv and fluorescence detection methods, as well as radioactive detection for trace-level analysis of fragments and hybridized species. These approaches will provide new and powerful methodologies for restriction mapping of a wide range of DNA fragment sizes.

We will be separating fragments into two size ranges, from 1 to 50 kb and 50 to 500 kb. Polyacrylamide gel columns will be used with pulsed-field electrophoresis for the smaller fragments; specific fragments will be collected by electric field programming. Agarose and agarose/polyacrylamide columns will be developed for the larger fragments; injection will be from agarose plugs. Well-defined portions of the *Escherichia coli* genome will be used to validate the methodology.

Image Acquisition and Analysis

William F. Kolbe, **Joseph E. Katz**, and **Joseph M. Jaklevic**
Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory,
Berkeley, CA 94720
510/486-7199, FTS 451-7199, Fax 510/486-5857, FTS Fax 451-5857, Internet: "wfkolbe@lbl.gov"

A number of procedures in genome mapping and sequencing require image acquisition and processing. These include the visualization of (1) DNA fragments in gel-electrophoresis separations either directly or following hybridization with specific probes; (2) hybridization reactions in high-density dot-blot arrays; (3) microscopic images of in situ hybridization experiments; and (4) patterns such as colony distribution in petri dishes for incorporation into automatic procedures for colony picking and other repetitive laboratory procedures. We are developing the technology for acquiring digital images amenable to computer-processing applications and for the mapping and sequencing procedures described above.

A conventional charged-coupled device (CCD) camera has been integrated into an image database system to capture autoradiogram images acquired with conventional film techniques. We have also evaluated a commercial phosphor image plate reader for direct digitization of autoradiographic images acquired using storage phosphor technology. This method was superior to conventional film in both linearity of response and dynamic range. We applied it to autoradiogram images of gel-electrophoresis separations and to dot-blot hybridization patterns. The advantages of dynamic range, linearity, and direct digitization were demonstrated for both applications.

The CCD system has also been used to provide a digital vision capability for robotics applications. Enhanced images of colony distributions in petri dishes were processed using commercial software to identify and locate individual colonies. These data were then transferred to the robot's coordinate system before automated picking and arraying of the colonies.

Since the conventional CCD camera is limited in its ability to respond to low light levels, a cooled CCD adapted for low-level applications is being evaluated. Preliminary applications that have been demonstrated include direct digitization of ethydium-bromide–stained gels and pattern detection using chemiluminescent probes. Longer-term applications include the optimization of multiplex fluorescence methods for imaging labeled probes with either direct CCD imaging or scanning methods.

Cloning-Independent Mapping Technology for Genomic Fidelity, Contig Linking, cDNA Site Analysis, and Gene Detection

Leonard Lerman
Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139
617/253-6658, Fax 617/253-8699, Internet: "lerman@fang.mit.edu"

We will construct as a model system a thermal-stability map of bacteriophage lambda DNA, both as it exists in the virus and in the prophage. The model system will be based on two-dimensional (2-D) separation of random fragments of lambda DNA, distributed according to their length and local thermal stability.

Specific detailed map information is obtained by applying appropriate probes to the 2-D random-fragment distribution. Results from this application should prove the usefulness of random-fragment or thermal-stability mapping as an alternative to restriction-fragment mapping. The procedure is expected to be valuable where cloning is difficult or impossible, as between unconnected contigs, and in testing the precision of correspondence between clones of extremely long human sequences and their presumed counterparts in the actual genome.

As time permits, we will construct a thermal-stability map of one or more regions in the human genome that contain long, continuous, known sequences, such as the betaglobin gene cluster or the T-cell receptor. High-sensitivity probe techniques may be studied as necessary for detecting particular sequence distributions from complex genomes.

Automated Methods for Large-Scale Physical Mapping

Patricia A. Medvick, Robert M. Hollen, Tony J. Beugelsdijk, Randy S. Roberts, David M. Trimmer, Leonard A. Stovall, and Mark A. Kozubel
Los Alamos National Laboratory, Los Alamos, NM 87545
505/667-2676, FTS 843-2676, Fax 505/665-3911, FTS Fax 855-3911, Internet: "pm@lanl.gov"

Developing and characterizing an ordered clone collection from human chromosome–specific DNA libraries, necessary for constructing both low- and high-resolution physical maps, offers the next challenge in completing the map of the entire human genome. This project focuses on instrumentation for the front-end processes required to construct a low-resolution physical map, specifically, the technology to array human DNA fragments automatically for subsequent analysis and use.

Our efforts in technology development have concentrated on identifying petri plate colonies, gridding hybridization membranes from microtiter well plates, and developing a database for robotic control and initial storage of hybridization results. An imaging system has been developed for the selection of bacterial colonies for sampling. A robotic arm will use this information to select colonies for introduction into microtiter well plates. We have assembled prototype hardware that delivers a large number of small-volume samples onto nylon filters. The samples are placed on this support in a precisely indexed array for multistage hybridization analysis and automated data acquisition.

The current prototype system will be delivered this year to life science personnel at Los Alamos National Laboratory (LANL) for testing, and their suggestions for improvement will be incorporated into the system. Plans include the implementation of a computer workstation to integrate automated tasks into one unit. Gathered information will be stored in an object-oriented database for perusal before being entered into the Laboratory Notebook database software at LANL.

Based on extensive experience in designing robotic and automated equipment, we are anticipating that practical problems in large-scale mapping programs will become evident only after attempts are made to apply new methods to actual map production. For this reason, the development of instruments for the construction of chromosome-specific physical maps has been evolving through a multidisciplinary program. The resulting automated devices are tools for research and production and will be appropriate targets for technology transfer.

Partial Bibliography

T. J. Beugelsdijk, "Engineering the Human Genome Project," *IEEE Potentials* **2**(1), 34–38 (1990).

T. J. Beugelsdijk, J. M. Hollen, and K. T. Snider, "Development of a Small Gantry Robotic Workcell for DNA Filter Array Construction," 8th International Symposium on Laboratory Robotics, Boston, September, 1990.

P. A. Medvick, R. M. Hollen, and R. S. Roberts, "Development of an Automated Workcell for DNA Hybridization Array Construction," *J. Lab. Rob. Autom.* (submitted).

P. A. Medvick, R. M. Hollen, R. S. Roberts, D. Trimmer, and T. J. Beugelsdijk, "Automated DNA Hybridization Array Construction and Database Design for Robotic Control and for Source Determination of Hybridization Responses," *Int. J. Genome Res.* (submitted).

Robotics and Automation

Donald C. Uber, Joseph M. Jaklevic, and Edward H. Theil
Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720
510/486-6378, FTS 451-6378, Fax 510/486-6816, FTS Fax 451-6816

We have initiated a series of projects with two general goals: to automate labor-intensive steps in certain laboratory procedures and to provide automatic capture of data and experimental parameters for computer storage and analysis. Initial steps completed at Lawrence Berkeley Laboratory (LBL) are automated procedures for clone library replication and automatic colony picking. For the former, we developed a 96-pin tool and sterilizer station to work with the 8 by 12 microtiter plate format. In the latter application, we built specialized petri dish holders and devised methods to calibrate the robot's interaction with the local coordinate system relative to fiducials on the dish. The colony picker also combines image acquisition and analysis with robotics to provide a simple form of machine vision for locating and picking colonies; a 5000-clone *Schizosaccharomyces pombe* library was picked using this system. Robotic methods for replicating libraries should decrease by about 40% the time and cost relative to fully manual methods. We are currently using this technique to produce four copies of the Olson yeast artificial chromosome library.

These initial applications have increased our understanding of the proper interactions between general-purpose robots and the tasks at hand. Improvements now under development include (1) a specialized side station for second-generation high-speed colony picking (using a fast-moving and accurate-position stage); (2) automatic handling and analysis of dot-blot hybridization experiments; and (3) new polymerase chain reaction technology. We are also building an integrated robotic and laboratory instrument environment that includes a clone inventory database and a high-level robot control language. These projects are designed to provide short-term support for LBL molecular biologists (within 1 year) while we are creating modules for a laboratory automation system that will take 3 to 5 years to develop fully.

Partial Bibliography

D. C. Uber, J. M. Jaklevic, E. H. Theil, A. Lishanskaya, and M. R. McNeeley, "Application of Robotics and Image Processing to Automated Colony Picking and Arraying," *BioTechniques* **11**, 642 (1991).

Quantitation in Electrophoresis Based on Lasers

Edward S. Yeung
Ames Laboratory, Department of Chemistry, Iowa State University, Ames, IA 50011
515/294-8062, Fax 515/294-0266

The goal of this project is to develop novel detection and quantitation techniques in gel and capillary electrophoresis to increase the cost-efficiency, reliability, convenience, sensitivity, and speed of processing DNA fragments. One technique used is indirect fluorometry, in which a fluorescing ion is used to elute the sample, resulting in a large fluorescence background signal throughout the gel. When a component of the sample appears, the fluorescing ion is "displaced," and a lower fluorescence signal is observed. This negative signal allows nonfluorescing species, such as DNA fragments, to be detected with the high sensitivity normally associated only with fluorescing species. Migration errors are eliminated because of the absence of tags. Sample preparation for electrophoresis and sample collection for subsequent sequencing are simplified also.

We have demonstrated that restriction fragments in the 0.1- to 23-kb range can be separated by this method and detected down to the pg level in capillary electrophoresis and down to the ng level in slab gel electrophoresis. A complete electrophoretogram can be obtained in a few minutes using capillary electrophoresis.

Abstracts: Mapping Instrumentation

We are also exploring novel ways to improve the speed and reliability of imaging schemes for use in mapping and sequencing. A charged-coupled detection device and a unique background correction algorithm can be employed to image native DNA in slab gels, using the intrinsic uv absorption of the bases at concentration levels comparable to those in standard protocols. This technique will be applied to the real-time monitoring of pulsed-field gel electrophoresis, so that interactive control can be implemented by field programming during the separation.

For sensitive laser-excited fluorescence detection in capillary electrophoresis, we are developing a multiplex approach that does not require critical optical alignment. Hundreds of parallel capillaries can be monitored at the same time, thus speeding up DNA sequencing.

Sequencing by Hybridization: Methods to Generate Large Arrays of Oligonucleotides

Thomas M. Brennan
Engineering Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
On site at Stanford University, Palo Alto, CA 94301
415/723-6277

This work uses a method of producing very large, high-density arrays of oligonucleotides for use in hybridization sequencing. These arrays of oligonucleotides are synthesized by massively parallel chemical reactions on glass plates. Each element in the array is separated from its nearest neighbors by a surface tension wall. The synthetic reactions are carried out on a picoliter scale. The specific nucleotides are delivered to each array element by arrays of piezoelectric pumps similar to an ink jet printer. A stable hydroxyalkyl group bound to the plate acts as the 5′-OH surrogate on which to initiate strand synthesis. This linker arm enables removal of purine and pyrimidine blocking groups without cleavage of the oligonucleotide from the support.

A relatively simple machine should be able to produce the full 1024 by 1024 array of 10-mers at the rate of 1 plate or 10 oligos/h.

Detection of Luminescence from Lanthanide Ions as Labels for DNA Sequencing

Gilbert M. Brown, Robert S. Foote, K. Bruce Jacobson, Frank W. Larimer, Roswitha S. Ramsey, Richard A. Sachleben, and Richard P. Woychik
Oak Ridge National Laboratory, Oak Ridge, TN 37831-6119
615/576-2756, FTS 626-2756, Fax 615/576-5235, FTS Fax 626-5235

The goal of this research project is the development of a sensitive and readily automated method of detecting labeled DNA fragments that can be applied to sequencing. We are developing a new luminescent labeling system that can be used with capillary electrophoresis.

The lanthanide ions [Ln (III)] used for labeling include Eu(III), Tb(III), Sm(III), and Dy(III). These ions will be attached to DNA with a derivative of the macrocyclic chelating agent 1,4,7,10-tetraazacyclododecane-1,4,7,10-tetraacetic acid (DOTA). Naphthalene, attached to DOTA, functions as an antenna to relay excitation energy to the excited states of the Ln(III) ions. The luminescent states of these ions are f-f in origin, and the narrow emission line widths will allow multiple labels to be detected simultaneously with minimal interference. Furthermore, the Ln(III) excited states have long lifetimes that will allow the use of gated detection for high background discrimination.

Energy transfer from naphthalene to the Ln(III) ions has been demonstrated in a naphthyl derivative of complexes of Eu(III), Tb(III), Sm(III), and Dy(III) with the chelating ligand diethylenetriaminepentaacetic acid (DTPA). We will test oligonucleotides labeled with Ln(III)DOTA complexes as primers for DNA polymerase in the Sanger sequencing procedure and in polymerase chain reaction amplification. Detection limits for labeled oligonucleotides will be determined with capillary gel electrophoresis on-column detection.

Partial Bibliography

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, G. M. Brown, R. S. Foote, F. V. Sloop, J. R. Peterson, and K. B. Jacobson, "Comparison of Sputter-Initiated Resonance Ionization Spectroscopy (SIRIS) and Laser Atomization RIS (LARIS) to Localize Tin-Labeled DNA," *J. Vac. Sci. Technol.* **A9**, 1312–19 (1991).

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, F. W. Larimer, R. S. Foote, R. P. Woychik, G. M. Brown, F. V. Sloop, and K. B. Jacobson, "Potential Application of Sputter-Initiated Resonance Ionization Spectroscopy for DNA Sequencing," *Anal. Chem.* **63**, 402–7 (1991).

K. B. Jacobson, H. F. Arlinghaus, H. W. Schmitt, R. A. Sachleben, G. M. Brown, N. Thonnard, F. V. Sloop, R. S. Foote, F. W. Larimer, R. P. Woychik, M. W. England, K. L. Burchett, and D. A. Jacobson, "An Approach to the Use of Stable Isotopes for DNA Sequencing," *Genomics* **9**, 51–59 (1991).

R. A. Sachleben, G. M. Brown, F. V. Sloop, H. F. Arlinghaus, M. W. England, R. S. Foote, F. W. Larimer, R. P. Woychik, N. Thonnard, and K. B. Jacobson, "Resonance Ionization Spectroscopy of Tin-Labeled DNA: Application to Multispectral/Multiplex DNA Sequencing," *Genet. Anal. Tech. Appl.*, in press (1991).

Vacuum Ultraviolet Ionizer Mass Spectrometer for Genome Sequencing

C. H. Winston Chen, **Marvin G. Payne**, and **K. Bruce Jacobson**
Health and Safety Research Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831
615/574-5895, FTS 624-5895, Fax 615/576-2115, FTS Fax 626-2115

We plan to construct a new high-resolution mass spectrometer equipped with a vacuum ultraviolet (vuv) photonionizer for use in human genome sequencing. Potential advantages of this method include eliminating the need for gel electrophoresis and radioactive labels and speeding up the DNA sequencing rate. This method conceivably could be used to sequence DNA segments up to 3000 nucleotides long.

Major steps for DNA sequencing by this approach include (1) incorporating organometallic compounds such as ferrocene into DNA molecules by means of the primer sequence, using the standard dideoxy terminator method; (2) developing methods to deliver ferrocene-tagged DNA segments to a mass spectrometer ionization zone without any DNA-segment decomposition; (3) using a vuv or uv coherent beam to ionize DNA segments without producing any fragmented ions; and (4) sending the ions produced into a mass spectrometer that can measure molecular weights higher than 100,000 Da.

During the first 3 years of this program, most of the effort will be concentrated on resolving several critically important questions. If results are favorable, a complete facility will be built to do routine DNA sequencing.

Partial Bibliography

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, G. M. Brown, R. S. Foote, F. V. Sloop, J. R. Peterson, and K. B. Jacobson, "Comparison of Sputter-Initiated Resonance Ionization Spectroscopy (SIRIS) and Laser Atomization RIS (LARIS) to Localize Tin-Labeled DNA," *J. Vac. Sci. Technol.* **A9**, 1312–19 (1991).

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, F. W. Larimer, R. S. Foote, R. P. Woychik, G. M. Brown, F. V. Sloop, and K. B. Jacobson, "Potential Application of Sputter-Initiated Resonance Ionization Spectroscopy for DNA Sequencing," *Anal. Chem.* **63**, 402–7 (1991).

K. B. Jacobson, H. F. Arlinghaus, M. V. Buchanan, C. H. Chen, G. L. Glish, R. L. Hettich, and S. A. McLuckey, "Applications of Mass Spectrometry to DNA Sequencing," *Genet. Anal. Tech. Appl.*, in press (1991).

K. B. Jacobson, H. F. Arlinghaus, H. W. Schmitt, R. A. Sachleben, G. M. Brown, N. Thonnard, F. V. Sloop, R. S. Foote, F. W. Larimer, R. P. Woychik, M. W. England, K. L. Burchett, and D. A. Jacobson, "An Approach to the Use of Stable Isotopes for DNA Sequencing," *Genomics* **9**, 51–59 (1991).

R. A. Sachleben, G. M. Brown, F. V. Sloop, H. F. Arlinghaus, M. W. England, R. S. Foote, F. W. Larimer, R. P. Woychik, N. Thonnard, and K. B. Jacobson, "Resonance Ionization Spectroscopy of Tin-Labeled DNA: Application to Multispectral/Multiplex DNA Sequencing," *Genet. Anal. Tech. Appl.*, in press (1991).

Development of a Fully Integrated Technology to Facilitate Sequencing the Human Genome

George M. Church
Department of Genetics, Harvard University, Boston, MA 02115
617/732-7562, Fax 617/732-7663, Internet: "church@rascal.bwh.harvard.edu"

Our major commitment is to develop fully integrated technology to enable the sequencing of the human genome. This goal is best achieved by focusing first on sequencing small, highly informative genomes, emphasizing the development of techniques that can be generalized for sequencing larger genomes. The complete sequencing of the eubacterial, archaebacterial, and simple eukaryotic genomes must be accomplished to determine the basic components of life in the three kingdoms. Completely sequencing the small genomes of *Escherichia coli* (4.7 Mb) and *Saccharomyces cerevisiae* (15 Mb), two of the most thoroughly studied and genetically useful organisms, allows a best-fit alignment of corresponding genes among the microbes and provides a valuable database for analyzing larger genomes.

A specific aim is to optimize multiplex sequencing, computer-assisted film reading, and interactions among individual steps. This will entail the development of methods to combine and compare automatic base assignment and the study of how systematic changes in laboratory protocols and base assignment algorithms affect final sequencing accuracy.

Another objective will be to develop multiplexed directed sequencing strategies, including oligonucleotide walking and hybridization selection methods.

These technologies will be developed and tested using high-throughput sequencing of portions of the *E. coli* genome and related microbial DNA and protein sequences.

Sequencing by Hybridization

Radomir Crkvenjakov and **Radoje Drmanac**
Biological and Medical Research Division, Argonne National Laboratory, Argonne, IL 60439-4833
708/972-3161 or -3175, Fax 708/972-3387, Internet: "crkve@mcs.anl.gov"

We proposed DNA sequencing by hybridization (SBH) in 1987. Steady progress in research and theory, including the sequencing of unknown short (100 bp) DNAs by this method, has opened the way for rapid development and laboratory-scale implementation of the SBH approach. To achieve our current research objective of developing potential SBH speeds of over 1 Mb per day per laboratory, we will explore the "sequencing chip" concept.

The chip is envisioned as a microhybridization surface with up to 100,000 known microlocations, each harboring a specific oligomer. The technologies to be developed include fluorescent labeling suitable for sequence-grade oligomer hybridization; microscopic detection of 100-m hybridization dots; mixed synthesis of large numbers of oligonucleotides on microbeads; monolayer formation from beads: bead position decoding by hybridization; and fast microscopy imaging.

A second project aim is the partial sequencing of a human chromosome and a mixed human cDNA library, which together comprise 400,000 clones. The clone-containing membranes will be hybridized with up to 5000 oligomer probes of 6 to 9 nucleotides in length. The one billion bits of hybridization information will be collected with a throughput goal of ten million bits per day. The data will be acquired, stored, and analyzed by massive computation.

This partial-sequencing experiment aims to furnish a rough chromosome overview that will be sufficient to locate and define most genes as well as regulatory and other interesting sequence motifs. In addition, it will provide the 10% of the data required for obtaining the complete chromosome sequence as well as sequence information on expressed genes represented in the cDNA library.

Partial Bibliography

R. Drmanac, I. Labat, I. Brukner, and R. Crkvenjakov, "Sequencing of Megabase Plus DNA by Hybridization: Theory of the Method," *Genomics* **4**, 114–28 (1989).

R. Drmanac, I. Labat, and R. Crkvenjakov, "An Algorithm for the DNA Sequence Generation from k-Tuple Word Contents of the Minimal Number of Random Fragments," *J. Biomol. Struct. Dyn.* **5**, 1085–1102 (1991).

R. Drmanac, Z. Strezoska, I. Labat, S. Drmanac, and R. Crkvenjakov, "Laboratory Methods: Reliable Hybridization of Oligonucleotides as Short as Six Nucleotides," *DNA Cell Biol.* **9**, 527–34 (1990).

Genomic Instrumentation Development: Detection Systems for Film and High-Speed Gel-Less Methods

Jack B. Davidson and Robert S. Foote*
Instrumentation and Controls Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6010
615/574-5599, FTS 624-5599, Fax 615/574-4058, FTS Fax 624-4058
*University of Tennessee Graduate School of Biomedical Sciences, Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-8077

We are taking a two-pronged approach to improve the tools needed for progress in the Human Genome Project. We will continue to seek ways to make conventional gel and film methods easier and more efficient because we believe that these techniques will be used in mapping and sequencing for many years, especially in smaller laboratories. We are also exploring new approaches to high-speed gel-less sequencing.

We have invented two simple but effective devices to aid in reading film—the Un-smiler and the Un-dimmer. The Un-smiler is an anamorphic optical viewer that straightens the tilt ("smiling") in sequencing gels and provides a horizontal reticle to aid in reading crowded gels. The Un-smiler can also be used for vertical magnification to give improved band separation. The Un-dimmer is an inexpensive contrast-enhancing viewer used to detect bands or spots that are almost invisible on a clear film background, nearly doubling contrast. In addition to sequencing and blotting applications, the device could be used in radiology (e.g., for bone imaging) and industrial radiography. The principles used in both hand-held viewers can be applied to improve electronic densitometry of film.

We are testing a new, simplified fast scanner for sequencing-films; it will provide sequence directly, without storage and manipulation of the image in computer. We expect one DNA sample to yield raw sequence in about 1 m after scanning starts. The scanner is expected to be inexpensive compared to software-intensive image analysis systems. Given "good" gels, multiplexing can be used to sequence several samples simultaneously.

The second part of our two-pronged program entails the development of new approaches to high-speed gel-less sequencing. We are collaborating on a project led by R. Foote, Oak Ridge National Laboratory (ORNL) Biology Division, to use our lensless radiation microscope (developed under the sponsorship of the Office of Health and Environmental Research) to contribute to the special detector design needed to implement microscale sequencing by hybridization.

We are continuing to develop a high-speed sequencing approach (initiated under ORNL Director's Fund sponsorship) involving high-speed vacuum electrophoresis and detection of biochemically produced fragments. In principle, the sequence of a 200- to 300-bp sample could be obtained in less than 1 s without gel or film. If successful, the new separation and detection method could be combined with a biochemical robot for continuous or quasi-continuous sequencing at an estimated rate of more than 10^6 bp/day. The ORNL biology group is providing expertise and will supply specially labeled DNA samples. Assistance in mass spectrometry has been provided by D. L. Donohue and other members of the Mass Spectrometry Section of the ORNL Analytical Chemistry Division.

Single-Molecule Detection Using Charge-Coupled Device Array Technology

M. Bonner Denton, **Richard Keller**,* Mark E. Baker, Colin W. Earle, and David A. Radspinner
Department of Chemistry, University of Arizona, Tucson, AZ 85721
602/621-8246, Fax 602/621-8272
*Los Alamos National Laboratory, Los Alamos, NM 87545

Successful single-molecule detection relies on a satisfactory signal-to-noise ratio to allow the specific determination of individual fluorescently tagged bases in a flowing stream. Current detection is limited by strong background emission, most notably Raman. Although the fluorescence signal of individual molecules is much stronger than the Raman background, the sheer number of solvent molecules produces noise that inhibits single-molecule detection. Indeed, the process is much like searching for a needle in a haystack.

Effective use of a charge-coupled device (CCD) promises significant enhancement of the signal-to-noise ratio. A special CCD mode of operation offers a unique means of detecting relatively bright sources in a large background field. This technique, known as time-delay integration (TDI) imaging, was first used in aerial surveillance by military satellites. In this process, the pixels in the CCD array are clocked or shifted synchronously with the flow of the analyte, much like a high-speed photographic technique in which the film is moved across the shutter in the same direction as the object being observed. In our system this synchronization effect collects the emission from a single species into a single, moving charge packet.

The fluorescent signal is enhanced over the background because it is condensed into a sector as small as a one- or two-pixel array, while the background emission is distributed over the entire photoactive surface of the CCD array. Thus, the entire fluorescent signal can be integrated into a charge packet containing background noise only from the immediate vicinity of the fluorescent species. With this technique, the signal-to-noise ratio can be improved by about 200- to several thousandfold.

Another advantage is that several fluorescent molecules may be present simultaneously in the observation region, because each molecule has its own individual pixel array that is separated from its neighbor by several pixels. Accordingly, observation rates of several hundred to many thousand molecules per second may be observed.

In addition, CCD possesses several special characteristics that make it ideally suited to the application of low-light-level detection. These array detectors offer high quantum efficiencies, low dark counts, and low read noise. The multiplex advantage gained by the array format of these detectors and the availability of several novel readout schemes make them very attractive alternatives to more-conventional single-element detectors.

We have investigated several different CCD formats to implement TDI mode imaging. A conventional square-format CCD (512 by 512 pixels) is being used initially with both sides masked off to produce a rectangular format of approximately 10 by 512 pixels; this design enables a long viewing region with few excess pixels requiring readout. Ultimately, a custom CCD with a long, narrow format will be fabricated to our specifications. Readout is accomplished by shifting the parallel registers down one row (at the same rate as the flow velocity) followed by shifting the entire serial register (at the bottom) to the readout preamplifier. At a continuous readrate in the 20- to 500-kHz range, the read noise of a cryogenically cooled CCD is negligible. Feasibility calculations indicate that a CCD operated in this fashion should produce sufficient signal to noise for very effective single-molecule detection.

Multicolumn Gel Electrophoresis and Laser-Induced Fluorescence Detection for DNA Sequencing at 64,000 Bases/Hour

Norman J. Dovichi
Department of Chemistry, University of Alberta, Edmonton, Alberta, Canada, T6G 2G2
403/492-3254 or -2845, Fax 403/492-8231

The goal of this research project is the development of a laser-induced fluorescence detector for multiple capillaries in gel electrophoresis separation of DNA sequencing samples. Capillary gel electrophoresis offers higher-speed and higher-resolution DNA fragment separations than those achievable with low electric field slab gel technology. Currently, we can sequence 2000 bases/h using a single capillary system; by operating 32 capillaries simultaneously, sequencing rates of 64,000 bases/h may be produced with a single instrument. Because this instrument requires only a modest power laser, a single electrophoresis power supply, and one computer for data collection, its cost should be quite modest.

The small capillary diameter will permit only a minute amount of fragment to be loaded onto the capillary without degradation of the separation. Smith[1] estimated that only 1 to 10 attomole (1 attomole = 10^{-18} mol) of each DNA sequencing fragment can be loaded before column overload effects become significant. For accurate measurement of fluorescence intensity, the signal-to-noise ratio must be as high as possible; detection limits well below an attomole are required in capillary gel electrophoresis. We applied three different sequencing systems to DNA samples in capillary gel electrophoresis: (1) a single laser and single emission spectral channel system for the Richardson/Tabor sequencing technique, (2) a single laser excitation and dual emission spectral channel detector for the DuPont sequencing technique, and (3) a dual laser excitation and four emission spectral channel detector for the Smith and Hood sequencing technique. The sequencing rate routinely exceeded 1000 bases/h/capillary and was occasionally pushed to 2000 bases/h. The best detection limits (1400 molecules) were obtained using a tetramethylrhodamine-labeled primer with the Richardson/Tabor sequencing technique, a 750-μW helium-neon laser (544-nm wavelength), and a cooled photomultiplier tube.

We are designing a five-capillary, laser-induced fluorescence detector based on a sheath flow cuvette. In this system, a single laser excites fluorescence from the capillaries, a single optic collects fluorescence, a single spectral filter rejects scatter light, and five high-sensitivity photodetectors measure fluorescence. A 32-capillary system will be constructed for high-throughput sequencing based on our experience with the 5-capillary system.

[1] H. Drossman, J. A. Luckey, A. J. Kostichka, J. D'Cunha, and L. M. Smith, "High-Speed Separations of DNA Sequencing Reactions by Capillary Electrophoresis," *Anal. Chem.* **62**, 900–903 (1990).

Partial Bibliography

D. Y. Chen, H. P. Swerdlow, H. R. Harke, J. Z. Zhang, and N. J. Dovichi, "Low-Cost, High Sensitivity Laser-Induced Fluorescence Detection for DNA Sequencing by Capillary Gel Electrophoresis," *J. Chromatogr.*, in press.

D. Y. Chen, H. Swerdlow, H. R. Harke, J. Z. Zhang, and N. J. Dovichi, "Single-Color Laser-Induced Fluorescence Detection and Capillary Gel Electrophoresis for DNA Sequencing," *Optical Methods for Ultrasensitive Detection and Analysis*, Proc. Soc. Photo-Opt. Instrum. Eng. **1435**, 161–67 (1991).

H. Swerdlow, J. Z. Zhang, D. Y. Chen, H. R. Harke, R. Grey, S. Wu, C. Fuller, and N. J. Dovichi, "Three DNA Sequencing Methods Based on Capillary Gel Electrophoresis and Laser-Induced Fluorescence," *Anal. Chem.*, in press.

Using Scanning Tunneling Microscopy to Sequence the Human Genome

Thomas L. Ferrell, Robert J. Warmack,[†] **David P. Allison**, K. Bruce Jacobson, Gilbert M. Brown, and Thomas G. Thundat
Oak Ridge National Laboratory, Oak Ridge, TN 37831-6123
[†]615/574-6215, FTS 624-6215, Fax 615/574-6210, FTS Fax 624-6210, BITNET: "rjw@ornlstc"

This project involves the operation and continued development of scanning tunneling microscopes (STMs) for basic physics research related to health and environmental problems. Methodologies are being developed to facilitate atomic and molecular studies important in sequencing the human genome and to provide other technology applications in surface science. An STM with single-atom resolution is used to image atomic structure on surfaces, to alter atomic positions, and to probe the dynamic phenomena caused by collective electron motion and motion of ions. Objectives include extending STM capabilities to a wider range of materials, identifying atomic species, and studying biological samples.

Partial Bibliography

D. P. Allison, J. R. Thompson, K. B. Jacobson, R. J. Warmack, and T. L. Ferrell, "Scanning Tunneling Microscopy and Spectroscopy of Plasmid DNA," *Scanning Microsc.* **4**(3), 517–22 (1990).

G. M. Brown, D. P. Allison, R. J. Warmack, K. B. Jacobson, F. W. Larimer, R. P. Woychik, and W. L. Carrier, "Electrochemically Induced Adsorption of Radio-Labeled DNA on Gold and HOPG Substrates for STM Investigations," *Ultramicroscopy*, in press (1991).

T. Thundat, D. P. Allison, R. J. Warmack, and T. L. Ferrell, "Imaging Isolated Strands of DNA Molecules by Atomic Force Microscope," Scanning Tunneling Microscopy '91, Interlaken, Switzerland, August 12–16, 1991.

R. J. Warmack, T. G. Thundat, D. P. Allison, and T. L. Ferrell, "Electrostatic Spraying of DNA Molecules for Investigation by Scanning Tunneling Microscopy," Scanning Tunneling Microscopy '91, Interlaken, Switzerland, August 12–16, 1991.

DNA Sequence Analysis by Solid-Phase Hybridization

Robert S. Foote,* Richard A. Sachleben,** and K. Bruce Jacobson*
University of Tennessee Graduate School of Biomedical Sciences, *Biology Division, and **Chemistry Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-8077
615/574-0801, FTS 624-0801, Fax 615/574-1274, FTS Fax 624-1274

The goal of this project is to develop the use of arrays of oligonucleotide sequences on solid supports for use in DNA sequence analysis. Complete sets of all possible sequences of a given length or any combination of selected sequences may be prepared, with each sequence being identified by its position in the array. While covalently attached to the support, the probes are hybridized with target DNA under conditions designed to optimize mismatch-free duplex formation. The target DNA fragments will carry labels detectable by optical or mass spectral characteristics. Arrays containing complete sets of sequences 8 or more nucleotides in length are used to implement DNA sequencing by hybridization (SBH). Smaller arrays of selected sequences can be used for mapping, fingerprinting, and other DNA diagnostic applications. Techniques are being developed for the synthesis of microscale arrays containing complete sets of sequences in a few square centimeters.

Partial Bibliography

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, G. M. Brown, R. S. Foote, F. V. Sloop, J. R. Peterson, and K. B. Jacobson, "Comparison of Sputter-Initiated Resonance Ionization Spectroscopy (SIRIS) and Laser Atomization RIS (LARIS) to Localize Tin-Labeled DNA," *J. Vac. Sci. Technol.* **A9**, 1312–19 (1991).

Abstracts: Sequencing Technologies

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, F. W. Larimer, R. S. Foote, R. P. Woychik, G. M. Brown, F. V. Sloop, and K. B. Jacobson, "Potential Application of Sputter-Initiated Resonance Ionization Spectroscopy for DNA Sequencing," *Anal. Chem.* **63**, 402–7 (1991).

K. B. Jacobson, H. F. Arlinghaus, H. W. Schmitt, R. A. Sachleben, G. M. Brown, N. Thonnard, F. V. Sloop, R. S. Foote, F. W. Larimer, R. P. Woychik, M. W. England, K. L. Burchett, and D. A. Jacobson, "An Approach to the Use of Stable Isotopes for DNA Sequencing," *Genomics* **9**, 51–59 (1991).

R. A. Sachleben, G. M. Brown, F. V. Sloop, H. F. Arlinghaus, M. W. England, R. S. Foote, F. W. Larimer, R. P. Woychik, N. Thonnard, and K. B. Jacobson, "Resonance Ionization Spectroscopy of Tin-Labeled DNA: Application to Multispectral/Multiplex DNA Sequencing," *Genet. Anal. Tech. Appl.*, in press (1991).

Advanced Sequencing Technology

Raymond F. Gesteland and Robert Weiss
Department of Human Genetics, University of Utah, Salt Lake City, UT 84132
801/581-5190, Fax 801/585-3910, Internet: "rayg@genetcs.med.utah.edu"

Our objective is the further development of the technology needed for multiplex sequencing, including an automated sequence-reading system for data collection and a front-end system for organizing the input clones to avoid shotgun approaches. The gel-reading system is based on a new algorithm using communication and signal-processing theory. The front-end system employs transposons that carry in the sequences necessary for multiplex sequencing. We are developing a rapid mapping method to identify the minimal set of DNA fragments.

These projects are being pursued in parallel with projects sponsored by the Utah Center for Human Genome Research, including the development of capillary gel sequencing and of comprehensive computer systems for accumulating and analyzing DNA sequence information.

Megabase Sequencing of Human Immune Receptor Loci

Leroy E. Hood
Science and Technology Center, Division of Biology, California Institute of Technology, Pasadena, CA 91125
818/397-2762, Fax 818/793-4627, Internet: "lee@hood.caltech.edu"

Our initial objective is the complete sequencing of the three human T-cell receptor gene families, which comprise more than 3 Mb. In the later stages of this research, we will initiate studies on the human HLA locus, which also covers about 3 Mb. Three major goals for this proposal include (1) evaluating new strategies, techniques, and instrumentation for large-scale DNA sequence analysis; (2) determining the base sequence for human T-cell receptor and HLA loci; and (3) using this information to create new approaches to studying the molecular biology of the immune response. Our efforts are based on using four-color, automated fluorescent DNA sequencers and a mixed shotgun/directed approach with Sanger sequencing methods. The enormous amount of sequence data generated in this project will provide us with an invaluable tool for dissecting the fundamentals of the human immune system and for constructing a paradigm for understanding basic molecular biological systems.

DNA Sequencing Using Stable Isotopes

K. Bruce Jacobson, Heinrich F. Arlinghaus,* Gilbert M. Brown,** Robert S. Foote, Frank W. Larimer, Richard A. Sachleben,** Norbert Thonnard,* and Richard P. Woychik
Biology Division and **Chemistry Division, Oak Ridge National Laboratory,
Oak Ridge, TN 37831-8077
615/574-1204, FTS 624-1204, Fax 615/574-1274, FTS Fax 624-1274, BITNET: "bru@ornl.stc"
*Atom Sciences, Inc., Oak Ridge, TN 37830

This project involves developing a new DNA sequencing approach to determine the sequence of 1 x 10^7 to 6 x 10^8 nucleotides per day. In this procedure, stable isotopes of iron, tin, and several lanthanides can be used to label either DNA or the oligonucleotide probes that locate DNA fragments after electrophoresis. The use of stable isotopes eliminates the problems associated with radioisotopes, including radiation exposure to personnel, the costs of radioactive waste disposal, and the need to cope with short half-lives of reagents that contain radioisotopes.

Resonance ionization spectroscopy (RIS) will be used to localize and quantify these elements and all their isotopes. The sensitivity and selectivity of RIS should be comparable to that of the radioisotope and fluorescent methods currently used. Furthermore, the multiplex method of DNA sequencing will be adapted for use with stable isotopes so that 20 to 40 labels can be used simultaneously. This will require that the maximum number of stable isotopes of a given element be attached to a series of oligonucleotides in a stable configuration, so the label does not interfere with accurate hybridization.

One section of this study involves the development of chemical methods to incorporate iron, tin, and the lanthanides into organometallic compounds that can be attached to DNA. A second section will consist of an evaluation of two analytical methods used to detect and quantify stable isotopes: sputter-initiated resonance ionization spectroscopy (SIRIS) measures fractions of a monolayer, and laser atomization resonance ionization spectroscopy (LARIS) measures several monolayers of the sample surface. Other properties of these two methods are also important and will be considered in the evaluation process.

A third aspect of this study will be modification of the gel electrophoresis procedure to optimize the amount of DNA that appears at the gel surface after the electrophoretic separation of the DNAs and the subsequent drying of the gel. Ultrathin horizontal gels and capillary gel columns may be evaluated for their ability to accomplish electrophoretic DNA separation and detection. A fourth component of this project is the development of data management methods that will enable the large volumes of nucleotide data obtained by RIS to be assembled into sequence information. This project will require the collaboration of physicists, chemists, molecular biologists, and computer scientists.

Partial Bibliography

D. P. Allison, J. R. Thompson, K. B. Jacobson, R. J. Warmack, and T. L. Ferrell, "Scanning Tunneling Microscopy and Spectroscopy of Plasmid DNA," *Scanning Microsc.* **4**, 517–22 (1990).

H. F. Arlinghaus, M. T. Spaar, N. Thonnard, A. W. McMahan, and K. B. Jacobson, "Use of RIS to Significantly Increase the Speed of Sequencing the Human Genome," pp. 26–35 in *Optical Methods for Ultrasensitive Detection and Analysis: Techniques and Application,* Vol. **1435**, ed. B. L. Fearey, Soc. Photo-Opt. Instrum. Eng. (1991).

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, G. M. Brown, R. S. Foote, F. V. Sloop, and K. B. Jacobson, "Comparison of Sputter-Initiated Resonance Ionization Spectroscopy (SIRIS) and Laser Atomization RIS (LARIS) to Localize Tin-Labeled DNA," *J. Vac. Sci. Technol.* **A9**, 1312–19 (1991).

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, F. W. Larimer, R. S. Foote, R. P. Woychik, G. M. Brown, F. V. Sloop, and K. B. Jacobson, "Potential Application of Sputter-Initiated Resonance Ionization Spectroscopy for DNA Sequencing," *Anal. Chem.* **63**, 402–7 (1991).

Abstracts: Sequencing Technologies

K. B. Jacobson, H. F. Arlinghaus, M. V. Buchanan, C. H. Chen, G. L. Glish, R. L. Hettich, and S. A. McLuckey, "Applications of Mass Spectrometry to DNA Sequencing," *Genet. Anal. Tech. Appl.*, in press (1991).

K. B. Jacobson, H. F. Arlinghaus, H. W. Schmitt, R. A. Sachleben, G. M. Brown, N. Thonnard, F. V. Sloop, R. S. Foote, F. W. Larimer, R. P. Woychik, M. W. England, K. L. Burchett, and D. A. Jacobson, "An Approach to the Use of Stable Isotopes for DNA Sequencing," *Genomics* **9**, 51–59 (1991).

R. A. Sachleben, G. M. Brown, F. V. Sloop, H. F. Arlinghaus, M. W. England, R. S. Foote, F. W. Larimer, R. P. Woychik, N. Thonnard, and K. B. Jacobson, "Resonance Ionization Spectroscopy of Tin-Labeled DNA: Application to Multispectral/Multiplex DNA Sequencing," *Genet. Anal. Tech. Appl.*, in press (1991).

Advanced Detectors for Mass Spectrometry

Joseph M. Jaklevic, **W. Henry Benner**, and Joseph Katz
Human Genome Center and Engineering Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-5647, FTS 451-5647, Fax 510/486-5857, FTS Fax 451-5857, BITNET: "jmj@lbl"

The application of mass spectrometry to the separation of macromolecules is currently limited by the maximum size that can be efficiently measured, a limitation that will seriously affect future DNA mapping and sequencing efforts. Although methods exist for mobilizing and accelerating intact, charged molecular species with a mass of 200,000 Da, conventional ion detectors are inefficient for these large ions. These detectors rely primarily on the multiplication of secondary electrons generated by the primary molecular ion's high-velocity impact—a process that becomes less efficient for heavier molecular ions.

This project will investigate the fundamental parameters that limit the use of secondary electron emission for efficiently detecting large ions; the goal is to extend this approach to the limits of applicability. In addition, we will explore alternative methods for efficiently detecting large molecules. These methods include hybrid detectors, in which the primary ion is used to produce secondary, smaller ions through collision cascades; cryogenic bolometric devices, in which the energy of the ion is determined directly without reference to ionization processes; and semiconductor diode detectors, in which the thermal spike generated by the heavy ion impulse is electronically measured. In each case, feasibility studies will be undertaken, followed by practical implementation and testing where appropriate. The long-term goal of the project is to design a mass spectrometer capable of efficiently operating in the mass range above 100,000 Da and optimized for measuring large DNA fragments.

Sequencing of Linear Molecules

Joseph M. Jaklevic and **W. F. Kolbe**[†]
Engineering Division, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720
[†]510/486-7199, FTS 451-7199, Fax 510/486-5401, FTS Fax 451-5401, Internet: "wfkolbe@lbl.gov"

This project involves the adaptation of innovative physical analytical methods for use in determining the base sequence of DNA. The initial approach focuses on a technique for manipulating individual DNA molecules or ordered arrays of identical molecules to allow sequencing by direct spectroscopic methods. Next will be an attempt to combine the manipulation of ordered arrays with solid-phase restriction enzyme chemistry to achieve an advanced method for physical mapping of intermediate-sized fragments. Since the spatial resolution and analytical sensitivity required for DNA sequencing are significantly beyond existing capabilities, the feasibility of combining several innovative technologies will be explored. The use of electromagnetic fields to align labeled DNA strands will be investigated through direct imaging methods such as scanning tunneling microscopy, atomic force microscopy, and fluorescence microscopy. Other objectives are to study methods for attaching cloned DNA fragments to appropriate substrates and to explore applicable spectroscopic methods for detecting DNA molecules at the required level of sensitivity.

Rapid DNA Sequencing Based on Fluorescence Detection of Single Molecules

James H. Jett, **Richard A. Keller**,[†] **John C. Martin**, and **E. Brooks Shera**
Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545
[†]505/667-3018, FTS 843-3018, Fax 505/665-3024, FTS Fax 855-3024

We are developing a laser-based technique for sequencing 40-kb DNA fragments at rates up to several hundred bases per second. This approach relies on fluorescent labeling of all the bases in a single DNA fragment, attaching the labeled fragment to a support, moving it into a flowing sample stream, and detecting the individual labeled bases as they are cleaved from the DNA fragment by an exonuclease. The ability to sequence such large DNA fragments will reduce significantly the subcloning and sequence assembly steps now required to acquire megabase segments of sequence information.

Significant progress has been made in the detection of individual molecules. Time-resolved photon counting in a system using a mode-locked, frequency-doubled Nd:YAG laser has demonstrated the ability to detect individual Rhodamine-6G molecules as they pass through the laser beam. We are now detecting individual molecules of Rhodamine-6G and Texas Red simultaneously in a two-color experiment. Detailed studies of the photophysics of fluorescent molecule decomposition are under way to determine the conditions under which a molecule will emit the maximum number of photons. In addition, we are developing techniques to determine the fluorescence lifetimes of individual molecules.

We are using flow-cytometric sorting techniques to identify and separate microspheres attached to individual lambda-phage DNA molecules. The microspheres will be manipulated and suspended in our apparatus by using a microinjection pipette or a laser-based optical trap.

The synthesis of a completely labeled DNA strand is necessary for the technique to work. Studies are under way to determine the limits of labeling with several compounds. Initial efforts used commercially available biotinylated nucleotide precursors as the first model compounds for incorporation into the synthesized second strand. The complete replication of an M13 phage with all its incorporated cytosine and thymidine bases tagged with biotin has been achieved. We are now investigating the replication of M13 using fluorescently labeled nucleotides. Replication out to 3 or 4 kb with 1 labeled base and out to about 500 bases with 2 labeled bases has been demonstrated. We expect that modifications in the structure of the linker arm connecting the fluorescent dye to the nucleic acid will improve the replication results.

A Cooperative Research and Development Agreement has been signed with Life Technologies, Inc., (LTI) of Rockville, Maryland. LTI will be responsible for the nucleic acid chemistry and enzymology associated with this project. The principal investigator for the LTI effort is Jack D. Harding, P.O. Box 6009, Gaithersburg, MD 20877 (301/670-7649, Fax 301/948-8977).

Partial Bibliography

L. M. Davis, E. R. Fairfield, C. A. Harger, J. H. Jett, J. H. Hahn, R. A. Keller, L. A. Krakowski, B. L. Marrone, J. C. Martin, H. L. Nutter, R. L. Ratliff, E. B. Shera, D. J. Simpson, and S. A. Soper, "Rapid DNA Sequencing Based Upon Single Molecule Detection," *Genet. Anal.: Tech. Appl.* **8**, 1 (1991).

J. H. Hahn, S. A. Soper, J. C. Martin, H. L. Nutter, J. H. Jett, and R. A. Keller, "Laser-Induced Fluorescence Detection of Rhodamine-6G at $5X10^{-15}$ M," *Appl. Spectros.* **45**, 743 (1991).

Abstracts: Sequencing Technologies

J. H. Jett, R. A. Keller, J. C. Martin, B. L. Marrone, R. K. Moyzis, R. L. Ratliff, N. K. Seitzinger, E. B. Shera, and C. C. Stewart, "High-Speed DNA Sequencing: An Approach Based Upon Fluorescence Detection of Single Molecules," pp. 79 in *Proceedings of the Sixth Conversation: Biomolecular Stereodynamics, Structure and Methods, Vol. 1,* ed. R. H. Sarma and M. H. Sarma, Adeline Press, (1990).

J. H. Jett, R. A. Keller, J. C. Martin, R. K. Moyzis, R. L. Ratliff, E. B. Shera, and C. C. Stewart, "Method for Rapid Base Sequencing in DNA," U. S. patent 4,962,037, October 9, 1990.

E. B. Shera, N. K. Seitzinger, L. M. Davis, R. A. Keller, and S. A. Soper, "Detection of Single Fluorescent Molecules," *Chem. Phys. Lett.* **174**, 553 (1990).

S. A. Soper, L. M. Davis, F. R. Fairfield, M. L. Hammond, C. A. Harger, J. H. Jett, R. A. Keller, B. L. Marrone, J. C. Martin, H. L. Nutter, E. B. Shera, and D. J. Simpson, "Rapid Sequencing of DNA Based on Single Molecule Detection," *Proc. Soc. Photo-Opt. Instrum. Eng.* **1435,** 168 (1991).

S. A. Soper, J. H. Hahn, H. L. Nutter, E. B. Shera, J. C. Martin, J. H. Jett, and R. A. Keller, "Single Molecule Detection of Rhodamine-6G in Ethanolic Solutions Utilizing cw Excitation," *Anal. Chem.* **63**, 432 (1991).

Scanning Molecular Exciton Microscopy: A New Approach to Gene Sequencing

Raoul Kopelman, **John Langmore**,* **Bradford Orr**,** Zhong You Shi,** Steven Smith,** Weihong Tan,** and Vladimir Makarov*
Department of Chemistry, *Biophysics Research Division, and **Department of Physics, The University of Michigan, Ann Arbor, MI 48109-1065
313/764-7541, Fax 313/747-4865, Internet: "usergb2q@ub.cc.umich.edu"

Molecular exciton microscopy (MEM) is a new principle of imaging, based on short-range optical interactions between a specimen and a microscopic optical probe that is used to scan the specimen surface. The technique is based on the mechanics of other scanned-tip microscopies such as scanning tunneling microscopy, but MEM is designed to take advantage of well-defined optical interactions that occur between molecules separated by 0.1 to 5 nm.

Recent developments with excitons have overcome problems caused by subwavelength apertures of the glass capillaries used for photon-emitting tips. In MEM, a small exciton-conducting organic or inorganic crystal is grown in the aperture region and illuminated with laser light. Energy in the form of excitons is conducted through the crystal to the microcapillary tip, where it is converted to an extremely small, bright source of light. Our intention, however, is to take advantage of nonradiative interactions between molecules on the capillary tip and those on the specimen. The simplest specimens are individual molecules bound to atomically smooth substrates. As the tip is scanned along a flat surface parallel to the substrate, optical interactions between the tip and the specimen will be detected.

Microscope resolution will be limited by the size of the optical source (dependent upon the geometry of the donor molecule or molecules at the tip), the distance between source and specimen ($\geq$0.2 nm), and the range of optical interaction. A useful low-resolution interaction would be Foerster-Dexter energy transfer, which would change with the intensity and wavelength of photons emitted from both the tip and the sample. A much higher resolution interaction would be spin-orbit coupling between specific molecules bound to the capillary tip and heavy atoms in the sample (known in molecular optics as the external heavy-atom effect). In principle, this interaction has a range of less than ~0.5 nm and is very effective at stimulating emission of light at a new wavelength. Photomultipliers placed above or below the specimen will detect specific emissions from the tip or the specimen, using optical filters. As the tip is scanned, detected signals will be stored on a computer.

The many possibilities of MEM offer improved gene mapping and sequencing techniques. For example, if single-stranded DNA with mercury atoms on specific bases can be bound to flat surfaces, the external heavy-atom effect could be used to localize the heavy atoms to high resolution for reading specific bases along the strands. Use of fluorescent end labels and intelligent scanning algorithms might allow the labeled bases to be mapped at high speed. These and other optical interactions could be exploited to extend their usefulness to near-atomic resolution in a number of potential applications to cell and molecular biology.

Partial Bibliography

R. Kopelman, A. Lewis, and K. Lieberman, "Nanometer Light Source and Molecular Exciton Microscopy," *J. Lumin.* **45**, 298–99 (1990).

R. Kopelman, K. Lieberman, A. Lewis, and W. Tan, "Evanescent Luminescence and Nanometer-Size Light Source," *J. Lumin.* **48/49**, 871–75 (1991).

A. Lewis, K. Lieberman, S. Haroush, V. Habib, R. Kopelman, and M. Isaacson, "Light Microscopy Beyond the Limits of Diffraction and to the Limits of Single Molecule Resolution," pp. 615–39 in *Optical Microscopy for Biology*, ed. B. Herman and K. Jacobson, Wiley-Liss, Inc., New York, 1990.

K. Lieberman, S. Harush, A. Lewis, and R. Kopelman, "A Light Source Smaller than the Optical Wavelength," *Science* **247**, 59–61 (1990).

Transposon-Based Genomic Sequencing

Christopher H. Martin,[†] Michael Strathmann, Carol A. Mayeda, and **Michael J. Palazzolo**[†]
Human Genome Center, Cell and Molecular Biology Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
[†]510/486-5909, FTS 451-5909, Fax 510/486-6816, FTS Fax 451-6816, Internet: "michaelp@lbl.gov" or "chrism@lbl.gov"

One of the most difficult aspects of genomic sequencing is the disparity between the size of the clones used to generate the physical map (typically 40 kb for cosmids to 500 kb for yeast artificial chromosomes) and the size of the region that can be sequenced from a single primer site (about 0.5 kb). Our strategies involve two steps for breaking down the mapping clone inserts. First, the inserts will be fragmented into 5-kb pieces and subcloned. Next, the clones will be ordered using at least three strategies, from which the most efficient methods will be chosen. The ordered-subclone inserts will then be sequenced using transposon-generated priming sites.

A number of strategies have been proposed that use transposons to generate priming sites. To make such a strategy cost-effective (1) the mobilization of the transposon must be simple and efficient; (2) some method must allow site selection for transposon insertion into target sequences and not into vector elements or the *Escherichia coli* genome; and (3) an efficient method must be available to map the insertion sites to minimize the number of required sequencing reactions. We have developed, tested, and published one such strategy based on the gamma-delta transposable element and hope to use it to sequence the 300-kb bithorax complex.

These experiments are being done in collaboration with Ed Lewis and Susan Celniker of the California Institute of Technology.

Partial Bibliography

M. Strathmann, B. A. Hamilton, C. A. Mayeda, M. I. Simon, E. M. Meyerowitz, and M. J. Palazzolo, "Transposon-Facilitated DNA Sequencing," *Proc. Natl. Acad. Sci. USA* **88**, 1247–50 (1991).

Ultrasensitive Fluorescence Detection of DNA

Richard A. Mathies, Mark A. Quesada, Hays S. Rye,* Xiaohua Huang, Jiun W. Chen, and **Alexander N. Glazer***
Departments of Chemistry and *Molecular and Cell Biology, University of California,
Berkeley, CA 94720
510/642-4192, Fax 510/642-3599

This research focuses on developing ultrasensitive fluorescence methods, reagents, and a laser-excited confocal fluorescence gel scanner that provide dramatically improved detection of DNA on gels.[1] Further improvements to the scanning system will incorporate the simultaneous detection of two or more fluorescent probes.[2] We will also develop technology for micro-DNA-sequencing to reduce electrophoresis time and increase information density on gels. Our initial goal is to perform reliable sequencing with lanes that are ~1 mm wide and 15 to 30 cm long, with plans for new techniques that will reduce the lane width to below 1 mm.

We will continue to develop stable double-stranded DNA-dye complexes, a new class of macromolecular assemblies for fluorescence labeling.[2,3] These complexes can be detected with high sensitivity directly on agarose gels, eliminating the need for staining with large quantities of mutagenic dyes or for using radioactive labels and autoradiography. Simultaneous two-color labeling and detection will facilitate high-resolution mapping experiments.

An ultrasensitive stage-scanned confocal fluorescence microscope is being constructed for use in developing new chromosome-staining dyes and procedures. This improved sensitivity will be vital as new probes are devised to locate single-copy sites such as sequence tagged sites on chromosomes.

We will continue to explore methods for single-molecule and single-DNA-particle fluorescence detection.[4] Fluorescence burst detection will be used to count labeled DNA molecules separated by capillary electrophoresis. Single-molecule detection of specifically labeled fragments separated by capillary electrophoresis will make possible the ultratrace detection of specific DNA sequences and dramatically advance diagnostics.

1. M. A. Quesada, H. S. Rye, J. C. Gingrich, A. N. Glazer, and R. A. Mathies, "High-Sensitivity DNA Detection with a Laser-Excited Confocal Fluorescence Gel Scanner," *BioTechniques* **10**, 616–25 (1991).
2. H. S. Rye, M. A. Quesada, K. Peck, R. A. Mathies, and A. N. Glazer, "High-Sensitivity Two-Color Detection of Double-Stranded DNA with a Confocal Fluorescence Gel Scanner Using Ethidium Homodimer and Thiazole Orange," *Nucleic Acids Res.* **19**, 327–33 (1991).
3. A. N. Glazer, K. Peck, and R. A. Mathies, "A Stable Double-Stranded DNA-Ethidium Homodimer Complex: Application to Picogram Fluorescence Detection of DNA in Agarose Gels," *Proc. Natl. Acad. Sci. USA* **87**, 3851–55 (1990).
4. R. A. Mathies, K. Peck, and L. Stryer, "Optimization of High-Sensitivity Fluorescence Detection," *Anal. Chem.* **62**, 1786–91 (1990).

Automation of the Front End of DNA Sequencing

David Mead and **Lloyd M. Smith**[†]
Department of Chemistry, University of Wisconsin, Madison, WI 53706
[†]608/263-2594, Fax 608/262-0381, Internet: "smith@bert.wisc.edu"

The goal of this research is the development of an automated front-end system for large-scale DNA sequence analysis. This system will perform purification, mapping, fragmentation, fractionation, cloning, and enzymatic extension steps. When fully implemented, it will be capable of producing extension reactions from 1000 DNA templates per day, a rate that translates into approximately 500,000 bases of raw sequence data per day. To achieve this goal, we plan to develop technologies for (1) solid-phase purification of bacteriophage lambda or cosmid DNA, as well as the recombinant inserts; (2) simple and efficient DNA fragmentation and fractionation; (3) "automatic" cloning and mapping vectors; (4) microsequencing instrumentation; and (5) a directed end-sequencing strategy. The methodology will include an automated flexible robotic liquid-dispensing instrument capable of controlling multiple parameters and large numbers of samples.

Partial Bibliography

J. D'Cunha, B. J. Berson, B. L. Brumley Jr., P. R. Wagner, and L. M. Smith, "An Automated Instrument for the Performance of Enzymatic DNA Sequencing Reactions," *BioTechniques* **9**(1), 80–90 (1990).

J. A. Luckey, H. Drossman, A. J. Kostichka, D. A. Mead, J. D'Cunha, T. B. Norris, and L. M. Smith, "High Speed DNA Sequencing by Capillary Electrophoresis," *Nucleic Acids Res.* **18**(15), 4417–21 (1990).

Preparation of Oligonucleotide Arrays for Hybridization Studies

Michael C. Pirrung
Department of Chemistry, Duke University, Durham, NC 27706
919/660-1556, Fax 919/660-1591

The long-term goal of this research is to provide all the chemistry needed to prepare complete miniaturized arrays of decanucleotides [4^{10} or 10^{6} oligonucleotides (1 million)]. Such arrays can be used in a variety of techniques in molecular biology and medicine, including mapping clones in large genomic libraries, sequencing DNA larger than a kilobase by hybridization, and diagnosing diseases via DNA probes.

Very Large Scale Immobilized Polymer Synthesis (VLSIPS, Affymax Technologies N.V.), the technique best suited to preparing sequence arrays, will be used. VLSIPS relies on standard polymer synthesis techniques modified by the addition of a photoremovable group. This project will therefore aim first at developing such protecting groups for nucleotides and nucleosides, groups that must have very specific properties that will be optimized for the particular polymer synthesis method used. These properties include wavelength of deprotection, extinction, coefficient, quantum yield, by-products, chemical yield, and rate of deprotection.

The value of different supports in both the synthesis and hybridization steps will also be investigated. By standard nonphotochemical means, arrays of oligonucleotides will be prepared on rigid and nonrigid supports and supplied to groups around the world that are developing oligonucleotide hybridization as a tool in studying the human genome. Finally, optimal supports, chemistry, and photoremovable groups will be combined to permit the synthesis of large oligonucleotide arrays.

Thioredoxin-Gene 5 Protein Interactions: Processivity of Bacteriophage T7 DNA Polymerase

Jeff Himawan, Stanley Tabor, and **Charles C. Richardson**
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115
617/432-3129, Fax 617/432-3362

Bacteriophage T7 encodes its own DNA polymerase, the product of gene 5. This 80-kDa protein has low processivity, dissociating from a primer template after incorporating only a few nucleotides. Thioredoxin, a 12-kDa protein encoded by the *Escherichia coli trxA* gene, binds tightly to the gene 5 protein in a 1:1 stoichiometry and confers a high degree of processivity by increasing the affinity of gene 5 protein for a primer template.[1,2]

In addition to its DNA polymerase activity, the gene 5 protein has a 3′ to 5′ exonuclease activity. We inactivated this exonuclease activity chemically by a localized oxidation reaction[3] and genetically by site-specific mutagenesis.[4] The resulting T7 DNA polymerase-thioredoxin complex is useful for DNA sequencing by the dideoxy chain termination method.[5,6] The use of Mn^{2+} instead of Mg^{2+} for catalysis eliminates discrimination against dideoxynucleoside triphosphates and generates bands of uniform intensity.[7]

We explored the thioredoxin-gene 5 protein interaction biochemically by examining the ability of mutant thioredoxins isolated by Russel and Model[8] to interact with wild-type gene 5 protein in vitro.[9] We found that a thioredoxin mutant replacing both active site cysteines with serine residues can stimulate DNA polymerase activity of the gene 5 protein to levels comparable to the native complex; however, it can do so only at a 100-fold higher concentration than wild-type thioredoxin. Another thioredoxin mutant (gly-74 to asp-74) can restore nearly full polymerase activity but only at a concentration that is several hundredfold higher than wild type. A third thioredoxin mutant (gly-92 to asp-92) does not appear to bind gene 5 protein, even at extremely high concentrations. The *E. coli* asp-92 thioredoxin mutant cannot support the growth of wild-type T7 phage; furthermore, we were unable to isolate any T7 mutants (revertants) that could grow within the asp-92 thioredoxin host.

We also investigated the interaction between gene 5 protein and thioredoxin genetically. We used the gly-74 to asp-74 thioredoxin mutant to select for T7 revertants and characterized the nature of the revertant mutations. Every T7 revertant contains a gene 5 mutation, whose products presumably can compensate for the defect created by the gly-74 to asp-74 alteration in host thioredoxin and can interact productively with the asp-74 thioredoxin mutant to form a DNA polymerase that can efficiently replicate the phage genome. Specifically, the gly-74 to asp-74 mutation in the thioredoxin gene is suppressed by replacing glu-319 in the gene 5 protein with either a valine or lysine residue.

The interaction site between thioredoxin and gene 5 protein is defined, as least in part, by the parental mutations in thioredoxin and the revertant or suppressor mutations in gene 5 protein. The crystal structure of *E. coli* thioredoxin positioned the gly-74 residue on an exposed surface of thioredoxin,[10] implicated in protein-protein interaction because of its hydrophobicity. The genetic evidence suggests that gly-74 of thioredoxin and glu-319 of gene 5 protein represent a contact point between the two proteins. More specifically, we speculate that the side groups of these two amino acids are physically adjacent to one another.

The crystal structure of the large Klenow fragment of *E. coli* DNA polymerase I has been determined.[11] The structure reveals a striking groove or crevice whose size and shape are compatible with its being the binding site for double-stranded DNA. The amino acid sequences of the large Klenow fragment of *E. coli* DNA polymerase I and T7 gene 5 protein show extensive homologies, especially in this DNA-binding domain.[12] On the basis of tertiary structure inference, we suggest that the glu-319 residue of gene 5 protein is located at the edge of the postulated DNA-binding groove within the polymerization domain. We further speculate that thioredoxin binds gene 5 protein at the edge of this crevice to lock the duplex DNA into position.

1. S. Tabor, H. E. Huber, and C. C. Richardson, "*Escherichia coli* Thioredoxin Confers Processivity on the DNA Polymerase Activity of the Gene 5 Protein of Bacteriophage T7," *J. Biol. Chem.* **262**, 16212–23 (1987).
2. H. E. Huber, S. Tabor, and C. C. Richardson, "*Escherichia coli* Thioredoxin Stabilizes Complexes of Bacteriophage T7 DNA Polymerase and Primed Templates," *J. Biol. Chem.* **262**, 16224–32 (1987).
3. S. Tabor and C. C. Richardson, "Selective Oxidation of the Exonuclease Domain of Bacteriophage T7 DNA Polymerase," *J. Biol. Chem.* **262**, 15330–33 (1987).
4. S. Tabor and C. C. Richardson, "Selective Inactivation of the Exonuclease Activity of Bacteriophage T7 DNA Polymerase by *In Vitro* Mutagenesis," *J. Biol. Chem.* **264**, 6447–58 (1989).
5. S. Tabor and C. C. Richardson, "DNA Sequence Analysis with a Modified Bacteriophage T7 DNA Polymerase," *Proc. Natl. Acad. Sci. USA* **84**, 4767–71 (1987).

6. S. Tabor and C. C. Richardson, "DNA Sequence Analysis with a Modified Bacteriophage T7 DNA Polymerase: Effect of Pyrophosphorolysis and Metal Ions," *J. Biol. Chem.* **265**, 8322–28 (1990).

7. S. Tabor and C. C. Richardson, "Effect of Manganese Ions on the Incorporation of Dideoxynucleotides by Bacteriophage T7 DNA Polymerase and *Escherichia coli* DNA Polymerase I," *Proc. Natl. Acad. Sci. USA* **86**, 4076–80 (1989).

8. M. Russel and P. Model, "The Role of Thioredoxin in Filamentous Phage Assembly," *J. Biol. Chem.* **261**, 14997–15005 (1986).

9. H. E. Huber, M. Russel, P. Model, and C. C. Richardson, "Interaction of Mutant Thioredoxins of *Escherichia coli* with the Gene 5 Protein of Phage T7: The Redox Potential of Thioredoxin is Not Required for Stimulation of DNA Polymerase Activity," *J. Biol. Chem.* **261**, 15006–12 (1986).

10. A. Holmgren, B.-O. Soderberg, H. Eklund, and C.-I. Branden, "Three-Dimensional Structure of *Escherichia coli* Thioredoxin-S_2 to 2.8 Å Resolution," *Proc. Natl. Acad. Sci. USA* **72**, 2305–9 (1975).

11. D. L. Ollis, P. Brick, R. Hamlin, N. G. Xuong, and T. A. Steitz, "Structure of Large Fragments of *Escherichia coli* DNA Polymerase I Complexed with dTMP," *Nature* **313**, 762–66 (1985).

12. D. L. Ollis, C. Kline, and T. A. Steitz, "Domaine of *E. coli* DNA Polymerase I Showing Sequence Homology to T7 DNA Polymerase," *Nature* **313**, 818–19 (1985).

Improvement and Automation of Ligation-Mediated Genomic Sequencing

Arthur D. Riggs and Gerd P. Pfeifer
Department of Biology, Beckman Research Institute of the City of Hope, Duarte, CA 91010
818/301-8352, Fax 818/358-7703

Ligation-mediated genomic sequencing (LMGS) is a newly developed technique that allows high-quality sequence information to be obtained from mammalian cells without cloning. The overall aim of this project is to improve and automate LMGS. Our first goals are to improve the chemistry and simplify the procedures; automation is a more distant goal.

The first step in LMGS is to phosphorylate the 5′ ends of DNA using Maxam-Gilbert cleavage, which also generates sequence information that can be amplified by LMGS. A primer extension reaction is performed using a gene-specific oligonucleotide (primer 1) to generate molecules having a blunt end on one side. Linkers are ligated to the blunt ends; the resulting linker-ligated molecules are amplified by an exponential polymerase chain reaction using the longer oligonucleotide of the linker (linker-primer) and a second, nested, gene-specific primer (primer 2). This method amplifies all molecules that have undergone complete primer extension and linker ligation. Electrophoretic separation of the amplified fragments on DNA sequencing gels yields high-quality sequence ladders that can be visualized by hybridization with an appropriate probe. Ladder quality is the same for genomic mammalian DNA and cloned DNA. We transfer the separated fragments to a nylon membrane before hybridization. This procedure has several advantages, including the possibility of a type of "multiplexing." LMGS will be useful for sequences that are hard to clone or when directional sequencing is desirable.

A High-Speed Automated DNA Sequencer

Lloyd M. Smith
Department of Chemistry, University of Wisconsin, Madison, WI 53706
608/263-2594, Fax 608/262-0381, Internet: "smith@bert.wisc.edu"

The goal of this project is the development of a high-speed automated instrument capable of sequencing 500 bases of DNA per hour from 50 different samples. The theoretical daily throughput of such an instrument would be 500 x 50 x 24 = 600,000 bases of raw sequence data. The instrument

Partial Bibliography

C. M. Berg, G. Wang, L. D. Strausbaugh, and D. E. Berg, "Transposon-Facilitated Sequencing of DNAs Cloned in Plasmids," in *Methods in Enzymology: Recombinant DNA*, ed. R. Wu, Academic Press, New York, 1991 (in press).

L. D. Strausbaugh, M. T. Bourke, M. T. Sommer, M. E. Coon, and C. M. Berg, "Probe Mapping to Facilitate Transposon-Based DNA Sequencing," *Proc. Natl. Acad. Sci. USA* **87**, 6213–17 (1990).

Large-Scale DNA Sequencing with a Primer Library

F. William Studier and **John J. Dunn**
Biology Department, Brookhaven National Laboratory, Upton, NY 11973
516/282-3390, FTS 666-3390 or -3012, Fax 516/282-3407, FTS Fax 666-3407

Our aim is the development of a DNA sequencing capacity that can contribute significantly to the 15-year goal of sequencing the human genome. The strategy is to sequence 40-kbp clones (such as cosmids) directly by primer walking, using only primers from a library. If successful, this approach would improve efficiency and reduce costs by at least an order of magnitude over current practice. It would also provide the basis for developing automated sequencing machines capable of generating perhaps 100,000 nucleotides of sequence/h. For this strategy to succeed, either primers as short as nonamers or decamers must be able to prime sequencing reactions specifically and reliably or some simple means must be devised to join shorter primers to make longer ones. Our initial goal is to find conditions or to determine rules that make the primer-library approach practical and to develop primer walking with longer primers. Even if the primer-library approach proves impractical, conventional primer walking, together with a machine that simultaneously synthesizes small amounts of many primers, could result in a substantial improvement over current practice.

Partial Bibliography

F. W. Studier, "A Strategy for High-Volume Sequencing of Cosmid DNAs: Random and Directed Priming with a Library of Oligonucleotides," *Proc. Natl. Acad. Sci. USA* **86**, 6917–21 (1989).

Time-of-Flight Mass Spectrometry of DNA for Rapid Sequence Determination

Peter Williams and **Neal Woodbury**
Department of Chemistry, Arizona State University, Tempe, AZ 85287-1604
602/965-4107, Fax 602/965-2747

We are developing a time-of-flight mass spectrometry technique for rapid and accurate analysis of DNA sequencing mixtures. Intact DNA molecules are volatilized by pulsed-laser ablation of frozen aqueous DNA solutions at 532, 578, or 589 nm. The ablated DNA molecules are ionized by attachment of sodium ions, which are produced during the ablating laser pulse by resonant multiphoton ionization of sodium atoms from the vaporized solution. Intact DNA up to molecular weight ~6 MDa (~9 bp) has been volatilized, and DNA molecular ions up to 18,500 Da (28 bp) have been detected using this combination of techniques.

The upper mass limit of the mass spectrometer will be extended to at least 300 kDa by increasing the ion impact energy on the electron multiplier detector. Secondary electron production by ions that are more massive than ~30 kDa is inefficient or nonexistent. We are exploring new technology for detecting massive ions, using conversion dynode surface coatings that efficiently emit secondary ions. We will develop a capability for high-speed DNA sequencing with time-of-flight mass spectrometric ordering of dideoxynucleotide-terminated DNA mixtures produced by the Sanger process. Accurate (3-Da) mass-difference measurement techniques will be developed to facilitate error detection in sequence ordering. Simplifications in the biochemical production of sequencing mixtures, made possible by such mass accuracy, will be explored.

Computational Support for the Human Genome Center: Statistical and Mathematical Analysis, Data Processing, and Databasing

Elbert Branscomb, Tom Slezak, David Nelson, and Anthony V. Carrano
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
510/422-5681, FTS 532-5681, Fax 510/423-3608, FTS Fax 543-3608, Internet: "elbert@alu.llnl.gov"

This project provides mathematical and computational support for the genomic mapping efforts under way at Lawrence Livermore National Laboratory. Our primary goals have been to develop the following:

1. Computer-automated technology capable of turning raw restriction fingerprint data into an assembled contig map whose quality exceeds that achievable by manual data analysis methods;
2. A fully integrated, high-quality database system that stores all relevant experimental and derived data along with associated laboratory notebook data and that supports both standard query interfaces and direct network access; and
3. An interactive visual data interface that has highly logical power and visual informativeness; provides for integrated map constructs; exploits the trans-network server/client model; uses standard graphical, OS, and windowing environments; and is fully integrated with the total-project database system.

The first primary goal is essentially complete, although significant tuning and optimization efforts are continuing. Raw restriction fingerprint data is captured directly from the Applied Biosystems, Inc., sequencer machine and is processed for fragment recognition and length assignment. A clone overlap LOD score is computed for each pair of cosmid fingerprints. In the final step, the fingerprinted cosmids (now over 8000) are assembled into contigs, and minimal spanning subsets are determined for each contig, all automatically.

We have made significant progress toward the second goal with the development of an operational database consisting of some 100 tables. Approximately 2000 data items, aggregating 60 Mbytes of storage, are currently held in the widely used database. We have produced a wide variety of tools for data entry and data query, and all fingerprint data-acquisition and data-processing tools interact directly and automatically with the database. We are continuing to expand and develop the database aggressively.

Work on developing an interactive visual interface, our third goal, is well advanced. The present version of the contig browser (1) displays minimal spanning subsets with the predicted position of all other cosmids in the contigs; (2) displays the confidence of overlap (overlap LOD scores) for each cosmid pair in a contig; (3) allows visualization of overlaps between members in separate contigs; and (4) allows display of all data on cosmid attributes. These include probing results and fluorescent in situ hybridization mapping results, as part of the graphical presentation and in text windows. This tool is used heavily by biologists and occasionally runs simultaneously on 4 or 5 workstations. It also works well on remote computers and over Internet.

New Goals. Immediate new objectives include the development of (1) analytical, database, and data-visualization tools for assembling, managing, and statistically characterizing integrated maps and (2) mechanisms for managing, processing, and analyzing sequence data for projects such as sequence contig assembly and homology searches.

Current development of the database and the graphical interface browser is focused on integration of a diverse array of additional physical mapping data. These include high-resolution (e.g., interphase and prophase) fluorescence in situ hybridization ordering and distance-estimation data, yeast artificial chromosome (YAC) contig data, YAC and cosmid probing data (with many different classes of probes), YAC-to-cosmid containment data, and *Alu*-polymerase chain reaction cross-hybridization data. To handle these highly fluid and diverse interrelated data types appropriately, a new browser based on fundamentally different principles has been designed and is now being implemented. Corresponding database extensions, including those necessary to support graphical map integration, have also been designed and are being implemented.

Relevant to this effort, we are beginning a study of statistical techniques for calculating LOD scores to entire map constructs (i.e., complete contigs) including integrated maps. These techniques are based on "minimum description length" methods for computing likelihoods.

GenBank®: The Genetic Sequence Data Bank

James Cassatt
National Institute of General Medical Sciences, National Institutes of Health, Bethesda, MD 20892
301/496-7463, Fax 301/402-0019, BITNET: "czj@nihcu.bitnet"

The Genetic Sequence Data Bank (GenBank) is an internationally available repository of all reported nucleotide sequences 50 nucleotides or longer that have been annotated for sites of biological interest. GenBank is connected to a similar databank at the European Molecular Biology Laboratory in Heidelberg and to the DNA Data Base of Japan in Mishima; the three groups share responsibility for collecting sequences worldwide.

This resource is administered by NIH and cosponsored by DOE. The databank is operated under contract to IntelliGenetics of Mountain View, California, which provides overall management, distribution, and user-support services. Under the direction of Christian Burks, the Los Alamos National Laboratory collects, enters, verifies, and annotates the sequences. GenBank data are available both online (by modem) and as a computer-readable magnetic tape.

GnomeView: A Graphical Interface to the Human Genome

Richard J. Douthart, Joanne E. Pelkey, and David A. Thurman
Life Sciences Center, Pacific Northwest Laboratory, Richland, WA 99352
509/375-2653, Fax 509/375-3649, Internet: "dick@gnome.pnl.gov"

GnomeView is a graphical user interface that displays color representations of genomic maps for review, analysis, and manipulation. GnomeView's local network-model database contains chromosome maps as well as high-level descriptive information taken from both GenBank® and the Genome Data Base (GDB). GnomeView locates and presents maps and other queried information in an organized and easily used format. Displays are available for one chromosome or any combination of chromosomes, including the entire human karyotype. GnomeView displays GenBank tables as color-coded objects mapped to a sequence representation. All maps are displayed in windows with full zoom and pan control.

GnomeView architecture was designed to represent many different types of genomic maps. An internal database management system stores map objects and information about their linkages; however, GnomeView is primarily an interface and not a database repository. In the future we plan for remote on-line access to GenBank, GDB, and other genome databases as primary map information sources.

Robust Contig Construction

Michael Cinkosky, Randall Dougherty, **Vance Faber**, Mark Goldberg,* Mark Mundt, Robert Pecherer, Doug Sorenson, and **David Torney**†
Los Alamos National Laboratory, Los Alamos, NM 87545
†505/667-7510, FTS 843-7510, Fax 505/665-3493, FTS Fax 855-3493, Internet: "dct@life.lanl.gov"
*Rensselaer Polytechnic Institute, Troy, NY 12181

Physical mapping will produce megabase-size contigs, sets of clones with delineated overlaps. Our first priority is to develop techniques for contig construction employing fingerprint data, maintaining contig integrity despite false positive overlaps due perhaps to repeated sequences on chimeric clones. We employ interval graph algorithms to reject postulated clone overlaps inconsistent with a linear order of the clones and to add nonpostulated clone overlaps required by a linear order of the clones. Clone localizations (e.g., via the genetic map, hybrid panels, and in situ hybridization) will be merged with graph algorithms to achieve map closure.

We do not initially construct contigs using the graph algorithms described above. Instead, we use a data structure described by Tarjan[1] that allows contigs to be constructed at all thresholds of clone overlap probability. The threshold is then lowered, starting with a probability of unity; at points where contig merging would occur (at the critical thresholds), the new overlap is allowed only if it connects clones at the ends of contigs. In this way, a rough ordering of the clones is derived, and false-positive overlaps are rejected. We are attempting to optimize this approach by analyzing data from human chromosome 16, and we will soon focus on the interval graph algorithms likely to be the most effective.

There are a number of ancillary goals of this project. We will derive and incorporate estimates of clone length and extent of overlap. We are developing approaches for measuring the reliability of a contig—determining the probability that a subset of clones in a contig spans the contig and selecting a useful subset (with a high probability that no gaps occur). The probabilities of clone overlap in the spanning set reflect "redundant" clones not retained.

Most of these contig construction algorithms apply to any mapping protocol because the work builds on overlap probabilities derived from the primary data for pairs of clones.

1. R. E. Tarjan, "An Improved Algorithm for Hierarchial Clustering Using Strong Components," *Inf. Process. Lett.* **17**, 37–41 (1983).

HGIR: Information Management for a Growing Map

James W. Fickett, Michael J. Cinkosky, Michael A. Bridgers, Henry T. Brown, Christian Burks, Philip E. Hempfner, Tran N. Lai, Debra Nelson, Robert M. Pecherer, Doug Sorenson, Peichen H. Sgro, Robert D. Sutherland, Charles D. Troup, and Bonnie C. Yantis
Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545
505/665-5340, FTS 855-5340, Fax 505/665-3493, FTS Fax 855-3493, Internet: "jwf@life.lanl.gov"

A goal of the Human Genome Project is to acquire, organize, and distribute a large and complex body of data; developing tools for information management will thus play a central role. The objective of the Human Genome Information Resource (HGIR) is to provide information management systems to support map construction. This includes developing tools to construct, store, and display an integrated, multilevel map; capture, store, and provide access to the experimental evidence for the map; and aid in project management.

Central storage of all project data at Los Alamos National Laboratory is provided by the HGIR database, managed by the Sybase relational database management system or RDBMS on a network of Sun workstations. The core purpose of the database is to represent the relative positional information that makes up chromosome maps. In addition, the database serves as a Laboratory Information Management System or LIMS, tracking the experimental information from which the map is produced and the administrative information necessary for running the project.

Access to the database is currently provided by the Laboratory Notebook software, a collection of intuitive, on-line forms. Schema-driven interface software based on the GenBank® model is being developed in a way that will lead to multidatabase access software in 2 to 3 years.

Large volumes of cosmid clone fingerprint and overlap data have been handled smoothly this past year. After finding that Southern blot scoring constituted a major bottleneck in the data capture process, we developed a program to assist with the scoring. This program, called SCORE, allows the user to align the restriction digest gel image with the autoradiogram image on the workstation screen; as the user scores each band on the screen, noting lane and band location, the hybridization data is stored in the HGIR database.

Taking the place of cosmid fingerprinting as the primary tools for physical mapping are YAC- and sequence tagged site (STS)-based methods; we have thus extended our database schema to handle sequence and STS data as well as information on polymerase chain reaction (PCR) protocols and results. New interface forms are being implemented to store and retrieve this data. We are integrating HGIR management of sequence data with the work of Chris Fields and coworkers (see Fields abstracts) in the automated selection of PCR primer pairs.

The integration of sequence, physical maps, and genetic linkage maps is increasingly important. We have developed a prototype tool for computer-assisted map assembly that handles clone overlap data derived from several kinds of experiments. We can now store all positional information in one uniform and flexible manner. Work is under way to generalize our map assembly tool for use with multilevel maps.

Partial Bibliography

T. M. Cannon, R. J. Koskela, C. Burks, R. L. Stallings, A. A. Ford, P. E. Hempfner, H. T. Brown, and J. W. Fickett, "A Program for Computer-Assisted Scoring of Southern Blots," *BioTechniques* **10**, 764–67 (1991).

M. J. Cinkosky, J. W. Fickett, P. Gilna, and C. Burks, "Electronic Data Publishing and GenBank," *Science* **252**, 1273–77 (1991).

Identification of Genes in Anonymous DNA Sequences

Christopher A. Fields and **Carol A. Soderlund***
National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892 (On leave from New Mexico State University)
301/496-8800, Fax 301/480-8588, Internet: "cfields@loglady.ninds.nih.gov"
*Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001

The **gm** software system was developed to automate the identification of genes in anonymous DNA sequences. This system predicts the exon-intron gene organizations from genomic DNA sequence data and displays the resulting exon maps and predicted amino-acid sequences via a graphic user interface.[1] **gm** is written in the C programming language, with the graphics interface in X-windows, and runs on a wide variety of Unix machines. Designed for laboratory use, it supports incremental, exploratory analysis of new sequences as they are obtained. Several dozen user sites, including major DOE and NIH laboratories, have provided feedback on the system's utility in analyzing DNA from a number of organisms.

We have spent the last year testing **gm**, version 1, on sequences of known genes from *Caenorhabditis elegans*, using **gm** to analyze new sequences obtained by collaborators and designing and implementing new functions. The newly released **gm**, version 2.0, includes several new functions listed below.

1. Partial 3′-end cDNA data may be used to initiate exon map construction, allowing maps consistent with a partial cDNA sequence to be examined alone or with all other possible maps. By predicting only exons that consistently extend a known cDNA, **gm** can be used for efficient design of polymerase chain reaction primers to amplify sequences from a total cDNA library.

2. Repetitive elements must be contained either in introns or intergenic regions. An input file, which can be generated by **1fasta** or other similarity searching programs, specifies repetitive element positions. Screening the sequence for repetitive elements also decreases running time by decreasing the number of maps that are evaluated by the compositional analysis.

3. Markov matrix analyses are used to measure single and up to hexanucleotide compositions of introns and exons. Tests applied to these matrix elements to identify exons and introns can be varied for use with sequences from different organisms.

4. Predicted exon maps are now ranked in order of increasing protein-coding capacity. Users may choose to view only the highest-ranked nonoverlapping maps, each of which has the maximal coding capacity for the region spanned. This procedure generates only the maps most likely to produce hits in protein database searches, greatly decreasing the number of maps that the user must examine.

5. The graphic interface now has functions for displaying microrestriction maps, sequence tagged site locations, cDNAs, and repetitive DNA elements at the same scale as the predicted exon maps. This scaled display allows predicted genes to be aligned quickly with physical maps and facilitates restriction-fragment selection for use as probes in Northern blots.

6. The graphic interface supports multiple system runs with different parameter settings, facilitating the use of **gm** as an exploratory analysis tool and as an interface for displaying exon maps of known genes as well as cDNA and physical mapping data. The interactive **menu** tool, which now includes a function that builds exon maps from known exon coordinates, can also be accessed directly from the graphic interface.

Our current efforts are focused on initiating exon map construction with arbitrary cDNA data and on developing and testing methods for dealing with sequencing errors.

[1] C. A. Fields and C. A. Soderlund, "**gm**: A Practical Tool for Automating DNA Sequence Analysis," *Comput. Appl. Biosci.* **6**, 263–70 (1990).

BISP: VLSI Solutions to Sequence-Comparison Problems

Tim Hunkapiller, Leroy Hood, Ed Chen,* and Michael Waterman**
Science and Technology Center, Division of Biology, California Institute of Technology,
Pasadena, CA 91125
818/356-6417, Fax 818/793-4627, Internet: "tim@hood.caltech.edu"
*Jet Propulsion Laboratory, Pasadena, CA 91109
**University of Southern California, Los Angeles, CA 90089

Our research focuses on redefining the dynamic programming paradigm to enable the effective implementation of biological information signal processor (BISP) and very large system integrated (VLSI) methods in silicon and to provide the flexibility required of an effective biological tool. BISP represents a systolic implementation of a dynamic programming algorithm (based on that of Smith and Waterman), optimized with its ability to define local similarities. Fundamental functional differences between BISP and previous efforts include the following: (1) BISP is optimized for determination of any number of local similarities between pairs of sequences (the algorithm is also capable of returning global comparison values); (2) BISP returns values that will allow for alignment reconstruction; and (3) BISP is specifically designed to employ complex, user-definable similarity rules. Most significantly, BISP supports user-selectable alphabets of up to 128 characters, complete indel penalty definition, and individually defined similarity values between characters in the chosen character set.

Efficient Identification and Analysis of Low- and Medium-Frequency Repeats

Jerzy Jurka, Aleksandar Milosavljevic, Jolanta Walichiewicz, and Sherman Yang
Linus Pauling Institute of Science and Medicine, Palo Alto, CA 94306
510/327-4064, Fax 510/327-8564, Internet: "jurek@jmullins.stanford.edu"

Understanding the biology of repetitive elements is becoming a practical issue for genomic studies as more and more repetitive sequences accumulate in the rapidly growing databases of sequenced DNA. To resolve complex questions, we need well-organized databases of repetitive DNA and adequate tools for computer-assisted analyses.

The specific goals of this project are (1) to organize and maintain databases of repetitive elements from primate genomes, (2) to develop an algorithm for rapidly identifying and extracting novel low- and medium-reiteration (LOR and MER) frequency sequences from available data, (3) to find and characterize novel primate LOR and MER sequences in the available genomic sequences, (4) to generate probes for identified repeats using the polymerase chain reaction and to estimate the number of repeats in the human genome by Southern blotting, and (5) to expand the analysis to nonprimate DNA.

Several probes for the recently identified MER sequences are in general use for mapping purposes.

Partial Bibliography

J. Jurka, "Novel Families of Interspersed Repetitive Elements from the Human Genome," *Nucleic Acids Res.* **18**, 137–41 (1990).

D. J. Kaplan, J. Jurka, J. F. Solus, and C. H. Duncan, "Medium Reiteration Frequency Repetitive Sequences in the Human Genome," *Nucleic Acids Res.*, in press (1991).

Efficient Algorithms and Data Structures in Support of DNA Mapping and Sequence Analysis

Eugene Lawler and Daniel Gusfield
Electronics Research Laboratory, University of California, Berkeley, CA 94720
510/642-4019, Fax 510/642-5775, Internet: "lawler@arpa.berkeley.edu"

Our research is concentrated on algorithms for physical map construction, sequence matching and alignment, detection of specific biologically significant motif patterns in DNA, and sensitivity analysis of these problems. We are developing practical algorithms that are faster than existing methods; in particular, we are looking for alternatives to the dynamic programming approach that has been the method of choice for most molecular biology applications. Maintaining close ties with the Lawrence Berkeley Laboratory Human Genome Center will enable us to focus on problems relevant to the Human Genome Project.

Genome Assembly Manager

Charles B. Lawrence, **Eugene W. Myers**,* and Sandra Honda
Department of Cell Biology, Baylor College of Medicine, Houston, TX 77030-3498
713/798-6226, Fax 713/790-1275, Internet: "chas@mbir.bcm.tmc.edu"
*Department of Computer Science, University of Arizona, Tucson, AZ 85721

As part of our long-term research goal of developing an integrated software environment to support the activities of laboratories engaged in a large-scale DNA sequencing, we will address the problem of sequence reconstruction when using a "random" sequencing strategy. One component of this integrated software environment is the ***Genome Assembly Manager*** **(*GAM*)**, an X Window–based application that will allow technicians, molecular biologists, and project managers to interact efficiently with the data and sequencing strategies involved in large-scale sequencing. Underlying

GAM will be a Fragment Assembly Engine based on a rigorous theoretical foundation devised by the Arizona group. ***GAM*** will be developed in collaboration with the Baylor College of Medicine Human Genome Center, which is also formulating other strategies for large-scale sequencing.

BIOPIX: Imaging for Molecular Biology

Suzanna Lewis, **Frank Olken**, Kevin Gong, and Marge Hutchinson*
Human Genome Center and *Data Management Group, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-7370, FTS 451-7370, Fax 510/486-4004, FTS Fax 451-4004, Internet: "selewis@lbl.gov"

BIOPIX is an ensemble of programs that have been imported, adapted, or developed by the Lawrence Berkeley Laboratory (LBL) Genome Computing and Imaging Technologies groups to support two-dimensional imaging applications for the LBL Human Genome Center. LBL biologists will use this software for image acquisition, analysis, archiving, and retrieval. BIOPIX includes:

- ANGEL - a program to identify lanes and bands in gel (probed filter) images.
- TAP - a program to analyze images of polymerase chain reaction transposon mapping gels and to find minimal covering sets of transposons for sequencing.
- HIPS - an imaging toolkit from New York University.
- HIPSTOOL - a graphical interface to HIPS.
- GENIAL - a general-purpose, image-analysis package for handling very large high-resolution images (e.g., from image plate readers).
- IMAGEGRID - a computer-aided system for dot-blot image analysis.
- XIMAGEVIEW - an image-viewing program from the University of California, Berkeley (UCB).
- XIMAGEQUERY - an image database program from UCB.
- BIOPIXDB - an annotated database of images and related information.

All the software runs under X-windows on Sun SPARCstations. Several image-analysis software packages, such as GENIAL and HIPSTOOL, were developed by the LBL Imaging Technologies group. All image information (essentially a subset of an electronic laboratory notebook) is stored in a Sybase relational database, and actual images are stored in Unix files, either locally or on the LBL Mass Storage System.

Query facilities are provided by Image Query (from UCB) and locally developed forms-based queries. Data entry is forms based. Image acquisition is from either an image plate reader or a cooled charged-coupled display camera, both presently interfaced to the LBL network via International Business Machines personal computers. Image files are transferred to the Suns for analysis.

Genomic Information Management System

Suzanna Lewis, **Manfred Zorn**, **John McCarthy**, Victor Markowitz, and Frank Olken
Human Genome Center, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-7370, FTS 451-7370, Fax 510/486-4004, FTS Fax 451-4004, Internet: "selewis@lbl.gov"

The development of a comprehensive general information system for molecular biologists is a major part of the computing effort at the Lawrence Berkeley Laboratory Genome Center. To develop such a system, we focused first on genomic maps that represent a consensus of experimental knowledge, and we chose to represent this knowledge directly rather than to infer maps from the experimental results for individual markers. Marker data are structured to record the relationships that constitute the map.

We designed the Chromosome Information System (CIS) to enable biologists to

- search and navigate through maps derived from current experimental results, collaborating laboratories, and public databases;
- edit and compare maps, through direct manipulation, for both tabular or graphical map presentations (analogous to computer-aided design systems); and
- interact with other programs by either accepting data (e.g., from map analysis programs) or providing data (e.g., to page layout programs).

This initial version of CIS includes data on all varieties of maps, including Human Genome Mapping Workshop, in situ, genetic, physical, radiation, and specific cell-line maps. Also included are data on (1) different varieties of genomic markers, such as loci, probes, primers, and sequence tagged sites and (2) related reference information, such as people, citations, and database references.

Molecular biologists interact with CIS through graphical user interfaces that make the underlying databases transparent. Thus, users can access and manipulate biological data without needing to know the database structure and implementation. Queries for retrieving maps can be formulated by selecting a chromosomal region from the screen or by choosing desired map attributes from a list of available data.

The existing version of CIS has a three-layer architecture: (1) a graphical user interface, (2) a database, and (3) an intermediate software layer that implements the translation from the user-interface biological perspective to the relational database model. Separating the user level (where actions on the biological data are expressed) and the database level (where data are stored) enhances the system's flexibility and functionality.

To model biological data at a level of abstraction close to the biologists' view of this information, we used an extended entity relationship model to analyze and organize different types of biological objects and their interrelationships. ERDRAW, a specialized graphical editor, was used directly by the information designer to describe biological objects and to construct the database's conceptual schema. SDT, a schema translator, automatically produced the lower-level schema definitions used by commercial relational database management systems. The current implementation uses the Sybase database management system on a Sun SPARCstation.

Database Tools Development

Victor M. Markowitz,*,** Arie Shoshani,* and Ernest Szeto
*Data Management Group and **Human Genome Center, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-6835, FTS 451-6835, Fax 510/486-4004, FTS Fax 451-4004, Internet: "v_markowitz@lbl.gov"

Database tools developed at Lawrence Berkeley Laboratory address two important problems regarding the use of commercial database management systems (DBMS). First, the cost of database development and maintenance increases with the number of code lines required to define and query the database. For commercial DBMSs the amount of code is usually very large; definition of the Chromosome Information System (CIS) database, for example, consists of over 3500 code lines for the SYBASE DBMS. Only 150 code lines were needed to define the CIS database with the database tools, thus significantly increasing database design productivity.

Second, DBMS database definitions are incomprehensible to most users. Database tool language is closer to the way users describe applications, thus facilitating application design as well as sharing these designs among users. In addition to increased productivity and improved clarity, a third benefit of using database tools is their independence from a particular DBMS vendor. This independence permits the use of several commercially available relational DBMSs and subsequent transfer to object-oriented DBMSs as they become available.

Two database tools, SDT and ERDRAW, are currently available. SDT[1,3] is a database schema definition and translation tool for relational DBMSs. The schema definition interface is based on a version of the Extended Entity-Relationship (EER) model. EER schemas can be specified using either a regular text editor or ERDRAW,[2] a graphical schema editor. Once an EER schema is specified, SDT is used to generate the DBMS schema definition and procedures for maintaining referential integrity constraints. SDT and ERDRAW, available on Sun 4 workstations, generate code for SYBASE 4.0, INGRES 6.3, and INFORMIX 4.0 DBMSs and can be easily extended to additional DBMSs.

We are currently developing database query specification and translation tools that will assist users in specifying interactive queries in terms of objects and then translate them into DBMS (SQL).[4] A first version of these tools is expected to be released in spring 1992.

1. V. M. Markowitz and W. Fang, *SDT 4.1. Reference Manual*, Technical Report LBL-27843, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, Calif., May 1991.

2. E. Szeto and V. M. Markowitz, *ERDRAW 2.2. Reference Manual*, Technical Report LBL-PUB-3084, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, Calif., May 1991.

3. V. M. Markowitz and A. Shoshani, "Representing Extended Entity-Relationship Structures in Relational Databases: A Modular Approach," *Assoc. Comput. Mach. Trans. Database Syst.* (submitted).

4. V. M. Markowitz and A. Shoshani, *Query Specification and Translation Tools. Design Document 1.7*, Technical Report LBL-31155, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, Calif., October 1991.

Laboratory Information Management System (LIMS) For Megabase Sequencing

Victor Markowitz,[*, **] **Tim Hunkapiller,**[***] and **Suzanna Lewis**[**]
*Data Management Group and **Human Genome Center, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-6835, FTS 451-6835, Fax 510/486-4004, FTS Fax 451-4004, Internet: "v_markowitz@lbl.gov"
***Science and Technology Center, California Institute of Technology, Pasadena, CA 91125

A laboratory information management system (LIMS) is being developed at Lawrence Berkeley Laboratory (LBL) to support large-scale DNA sequencing efforts. This system is intended to work with a variety of sequencing protocols and will be readily adaptable as protocols are modified. This project is a collaborative effort between LBL and Leroy Hood's sequencing project at California Institute of Technology.

The proposed LIMS differs dramatically from earlier efforts at other laboratories that relied on relational schema design closely tailored to local protocols and handcrafted forms design. Such systems have proven difficult to adapt as needs changed. Instead, this system will use a protocol editor that permits the biologist-designer to specify graphically the protocol process flow. The process flow diagrams will be mapped automatically into Extended Entity-Relationship (EER) schemas and then into relational database schema definitions.

The protocol specification (and derived database schema) will be used to generate data entry forms for each step. A query interface will allow users to pose queries in terms of protocols. The system will have a library of standard molecular biology protocol procedures and corresponding derived EER schemas.

While our approach entails substantial initial software tool development, the result should be a LIMS that is easily adaptable to a variety of protocols and easy to query, modify, and maintain. LIMS is expected to be useful to a wide range of molecular biology laboratories, including genome centers and other DNA sequencing laboratories.

A Computer System for Access to Distributed Genome Mapping Data

Thomas G. Marr and **Andrew Reiner**
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
516/367-8393, Fax 516/367-8416, Internet: "marr@cshl.org"

A major focus of this project is to develop software that can access multiple databases to retrieve data useful for accelerating human chromosome mapping. These data sources include the molecular sequence databases (Genbank® and European Molecular Biology Laboratory Data Library), which contain information required for the development of (1) oligonucleotides for probing DNA and (2) primer pairs used in polymerase chain reaction–based methods. The software will also provide access to the Genome Data Base and to centers involved in large-scale mapping of the human genome.

Another objective is the development of software that qualitatively integrates mapping data such as (1) markers regionally localized by cytogenetic methods; (2) polymorphic markers ordered by genetic linkage analysis; (3) clones ordered by various fingerprinting methods; (4) fragments ordered by long-range restriction mapping; (5) single genomic fragments or clones that have sequence tagged sites assigned to them; (6) nucleotide sequences; and (7) the associated metadata with available detail, including the submitting investigator's name and location, the source organism, the element's chromosomal source, and the chromosomal location.

Shotgun Sequence Assembly Project

Frank Olken, Eugene Lawler,* Daniel Gusfield,** Terence Speed,* and Tim Hunkapiller***
Human Genome Center, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-5891, FTS 451-5891, Fax 510/486-4004, FTS Fax 451-4004, Internet: "olken@lbl.gov"
*University of California, Berkeley, CA 94707
**University of California, Davis, CA 95616
***California Institute of Technology, Pasadena, CA 91125

We will be developing improved software for assembling shotgun DNA sequence data into contiguous sequences. A central concern of the project is to develop robust, computationally practicable methods of dealing with common errors in sequence data, including chimeric clones.

A Human Genome Database

Peter L. Pearson, Robert J. Robbins, Nina W. Matheson,[†] Victor A. McKusick, and Clare Francomano
Genome Data Base, Johns Hopkins University School of Medicine, Baltimore, MD 21205
[†]410/955-7058, Fax 410/614-0434, Internet: "nwm@welch.jhu.edu"

The Genome Data Base (GDB) stores and distributes nonsequence data relevant to the Human Genome Project. Included in the database are information about human genetic loci (and their polymorphisms, alleles, and allele frequencies), probes and other reagents (such as yeast artificial chromosomes, cosmids, sequence tagged sites, and polymerase chain reaction primer sets), maps (genetic and physical), clinical disorders, citations (including abstracts), and contacts (names, addresses, telephone numbers, and e-mail addresses).

The ability to store varied map data and to compare and integrate different maps will form the intellectual focus of GDB within the genome community for the coming years. Because GDB "publishes" both individually authored maps and consensus maps for each chromosome, GDB contains many maps and map fragments for each chromosome.

The GDB program is divided into three core activities and two research projects, as described below:

Core Activities

1. The Administration Core provides general financial and policymaking support structure for the program and coordinates external advisory group activities.
2. The Informatics Core consists of three sections: product development, computing and network services, and product services. The last activity, which is expanding rapidly as the user base increases, includes a training program, the production of user and technical documentation, and a user help service.
3. The Data Acquisition Core maintains three editorial levels: a central support group (located at Johns Hopkins University), a group of five core editors (located at different sites in the United States) with responsibility for coordinating aspects of nomenclature, DNA polymorphisms, Mendelian phenotypes, and physical and genetic maps; and 50 editors (located worldwide) who are responsible for maintaining information on individual chromosomes. All editorial activities are accomplished by direct online access to GDB.

Research Activities

1. The Map Integration Project develops tools for storing, retrieving, and manipulating map information. In collaboration with Lawrence Livermore National Laboratory, we are exploring the relationship between the type of information needed to construct component maps at the genome center level and consensus maps at the GDB level.
2. The Online *Mendelian Inheritance in Man* (OMIM) Integration Project restructures genetic disease information in OMIM to make it more easily accessible to the scientific community and to establish efficient links with the map information.

GDB provides worldwide online access to the system in Baltimore via Internet and SprintNet (X.25). To facilitate access from sites around the world, read-only nodes have been established in the United Kingdom, Germany, Australia, and Japan. The establishment of nodes at additional locations is under way.

All the data in GDB and OMIM, as well as a variety of standard reports, system documentation, and other materials, are available for direct downloading from our anonymous ftp server "mendel.welch.jhu.edu". The data in OMIM may also be obtained via our WAIS server. Other forms of data distribution are planned, including flat file on CD ROM and tapes and direct database-to-database transfer. Those wishing to develop their own interfaces into GDB data may obtain direct, networked access to our ISQL server. Contact "rrobbins@welch.jhu.edu" for further information.

A set of advisory bodies ensures that the user community and funding agencies have adequate input in determining the policies of GDB.

Information regarding GDB is published regularly in *Human Genome News*, which is distributed without charge six times a year by the Human Genome Management Information System at Oak Ridge National Laboratory (ORNL). To subscribe, contact ORNL at 615/576-6669. Inquiries about GDB may be sent directly to "help@welch.jhu.edu".

GDB was developed using SYBASE (a commercially available relational database system) running on Sun servers and workstations.

Abstracts: Informatics

Foundations for a Syntactic Pattern-Recognition System for Genomic DNA Sequences

David B. Searls
Department of Human Genetics, University of Pennsylvania, Philadelphia, PA 19104-6145
215/573-3107, Fax 215/573-5892, Internet: "dsearls@cis.upenn.edu"

DNA is a language, and as such one should be able to analyze it using the tools and techniques of *computational linguistics*. This project is proceeding along two paths to accomplish this analysis: on the theoretical side, the linguistic nature of biological sequences is being studied using the mathematical discipline of *formal language theory*; on the practical side, tools based on these linguistic principles are being built for sophisticated pattern-matching search and other forms of genomic analysis.

In formal language theory, the *Chomsky hierarchy* is often used to classify languages according to their complexity. Many search algorithms now used for DNA are based essentially on simple pattern specifications called *regular expressions*, which are capable of denoting languages only at the very bottom of the Chomsky hierarchy. More-powerful pattern specification tools called *grammars* have greater range. The so-called *context-free* grammars, for instance, can express the biological notion of an arbitrarily large inverted repeat, which is beyond the scope of a pure regular expression. Phenomena such as direct repeats and RNA pseudoknots require even more powerful *context-sensitive* grammars. Evidence indicates that genes themselves reside at a similar level on the Chomsky hierarchy—similar, in fact, to the formal complexity of human languages in terms of mathematical properties such as *nonlinearity, nondeterminism,* and *inherent ambiguity.* Understanding the linguistic nature of genomic DNA is important because it helps to determine what kinds of questions can be answered about such languages computationally and at what cost; the higher a language in the Chomsky hierarchy, the less manageable in the most general case and the more computational power required to recognize it.

We have found that genomic rearrangements may have marked effects on the fundamental nature of any underlying language. The lower levels of the Chomsky hierarchy, for example, are not *closed* under duplication, inversion, or transposition—that is to say, when these evolutionary operations are applied to sequences in such a language, the resulting language may possibly be promoted in the hierarchy. (This is not the case with simple recombinatory or replicational events.) Thus, evolution by its nature may provide pressure toward increasing linguistic complexity. Another such pressure may arise from mathematical consequences of the *superposition* of multiple levels of information, for example, signals for successive steps in gene expression on the same region of DNA.

On the practical side, we have found that grammars form an excellent basis for computational analysis of the genome. Not only can grammars specify more-sophisticated patterns in theory, but they are well suited to hierarchical, abstracted, structural descriptions of complex features—particularly when the objects being described are model-based and involve dependencies among distant sites. Moreover, grammars represent an intensively studied, well-founded methodology appropriate as a "common language" for declarative descriptions of biological features. Most importantly, an extensive technology exists for using computational tools called *parsers* to recognize and understand genomic sequences in terms of these grammars.

We have identified a class of languages, lying between context-free and context-sensitive in the Chomsky hierarchy, that is adequate to describe all the features we have encountered in DNA. Using the logic programming language Prolog, we have implemented a special-purpose parser based on these grammars to perform pattern-matching search of DNA. With the resulting system we have been able to "rediscover" globin and tRNA genes in genomic sequence data, based not on sequence similarity but on first principles of gene structure, as captured in the grammars. Currently, we are augmenting the grammar system with a graphical interface and new domain-specific features to simplify its use, and we are working to increase the efficiency of the parser. We are also establishing a library of grammars specifically for biological features. Finally, we are collaborating with several other groups to incorporate additional techniques, such as dynamic programming and connectionist algorithms, into the grammar framework; such algorithms are more suitable than grammars for detecting certain types of consensus sequences or rough similarities. However, they can be easily embedded in grammars, which can then provide a high-level framework specifying overall structure and coordinating the application of the lower-level algorithms.

Applying Machine Learning Techniques to DNA Sequence Analysis

Jude W. Shavlik, **Michiel O. Noordewier**,* Geoffrey Towell, Mark Craven, Andrew Whitsitt, Kevin Cherkauer, and Lorien Pratt*
Department of Computer Science, University of Wisconsin, Madison, WI 53706
608/262-7784, Fax 608/262-9777, Internet: "shavlik@cs.wisc.edu"
*Department of Computer Science, Rutgers University, New Brunswick, NJ 08903

The Human Genome Project will produce large amounts of uncharacterized DNA sequences and must develop automated techniques for efficient analysis of this data. Algorithms must be designed to recognize important intragenic regions of DNA sequences and categorize them, allowing researchers to focus their attention on interesting subsets of the data. Objectives of this project are (1) to develop useful automated devices that will recognize important entities in DNA sequences, (2) to produce tools for extending the usefulness of such devices to include additional sequence features, and (3) to advance the state of the art in machine learning.

This research project applies and extends a machine learning algorithm that modifies existing knowledge about biological sequences by considering sample members and nonmembers of the sequence motif being learned. Using this information, the learning algorithm produces a more accurate representation of the knowledge needed to categorize future sequences. Specifically, the Knowledge-Based Artificial Neural Network (KBANN) algorithm maps inference rules, such as consensus sequences, into a neural (connectionist) network. Neural-network training techniques then use the training examples to refine these inference rules. Neural networks represent knowledge in an opaque form; this project will develop techniques for converting the results of the neural-learning algorithm into a form interpretable by humans.

In preliminary investigations, the system learned to recognize promoter regions in *Escherichia coli* DNA sequences; the approach appears more powerful than alternative machine learning algorithms and standard techniques reported in the biological literature. This project will evaluate further the applicability of the KBANN approach to molecular biology, systematically comparing it to alternate techniques and investigating its dependence on the amount of training data and sensitivity to sequence errors. Systems will be trained to recognize additional features of DNA sequences, including complete prokaryotic and eukaryotic genes. These recognizers will be distributed to molecular biologists in a form requiring no knowledge of neural networks for their use.

The project will also extend the existing KBANN algorithm in biology-specific ways. Progress in machine learning can best occur by applying existing techniques to practical tasks, thereby allowing strengths and shortcomings of current technology to become apparent.

Partial Bibliography

M. Noordewier, G. Towell, and J. Shavlik, "Training Knowledge-Based Neural Networks to Recognize Genes," pp. 530–36 in *Advances in Neural Information Processing Systems 3,* ed. R. Lipmann, J. Moody, and D. Touretsky, Morgan Kaufmann, San Mateo, Calif., 1991.

Abstracts: Informatics

Computational Analysis and Support for Extensive Physical Mapping of Genomes

Tom Blackwell, David Balding, Frederic Fairfield, Jim Fickett, Catherine Macken, Karen Schenk, **David Torney**,[†] Burton Wendroff, and Clive Whittaker
Los Alamos National Laboratory, Los Alamos, NM 87545
[†]505/667-7510, FTS 843-7510, Fax 505/665-3493, FTS Fax 855-3493, Internet: "dct.@life.lanl.gov"

Our work has focused on developing optimal statistical methods for mapping with fingerprinted clones, allowing the selection of efficient mapping protocols. We advocate a Bayesian framework for statistical analysis, incorporating available information about fingerprint experiments and the region to be mapped.[1] For the central problem of identifying overlapping pairs of clones, Bayes's rule gives the probabilities of overlap for each pair of fingerprinted clones, allowing flexible construction of contigs. To calculate an overlap probability we require (1) a prior ("no data") probability of overlap and (2) likelihoods of the data under the hypotheses of overlap and nonoverlap. Part 1 depends on the method of clone selection. Part 2 is complex, but we have made useful progress in explicitly writing down likelihoods for data consisting both of fragment lengths and of Southern blot hybridization to restriction fragments. The likelihoods of length data from agarose gels and of hybridization data were estimated from the fingerprint data of 3763 cosmid clones from human chromosome 16. (Note that in overlapping clones the likelihood includes reproducibility.)

Features of the data are used in a computer program evaluating reasonable approximations to the likelihoods given above (yielding the overlap probabilities based on the fingerprint data). Our computer program can be used to determine overlap probabilities for a variety of fingerprint data. We have used this program to predict the efficacy of using different fingerprints when mapping using yeast artificial chromosomes (YACs). To accommodate the latter, we have studied the likelihoods of the fingerprint data when not all the restriction fragments are reliably detected.[1] In the process of mapping human chromosome 16, we have observed that some regions have a characteristic repeat sequence interspersed with unique sequences. For these regions it is efficacious to remove fingerprint fragments from the repeat. Modeling was used to determine the effects of a variety of sequence heterogeneities on the predicated progress for a variety of mapping strategies; the computer programs developed for this modeling have been documented and disseminated.[2]

We have studied the compositional heterogeneities present in GenBank® sequences, thereby refining the isochore model of genomes.[3] The model appears, perhaps surprisingly, to apply to *Escherichia coli* as well as to higher eucaryotes. Since there are four A+Ts in the six-base recognition sites for *ECO*R I and *Hin*d III restriction enzymes used to fingerprint cosmid clones from human chromosome 16, we expect some contigs from A+T–"poor" regions will have larger restriction fragments and these from A+T–"rich" regions will have smaller restriction fragments—clues for map completion. We are exploring new approaches for finding sequence features unlikely to have occurred in a long sequence generated by a simple statistical model. Such features are intrinsically interesting, and cognizance of these features will be useful for mapping. We are studying the process of hybridizing *Alu*-PCR products from YACs to gridded clones, so we can effectively design experiments to achieve closure.

Finally, theoretical predictions for mapping via sequence tagged sites (STSs) were derived.[4] In this protocol, clones are linked together by their STS (unique sequence) content. The predicted mapping progress is seen to depend principally on dimensionless counterparts of the number of clones and STSs used and to depend also on the dispersion of clone length.

1. D. J. Balding and D. C. Torney, "Statistical Analysis of Fingerprint Data for Ordered Clone Physical Mapping of Human Chromosomes," *Bull. Math. Biol.* **53**, 853–79 (1991).

2. K. Sirotkin and J. J. Loehr, "Simulation and Analysis of Physical Mapping," pp. 158–241 in *Computers and DNA*, ed. G. I. Bell and T. G. Marr, Addison-Wesley, Reading, Mass., 1989 (*Santa Fe Institute Studies in the Science of Complexity*, Vol. VII).

3. J. W. Fickett, D. C. Torney, and D. R. Wolf, "Base Compositional Structure of Genomes," *Genomics* (submitted).
4. D. C. Torney, "Mapping Using Unique Sequences," *J. Mol. Biol.* **217**, 259–64 (1991).

Partial Bibliography

R. L. Stallings, D. C. Torney, A. F. Ford, C. E. Hildebrand, and R. K. Moyzis, "Evolution and Distribution of (GT) Repetitive Sequences in Mammalian Genomes," *Genomics* **10**, 807–15 (1991).

R. Stallings, D. C. Torney, C. E. Hildebrand, J. Longmire, L. Deaven, J. Jett, N. Doggett, and R. Moyzis, "Physical Mapping of Human Chromosomes by Repetitive Sequence Fingerprinting," *Proc. Natl. Acad. Sci. USA* **87**, 6218–22 (1990).

D. C. Torney, S. W. White, C. C. Whittaker, and K. R. Schenk, "Computational Methods for Physical Mapping of Chromosomes," pp. 268–78 in *Proceedings of the First International Conference on Electrophoresis, Supercomputing, and the Human Genome*, ed. C. R. Cantor and Hwa A. Lim, World Scientific Publishing, New York, 1991.

S. W. White, D. C. Torney, and C. C. Whittaker, "A Parallel Computational Approach Using a Cluster of IBM ES/3090 600Js for Physical Mapping of Chromosomes," *Proceedings of Supercomputing '90: IEEE Computing Society and ACM SIGARCH, Manhattan, NY, Nov 12–16, 1990.*

Informatics Support for Mapping in Mouse-Human Homology Regions

Edward Uberbacher, Richard Mural,* Eugene Rinchik,* and Richard Woychik*
Engineering Physics and Mathematics Division and *Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6364
615/574-6134, FTS 624-6134, Fax 615/574-7860, FTS Fax 624-7860, Internet: "uber@msr.epm.ornl.gov"

This project provides computational and analytical support for mapping in mouse-human homology regions at Oak Ridge National Laboratory (ORNL). The effort has three immediate priorities:

(1) **Laboratory Information Management System (Electronic Laboratory Notebook)**: A laboratory information management system will be implemented to provide the necessary mechanisms for recordkeeping, sample tracking, routine experimental analysis, data archiving, error tracking, and quality control at individual laboratories involved in the mouse-human mapping effort.

(2) **Mapping Database**: It will be necessary to construct a fully integrated relational database system for mouse chromosome 7 to be used by the ORNL mapping group and to serve as a resource for the genome community. This database will use SYBASE, a standard commercial relational database system. Applications software will be constructed or imported to support forms-based data entry, display, editing, browsing, and printing of the many types of information relevant to the mouse-human physical- and functional-mapping effort. A visually informative X-Windows graphical interface will be designed to allow simultaneous display of multiple types of mapping information, clones, and map objects (e.g., synteny regions, loci, and genes), as well as the ability to reference ES-cell lines, mouse lines, phenotypes, annotations, and information related to available map data. The system will run on a small local area network of UNIX workstations with a central database server and will support outside access to map information, the transgenic mouse database, and ES-cell line data.

(3) **Data Analysis Tools**: We will implement locally a wide variety of analysis tools for such tasks as sequence and contig assembly, primer evaluation, sequence comparison (GCG, Blast), fluorescence in situ hybridization analysis, and interspecific backcross characterization. Additionally, we will provide support for access to the major sequence and mapping database resources in the human and mouse communities. The Gene Recognition and Analysis Internet Link (GRAIL) system, currently under development at ORNL, will be (1) adapted as necessary for characterizing mouse

exons for the proposed phenotypic analysis using targeted gene "knockouts" and (2) used to facilitate the localization and characterization of genes in regions of human-mouse homology as well as potential sequence tagged sites for mapping.

Future Goals: Several less-immediate goals will also be considered. Construction of an image-storage and analysis database system may be particularly useful for archiving, analyzing, and exporting the unique phenotypic images that will be created from gene "knockouts." Systems for storing and manipulating more-standard image types such as gels will also be explored. Additional computer support for instrumentation, automation, and robotics needs will be addressed at the appropriate time.

The ORNL mouse-human mapping effort is in its initial startup phase and in a position to benefit from software developments and experiences at the DOE genome centers. Part of our mission will be to take advantage where possible of existing approaches, standards, and technologies for constructing laboratory information management systems and mapping database and interface systems that have been developed at Los Alamos National Laboratory, Lawrence Berkeley Laboratory, Lawrence Livermore National Laboratory, and the NIH centers. By utilizing existing technologies, we intend to limit de novo development to the more-unique aspects of the mouse-human mapping project.

For more information on the ORNL mouse-human mapping project, refer to the abstracts by L. Stubbs and E. Rinchik and by E. Rinchik and R. Woychik.

An Intelligent System for High-Speed DNA Sequence Pattern Analysis and Interpretation

Edward Uberbacher, **Richard Mural,*** Ralph Einstein, and **Reinhold Mann**
Engineering Physics and Mathematics Division and *Biology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831
615/574-6134, FTS 624-6134, Fax 615/574-7860, FTS Fax 624-7860, Internet: "ube@stc10.ctd.ornl.gov", BITNET: "ube@ornlstc.bitnet"

We are developing systems using artificial intelligence and machine learning to recognize important features in genomic sequence data. We will use these approaches to construct an integrated system for the reliable recognition and assembly of genes from sequence data. To accomplish these goals, we will focus our efforts in two central areas of DNA sequence analysis: (1) the use of machine-learning techniques to improve basic pattern-recognition technology, which is the foundation for identifying sequence regions that correspond to genes and other biologically important features, and (2) the development and implementation of an expert system to facilitate construction of hypothetical models of gene structure and function through the application of recognition tools and the intelligent automated assembly of recognized genetic features.

Our initial application of those methods has been to the problem of locating protein-coding exons in anonymous DNA sequences.[1] The neural-network–based coding-recognition module (CRM) has been extensively tested on both fully characterized sequences from GenBank® as well as anonymous DNA sequences from both mouse and human sources. The output of CRM clearly distinguishes coding and noncoding regions and has a very low noise level. In our initial test set, the system located 90% (71/79) of coding exons of 100 or more bases. Only 20 of the 105 indicated coding exons did not correspond to known coding regions.

As the first step in the construction of our integrated system, the coding-recognition capabilities of CRM have been combined with a rule-based interpreter and user interface to allow the exchange of sequence data over the Internet. Through the Gene Recognition and Analysis Internet Link (GRAIL), users e-mail DNA sequence files to the system and have the analysis returned automatically by e-mail. The current analysis includes potential exon positions, with strand assignment and preferred reading-frame determination, and an evaluation of the statistical quality of each potential exon. Turnaround time for the analysis (depending on network traffic) is generally only a few minutes for a sequence less than 100 kb; either single-or multiple-sequence entries may be submitted in one e-mail message. GRAIL is implemented on a UNIX workstation with certain processes, such as an optional protein database search of the translation of potential exons, automatically ported to an IPSC/860 hypercube. The GRAIL system currently has about 200 users worldwide.

A prototype framework for an expert system for automated gene assembly has been constructed and is being evaluated on characterized DNA. The system automatically uses recognition tools developed for coding regions and splice junctions and performs a preliminary gene assembly. Several tests have been performed to evaluate rules for overcoming inconsistencies in the initial set of recognized features such as missing coding regions or splice junctions. We are now integrating the assembly modules and other components into a system based on the blackboard CLIPS control architecture.

[1] E. C. Uberbacher and R. J. Mural, "Locating Proton-Coding Regions in Human DNA Sequences by a Multiple Sensor–Neural Network Approach," *Proc. Natl. Acad. Sci. USA* **88,** 11261–65 (1991).

Partial Bibliography

J. R. Einstein, E. C. Uberbacher, X. Guan, R. J. Mural, and R. C. Mann, *GAP—A Computer Program for Gene Assembly,* ORNL/TM-11934, Martin Marietta Energy Systems, Inc., Oak Ridge National Laboratory, September 1991.

X. Guan, R. J. Mural, J. R. Einstein, R. C. Mann, and E. C. Uberbacher, "GRAIL: An Integrated Artificial Intelligence System for Gene Recognition and Interpretation," in *Proceedings of the Eighth IEEE Conference on Artificial Intelligence Application,* ed. Bob Werner, IEEE Computer Society Press, Los Alamitos, Calif., 1992 (in press).

Analysis of Sequence Data

Manfred D. Zorn and Marge S. Hutchinson*
Human Genome Center and *Data Management Group, Information and Computing Sciences Division, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-5041, FTS 451-5041, Fax 510/486-4004, FTS Fax 451-4004, Internet: "mdzorn@lbl.gov", BITNET: "mdzorn@lbl"

Data analysis is an important step in the sequencing process at the Lawrence Berkeley Laboratory (LBL) Human Genome Center. To facilitate this step, we provide a software program suite that includes (1) a trace editor (part of the Staden package) to analyze raw sequencing data from automated sequencing machines; (2) project management software for organizing a sequencing project; (3) sequence analysis programs such as BLAST and FASTA; and (4) motif searching routines (prosite) that scan an amino acid sequence for common sequence motifs. The software may be developed locally and modified to meet specific user requirements; derived from academic contributions; or part of commercial packages.

Software support services will be extended to provide a computing resource to the University of California, Berkeley, through a collaboration with the Department of Molecular and Cell Biology and other possible interested parties.

In addition to software, we maintain a variety of different nucleic acid and protein sequence databases, such as GenPept, Swissprot, Protein Information Resource, and Enzyme. We also participate in the evaluation of the Entrez database developed at the National Center for Biotechnology Information, and, in collaboration with Los Alamos National Laboratory, we host a relational GenBank® satellite database.

We also provide user training and support at LBL and at the Human Genome Center and manage the center's computing hardware in cooperation with Computing Services personnel.

will greatly increase the knowledge necessary for adequately teaching the science student in the 1990s. A working understanding of the implications of the Human Genome Project will be required, not only for the sake of pure science but for the social, ethical, and legal issues surrounding the deluge of new information.

The University of Kansas Medical Center, in conjunction with area science teachers, Science Pioneers (a local nonprofit organization to promote science education), the Midwest Bioethics Center, and voluntary genetics organizations, will conduct a 3-year, four-phase national program to prepare selected science teachers to become state "resource" teachers. Teacher participants will be recruited from public, parochial, private, and special schools (such as those for visually or hearing impaired). Resource teachers (50 per year for 3 years) will be chosen for their knowledge, experience, and links with existing teacher organizations.

Using an inquiry orientation and hands-on materials, participants in a series of workshops will update and expand the use of human genetics curriculum materials. These workshops will involve clinical geneticists and genetics counselors certified by the American Board of Medical Genetics as well as individuals with genetic conditions. Discussions will include the social, legal, and ethical implications of the Human Genome Project.

Pathways to Genetic Screening: Patient Knowledge—Patient Practices

Troy Duster and Diane Beeson*
Institute for the Study of Social Change, University of California, Berkeley, CA 94720
510/642-0813, Fax 510/642-8674
*Department of Sociology, California State University, Hayward, CA 94542

The near future will witness the screening and counseling of scores of thousands who have had no previous experience with a genetic disorder. The purpose of this project is to illuminate the processes by which genetic screening and genetic concepts of health and illness penetrate two contrasting communities and become integrated into the health concerns of high-risk families. Our goal is to clarify the cultural frames used to process new genetic information and to explore barriers and bridges to successful genetic intervention.

Because the vast majority of at-risk individuals are not part of clinic populations, this project is designed to reach beyond the clinic into high-risk groups that have not yet taken advantage of screening options. We propose to study high-risk men and women who are in their early reproductive years, compare those who have used genetic services with those who have not, and examine their social networks and extended families. We will focus on issues of privacy, stigma, and discrimination as identified in our earlier research and how these issues are managed within family and institutional networks. Understandings, interest, and responses will be analyzed in two cultural contexts: one in which the disorder is generally recognized as race-linked, and the other in which this association is not part of the popular consciousness. We will explore barriers and potential conflicts as they vary by distance from the experience of genetic disorder.

Findings from this study will assist health policymakers in anticipating issues and in preventing the potential adverse effects of new genetic interventions. This assistance will contribute to culturally sensitive genetic screening and counseling that accomplishes medical objectives and responds to privacy concerns and other interests of the patient population.

Lawful Uses of Knowledge from the Human Genome Project

Frank Grad, Neil Holtzman,* Dorothy Warburton,** and Ilise Feitshans
Legislative Drafting Research Fund, Columbia University Law School, New York, NY 10027
212/854-2685, Fax 212/854-7946
*Johns Hopkins School of Medicine, Baltimore, MD 21205
**School of Physicians and Surgeons, Columbia University, New York, NY 10032

"Lawful Uses of Knowledge from the Human Genome Project," a study of the project's social and legal implications, will be carried out by the Legislative Drafting Research Fund (LDRF).

Part I: "The Right To Know or Not To Know about Personal Genomic Information" will involve the study of available legal protections of these rights of confidentiality. The study will examine the rationale for legal protection of information and seek to balance confidentiality against the value of disclosure for the public or individuals. The risk of discrimination will be weighed against the risk of nondisclosure to spouses and other family members, intended spouses, and appropriate government agencies that need information for sound public health planning and other policy purposes. Should physicians be subject to less restriction when information can protect a spouse or other family member?

Part II: "The Uses of Genomic Information in Public Health Planning" will study the availability of and need to collect such data for planning public health and therapeutic services and for program development. The feasibility of incorporating these findings into legislation to ensure the future flow of information on the prevalence, treatment, and outcome of genetic disease and the development of programs will also be addressed.

Mapping and Sequencing the Human Genome: Science, Ethics, and Public Policy—Development and Distribution of Educational Materials for Use in High School Biology Classes

Joseph D. McInerney, Jenny Stricker, and Katherine Winternitz
Biological Sciences Curriculum Study, The Colorado College, Colorado Springs, CO 80903
719/578-1136, Fax 719/578-9126

The Biological Sciences Curriculum Study, in conjunction with the American Medical Association, will conduct a 16-month project to develop and distribute a 75-page module entitled "Mapping and Sequencing the Human Genome: Science, Ethics, and Public Policy." The module will consist of background material for the teacher (25 pages) and at least 5 instructional activities for the classroom (50 pages). The materials will be designed for use by average students in first-year biology courses (generally taught in the 10th grade), with extensions and elaborations appropriate for advanced students. About 80% of all U.S. high school students, some 2.25 million annually, take introductory biology.

Module content will be determined by a ten-person project advisory committee, including experts in science education, classroom teaching, human genetics and molecular biology, ethics, public policy, and computing. This committee will meet in the first month to help set the conceptual framework for the module, propose specific content for teachers' background material, and suggest classroom instructional activities. The planning meeting summary will be reviewed by the education committees of the American Society of Human Genetics, the National Society of Genetic Counselors, the Council of Regional Networks for Genetic Services, and two independent reviewers (months 2–4). A six-member writing team will convene at BSCS for 2 weeks to prepare experimental materials; BSCS and AMA project staff will supervise the writers and review material (month 5). The experimental materials (months 6-8) will be field-tested and evaluated (months 9-13). The advisory committee will review the field-test data and make final recommendations for revision (month 14), with the final revision to be completed during months 14–16. The module will be distributed free by mail to all 55,000 high school biology teachers in the United States (month 16).

Ward's Natural Science Establishment, Inc., will become the official supplier of the published laboratory and kit materials.

Studies of Genetic Discrimination

Marvin Natowicz
Division of Medical Genetics, Shriver Center, Waltham, MA 02254
617/642-0176, Fax 617/894-9968

The major objective of this study is to assess the significance in our society of discrimination directed against individuals and family members because of real or perceived differences in their genetic constitution. Few studies have been done, and limited data exist regarding this issue. Our preliminary study indicated that genetic discrimination does occur, and it may have a significant effect on the lives of affected people.

The specific aims of the present study are to (1) determine the particular entities, such as insurance companies, government agencies, employers, educational institutions, and the military, that might engage in discriminatory practices and (2) evaluate the nature of the discrimination and determine the underlying basis for it. We are interested in learning whether this discrimination is the result of ignorance or of systematic policy.

This study will be carried out using a case history analysis of discriminatory practices against individuals and families in the areas listed. Questionnaires will be distributed to persons who have or are at risk for certain single-gene disorders and to asymptomatic heterozygotes for other conditions; detailed follow-up interviews will be conducted when appropriate. Specific disorders were selected for the following reasons: (1) the genetic basis of the chosen conditions is well established; (2) the chosen conditions are not associated with disfigurement or disability, so discrimination would most likely be based on genetic concerns; and (3) individuals with these disorders can easily be contacted through existing support groups.

Genetic information and techniques discovered through the Human Genome Project will raise difficult and important ethical and social policy issues. The successful completion of this study will provide substantial new data regarding the little-studied problem of genetic discrimination.

"Medicine at the Crossroads"

George Page and **Stefan Moore**
WNET/Thirteen, New York, NY 10019
212/560-2767, Fax 212/582-3297

"Medicine at the Crossroads," an important public television series relevant to the DOE Human Genome Program, is being developed in New York by WNET/Thirteen, the major producer of scientific documentaries for the Public Broadcasting System (PBS) and more than 300 member stations. Planned as the PBS primetime centerpiece for the winter of 1992–93, the eight-part series will explore tensions between advances in bioscience and technology and the needs of people and society. The series objective is to illuminate the scientific, social, and philosophical meaning of medicine in people's lives.

Because human genome research will profoundly affect the practice of medicine, public accessibility to the scientific basis of the Human Genome Project and its ethical and legal implications is critical. A key element in project success is the creation of public awareness, understanding, and support for mapping and sequencing the human genome. "Medicine at the Crossroads" can play an integral role in educating the public about genetic knowledge and how its application can aid in coping with illness in relation to environmental risk and workplace safety.

Genetic Data and Privacy: A Search for Model Legislation

Phillip Reilly
Shriver Center, Waltham, MA 02254
617/642-0222, Fax 617/894-9968

The Human Genome Project holds enormous promise for scientific and medical progress. The public, however, is largely unprepared for the anticipated avalanche of new genetic information and its myriad applications. At present there are no safeguards to discourage injudicious applications. A public policy must be formulated to maximize the benefits and minimize the risks of this project.

The purposes of our project are (1) to conduct a thorough survey of existing state and federal legislation (and selected local laws) to identify statutes and ordinances that attempt to establish genetic confidentiality or that could be interpreted as so intended by a judge; (2) to survey the 50 state legislatures to identify, compile, and analyze all bills regulating access to or use of genetic data; (3) to survey major insurers and state insurance commissioners concerning their views on the need to use genetic information and the need to protect it from abuse; and (4) to draft—or to argue against drafting, depending on the results of legislative analysis—model legislation on how to protect the privacy of personal genetic information.

Impact of Technology Derived from the Human Genome Project on Genetic Testing, Screening, and Counseling: Cultural, Ethical, and Legal Issues

Ralph W. Trottier, **Lee A. Crandall**,** David Phoenix,* Mwalimu Imara,* and Ray E. Mosley**
Department of Pharmacology and Toxicology and *Department of Community Health and Preventive Medicine, Morehouse School of Medicine, Atlanta, GA 30310
404/752-1711, Fax 404/755-7318
**Department of Community Health and Family Medicine, University of Florida, Gainesville, FL 32610

This research project will explore a variety of complex and interrelated issues involving differences among state-supported genetic testing, screening, and counseling programs, with particular concentration on Georgia and Florida. It will assess factors that influence legislation determining which conditions require mandatory screening. The project will also examine the potential implications and impact of genetic screening-policy variations on a mobile society of broad ethnic diversity. It will address issues of access and availability of genetic services in various locales, the nature of counseling provided through the state programs, and the preparation and training of counselors. The project will explore the extent to which "genetic counselors," nationally certified or not, are trained in the field of counseling and in the science and medicine of genetics. Of particular importance to this research is the core essence of confidentiality of genetic information.

Human Genetics and Genome Analysis: A Practical Workshop for Public Policymakers and Opinion Leaders

Jan Witkowski, David A. Micklos, and Margaret Henderson
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
516/549-0507, Fax 516/549-0672

Cold Spring Harbor Laboratory plans to conduct workshops for nonscientists concerned with the societal implications of data usage from human molecular genetics research. Participants will include public policymakers (e.g., congressional staff), civic leaders (e.g., representatives of communities or public interest groups with special concern about applications of DNA technology and related health or legal issues), program officers and planners from health-related foundations, and science journalists. Three-day workshops will be offered twice yearly for 3 years. Selection of participants will be based solely on their interest or involvement in the Human Genome Project and will include critics as well as proponents of the practical applications of human molecular genetics data. The program objective is to expand the knowledge base of community and public policy leaders to enable more effective analysis of social issues arising from data produced in human genome research.

The workshops will be modeled on the annual Banbury conferences held for congressional staff and science journalists and on the practical laboratory classes taught at the DNA Learning Center. They will include lecture/discussions on the scientific basis of human genetics and recombinant-DNA techniques and an in-depth review of past and future uses of genetic technology. Since the most effective way to convey the logic of human molecular genetics to nonscientists is with hands-on experience, participants will have the opportunity to perform some of the standard experiments from the molecular biologist's tool box.

The workshops will be held at the Banbury and DNA learning centers of Cold Spring Harbor Laboratory.

Functions of the DOE Human Genome Program Principal Scientist

Charles R. Cantor
Lawrence Berkeley Laboratory, Berkeley, CA 94620
510/486-6900, FTS 451-6900, Fax 510/486-5282, FTS Fax 451-5282, Internet: "crcantor@lbl.gov"

A Principal Scientist reports to the Human Genome Program Task Group regarding the responsibility of keeping the program at the leading edge of genome research. A Principal Scientist serves on the Human Genome Coordinating Committee, which makes recommendations on DOE human genome centers and broad scientific policy, including coordinating activities with NIH and foreign genome efforts and assisting DOE OHER in the evaluation of program direction.

The Human Genome Distinguished Postdoctoral Fellowships

Linda Holmes and **Alfred Wohlpart**
Science and Engineering Education Division, Oak Ridge Associated Universities, Oak Ridge, TN 37831-0117
615/576-3192, FTS 626-3192, Fax 615/576-0202, FTS Fax 626-0202

The Human Genome Distinguished Postdoctoral Fellowships were initiated in FY 1991 by the DOE Office of Health and Environmental Research (OHER) to support research by recent doctoral degree recipients on projects related to the DOE Human Genome Program. Fellowships of up to 2 years are tenable at laboratories having research programs supportive of the Human Genome Program. Fellows earn stipends of $35,000 for the first year and $37,000 for the second. Eligible applicants must be U. S. citizens or permanent resident aliens and must plan to receive their doctoral degrees within 3 years of starting their fellowships. Principal investigators at host laboratories must be supported by OHER in the amount of at least $150,000 per year.

Oak Ridge Associated Universities (ORAU), administrator of the fellowships, prepares and distributes program literature to universities and laboratories across the country, accepts applications and convenes a panel to make recommendations for awards, and issues stipend checks to fellows. During the first program cycle in FY 1991, 42 applications were received. The review panel identified ten finalists from which DOE chose award winners. The new fellows received their doctoral degrees in biophysical chemistry, molecular biology, computer science, and bioengineering and are serving their fellowships at four different host laboratories. Deadline for the FY 1993 fellowship cycle is February 1, 1993. For more information or an application packet, contact Linda Holmes at the Science/Engineering Education Division; ORAU, Rm. 45; P.O. Box 117; Oak Ridge, TN 37831-0117; 615/576-4805.

Support of Human Genome Program Proposal Reviews

Amanda Lumley and **James Wright**
Science and Engineering Education Division, Oak Ridge Associated Universities, Oak Ridge, TN 37831-0117
615/576-4811, FTS 626-4811, Fax 615/576-0202, FTS Fax 626-0202, Internet: "lumleya%orau2@cunyvm.cuny.edu"

Oak Ridge Associated Universities (ORAU) provides assistance to the Office of Health and Environmental Research in the technical review of proposals submitted in response to solicitations by the Human Genome Program. ORAU staff members create and maintain a database containing all proposal information, including abstracts, relevant names and addresses, and budget information. This information is compiled and presented to proposal reviewers. Prior to review meetings, ORAU staff members make appropriate hotel and meeting arrangements, provide each reviewer with proposal copies, and coordinate reviewer travel and honoraria payment. Other ORAU support includes assistance with program advertising and preparation of reviewer comments following each review.

Human Genome Management Information System

Betty K. Mansfield, Anne E. Adamson, Denise K. Casey, K. Alicia Davidson, Sheryl A. Martin, Donna B. Stinnett, **John S. Wassom**, Judy M. Wyrick, and Laura N. Yust
Health and Safety Research Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6050
615/576-6669, FTS 626-6669, Fax 615/574-9888, FTS Fax 624-9888, Internet: "bkq@ornl.gov", BITNET: "bkq@ornlstc"

The Human Genome Management Information System (HGMIS), sponsored by DOE at Oak Ridge National Laboratory, is charged with (1) helping to communicate genome-related issues and research to contractors, grantees, and the public and (2) providing a forum for information exchange among investigators in the Human Genome Project. To fulfill these communication goals, HGMIS produces contractor-grantee workshop reports, DOE Human Genome Program reports, a bimonthly newsletter (cosponsored by the NIH National Center for Human Genome Research), and a traveling exhibit on the DOE genome program. HGMIS also supplies support on demand to project administrators in preparing meeting minutes, conducting information searches, writing, and editing. To facilitate progress in the international Human Genome Project, HGMIS disseminates information to individual requestors and refers them to other sources. A textual database of genome-related material is under development.

Workshop reports feature (1) an overview of research presented at the meeting and (2) detailed abstracts of ongoing research. Program reports give background material on the project and on genomics and describe progress in DOE-supported human genome research and development. The newsletter *Human Genome News* features technical and general interest articles, meeting reports, project news, genome events and training calendars, and grant and fellowship announcements. Working with the international Human Genome Organization and Genome Data Base, HGMIS also reports international genome project news and major mapping database information in the newsletter.

HGMIS invites comments and suggestions from the research community and the public about its documents and services, which are available upon request and without charge.

Human Genome Coordinating Committee Administration

Sylvia Spengler
Human Genome Center, Lawrence Berkeley Laboratory, Berkeley, CA 94720
510/486-5874, FTS 451-5874, Fax 510/486-5717, FTS Fax 451-5717, Internet: "sylviaj@violet.berkeley.edu"

Sylvia Spengler is Executive Officer of the Human Genome Coordinating Committee (HGCC). With her staff she provides several services for the DOE Genome Program. These services include making arrangements for and covering Joint DOE-NIH Genome Program meetings; preparing materials for and arranging meetings of HGCC and its working groups and subcommittees; and preparing meeting and workshop reports.

Assistance for Ethical, Legal, and Social Issues Projects

Michael S. Yesley
Los Alamos National Laboratory, Los Alamos, NM 87545
505/665-2523, FTS 855-2523, Fax 505/665-4424, FTS Fax 855-4424, Internet: "yesley_michael_s@ofvax.lanl.gov"

Michael S. Yesley, working closely with Daniel W. Drell at the Human Genome Program Office, coordinates the DOE program on the ethical, legal, and social issues (ELSI) raised by the Human Genome Project. In this role, Yesley is involved in establishing the direction of the DOE ELSI program, serving as liäison with grant recipients and potential grant applicants, reviewing grant preapplications, arranging peer-review and technical evaluations of grant proposals, developing additional activities in support of the program, and representing DOE at meetings of the DOE-NIH Joint ELSI Working Group and other ELSI conferences and workshops.

Development of Micron to Sub-Micron Thickness Electrophoresis Gels to Optimize Resolution in DNA Sequencing Using Resonance Ionization Spectroscopy (RIS)

Heinrich F. Arlinghaus, Willliam A. Gibson, and Norbert Thonnard
Atom Sciences, Inc., Oak Ridge, TN 37830
615/483-1113, Fax 615/483-3316

DNA sequences are usually obtained using electrophoresis gels several hundred microns thick because these gels transport the necessary concentration of fluorescent or radioactive DNA-fragment markers and because the technology is well established. Recent work has indicated that thinner gels can provide better resolution, a reduction in migration time, and adequate fragment separation.

To benefit from these improvements, however, a fast, high- resolution, and sensitive method for detecting the markers on the fragments is necessary. Sputter-initiated resonance ionization spectroscopy (SIRIS) offers these features as well as the ability to use a series of stable isotopes as markers. Thus, all four nucleotides can be identified in a single lane, increasing analysis speed and reducing the need for gel uniformity. SIRIS can be used to measure subattomole quantities of isotope-labeled DNA with excellent lateral and mass resolution and has the potential for orders-of-magnitude improvement in the speed of DNA sequencing. Only the markers in the gel's first few molecular layers are detected; to obtain high sensitivity, therefore, electrophoresis gels must be developed in which the separated DNA fragments are concentrated at the gel surface. The objectives of this project are to (1) develop a method for producing microthin electrophoresis gels with subfemtomole concentrations of isotope-labeled DNA at the surface and (2) characterize and demonstrate the gel's performance using SIRIS.

Advances in microthin gel electrophoresis technology will have a major impact on the cost and effectiveness of both research and practical applications in a wide range of genetic topics. DNA sequencing using RIS to localize multiple stable-isotope-labeled fragments could allow detection of more than 10^7 bases per day. Once established, the technique will be used at national, university, and private laboratories; significant business could develop in supplying instrumentation, gels, and services for high-speed DNA sequencing.

Partial Bibliography

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, G. M. Brown, R. S. Foote, F. V. Sloop, and K. B. Jacobson, "Comparison of Sputter-Initiated Resonance Ionization Spectroscopy (SIRIS) and Laser Atomization RIS (LARIS) to Localize Tin-Labeled DNA," *J. Vac. Sci. Technol.* **A9**, 1312 (1991).

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, F. W. Larimer, R. S. Foote, R. P. Woychik, G. M. Brown, F. V. Sloop, and K. B. Jacobson, "Potential Application of Sputter-Inititated Resonance Ionization Spectroscopy for DNA Sequencing," *Anal. Chem.* **63**, 402–7 (1991).

K. B. Jacobson, H. F. Arlinghaus, M. V. Buchanan, C. H. Chen, G. L. Glish, R. L. Hettich, and S. A. McLuckey, "Applications of Mass Spectrometry to DNA Sequencing," *Genet. Anal. Tech. Appl.*, in press (1991).

K. B. Jacobson, H. F. Arlinghaus, H. W. Schmitt, R. A. Sachleben, G. M. Brown, N. Thonnard, F. V. Sloop, R. S. Foote, F. W. Larimer, R. P. Woychik, M. W. England, K. L. Burchett, and D. A. Jacobson, "An Approach to the Use of Stable Isotopes for DNA Sequencing," *Genomics* **9**, 51 (1991).

Development of an Ultrasensitive Detection System for DNA Sequencing

Frederic R. Furuya, James F. Hainfeld, and Richard D. Powell
Nanoprobes, Inc., State University of New York, Stony Brook, NY 11794-5208
516/632-9235, Fax 516/632-9234

The goal of this project is the development of an economical, efficient, ultrasensitive, and nonradioisotopic method for DNA sequencing with a newly developed water-soluble gold particle; the particle can be reacted with nucleic acids in a specific way and is highly sensitive to amplification with silver developer. This system has demonstrated the potential to detect minute quantities (10^{-20} mol) of target. We anticipate that this method will be adaptable to sequencing by hybridization, multiplex sequencing, and direct observation of DNA by scanning tunneling microscopy.

The gold particle will be applied to model DNA assay systems requiring high sensitivity and specificity and then amplified with silver developer to evaluate the particle's potential as a nucleic acid probe. The system will be evaluated in hybridization assays and gels to determine the best method for fragment detection. Its performance in multiplex sequencing models will also be examined.

Oligonucleotide Libraries for High-Throughput DNA Sequencing

Gerald D. Hurst
Genosys Biotechnologies, Inc., The Woodlands, TX 77381
713/363-3693, Fax 713/363-2212

A proprietary multiple DNA–synthesis instrument will be used to synthesize the library of 3314 nonamer oligonucleotides as specified by D. R. Siemieniak and J. L. Slightom[1] for use in high-throughput genome sequencing by a continuous primer walking approach. The oligonucleotides will be synthesized at a rate of 100 per day, affinity-purified by RP1 reverse-phase cartridge chromatography, and quantitated and aliquoted into a stock library (~99% of the material) and a test library (~1% of each nonamer, or ~1.5 μg/tube, sufficient for 30 sequencing reactions per primer). The test library will be used by Slightom to test his proposed high-throughput sequencing strategy.

[1] D. R. Siemieniak and J. L. Slightom, "A Library of 3342 Useful Nonamer Primers for Genome Sequencing," *Gene* **96**(1), 121–124 (1990).

Site-Specific Endonucleases for Human Genome Mapping

Robert G. Lowery, Kimberly Knoche, and Chuck Oehler
Promega Corporation, Madison, WI 53711
608/274-4330, Fax 608/273-6967

A major limitation in current large-scale genome-mapping methodology is that few tools exist for generating specific DNA fragments in the megabase-size range. Promega Corporation is conducting several Phase I studies to examine the feasibility of developing a family of intron-encoded site-specific endonucleases capable of generating 2- to 100-Mb DNA fragments. These studies include (1) developing methods for expressing the intron-encoded enzyme I-*Ppo* in *Escherichia coli* at levels that would allow further study and eventual commercialization of the enzyme; (2) purifying I-*Ppo* to determine its suitability as a molecular biology reagent; (3) developing a rapid, quantitative assay to determine the initial velocity of I-*Ppo*; (4) optimizing reaction conditions, examining stability, and characterizing other properties of I-*Ppo* relevant to megabase mapping methodology; and (5) assessing the enzyme's activity and specificity in an agarose matrix and its ability to cleave complex genomic DNA.

Phase II research will include studies to develop I-*Ppo* for the intended application, focusing on (1) cutting frequency and specificity, (2) partial digestion and star activity, (3) protein modification through chemical modification and limited proteolysis as needed, (4) isolation of intron-encoded endonucleases with different specificities from other mobile group-I introns.

Upon completion of Phase II research, Promega Corporation anticipates having developed (1) a family of site-specific endonucleases capable of generating 2- to 100-Mb fragments that can be produced on a commercial scale and immediately marketed to laboratories involved in human genome mapping, DNA sequencing, and related research activity; and (2) proprietary technology for developing additional intron-encoded endonucleases with a useful range of cutting frequencies.

SBIR Phase I (New)

The awards for the following projects, made as this report went to press, will take effect July 27, 1992.

Interactive DNA Sequence Processing for a Microcomputer
W. Holt Anderson, MCNC, Research Triangle Park, NC 27709; 919/248-1800, Fax 919/248-1455

Chemiluminescent Multiprimed DNA Sequencing
Irena Bronstein, Tropix, Inc., Bedford, MA 01730; 617/271-0045, Fax 617/275-8581

Rapid, High-Throughput DNA Sequencing Using Confocal Fluorescence Imaging of Capillary Arrays
Jay Flatley, Molecular Dynamics, Sunnyvale, CA 94086; 408/773-1222, Fax 408/773-8343

Low-Cost Massively Parallel Neurocomputing for Pattern Recognition in Macromolecular Sequences
John R. Hartman, Computational Biosciences, Inc., Ann Arbor, MI 48106; 313/426-9050, Fax 313/426-5311

Acoustic Plate Mode (APM) DNA Biosensor
Douglass J. McAllister, BIODE, Inc., Cape Elizabeth, ME 04107; 207/883-1492, Fax 207/883-1482

Spatially Defined Oligonucleotide Arrays
Gordon Ringold, Affymax Research Institute, Palo Alto, CA 94304; 415/496-2300, Fax 415/424-9860

High-Speed Electrophoresis in a Solid Substrate
Darlene B. Roszak-MacDonell, Ransom Hill Bioscience, Romona, CA 92065; 619/789-9483, Fax 619/789-6902

SBIR Phase II

Abstracts

Instrumentation for Automated Colony Processing

Norman G. Anderson and N. Leigh Anderson
Large Scale Biology Corporation, Rockville, MD 20850
301/424-5989, Fax 301/762-4892

The objective of this project is to develop a 70-mm filmstrip technology for automated colony processing that can be used to (1) clone very large numbers of phage, bacterial, or yeast cells; (2) locate clones and plaques by electronic scanning; (3) identify positive (non-blue) clones (where this technique applies); (4) transfer and reposition clones in large rectangular arrays on a secondary 70-mm filmstrip; and (5) transfer the entire array to 70-mm strips of nitrocellulose, nylon, or polyvinylidene difluoride. The technology will also enable processing of these strips to lyse cells or phage, attaching the released DNA to the strips, probing the arrays en masse with DNA probes under different stringency conditions, locating those clones to which the probes hybridize, and identifying and ordering overlapping clones.

Technology and methods have been developed to attach porous spacer strips to 100-ft perforated film and filter strips, allowing them to be wound in reels without contact between surfaces. A method for enabling agar gels to adhere to the film was also developed, as was a machine that allows casting of a gel along the entire film length and the even application of cells or phage over the entire length. The system will interface with a filmstrip-based mass sample storage system and eventually with a new polymerase chain reaction (PCR) machine now under development. The system is designed to clone more than 10^6 colonies or plaques automatically, do large matrix hybridization analyses at different and closely controlled stringencies, and recover filter material for multiple reprobing. Key elements of the system include positive clone identification and storage and methods for producing duplicate sets of large arrays. The system is being designed to run automatically in a remote environment under sterile conditions.

This automated colony processing and hybridization system (CLONEPIC), together with the matching film-based storage system (GENSTOR), the PCR system (GENSYN), and the multiple-parallel oligonucleotide synthesis system (CENTRISYN), will form the key large-scale systems that will help make completion of the genome project feasible.

Increased Speed in DNA Sequencing by Utilizing LARIS and SIRIS to Localize Multiple Stable Isotope Labeled Fragments

Heinrich F. Arlinghaus
Atom Sciences, Inc., Oak Ridge, TN 37830
615/483-1113, Fax 615/483-3316

Major developments are critically needed to improve current slow, labor-intensive DNA sequencing procedures. Advances in sequencing technology will have a serious impact on the cost and effectiveness of research and the practical applications taking place at national, university, and private laboratories. At Oak Ridge National Laboratory, methods have been developed for synthesizing iron and tin labels for oligonucleotides, with other element labels being formulated. Given a fast, sensitive, and selective detection method, large numbers of stable isotopes can be used to label DNA, thereby multiplexing the separation process and providing a new, much faster procedure for localizing DNA after electrophoresis.

In Phase I, laser atomization resonance ionization spectroscopy (LARIS) and sputter-initiated resonance ionization spectroscopy (SIRIS), which use lasers tuned to specific energy levels of the selected element, were used to demonstrate detection of subattomole quantities of iron- and tin-labeled DNA with excellent lateral and mass resolution. Both LARIS and SIRIS can make strong contributions to DNA sequencing and to any other procedures that require detection and localization of DNA or of an oligonucleotide that hybridizes to DNA. Since the ion beam and the laser beam used in the two atomization processes can be focused to 2 to 3 μm, the electrophoresis gel length needed for sequencing can be reduced to 5 cm or smaller from the current 50 to 100 cm. This reduction will lead to shorter electrophoresis times, smaller quantities of required DNA, less background noise, and faster analysis. The use of higher- repetition-rate lasers (to 1 kHz using excimers, to 10 kHz using Cu vapor) should eventually allow detection of more than 10^7 bases per 24 hours.

Partial Bibliography

H. F. Arlinghaus, M. T. Spaar, N. Thonnard, A. W. McMahon, and K. B. Jacobson, "Applications of Resonance Ionization Spectroscopy for Semiconductor, Environmental, and Biomedical Analysis, and for DNA Sequencing," p. 26 in *Optical Methods for Ultrasensitive Detection and Analysis: Techniques and Applications*, Vol. 1435, ed. B. L. Fearey, Soc. Photo-Opt. Instrum. Eng., 1991.

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, and K. B. Jacobson, "Use of RIS to Significantly Increase the Speed of Sequencing the Human Genome," p. 341 in *Resonance Ionization Spectroscopy 1990*, Proceedings of the Fifth International Symposium on Resonance Ionization Spectroscopy and Its Applications, *Inst. Phys. Conf. Series 114*, ed. J. E. Parks and N. Omenetto, Bristol: The Institute of Physics, 1991.

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, G. M. Brown, R. S. Foote, F. V. Sloop, and K. B. Jacobson, "Comparison of Sputter-Initiated Resonance Ionization Spectroscopy (SIRIS) and Laser Atomization RIS (LARIS) to Localize Tin-Labeled DNA," *J. Vac. Sci. Technol.* **A9**, 1312 (1991).

H. F. Arlinghaus, N. Thonnard, M. T. Spaar, R. A. Sachleben, F. W. Larimer, R. S. Foote, R. P. Woychik, G. M. Brown, F. V. Sloop, and K. B. Jacobson, "Potential Application of Sputter-Initiated Resonance Ionization Spectroscopy for DNA Sequencing," *Anal. Chem.* **63**, 402 (1991).

K. B. Jacobson, H. F. Arlinghaus, M. V. Buchanan, C. H. Chen, G. L. Glish, R. L. Hettich, and S. A. McLuckey, "Applications of Mass Spectrometry to DNA Sequencing," *Genet. Anal. Tech. Appl.*, in press (1991).

K. B. Jacobson, H. F. Arlinghaus, H. W. Schmitt, R. A. Sachleben, G. M. Brown, N. Thonnard, F. V. Sloop, R. S. Foote, F. W. Larimer, R. P. Woychik, M. W. England, K. L. Burchett, and D. A. Jacobson, "An Approach to the Use of Stable Isotopes for DNA Sequencing," *Genomics* **9**, 51 (1991).

R. A. Sachleben, G. M. Brown, F. V. Sloop, H. F. Arlinghaus, M. W. England, R. S. Foote, F. W. Larimer, R. P. Woychik, N. Thonnard, and K. B. Jacobson, "Resonance Ionization Spectroscopy of Tin-Labeled DNA: Application to Multispectral/Multiplex DNA Sequencing," *Genet. Anal. Tech. Appl.*, in press (1991).

High-Performance DNA and Protein Sequence Analysis on a Low-Cost Parallel-Processor Array

John R. Hartman and David L. Solomon
Computational Biosciences, Inc., Ann Arbor, MI 48106
313/426-9050, Fax 313/426-5311

The growth of DNA and protein sequence databases continues to accelerate, making methods for searching databases for homologies and biologically meaningful features increasingly impractical when carried out on traditional serial-architecture computers. Considerable work has been done recently to implement these methods on supercomputers and high-end massively parallel processors, but the expense and poor accessibility of these machines place them out of reach of most molecular biologists.

In Phase I of this project, several efficient algorithms for comparing and searching macromolecular sequence data were implemented on a relatively inexpensive single-instruction, multiple-data-stream, parallel-array computer (the AIS-4000, Applied Intelligent Systems, Inc.) The algorithms were evaluated carefully and rigorously with respect to their correctness and performance behavior, and one was found to offer impressive cost and performance advantages over functionally equivalent serial software. The system as presently configured consists of a Sun SPARCstation 1+ host in addition to the parallel computer, which also includes an MC68020 processor that can be used concurrently.

During Phase II, overall system throughput will be maximized using a detailed multiplex optimization strategy designed during Phase I. Also, an X-Window–based graphical user interface following the OpenLook specification will be designed and implemented, and significant new sequence analysis and data management functionality will be developed. An object-oriented design methodology will be employed, and the software will be coded in the C++ language.

The success of this project will result in the commercial introduction of a parallel-processing sequence-analysis workstation with robust capabilities and unprecedented cost and performance characteristics. We believe the availability of very high performance sequence-analysis capabilities at reasonable cost will facilitate a decentralized approach to the Human Genome Project and substantially improve the return realized on research dollars invested.

Development of Chemiluminescence-Based DNA Sequencing Kits and Systems

Christopher Martin, **Irena Bronstein**, John C. Voyta, Rouh-Rong Juo, Brooks Edwards, and Alison Varghese
Tropix, Inc., Bedford, MA 01730
617/271-0045, Fax 617/275-8581

This project's Phase II objective is to develop a set of DNA sequencing test kits that use chemiluminescent detection instead of radioisotopic or fluorescent methods. Phase I investigations demonstrated the feasibility of using the chemiluminescent signal originating from the alkaline phosphatase substrate disodium 3-(4 methoxyspirol[1,2-dioxetane-3,2´-tricyclo[3.3.1.$1^{3,7}$]decan]4-yl) phenyl phosphate, AMPPD, for detecting polynucleotides in blots. Two types of sequencing kits, both based on the Sanger dideoxy technique, will be developed. One of these kits incorporates biotinylated M13 sequencing primers for labeling the polynucleotide DNA sequencing products; the second will use biotinylated dideoxynucleotide triphosphate terminators. The various components of the kits include a polymerase, nylon membrane, buffers, streptavidin-alkaline phosphatase conjugate, dioxetane substrate (either AMPPD, or the newly developed CSPD), and other reagents, as well as detailed experimental protocols to facilitate reproducible chemiluminescence-based DNA sequence data.

In addition, a charged-coupled device–based image camera system will be evaluated for automation of the DNA sequencing data acquisition. Such a system will enable data to be acquired and processed directly from the chemiluminescent blots. Kits will be provided to allow users to assay reagents for alkaline phosphatase contamination, which may severely interfere with the chemiluminescent assay. Advantages of using the chemiluminescent kits include sensitive, rapid, nonisotopic detection; excellent DNA band resolution; stable, inexpensive reagents; and automated signal acquisition.

Projects in the Completed Projects section are either completed or no longer receiving support through the DOE Human Genome Program.

Resource Development

Optimizing Procedures for a Human Genome Repository

William C. Nierman and **Donna R. Maglott**
American *Type Culture* Collection, Rockville, MD 20852-1776
310/231-5559, Fax 301/770-1848

The cloned genes and DNA fragments identified during the Human Genome Project should be stored in a repository and made available to the research community. Such a repository would also establish a set of reference clones to facilitate comparison of data generated from different laboratories.

Repositories of well-characterized cloned human DNA fragments currently exist, but at a much smaller scale than necessary for the Human Genome Project. Procedures used in these repositories cannot be expanded without modification. Methods must be improved to automate DNA preparation; clone verification; data maintenance and analysis; and sample storage, recovery, and distribution. Procedures reducing the amount of sample needed for verification and storage must be perfected. The objective of this project is to establish a pilot repository to evaluate such protocols and instrumentation. Initial emphasis will be placed on automating clone verification by analyzing restriction fragments on a DNA sequencing machine and comparing fragment sizes to those already obtained by depositors. Methods will also be explored to use robotics for DNA preparation; to manage information effectively; to verify clones for which there is no restriction data; and to improve methods of sample storage, retrieval, and distribution. The procedures will be tested through the development and operation of a pilot repository using the contigs of lambda clones identified by Maynard Olson's laboratory for the *Saccharomyces cerevisiae* genome, and chromosome 16– and chromosome 19–specific contigs identified by the Los Alamos and Lawrence Livermore National Laboratories.

Normalized cDNA Libraries

Sherman M. Weissman, Rajendra P. Kandpal, Anand Swaroop, Satish Parimoo, Sankhavaram R. Patanjali, Hartwig Arenstorf, and **David C. Ward**
Department of Human Genetics, Yale University, New Haven, CT 06510
203/785-2677, Fax 203/785-3033

The overall objective of this project is to develop and demonstrate methods for preparing normalized cDNA libraries and using them for gene mapping and mutation detection. A phage vector has been prepared that can be used efficiently to generate single-stranded cDNA clones. A number of human sources have been used to prepare the cDNA libraries. Biotin avidin selection methods have been developed for convenient preparation of subtracted cDNA libraries and are currently being evaluated for their ability to generate selected cDNA libraries containing only cDNAs complementary to selected segments of the human genome. In addition, polymerase chain reaction methodology is being adapted to provide improved methods for preparing selected cDNA libraries and to make chromosome jumping a much more efficient general procedure for long-range genome mapping.

Partial Bibliography

R. P. Kandpal, H. Shukla, D. C. Ward, and S. M. Weissman, "A Polymerase Chain Reaction Approach for Constructing Jumping and Linking Libraries," *Nucleic Acids Res.* **18**, 3081 (1990).

S. R. Patanjali, S. Parimoo, and S. M. Weissman, "Construction of a Uniform-Abundance (Normalized) cDNA Library," *Proc. Natl. Acad. Sci. USA* **88**, 1943–47 (1991).

Mapping and Mapping Instrumentation

Human Chromosome 21: Linkage Mapping and Cloning DNA in Yeast Artificial Chromosomes

Stylianos E. Antonarakis and **Philip A. Hieter**
Center for Medical Genetics, The Johns Hopkins University School of Medicine, Baltimore, MD 21205
301/955-7872, Fax 301/955-0484, Internet: "sea@welchlab.welch.jhu.edu"

The goal of our research is to contribute to the cloning of human chromosome 21 DNA in yeast artificial chromosomes (YACs). Chromosome 21 is the smallest human chromosome and contains about 1.4% of the human genome. Cloning in YACs[1] allows large DNA fragments (100 to 1000 kb) to exist as additional chromosomes in *Saccharomyces cerevisiae*. We used new YAC cloning vectors that facilitate manipulation and mapping of the resulting YACs. DNA from cell line WAV-17 (a mouse-human hybrid with chromosome 21 as the only human material) and from flow-sorted chromosome 21 were used as the starting material. Size-selected DNA from complete *Not* I or partial *Eco*R I digestion was ligated to the vectors, and yeast spheroplasts were transformed in the presence of polyamines to eliminate a bias in favor of smaller DNA inserts. In our initial experiments, YACs were obtained from both DNA sources. The average fragment size was 400 kb from the WAV-17 cell line and 200 kb from the flow-sorted chromosome 21 material.[2]

Mapping YACs in somatic cell hybrids and constructing contigs is accomplished by several methods, including (1) subcloning the ends of the artificial chromosomes and using them as hybridization probes; (2) amplifying the ends of artificial chromosomes by polymerase chain reaction (PCR) using oligonucleotide primers from the vector and from an artificial sequence ligated to the first *Hin*d I, *Sau*3A or *Alu* I site and using the products as hybridization probes or sequence tagged sites (STSs); (3) amplifying by PCR between *Alu* sequences and using the products as hybridization probes or STSs. In addition, a large number of STSs have been developed by sequencing well-mapped cosmid, lambda, and plasmid probes and other polymorphic DNA markers. Finally, the YACs are used as hybridization probes in brain cDNA libraries in order to clone, map, and sequence cDNAs that map on chromosome 21.

[1] Burke et al., "Cloning of Large Segments of Exogenous DNA into Yeast by Means of Artificial Chromosome Vectors," *Science* **236**, 806–12 (1987).

[2] McCormick et al., "Construction of Human Chromosome 21–Specific Yeast Artificial Chromosomes," *Proc. Natl. Acad. Sci. USA* **86**, 9991–95 (1989).

Construction and Correlation of Genetic and Chromosome Breakpoint Maps for Chromosomes X and 17

David F. Barker and Pamela R. Fain
University of Utah Medical School, Research Park, Salt Lake City, UT 84108
801/581-5070, Fax 801/581-6052, Internet: "dfbarker@cc.utah.edu"

To construct high-density genetic maps of chromosomes X and 17, we have isolated restriction fragment length polymorphism (RFLP) markers from appropriate chromosome flow-sorted libraries. Over 75 markers specific for chromosome 17 and 80 markers specific for X have been identified in libraries of small (1 to 15 kb) segments, including the National Laboratory Gene Library project libraries LAOXNLO1, LA17NSO3, LL17NSO1, and LA17NLO1. Marker heterozygosities range from 10 to 60%, as is typical for RFLPs. Each marker's chromosomal localization was confirmed and, in most cases, original probe segments were cloned into plasmid vectors to produce probes with improved signal-to-noise ratios for use in family genetic mapping studies.

All markers have been regionally localized with respect to breakpoints associated with chromosome translocations or deletions as well as with "push-pull" hybrids for chromosome X (provided by Huntington Willard) and radiation hybrids for chromosome X (provided by Robert Nussbaum). Seven intervals were defined on chromosome 17, and 35 intervals on chromosome X.

Genetic mapping studies were performed using these sets of probes on DNA samples from the Centre d'Etude Polymorphisme Humain (CEPH) reference linkage families as well as on samples from families with several disease phenotypes studied in our laboratory, including NF1 and Alport syndrome. Marker maps have been constructed for 94 probes on the X chromosome (averaging 1 probe every 2 cM) and 89 probes on chromosome 17 (averaging 1 probe every 2 cM). The chromosome X map includes 50 probes from our set and 44 probes from the CEPH database. The chromosome 17 map includes 49 probes from our set and 40 probes from the CEPH database.

A number of collaborators have found the localized and mapped genetic marker probes useful for disease studies, as both the X and 17 chromosomes include several important disease-gene loci. Chromosome 17 probes have been useful in defining the region deleted in Miller-Dieker syndrome and in further refining the locations of the Charcot-Marie-Tooth neuropathy locus and an early-onset breast cancer locus recently identified on 17q. The X-chromosome probes have been used for mapping several loci: hypophosphatemic rickets, glycerol kinase deficiency, severe combined immunodeficiency, X-linked mental retardation, Alport syndrome, torsion dystonia, and Bruton's agammaglobulinemia.

Correlation of these genetic marker probe locations with physical chromosome maps and overlapping clone sets providing coverage of these chromosomes will allow rapid identification of DNA segments containing candidate genes for diseases mapped in proximity to the markers. The X-chromosome probes are being physically mapped (by David Ward) by fluorescent in situ hybridization. Probes in the Xq24-ter region have been received (by David Schlessinger) for inclusion in the set of overlapping yeast artificial chromosome clones being developed for that region. Work is also in progress to improve the genetic informativeness of these marker loci by identifying additional sequence microheterogeneity detectable by polymerase chain reaction methods.

A Real-Time Imaging System for Enhancing DNA Hybrization Technologies

Gerald Entine
Radiation Monitoring Devices, Inc., Watertown, MA 02172
617/926-1167, Fax 617/926-9743

One of the most active and challenging fields in molecular biology today is the genetic study of eukaryotic systems. Preliminary but often rate-limiting steps in this area are the (1) DNA transfer/hybridization technologies associated with gene mapping and (2) screening of DNA libraries for gene isolation. These procedures use radiolabeled hybridization techniques and autoradiography to achieve the required detection sensitivity and specificity.

This project addresses the measurement of radioactivity distribution on the transfer membrane. Normally, the autoradiographs taken of these membranes require an exposure of 16 to 24 hours. Because of the uncertainty in sample activity, the autoradiographic step is often iterated to bracket the exposure times and obtain the maximum image sharpness. A newly available, position-sensitive nuclear detector will be used to generate a real-time image of radioactivity distribution on the transfer membrane. This image-detection system will be coupled to a laboratory computer for data storage and analysis.

By this approach, the researcher will not only be able to optimize data without the delays associated with autoradiography but will also be able to carry out the procedure with a much higher level of confidence. This ability will significantly improve the efficiency of the entire procedure and allow more resources to be directed toward the analytical aspects of the work.

Abstracts: Completed Projects—Mapping

Technology Development for Physical Mapping of Human Chromosomes

Glen A. Evans, David McElligott, Gary Hermanson, Licia Selleri, Daniel Kaufman, Mary Saleh, Susanne Maurer, Marco Giovannini, Malek Djbali, Jun Zhao, Caryn Wagner, Greg Huhn, Anthony Romo, Grai Andreason, James Eubanks, Shizhong Chen, Ken Snider, Reece Hart, Maria Celicia Roman, Greg Toliver, and Lisa Leonard
Molecular Genetics Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037
619/453-4100 ext. 279 or 376, Fax 619/558-9513, Internet: "gevans@salk-sd2.sdsc.edu" or "gevans@molly.sdsc.edu"

The overall goal of this project is to develop new technologies that require minimal human effort to construct (1) large-scale physical maps of human chromosomes and (2) complete clone sets. Our mapping strategy includes high-resolution localization of chromosome-specific landmark cosmids by in situ hybridization, followed by the isolation of yeast artificial chromosome (YAC) clones spanning the landmarks. We will improve current techniques involving the individual analysis of each landmark clone and associated YAC by developing automated analysis methods and multiplex methods for YAC clone screening and mapping.

Our specific technological goals are to (1) develop automated chromosome mapping applications involving high-density arrays of YAC clones on nylon or nitrocellulose filters; (2) develop and improve techniques for polymerase chain reaction and hybridization-based YAC library screening for individual clone isolation; (3) develop hybridization-based techniques for detecting overlapping regions in YAC libraries to allow a multiplex approach to rapid, simultaneous screening for multiple YAC clones; (4) develop technology using combined YAC and cosmid clones for rapid construction of contigs greater than 2 Mb; (5) develop methodology to detect genes (expressed sequences in cDNA libraries) within physical maps and to isolate corresponding cDNA clones; (6) develop and improve techniques for determining isolated cDNA clone sequence based on map position; and (7) convert genome databases derived in this laboratory to a multiuser networked environment compatible with computer systems used at DOE national laboratories.

The Separation of Large DNA Fragments with Oscillating Electric and Magnetic Fields

Gunter A. Hofmann
BTX/Biotechnologies and Experimental Research, Inc., San Diego, CA 92109
619/270-0861, Fax 619/483-3817

The accurate and fast separation of large DNA fragments is a crucial technology needed for mapping the human genome. Existing methods such as pulsed-field gel electrophoresis appear to have severe limitations. This project uses a novel approach: Lorentz-force-mediated separation in the form of oscillating electric and magnetic fields.

DNA exhibits a large induced dipole moment at low frequencies, with a relaxation time dependent on the length of the fragment. The planned separation method makes use of the polarizability of DNA fragments by subjecting them to an oscillating electric field with a superimposed perpendicular oscillating magnetic field in a liquid without matrix. The resulting unidirectional Lorentz force moves the DNA fragments perpendicular to both the electric and the magnetic fields with a drift velocity dependent on the DNA polarizability, which depends in turn on the DNA size and configuration.

Phase I studies demonstrated that DNA fragments can be polarized by an oscillating electric field and that a superimposed magnetic field results in a unidirectional Lorentz force and a unidirectional drift velocity. The drift velocity increases with the size of the DNA fragment, in contrast to conventional electrophoresis. For large DNA fragments, the drift velocity is 2 to 3 orders of magnitude higher than the drift velocity observed in experiments in pulsed-field gel electrophoresis. A theoretical model with a porous sphere simulation of the DNA molecule showed promise in yielding some qualitative agreements with the observed experimental dependencies of the drift velocity.

Phase II studies will develop several prototype apparatuses of increasing complexity that will allow the rapid, accurate separation of large DNA fragments. A wider parameter range will be investigated, especially separation at higher frequencies, and optimum operating conditions will be determined. The limits of the crossed-field separation method will be investigated in experiments and theory.

Robust Detailed Mapping of the Human Genome

Leonard S. Lerman, Nashua Gabra, and Eric Schmitt
Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139
617/253-6658, Fax 617/253-8699, Internet: "lerman@fang.mit.edu"

This project continues a recently initiated effort to develop a new approach to human genomic mapping using variations in thermal stability to characterize long sections of DNA. The map topography is defined by domains (DNA sequences ranging from a few dozen bases up to several hundred base pairs) that reflect the local stability of the double helix toward melting equilibrium, with domains of both low and high stability serving as markers. Cloning and restriction endonuclease digestion are not required for this method. We expect the domain maps to be substantially more informative than restriction maps, less subject to confusion from minor sequence variations, independent of cloning difficulties, and capable of indicating similarities between genes and perhaps between genes of related species.

Partial Bibliography

S. G. Fischer and L. S. Lerman, "Separation of Random Fragments of DNA According to Properties of Their Sequences," *Proc. Natl. Acad. Sci. USA* **77**, 4420–24 (1980).

L. S. Lerman, S. G. Fischer, I. Hurley, K. Silverstein, and N. Lumelsky, "Sequence-Determined DNA Separations," *Annu. Rev. Biophys. Bioeng.* **13**, 399–423 (1984).

An Image-Acquisition and Processing System for the Analysis of Fluorescence from Stained DNA Gels

Ronald A. McKean
KMS Fusion, Inc., Ann Arbor, MI 48106
313/769-8500

Current methods that use fluorescence techniques to analyze stained DNA gels are inadequate because the test results (1) cannot be easily accessed by a computer and (2) cannot be precisely reproduced. The lack of available instrumentation to convert a fluorescing image into a digitized record for computer entry and the lack of any technique for standardizing analyses performed under varying conditions severely limit the usefulness of this analysis. Overcoming these problems is essential as the need increases for an efficient means of performing both the analysis and the statistical review of the results.

The goal of this project is to develop an instrument that will digitize stained DNA directly from agarose gels, standardize the results, and provide statistical analysis features. This instrument will quickly scan the gel with an optical system capable of high photometric resolution as well as very high spatial precision. The digitized data will then be electronically preprocessed (e.g., background subtraction, filtering, data compression) to ensure correct, noise-free data structured for storage with minimal memory requirements. The stored data, then available for use in imaging the gel, will produce statistical comparisons and generate graphical displays. The data will be archivable using media such as magnetic tapes or disks.

Phase I efforts have been highly successful. The results demonstrated (1) acquisition of high-quality image data directly from agarose gels without elaborate or high-cost components, (2) feasibility of standardizing DNA results under varying electrophoretic conditions, and (3) preliminary optical and electronic designs for this instrument. Phase I results thus show the feasibility of the instrument that will solve many problems associated with agarose gel analysis of DNA.

New Recording Media for an Automated Genome Mapping System

George M. Storti
Quantex Corporation, Rockville, MD 20850
301/258-2701, Fax 301/258-9871

New recording media for an automated genome-mapping system is the goal of this project. The media are different formulations of Quantex's electron-trapping (ET™) materials that have the capability of recording beta particle events and events produced by visible light of relatively short wavelength. Consequently, the presence of radioactively tagged and dye-tagged DNA fragments can be recorded, and event information can be integrated. Readout will be performed by an X-Y scanning system that allows for large signal-to-noise ratios. Compatibility of a scanning system with large-area DNA gels will be evaluated.

Sequencing

DNA Sequencing by Hybridization on Surfaces: Development of Ultrasensitive Two-Dimensional Detectors

Heinrich Arlinghaus
Atom Sciences, Inc., Oak Ridge, TN 37830
615/483-1113, Fax 615/483-3316

The need to map and sequence DNA in the human genome is of crucial importance for a better understanding of genetics and disease processes. Current DNA sequencing procedures almost universally require gel electrophoresis of DNA fragments, and several laboratories are devising procedures to automate as many steps of this process as possible.

Sequencing by hybridization (SBH), a technique planned for this project, does not employ electrophoresis but instead utilizes an array of short oligonucleotides representing all possible sequences in DNA that hybridize to the unknown fragment of DNA to be sequenced. The sequence of the unknown fragment can be reconstructed by computer methods. Attaching the array to a surface in a predetermined pattern, the DNA sequencing matrix, leads to a highly efficient means of sequencing large portions of genomic DNA. Developing this technique will require cost-effective methods for detecting surfacebound DNA with high sensitivity, selectivity, spatial resolution, and high analysis rate of the multiple sites within the array of oligonucleotides.

During Phase I, the most promising detection techniques for positively identifying hybridized and unhybridized sites on a DNA sequencing matrix will be evaluated experimentally. Test matrices onto which octomer (or larger) oligonucleotide sequences have been attached will be analyzed by SIRIS (sputter-initiated resonance ionization spectroscopy), LARIS (laser atomization resonance ionization spectroscopy), and LEF (laser-excited fluorescence). The effort will be directed toward defining an optimal detection system for achieving minimization of matrix size and analysis time while maintaining high accuracy. Recommendations will be made for the most fruitful DNA sequencing directions during Phase II, and a detection system dedicated to the study of DNA SBH will be specified.

Scanning Tunneling Microscopy of DNA

Rodney L. Balhorn, Wigbert Siekhaus,* Michael Allen, and Mehdi Balooch
Biomedical Sciences and *Chemistry and Materials Sciences Divisions, Lawrence Livermore National Laboratory, Livermore, CA 94550
510/422-6284, FTS 532-6284, Fax 510/422-2282, FTS Fax 532-2282, Internet: "rod_balhorn.bio_al@biomed.llnl.gov"

Researchers at Lawrence Livermore National Laboratory and other institutions have recently shown that scanning tunneling microscopy (STM) can be performed on the DNA molecule with angstrom resolution. In addition, STM in the spectroscopic mode (scanning tunneling spectroscopy) has been used to characterize the electronic structure of semiconductor substrates and their interaction with molecules adsorbed on such substrates. The goals of this project are to (1) develop the instrumentation and techniques required for imaging naked double- and single-stranded DNA fragments at or near atomic resolution using STM and (2) devise methods for obtaining spectroscopic information with STM that allow us to distinguish between the four bases and to sequence DNA.

To accomplish these goals, the four bases and various single- and double-stranded DNA fragments are being imaged after deposition onto graphite and other substrates. Images of the individual bases adenine and thymine have recently been obtained at atomic resolution, suggesting for the first time that a scanning probe microscopy technique can discriminate between purines and pyrimidines.
To identify individual bases, free bases are measured and analyzed. Synthetic single- and

double-stranded DNA sequences of known length are made with specific molecular tags so that various types of spectroscopy (work function, laser-enhanced vibration, and photon emission) can be performed on the molecules.

Our application of STM and scanning tunneling spectroscopy to DNA analysis and sequencing should directly impact the progress of the human genome effort by eventually providing a new, electronic method for sequencing DNA at a rate 3 orders of magnitude faster than existing methods. The techniques and instrumentation developed during the course of this project will also be directly applicable to the analysis of biological samples in general.

Transposon-Facilitated DNA Sequencing

Douglas E. Berg, Claire M. Berg,* and Henry V. Huang
Department of Molecular Microbiology, Washington University Medical School,
St. Louis, MO 63110-1093
314/362-2772, Fax 314/362-1232, Internet: "berg@ borcim.wustl.edu"
*Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269

Transposons facilitate DNA sequencing by serving as mobile binding sites for DNA sequencing primers, eliminating the need for traditional but relatively inefficient shotgun and primer-walking methods for analyses of cloned DNA fragments. In collaboration with Linda Strausbaugh (see Strausbaugh and Berg abstract), we are developing different transposon-based approaches for λ and cosmid clones to exploit the different biological properties of these two vector systems. In addition, polymerase chain reaction (PCR) methods are being optimized to increase the effectiveness of transposon-based sequencing strategies.

The strategy for sequencing λ phage clones entails intermolecular transposition of Tn5 and γδ derivatives that can be selected simply by plaque formation on an appropriate host strain. We are now using *supF* (suppressor tRNA) as a marker that can be selected on a *dnaB*-amber *Escherichia coli* host. Other markers that might allow transposon-containing phage to make larger, more easily detected plaques are also being investigated.

To facilitate DNA sequencing in cosmids, we are testing "deletion-factory" cosmid cloning vectors, whose critical features include a derivative of transposon γδ, adjacent contraselectable *sacB* (sucrose-sensitivity) and *rpsL* (streptomycin sensitivity) genes, a *cos* site, and an SP6 or T7 phage promoter. Intramolecular transposition of γδ, selected using *sacB* or *rpsL*, results in nested sets of deletions beginning at the γδ end and extending to random locations in the target cloned DNA. We will be investigating the feasibility of using size-fractionated DNA fragments from pools of deletion-containing plasmids to obtain a more efficient sequence of all cloned DNA regions of interest. We will also be constructing and testing a cassette containing γδ and *sacB* or *rpsL* that can be efficiently recombined in vivo into preexisting cosmid clones for conversion to deletion-factory plasmids.

To achieve transposition efficient enough for automation, we must be able to obtain many transposition products from individual microtiter plate wells and to reduce the fraction of "jackpots" reflecting transposition events early during culture growth. Multiplex derivatives of the Tn5 and γδ transposons are being designed to increase the efficiency of handling many transposon-derived template samples. In addition, the Tn5 and γδ transposase genes are being put under strong, highly regulated promoters; other conditions that increase the frequency of transposition are also being sought.

We are using PCR to map transposon insertion sites rapidly and to prepare high-quality sequencing templates by easily automated procedures. Amplifying phage DNA segments by PCR for distances of 5 to 8 kb is now routine in our laboratories, but there is a great need for amplification with fidelity for considerably longer distances. Accordingly, transposon insertions at appropriate λ DNA positions are being used to screen for enzymes and for conditions that allow the longest possible PCR amplification.

Conventional "direct PCR" is carried out on a single template molecule. In contrast, "crossover" or "biparental" PCR entails amplifying a DNA segment between primer binding sites in two different DNA molecules; crossover PCR had previously been considered an undesirable artifact. We are exploiting crossover PCR between phage that contain different transposon insertions to extend the reach of direct PCR amplification methods and to generate useful DNA sequencing templates. As in the case of direct PCR, transposon insertions at appropriate positions in λ DNAs are being used to screen enzymes and conditions that maximize the distance and fidelity of crossover PCR.

Our analyses showed that more than 90% of transposon Tn5*supF* insertions in λ clones are in one orientation relative to the λ arms. This unexpected pattern causes most crossover PCR fragments to have different transposon ends and thus different primer binding sites at their left and right termini. As a result, these crossover PCR fragments are immediately useful for DNA sequencing.

Our studies have also led to the development of a linear DNA-amplification method for generating readable sequence of about 200 bp directly from single phage plaques or colonies containing multicopy plasmids. This method, which does not require further bacterial or phage growth or DNA preparation steps, is expected to be useful in the generation of sequence tagged sites and in mutational analyses of structure and function of DNA regulatory sites and proteins.

Partial Bibliography

S. Kulakauskas, P. M. Wikström, and D. E. Berg, "Efficient Introduction of Cloned Mutant Alleles into the *Escherichia coli* Chromosome," *J. Bacteriol.* **173**, 2633–38 (1991).

New Dyes for DNA Sequencing

Hee-Chol Kang, James E. Whitaker, Peter C. Hewitt, and **Richard P. Haugland**
Molecular Probes, Inc., Eugene, OR 97402
415/486-5717

We have been actively synthesizing and evaluating sets of new fluorophores, which can be excited by the argon laser at 488 or 514 nm, for possible use in DNA sequencing. The objectives are to synthesize sets of four dyes whose emission spectra have relatively low overlap, whose fluorescence when excited with the argon laser is brighter than currently available fluorophores, and whose properties of ionic charge are uniform for minimum interference with electrophoretic separations.

Principal among the dyes prepared have been fluorescein-rhodamine bifluorophores, in which the energy absorbed by the fluorescein is emitted almost totally at the rhodamine emission wavelength. In examples of these dyes, the energy transfer has been >98% efficient, with pseudo-Stokes shifts of up to 100 nm. Several reactive versions of rhodamine and of rhodol dyes have been prepared that fluoresce when excited by the argon laser and have brighter emission than tetramethylrhodamine. The fourth class of new fluorophores with potential for use in DNA sequencing is reactive, boron dipyrromethene difluoride (Bodipy™) derivatives, which have been prepared in several reactive forms. Probes derived from this fluorophore have unusually narrow emission band width and have high absorbency and quantum yield. The prospects for preparation of new DNA sequencing dyes with higher detectability and spectral resolution will be presented.

DNA Sequence Analysis with Modified Bacteriophage T7 DNA Polymerase

Stanley Tabor, Hans E. Huber, John Rush, and **Charles C. Richardson**
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115
617/432-1864, Fax 617/432-3362

The 3′ to 5′ exonuclease activity of phage T7 DNA polymerase (gene 5 protein) can be inactivated selectively by reactive oxygen species. The chemically modified enzyme is highly processive in the presence of *Escherichia coli* thioredoxin and discriminates against dideoxynucleoside triphosphates (ddNTPs) only four- to sixfold. Consequently, dideoxynucleotide-terminated fragments have highly uniform radioactive intensity throughout the range of a few to thousands of nucleotides in length. There is virtually no background due to terminations at pause sites or secondary-structure impediments in the template. Chemically modified gene 5 protein, by virtue of having low exonuclease activity, has enzymatic properties that distinguish it from native gene 5 protein.

We have exploited these properties to show by a chemical screen that modification of a histidine residue reduces selectively the exonuclease activity. In vitro mutagenesis of histidine 123 and of the neighboring residues results in varying reduction of the exonuclease activity. A deletion of 28 amino acids that encompasses His123 eliminates all exonuclease activity (<10^{-6}%).

Incorporation of ddNTPs by T7 DNA polymerase and *E. coli* DNA polymerase I is more efficient when Mn^{2+} rather than Mg^{2+} is used for catalysis. Substituting Mn^{2+} for Mg^{2+} reduces the discrimination against ddNTPs approximately 100-fold for DNA polymerase I and 4-fold for T7 DNA polymerase. With T7 DNA polymerase and Mn^{2+}, ddNMPs and dNMPs are incorporated at virtually the same rate. Mn^{2+} also reduces the discrimination against other analogs with modifications in the furanose moiety, the base, and the phosphate linkage. The lack of discrimination against ddNTPs using the genetically modified T7 DNA polymerase and Mn^{2+} results in uniform terminations of DNA sequencing reactions, with the intensity of adjacent bands on polyacrylamide gels varying in most instances by less than 10%.

A novel procedure that exploits the high uniformity of bands can be used for automated DNA sequencing. A single reaction with a single labeled primer is carried out using four different ratios of ddNTPs to dNTPs; after gel electrophoresis in a single lane, the sequence at each position is determined by the relative intensity of each band.

Partial Bibliography

S. Tabor and C. C. Richardson, "DNA Sequence Analysis with a Modified Bacteriophage T7 DNA Polymerase," *Proc. Natl. Acad. Sci. USA* **84**, 4767–71 (1987).

S. Tabor and C. C. Richardson, "Effect of Manganese Ions on the Incorporation of Dideoxynucleotides by Bacteriophage T7 DNA Polymerase and *Escherichia coli* DNA Polymerase," *Proc. Natl. Acad. Sci. USA* **86**, 4076–80 (1989).

S. Tabor and C. C. Richardson, "Selective Inactivation of the Exonuclease Activity of Bacteriophage T7 DNA Polymerase by *In Vitro* Mutagenesis," *J. Biol. Chem.* **264**, 6447–58 (1989).

X-Ray Imaging for Intermediate- and High-Resolution Structural Analysis of Biological Objects

James Trebes, Joe Gray,[†*] David Birdsall, James Brase, Rodney Balhorn, and Thomas Yorkey
Human Genome Center, Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA 94550
[†]415/476-3461, Fax 415/476-8218, Internet: "gray@lcaquips.ucsf.edu"
[*]Now at Division of Molecular Cytometry, Department of Laboratory Medicine, University of California, San Francisco, CA 94143-0808

We are developing X-ray imaging for intermediate- and high-resolution structural analysis of biological objects. Initial work focused on high-resolution X-ray imaging for DNA sequence analysis using diffraction techniques with 1.54 Å X rays. This technique will enable us to determine label sequence along individual elements of crystalline DNA. The crystalline DNA in this application is comprised of a large number (~10^{12}) of identical DNA molecules, each treated so that all bases of one type are labeled with an efficient X-ray scatterer. Analysis of four crystals, each having a different base labeled, yields the complete sequence of the target DNA.

An experimental X-ray diffraction apparatus has been constructed, based on a rotating copper anode X-ray source. We have also designed and evaluated software for DNA sequence determination. In addition, arrays of linear DNA molecules have been produced by pulling DNA fibers and by growing liquid crystalline DNA in high-magnetic fields. X-ray diffraction patterns measured for the DNA fibers indicate that individual DNA molecules are oriented with their long axes within 5° of the fiber's central axis.

In the coming years, our goal will be to develop intermediate-resolution X-ray imaging for structural analysis of intact biological cells and organelles. This will be accomplished by applying three-dimensional X-ray holography and/or zone plate imaging for elucidation of such structures as chromatin fibers (300 Å), nuclear pores, and nucleic acid replication complexes in living cells. The X-ray source for these studies, a 44.5 Å X-ray laser, is now under development at Lawrence Livermore National Laboratory.

Informatics

A DNA/Protein Sequence-Analysis Database Accelerator

Peter Alexander
Numerix Corporation, Newton, MA 02164
617/964-2500

Homology searches, optimal sequence comparisons, and studies of RNA secondary structure are all important molecular biology research topics, requiring the availability of powerful database and computational facilities. This investigation will explore the relationships between popular and emerging computerized sequence-analysis algorithms, including DNA/protein sequence mapping and RNA folding programs, and their fundamental computer architectural requirements. A baseline system, configured from commercially available elements, will be analyzed. If implemented, it would provide 100 million instructions per second of equivalent processing power to networked research teams at the equivalent cost of a superminicomputer.

Using Logic Programming for Problems in Molecular Biology

Ross Overbeek
Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439
708/972-7856, FTS 972-7856, Fax 708/972-5986, FTS Fax 972-5986, Internet:
"overbeek@mcs.anl.gov"

The objective of this project is to create flexible, integrated databases and to support rapid prototyping. We are focusing on the utility of logic programming and parallel processing to aid biologists in gathering and interpreting genome sequence data. Three applications of logic programming to genome analysis are being pursued in close collaboration with biologists. In each case, a concerted attempt is made to understand and address the biologists' computational needs and to evaluate the effectiveness of logic programming technology.

1. R. Drmanac and R. Crkvenjakov (originally from Genetic Engineering Center, Belgrade, Yugoslavia) are developing a new technology for large-scale sequencing based on oligohybridization. Their team has now moved to Argonne, where we are conducting computer simulations to help determine appropriate experimental parameters. We are also designing and constructing robots to allow the acquisition of massive amounts of hybridization data. The objective is to reduce significantly the cost of large-scale sequencing while improving the accuracy.

2. The rapid acquisition of data relating to single chromosomes has introduced the need for integration and graphical representation of genetic maps, cytogenic maps, physical maps, and sequence data. We are working with biologists to help develop a general-purpose tool, based on logic programming, for easily incorporating several maps into a single map and graphically examining and manipulating the results. We plan to use this new tool in studies of the *Escherichia coli* chromosome and human chromosome 21.

3. We are supporting the Ribosomal Database Project, directed by Carl Woese and Gary Olsen at the University of Illinois at Urbana. The sequence data being integrated by this project represents one of the largest collections of highly structured and frequently analyzed sequence data in existence; it has been successfully used in studies of phylogeny and RNA secondary structure. We are supporting the effort by helping to design and implement the basic database and by developing tools for extraction and manipulation of the sequence data.

Partial Bibliography

I. Foster and R. Overbeek, "Bilingual Parallel Programming," *Proceedings 3rd Workshop on Parallel Programming and Compilers*, MIT Press, Cambridge, Mass., 1991.

R. Overbeek and I. Foster, "Aligning Multiple RNA Sequences," in *Essays for Bledsoe*, ed. R. S. Bledsoe, Kluwer Academic Press, Dordrecht, Holland, 1991.

S. Winker, R. Overbeek, C. R. Woese, G. J. Olsen, and N. Pfluger, "Structure Detection through Automated Covariance Search," *Comput. Appl. Biosci.* **6**, 365–71 (1990).

C. R. Woese, S. Winker, and R. R. Gutell, "The Architecture of Ribosomal RNA: Constraints on the Composition of Tetra-loops," *Proc. Natl. Acad. Sci. USA* **87**, 8467–71 (1990).

L. Wos, S. Winker, W. McCune, R. Overbeek, E. Lusk, R. Stevens, and R. Butler, "Automated Reasoning Contributes to Mathematics and Logic," pp. 485–99 in *Proceedings 10th International Conference on Automated Deduction*, in *Lecture Notes in Artificial Intelligence*, Vol. 449, ed. M. E. Stickel, Springer-Verlag, Berlin, 1990.

GENCORE: An Automatic Genetic Database Cross Correlator

Stanley Schwartz
CHI Systems, Inc., Spring House, PA 19477
215/542-1400, Fax 215/542-1412

This Phase I project will initiate research to develop an intelligent interface and cross-referencing tool addressing the unique needs of nucleotide/peptide sequence-database developers. GENCORE, a computer system that is an automatic genetic-database cross corrector, will automate the process of generating, implementing, and storing the results of cross-reference searches across multiple sequence databases with incompatible formats. Thus, future database users will have ready access to information about sequence and annotation homologies.

GENCORE provides several innovative features that will enhance the productivity of researchers involved in the Human Genome Project. These features include an expert system with an explanation facility that uses embedded molecular biology and database-specific knowledge to (1) generate search criteria based on user-selected source records and (2) implement a cross-reference search across multiple databases in a single pass. Other features include a specialized, portable, human-factored, user-intuitive interface design that aids in both the specification and modification of search criteria and the definition of fields to store cross-reference results. The design also allows downloading of search results into a local database for future access or convenient display.

Phase I will (1) use cognitive engineering techniques to model the usage environment for GENCORE and design an appropriate interface; (2) use knowledge-acquisition techniques to elicit procedural knowledge from domain experts in order to define and implement molecular database searches and to begin building the rule set and segmenting the knowledge bases; (3) specify a detailed functional component architecture; and (4) plan for the implementation of a full-scale prototype in Phase II.

PC Program for Automated Analysis of Stained DNA Gels

Jeffrey M. Stiegman
BioPhotonics Corporation, Ann Arbor, MI 48106
313/426-8299, Fax 313/426-5311

Electronic image processing of one-dimensional (1-D) or two-dimensional (2-D) electrophoretic gels requires specialized software for analysis. Although several commercially available packages provide generic image enhancement and spatial analysis, a high degree of user modification is necessary to produce useful data for 1-D or 2-D gels. Dedicated software for electrophoretic gels has focused primarily on 2-D protein separation. However, the application of electronic image analysis to 1-D separations of DNA fragments has received relatively little attention.

The goal of the BioPhotonics Corporation project is to develop computer software for capturing, enhancing, and documenting electronically acquired images of 1-D electrophoretic separations of fluorescently stained DNA fragments and for subsequently converting this data to molecular weight and concentration information to be used for statistical analysis and restriction fragment mapping. The emphasis of this project is to automate fully the procedures for locating band positions, determining relative DNA concentration, calculating molecular weights, and analyzing migration abnormalities. This approach will provide improvement in time and accuracy many times over that of manual methods.

A Workstation for Automated Recovery of Unstained DNA Fragments from Gels

Jeffrey M. Stiegman and **Angela Corona**
BioPhotonics Corporation, Ann Arbor, MI 48106
313/426-8299, Fax 313/426-5311

Many applications of electrophoresis involve separation of DNA restriction enzyme fragments, not only for analysis but also for sample isolation and purification. Recovering DNA from electrophoretic gels and purifying it is an initial step in tasks such as developing cloning libraries, DNA sequencing, or sample purification after amplification using the polymerase chain reaction (PCR). Present procedures for extracting DNA from a gel medium are tedious, produce inconsistent results, and consume many hours of a skilled technical staff. Such manually intensive procedures are clearly not suitable for applications requiring isolations of hundreds of DNA fragments per day.

In addition, present techniques for viewing DNA restriction fragments typically require staining with ethidium bromide dye. Ethidium bromide is an inhibitor of many enzymes that are commonly used in subsequent cloning steps. Removing such stains from DNA samples requires additional processing steps for extraction and dialysis.

BioPhotonics has proposed to address these issues by developing a robotic workstation to provide (1) direct visualization of electrophoretic patterns of unstained DNA using electric birefringence as a detection scheme; (2) machine vision technology to locate and excise DNA bands directly from the gel slab; and (3) automated extraction of DNA fragments from the gel matrix using a special electroelution system designed for simultaneous elution of 100 to 150 fragments.

We have demonstrated the use of electric birefringence for recording two-dimensional gel images of unstained DNA restriction fragments down to 4 kb in length and at concentrations of 25 ng/μl. A vacuum aspirator has also been demonstrated that provides clean and precise excision of DNA bands directly from a gel without damage to the fragment and with minimal cross contamination between sampling. We have proposed to extend these developments in a prototype instrument for semiautomated recovery of unstained DNA from electrophoretic gels. Improvements have been identified that should enable recovery of unstained DNA fragments down to 100 bp or less and at concentrations approaching 10 ng/μl with little or no cross contamination between samplings.

We expect the proposed DNA Recovery Workstation to find initial applications in those labs charged with preparing the many thousands of DNA probes required to develop physical maps. Such tasks will benefit from the time savings and quality control offered by the proposed system. Future applications will be found in research and industrial sectors that will take up the task of generating large numbers of DNA probes commercially distributed for diagnostic purposes.

Rapid Genome Analysis on a Workstation with an Associative Coprocessor

Charles D. Stormon, James Brule, and Hamid Bacha
Coherent Research, Inc., Syracuse, NY 13202
315/426-0929

The goal of this research project is to produce a high-performance, low-cost sequence-analysis workstation that will allow genetic researchers to perform complex analyses on genome data. Phase I research involves feasibility studies of processing genome data at a workstation with an associative coprocessor. Many of the DNA sequence analyses required by molecular biologists involve finding similarities between different molecules or within one molecule. Several different approaches, represented by dozens of programs, have been taken to solve this problem. With increasing sequence length (or larger numbers of fragments) to be analyzed, these methods can quickly become computationally impractical. Recently, parallel processors have become available, and they provide an avenue of speedup for sequence analysis.

By implementing a simplified computational model in VLSI (very large scale integrated), our associative Coherent Processor™ (the AP-DS) provides the same degree of parallel processing as either the Connection Machine™ (Thinking Machines, Inc.) or the DAP™ (Active Memory Technology) at a much lower cost. The research in this project will show the feasibility of processing complex genome queries on a workstation-class machine with attached AP-DS coprocessor.

A Novel, Low-Cost System for Automated DNA Sequence Reading

John S. West
BioAutomation, Inc., Bridgeport, PA 19405
215/275-4540

Analytical instruments for the automated measurement of DNA are just emerging. Automated sequence readers commercialized to date are expensive (over $90,000 per instrument). As a result, in more than 10,000 sequencing laboratories worldwide, fewer than 3% have acquired such instrumentation. Because of the high cost, adoption of automated DNA-sequencing instrumentation has been slow, even though interest in the technology is high. First-generation sequence readers also lack the throughput required for large-scale sequencing projects. Technological feasibility of an autoradiography-based approach incorporating the use of application-specific, highly integrated circuits and parallel processing was demonstrated in Phase I. During Phase II, full working prototypes will be developed using this approach. These prototypes will (1) provide a significant increase in accessibility of automated sequence-reading equipment to smaller laboratories, and (2) provide the quantum jump in throughput required by large-scale sequencing projects.

ELSI

Funding Young Investigators in the Biological and Biomedical Sciences

Peggy Fischer
National Academy of Sciences, Board on Biology, National Research Council, Washington, DC 20418
202/334-1552, Fax 202/334-1687, BITNET: "pfische1@nas"

A great deal of concern exists in academic, government, and industrial circles about the increasing numbers of young people who may be deterred by funding difficulties from pursuing careers in academic biological and biomedical research. Because committed young investigators are crucial to the expansion of the fundamental biological knowledge base that ranges from ecology to medicine and fuels academic, industrial, and economic growth, the National Research Council's Commission on Life Sciences proposes to convene a committee of experts to examine the funding issue.

Panel members will specifically address biological and biomedical funding for young investigators by major federal agencies and private organizations, as well as funding problems and constraints; they also will determine the impact of these issues on recruiting and retaining young academic researchers. While the intent of this activity is not to examine the flow of individuals through school and into careers in science, the committee will review this subject where relevant to the study. The resulting analysis and recommendations about ways to improve the funding situation will be included in a report, which will be available to the public in February 1992.

Conference: Justice and the Human Genome

Marc Lappé, **Timothy Murphy**, and Kenneth Vaux
Department of Medical Education, University of Illinois College of Medicine, Chicago, IL 60612
312/996-7528, Fax 312/413-2048

This conference will address the implications of applying the knowledge and data generated by the Human Genome Project. A central theme will be a review of justice issues that arise from the distribution of genomic data and the resulting potential benefits and risks.

The overall conference goal is to provide a basis for equitable distribution of the products of genomic research. Participants will also consider public and private choices made possible by the expansion of genetic knowledge. A legal scholar will address constitutional issues of rights, access, and discrimination. Other speakers will explore past episodes of discrimination and their relevance to possible misuse today. Additional topics are religious perspectives of reproductive ethics, the philosophical roots of entitlement, and the potential use and misuse of genetic information in the corporate and government sectors.

Other ELSI Conferences

Second conference on **Computers, Freedom, and Privacy**; March 18–20, 1992; Washington, D.C.
Lance Hoffman, George Washington University; 202/994-4955

Educational Forum to Focus on Science, Technology, and Ethical Responsibility; June 14–19, 1992; Atlanta, Ga.
Betsy Fader, Student Pugwash USA; 202/328-6555

Genes and Human Behavior: A New Era? October 26, 1992; Boston, Mass.
Jonathan Beckwith, Harvard Medical School; 617/432-1920

Preparing Yeast Artificial Chromosome (YAC) Arrays for Hybridization. *Using a Biomek 1000 robotic workstation, an investigator prepares high-density filter arrays of YAC clones from the NIH Genome Center at Washington University, St. Louis, or from Centre d' Etude du Polymorphisme Humain YAC libraries. The library, maintained as about 60,000 individual clones in 96-well microtiter plates, is manipulated by a robotic system. Some functions performed by the robot to grow YAC clones include arraying clones on filter grids, filling plates with media, and copying the clone library. The robot is also used to prepare high-density filters of cosmid clones for distribution to other investigators. (Photograph provided by Glen Evans, Salk Institute for Biological Studies. For more information, refer to Evans in the Index to Principal and Coinvestigators.)*

Appendices

Appendix A: Primer on Molecular Genetics

Appendix B: Conferences, Meetings, and Workshops Sponsored by DOE

Appendix C: Members of the DOE Health and Environmental Research Advisory Committee

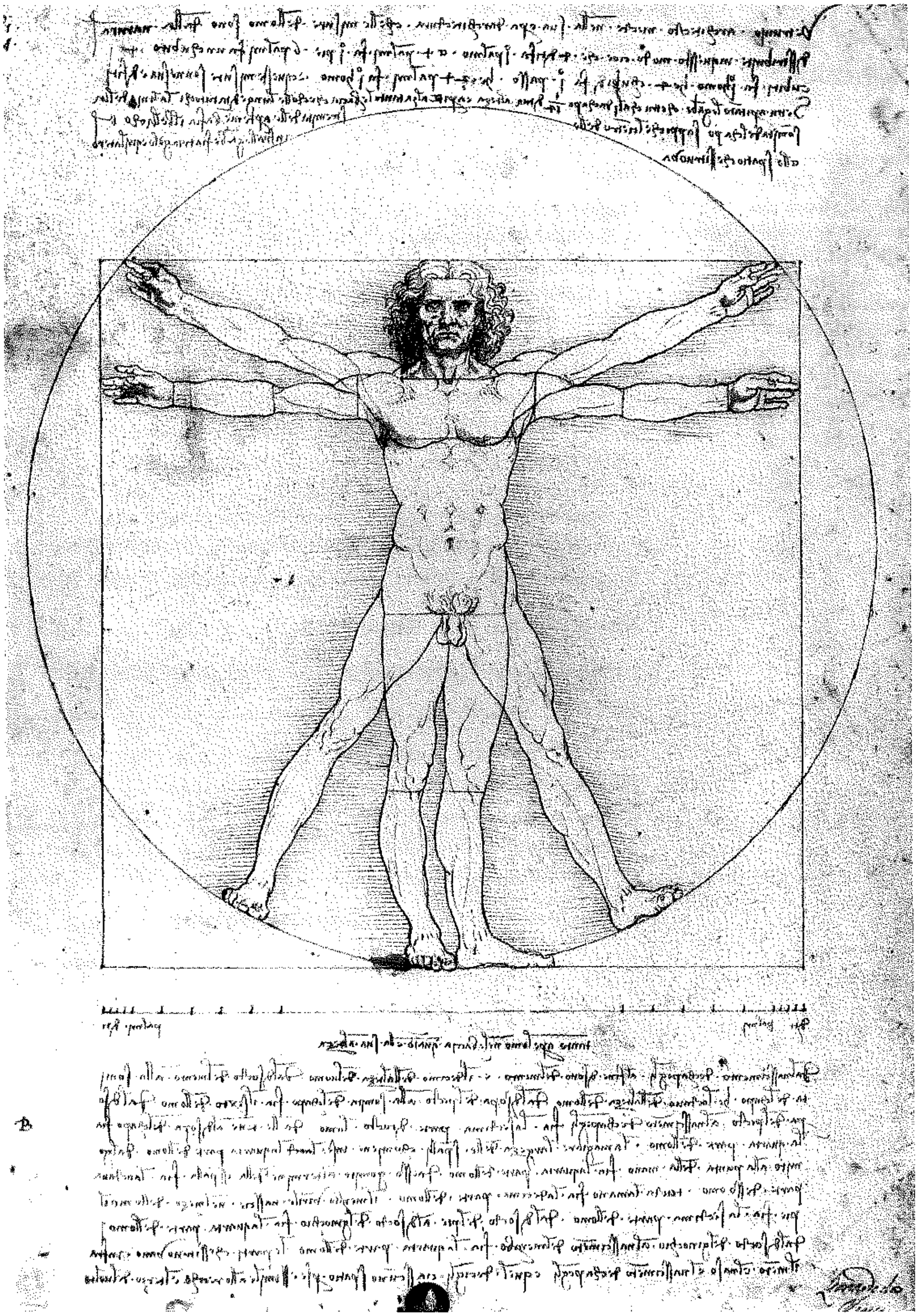
Appendix D: Members of DOE-NIH Joint Working Groups

Appendix E: Glossary

This drawing by Leonardo da Vinci symbolizes the quest for knowledge through exploration of the unknown. In his art, Leonardo concentrated on illustrating fundamental rules governing the physical world to reveal the unity underlying the diversity of nature. Just as the Renaissance brought broadened intellectural horizons and rapid advances in the natural sciences and technology, so will the 21st century, 500 years later, witness a revolution in many sciences as research unlocks the secrets of the molecular structure governing the human body, one of nature's masterpieces.

Primer on Molecular Genetics

Appendix A



Contents

Appendix A: Primer on Molecular Genetics

Revised and expanded by Denise Casey (HGMIS) from the primer contributed by Charles Cantor and Sylvia Spengler (Lawrence Berkeley Laboratory) and published in the *Human Genome 1989–90 Program Report*.

Introduction 193
DNA 194
Genes 195
Chromosomes 196
Mapping and Sequencing the Human Genome 198
Mapping Strategies 199
Genetic Linkage Maps 199
Physical Maps 201
Low-Resolution Physical Mapping 202
Chromosomal map 202
cDNA map 202
High-Resolution Physical Mapping 202
Macrorestriction maps: Top-down mapping 204
Contig maps: Bottom-up mapping 204
Sequencing Technologies 206
Current Sequencing Technologies 211
Sequencing Technologies Under Development 212
Partial Sequencing to Facilitate Mapping, Gene Identification 212
End Games: Completing Maps and Sequences; Finding Specific Genes 213
Model Organism Research 215
Informatics: Data Collection and Interpretation 215
Collecting and Storing Data 215
Interpreting Data 216
Mapping Databases 217
Sequence Databases 217
Nucleic Acids (DNA and RNA) 217
Proteins 218
Impact of the Human Genome Project 218

Introduction

The complete set of instructions for making an organism is called its genome. It contains the master blueprint for all cellular structures and activities for the lifetime of the cell or organism. Found in every nucleus of a person's many trillions of cells, the human genome consists of tightly coiled threads of deoxyribonucleic acid (DNA) and associated protein molecules, organized into structures called chromosomes (Fig. 1).

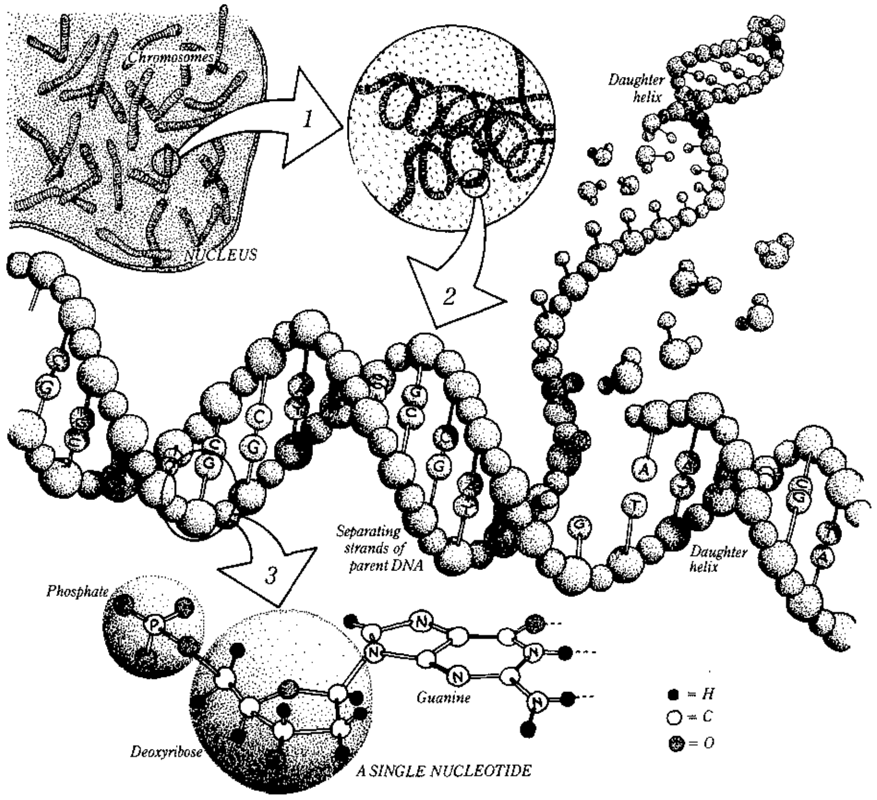


Fig. 1. The Human Genome at Four Levels of Detail. *Apart from reproductive cells (gametes) and mature red blood cells, every cell in the human body contains 23 pairs of chromosomes, each a packet of compressed and entwined DNA (1, 2). Each strand of DNA consists of repeating nucleotide units composed of a phosphate group, a sugar (deoxyribose), and a base (guanine, cytosine, thymine, or adenine) (3). Ordinarily, DNA takes the form of a highly regular double-stranded helix, the strands of which are linked by hydrogen bonds between guanine and cytosine and between thymine and adenine. Each such linkage is a base pair (bp); some 3 billion bp constitute the human genome. The specificity of these base-pair linkages underlies the mechanism of DNA replication illustrated here. Each strand of the double helix serves as a template for the synthesis of a new strand; the nucleotide sequence (i.e., linear order of bases) of each strand is strictly determined. Each new double helix is a twin, an exact replica, of its parent. (Figure and caption text provided by the LBL Human Genome Center.)*

Appendix A: Primer on Molecular Genetics

If unwound and tied together, the strands of DNA would stretch more than 5 feet but would be only 50 trillionths of an inch wide. For each organism, the components of these slender threads encode all the information necessary for building and maintaining life, from simple bacteria to remarkably complex human beings. Understanding how DNA performs this function requires some knowledge of its structure and organization.

DNA

In humans, as in other higher organisms, a DNA molecule consists of two strands that wrap around each other to resemble a twisted ladder whose sides, made of sugar and phosphate molecules, are connected by "rungs" of nitrogen-containing chemicals called bases. Each strand is a linear arrangement of repeating similar units called nucleotides, which are each composed of one sugar, one phosphate, and a nitrogenous base (Fig. 2). Four different bases are present in DNA—adenine (A), thymine (T), cytosine (C), and guanine (G). The particular order of the bases arranged along the sugar-phosphate backbone is called the DNA sequence; the sequence specifies the exact genetic instructions required to create a particular organism with its own unique traits.

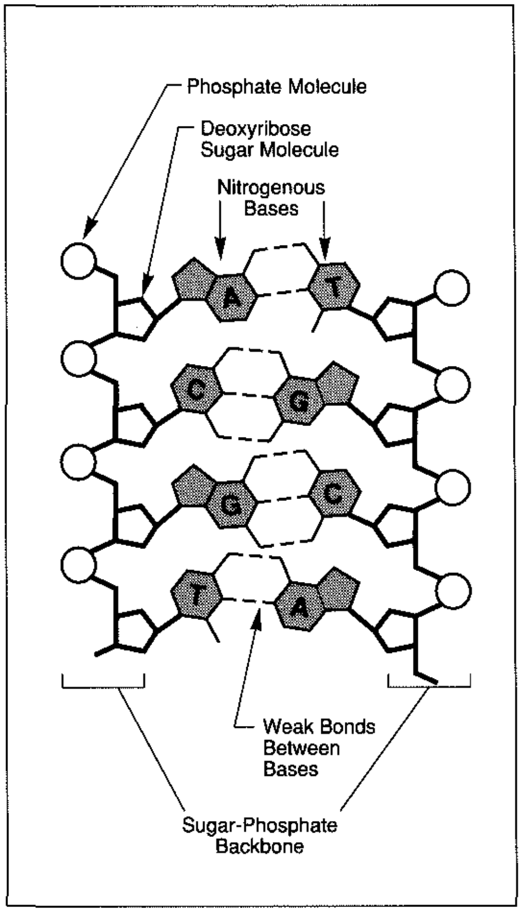


Fig. 2. DNA Structure. *The four nitrogenous bases of DNA are arranged along the sugar-phosphate backbone in a particular order (the DNA sequence), encoding all genetic instructions for an organism. Adenine (A) pairs with thymine (T), while cytosine (C) pairs with guanine (G). The two DNA strands are held together by weak bonds between the bases.*
A gene is a segment of a DNA molecule (ranging from fewer than 1 thousand bases to several million), located in a particular position on a specific chromosome, whose base sequence contains the information necessary for protein synthesis.

The two DNA strands are held together by weak bonds between the bases on each strand, forming base pairs (bp). Genome size is usually stated as the total number of base pairs; the human genome contains roughly 3 billion bp (Fig. 3, p. 195).

Each time a cell divides into two daughter cells, its full genome is duplicated; for humans and other complex organisms, this duplication occurs in the nucleus. During cell division the DNA molecule unwinds and the weak bonds between the base pairs break, allowing the strands to separate. Each strand directs the synthesis of a complementary new strand, with free nucleotides matching up with their complementary bases on each of the separated strands. Strict base-pairing rules are adhered to—adenine will pair only with thymine (an A-T pair) and cytosine with guanine (a C-G pair). Each daughter cell receives one old and one new DNA strand (Figs. 1 and 4). The cell's adherence to these base-pairing rules ensures that the new strand is an exact copy of the old one. This minimizes the incidence of errors (mutations) that may greatly affect the resulting organism or its offspring.

Genes

Each DNA molecule contains many genes—the basic physical and functional units of heredity. A gene is a specific sequence of nucleotide bases, whose sequences carry the information required for constructing proteins, which provide the structural components of cells and tissues as well as enzymes for essential biochemical reactions. The human genome is estimated to comprise at least 100,000 genes.

Human genes vary widely in length, often extending over thousands of bases, but only about 10% of the genome is known to include the protein-coding sequences (exons) of genes. Interspersed within many genes are intron sequences, which have no coding function. The balance of the genome is thought to consist of other noncoding regions (such as control sequences and intergenic regions), whose functions are obscure. All living organisms are composed largely of proteins; humans can synthesize at least 100,000 different kinds. Proteins are large, complex molecules made up of long chains of subunits called amino acids. Twenty different kinds of amino acids are usually found in proteins. Within the gene, each specific sequence of three DNA bases (codons) directs the cell's protein-synthesizing machinery to add specific amino acids. For example, the base sequence ATG codes for the amino acid methionine. Since 3 bases code for 1 amino acid, the protein coded by an average-sized gene (3000 bp) will contain 1000 amino acids. The genetic code is thus a series of codons that specify which amino acids are required to make up specific proteins.

The protein-coding instructions from the genes are transmitted indirectly through messenger ribonucleic acid (mRNA), a transient intermediary molecule similar to a single strand of DNA. For the information within a gene to be expressed, a complementary RNA strand is produced (a process called transcription) from the DNA template in the nucleus. This

Comparative Sequence Sizes	**Bases**
• Largest known continuous DNA sequence (yeast chromosome 3)	350 Thousand
• *Escherichia coli* (bacterium) genome	4.6 Million
• Largest yeast chromosome now mapped	5.8 Million
• Entire yeast genome	15 Million
• Smallest human chromosome (Y)	50 Million
• Largest human chromosome (1)	250 Million
• Entire human genome	3 Billion

Fig. 3. Comparison of Largest Known DNA Sequence with Approximate Chromosome and Genome Sizes of Model Organisms and Humans. *A major focus of the Human Genome Project is the development of sequencing schemes that are faster and more economical.*

mRNA is moved from the nucleus to the cellular cytoplasm, where it serves as the template for protein synthesis. The cell's protein-synthesizing machinery then translates the codons into a string of amino acids that will constitute the protein molecule for which it codes (Fig. 5). In the laboratory, the mRNA molecule can be isolated and used as a template to synthesize a complementary DNA (cDNA) strand, which can then be used to locate the corresponding genes on a chromosome map. The utility of this strategy is described in the section on physical mapping, p. 201.

Chromosomes

The 3 billion bp in the human genome are organized into 24 distinct, physically separate microscopic units called chromosomes. All genes are arranged linearly along the chromosomes. The nucleus of most human cells contains 2 sets of chromosomes, 1 set given by each parent. Each set has 23 single chromosomes—22 autosomes and an X or Y sex chromosome. (A normal female will have a pair of X chromosomes; a male will have an X

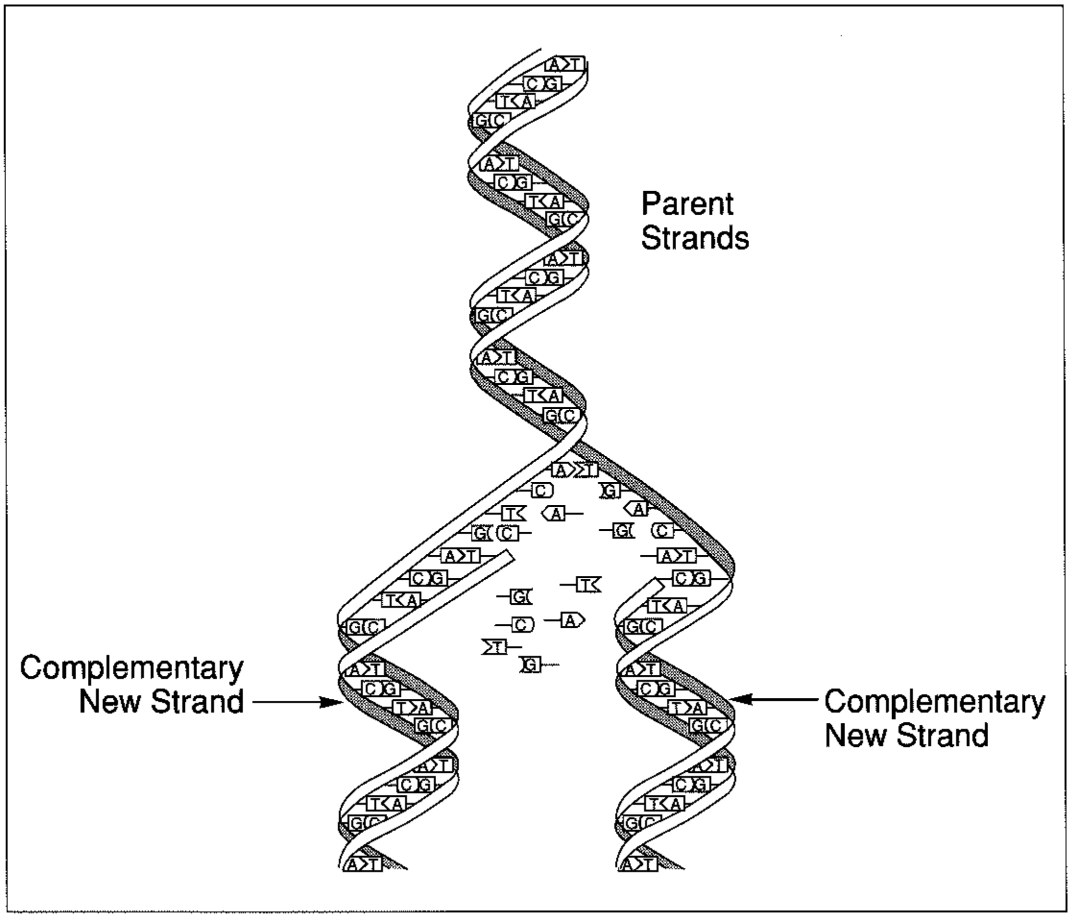


Fig. 4. DNA Replication. *During replication the DNA molecule unwinds, with each single strand becoming a template for synthesis of a new, complementary strand. Each daughter molecule, consisting of one old and one new DNA strand, is an exact copy of the parent molecule. [Source: adapted from* Mapping Our Genes—The Genome Projects: How Big, How Fast? *U.S. Congress, Office of Technology Assessment, OTA-BA-373 (Washington, D.C.: U.S. Government Printing Office, 1988).]*

and Y pair.) Chromosomes contain roughly equal parts of protein and DNA; chromosomal DNA contains an average of 150 million bases. DNA molecules are among the largest molecules now known.

Chromosomes can be seen under a light microscope and, when stained with certain dyes, reveal a pattern of light and dark bands reflecting regional variations in the amounts of A and T vs G and C. Differences in size and banding pattern allow the 24 chromosomes to be distinguished from each other, an analysis called a karyotype. A few types of major chromosomal abnormalities, including missing or extra copies of a chromosome or gross breaks and rejoinings (translocations), can be detected by microscopic examination; Down's syndrome, in which an individual's cells contain a third copy of chromosome 21, is diagnosed by karyotype analysis (Fig. 6). Most changes in DNA, however, are too subtle to be detected by this technique and require molecular analysis. These subtle DNA abnormalities (mutations) are responsible for many inherited diseases such as cystic fibrosis and sickle cell anemia or may predispose an individual to cancer, major psychiatric illnesses, and other complex diseases.

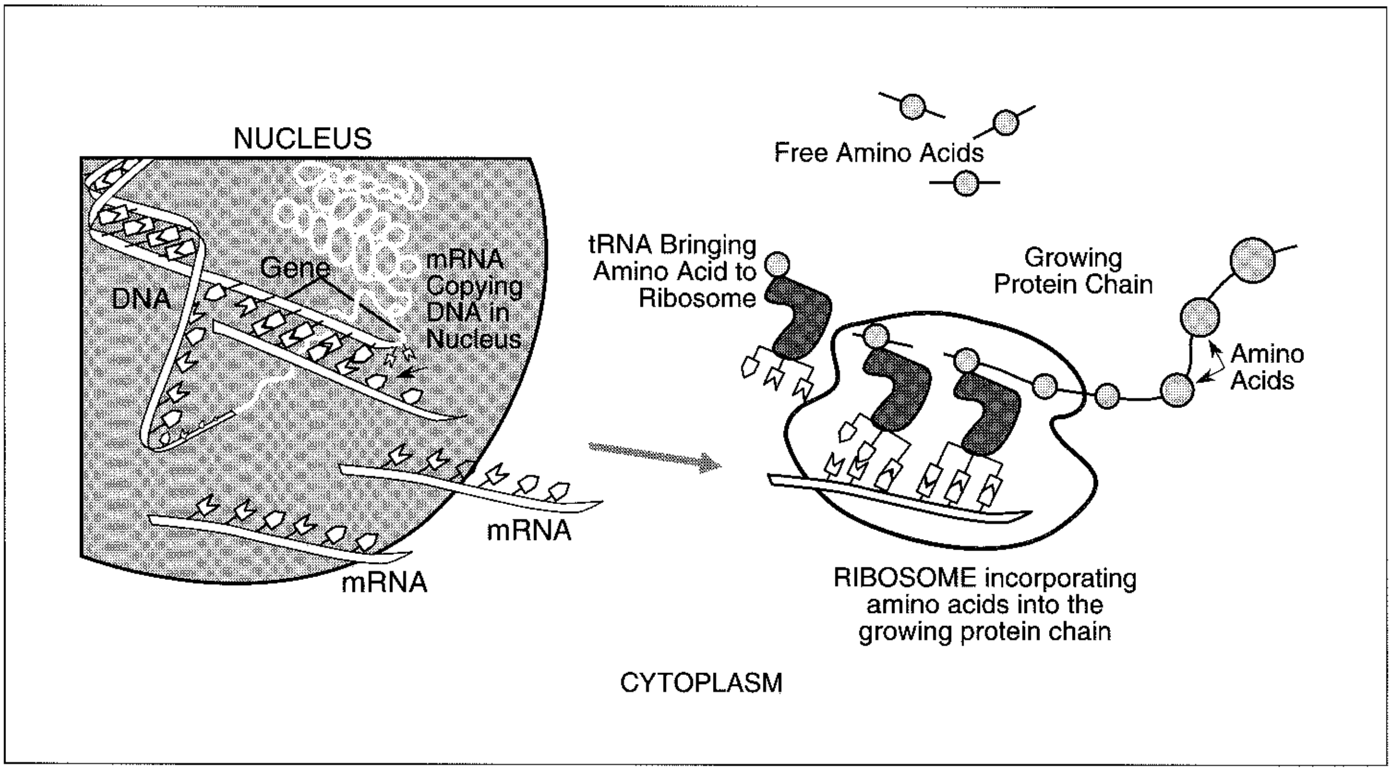

NUCLEUS
DNA
Gene
mRNA Copying DNA in Nucleus
mRNA
mRNA
Free Amino Acids
tRNA Bringing Amino Acid to Ribosome
Growing Protein Chain
Amino Acids
RIBOSOME incorporating amino acids into the growing protein chain
CYPTOPLASM

Fig. 5. Gene Expression. *When genes are expressed, the genetic information (base sequence) on DNA is first transcribed (copied) to a molecule of messenger RNA in a process similar to DNA replication. The mRNA molecules then leave the cell nucleus and enter the cytoplasm, where triplets of bases (codons) forming the genetic code specify the particular amino acids that make up an individual protein. This process, called translation, is accomplished by ribosomes (cellular components composed of proteins and another class of RNA) that read the genetic code from the mRNA, and transfer RNAs (tRNAs) that transport amino acids to the ribosomes for attachment to the growing protein. (Source: see Fig. 4.)*

Appendix A: Primer on Molecular Genetics

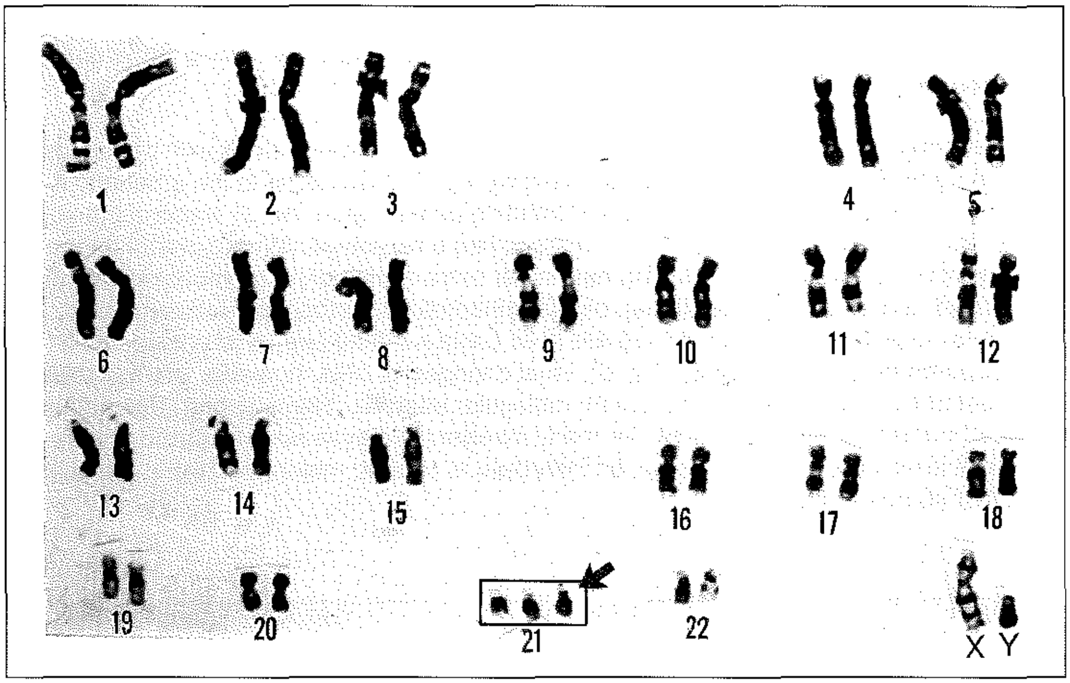


*Fig. 6. **Karyotype.** Microscopic examination of chromosome size and banding patterns allows medical laboratories to identify and arrange each of the 24 different chromosomes (22 pairs of autosomes and one pair of sex chromosomes) into a karyotype, which then serves as a tool in the diagnosis of genetic diseases. The extra copy of chromosome 21 in this karyotype identifies this individual as having Down's syndrome.*

Mapping and Sequencing the Human Genome

A primary goal of the Human Genome Project is to make a series of descriptive diagrams—maps—of each human chromosome at increasingly finer resolutions. Mapping involves (1) dividing the chromosomes into smaller fragments that can be propagated and characterized and (2) ordering (mapping) them to correspond to their respective locations on the chromosomes. After mapping is completed, the next step is to determine the base sequence of each of the ordered DNA fragments. The ultimate goal of genome research is to find all the genes in the DNA sequence and to develop tools for using this information in the study of human biology and medicine. Improving the instrumentation and techniques required for mapping and sequencing—a major focus of the genome project—will increase efficiency and cost-effectiveness. Goals include automating methods and optimizing techniques to extract the maximum useful information from maps and sequences.

A genome map describes the order of genes or other markers and the spacing between them on each chromosome. Human genome maps are constructed on several different scales or levels of resolution. At the coarsest resolution are genetic linkage maps, which depict the relative chromosomal locations of DNA markers (genes and other identifiable DNA sequences) by their patterns of inheritance. Physical maps describe the chemical characteristics of the DNA molecule itself.

Geneticists have already charted the approximate positions of over 2300 genes, and a start has been made in establishing high-resolution maps of the genome (Fig. 7). More-precise maps are needed to organize systematic sequencing efforts and plan new research directions.

Mapping Strategies

Genetic Linkage Maps

A genetic linkage map shows the relative locations of specific DNA markers along the chromosome. Any inherited physical or molecular characteristic that differs among individuals and is easily detectable in the laboratory is a potential genetic marker. Markers can be expressed DNA regions (genes) or DNA segments that have no known coding function but whose inheritance pattern can be followed. DNA sequence differences are especially useful markers because they are plentiful and easy to characterize precisely.

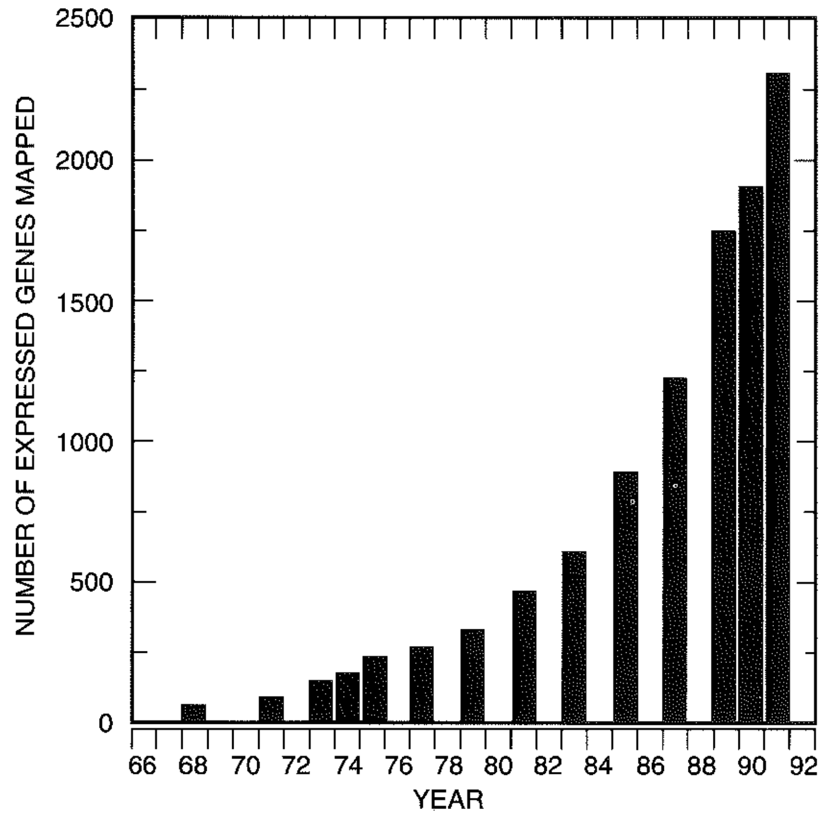


Fig. 7. Assignment of Genes to Specific Chromosomes. *The number of genes assigned (mapped) to specific chromosomes has greatly increased since the first autosomal (i.e., not on the X or Y chromosome) marker was mapped in 1968. Most of these genes have been mapped to specific bands on chromosomes. The acceleration of chromosome assignments is due to (1) a combination of improved and new techniques in chromosome sorting and band analysis, (2) data from family studies, and (3) the introduction of recombinant DNA technology. [Source: adapted from Victor A. McKusick, "Current Trends in Mapping Human Genes,"* The FASEB Journal ***5(1),*** *12 (1991).]*

Appendix A: Primer on Molecular Genetics

Markers must be polymorphic to be useful in mapping; that is, alternative forms must exist among individuals so that they are detectable among different members in family studies. Polymorphisms are variations in DNA sequence that occur on average once every 300 to 500 bp. Variations within exon sequences can lead to observable changes, such as differences in eye color, blood type, and disease susceptibility. Most variations occur within introns and have little or no effect on an organism's appearance or function, yet they are detectable at the DNA level and can be used as markers. Examples of these types of markers include (1) restriction fragment length polymorphisms (RFLPs), which reflect sequence variations in DNA sites that can be cleaved by DNA restriction enzymes (see box, p. 203), and (2) variable number of tandem repeat sequences, which are short repeated sequences that vary in the number of repeated units and, therefore, in length (a characteristic easily measured). The human genetic linkage map is constructed by observing how frequently two markers are inherited together.

Two markers located near each other on the same chromosome will tend to be passed together from parent to child. During the normal production of sperm and egg cells, DNA strands occasionally break and rejoin in different places on the same chromosome or on the other copy of the same chromosome (i.e., the homologous chromosome). This process (called meiotic recombination) can result in the separation of two markers originally on the same chromosome (Fig. 8). The closer the markers are to each other—the more "tightly linked"—the less likely a recombination event will fall between and separate them. Recombination frequency thus provides an estimate of the distance between two markers.

On the genetic map, distances between markers are measured in terms of centimorgans (cM), named after the American geneticist Thomas Hunt Morgan. Two markers are said to be 1 cM apart if they are separated by recombination 1% of the time. A genetic distance of 1 cM is roughly equal to a physical distance of 1 million bp (1 Mb). The current resolution of most human genetic map regions is about 10 Mb.

The value of the genetic map is that an inherited disease can be located on the map by following the inheritance of a DNA marker present in affected individuals (but absent in unaffected individuals), even though the molecular basis of the disease may not yet be understood nor the responsible gene identified. Genetic maps have been used to find the exact chromosomal location of several important disease genes, including cystic fibrosis, sickle cell disease, Tay-Sachs disease, fragile X syndrome, and myotonic dystrophy.

HUMAN GENOME PROJECT GOALS

	Resolution
• Complete a detailed human genetic map	2 Mb
• Complete a physical map	0.1 Mb
• Acquire the genome as clones	5 kb
• Determine the complete sequence	1 bp
• Find all the genes	

With the data generated by the project, investigators will determine the functions of the genes and develop tools for biological and medical applications.

One short-term goal of the genome project is to develop a high-resolution genetic map (2 to 5 cM); recent consensus maps of some chromosomes have averaged 7 to 10 cM between genetic markers. Genetic mapping resolution has been increased through the application of recombinant DNA technology, including in vitro radiation-induced chromosome fragmentation and cell fusions (joining human cells with those of other species to form hybrid cells) to create panels of cells with specific and varied human

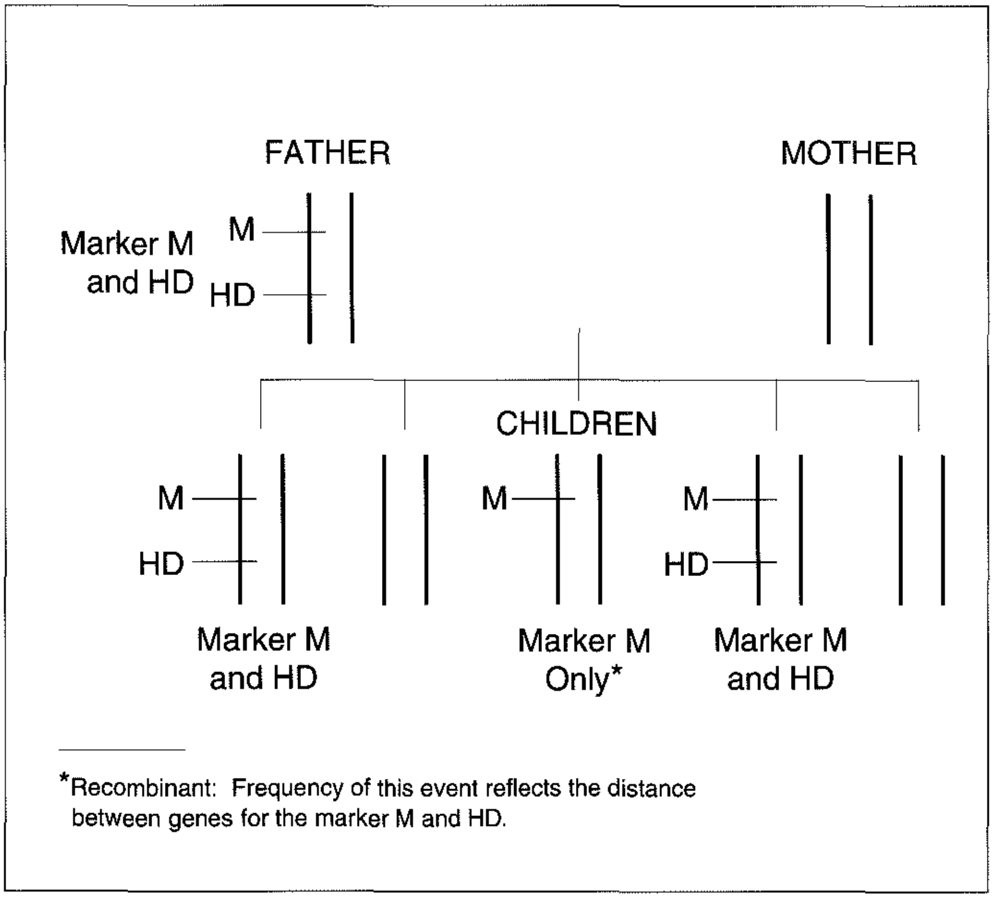

FATHER
MOTHER
Marker M and HD
M
HD
CHILDREN
M
HD
M
M
HD
Marker M and HD
Marker M Only*
Marker M and HD
*Recombinant: Frequency of this event reflects the distance between genes for the marker M and HD.

Fig. 8. Constructing a Genetic Linkage Map. *Genetic linkage maps of each chromosome are made by determining how frequently two markers are passed together from parent to child. Because genetic material is sometimes exchanged during the production of sperm and egg cells, groups of traits (or markers) originally together on one chromosome may not be inherited together. Closely linked markers are less likely to be separated by spontaneous chromosome rearrangements. In this diagram, the vertical lines represent chromosome 4 pairs for each individual in a family. The father has two traits that can be detected in any child who inherits them: a short known DNA sequence used as a genetic marker (M) and Huntington's disease (HD). The fact that one child received only a single trait (M) from that particular chromosome indicates that the father's genetic material recombined during the process of sperm production. The frequency of this event helps determine the distance between the two DNA sequences on a genetic map .*

chromosomal components. Assessing the frequency of marker sites remaining together after radiation-induced DNA fragmentation can establish the order and distance between the markers. Because only a single copy of a chromosome is required for analysis, even nonpolymorphic markers are useful in radiation hybrid mapping. [In meiotic mapping (described above), two copies of a chromosome must be distinguished from each other by polymorphic markers.]

Physical Maps

Different types of physical maps vary in their degree of resolution. The lowest-resolution physical map is the chromosomal (sometimes called cytogenetic) map, which is based on the distinctive banding patterns observed by light microscopy of stained chromosomes. A cDNA map shows the locations of expressed DNA regions (exons) on the chromosomal map. The more detailed cosmid contig map depicts the order of overlapping DNA fragments spanning the genome. A macrorestriction map describes the order and distance between enzyme cutting (cleavage) sites. The highest-resolution physical map is the complete elucidation of the DNA base-pair sequence of each chromosome in the human genome. Physical maps are described in greater detail below.

Low-Resolution Physical Mapping

Chromosomal map. In a chromosomal map, genes or other identifiable DNA fragments are assigned to their respective chromosomes, with distances measured in base pairs. These markers can be physically associated with particular bands (identified by cytogenetic staining) primarily by in situ hybridization, a technique that involves tagging the DNA marker with an observable label (e.g., one that fluoresces or is radioactive). The location of the labeled probe can be detected after it binds to its complementary DNA strand in an intact chromosome.

As with genetic linkage mapping, chromosomal mapping can be used to locate genetic markers defined by traits observable only in whole organisms. Because chromosomal maps are based on estimates of physical distance, they are considered to be physical maps. The number of base pairs within a band can only be estimated.

Until recently, even the best chromosomal maps could be used to locate a DNA fragment only to a region of about 10 Mb, the size of a typical band seen on a chromosome. Improvements in fluorescence in situ hybridization (FISH) methods allow orientation of DNA sequences that lie as close as 2 to 5 Mb. Modifications to in situ hybridization methods, using chromosomes at a stage in cell division (interphase) when they are less compact, increase map resolution to around 100,000 bp. Further banding refinement might allow chromosomal bands to be associated with specific amplified DNA fragments, an improvement that could be useful in analyzing observable physical traits associated with chromosomal abnormalities.

cDNA map. A cDNA map shows the positions of expressed DNA regions (exons) relative to particular chromosomal regions or bands. (Expressed DNA regions are those transcribed into mRNA.) cDNA is synthesized in the laboratory using the mRNA molecule as a template; base-pairing rules are followed (i.e., an A on the mRNA molecule will pair with a T on the new DNA strand). This cDNA can then be mapped to genomic regions.

Because they represent expressed genomic regions, cDNAs are thought to identify the parts of the genome with the most biological and medical significance. A cDNA map can provide the chromosomal location for genes whose functions are currently unknown. For disease-gene hunters, the map can also suggest a set of candidate genes to test when the approximate location of a disease gene has been mapped by genetic linkage techniques.

High-Resolution Physical Mapping

The two current approaches to high-resolution physical mapping are termed “top-down” (producing a macrorestriction map) and “bottom-up” (resulting in a contig map). With either strategy (described on pp. 204-5) the maps represent ordered sets of DNA fragments that are generated by cutting genomic DNA with restriction enzymes (see Restriction Enzymes box at right). The fragments are then amplified by cloning or by polymerase chain reaction (PCR) methods (see DNA Amplification, pp. 208–10). Electrophoretic techniques are used to separate the fragments according to size into different bands, which can be visualized by direct DNA staining or by hybridization with DNA probes of interest. The use of purified chromosomes separated either by flow sorting from human cell lines or in hybrid cell lines allows a single chromosome to be mapped (see Separating Chromosomes box at right).

A number of strategies can be used to reconstruct the original order of the DNA fragments in the genome. Many approaches make use of the ability of single strands of DNA and/or RNA to hybridize—to form double-stranded segments by hydrogen bonding between complementary bases. The extent of sequence homology between the two strands can be inferred from the length of the double-stranded segment. Fingerprinting uses restriction map data to determine which fragments have a specific sequence (fingerprint) in common and therefore overlap. Another approach uses linking clones as probes for hybridization to chromosomal DNA cut with the same restriction enzyme.

Restriction Enzymes: Microscopic Scalpels

Isolated from various bacteria, restriction enzymes recognize short DNA sequences and cut the DNA molecules at those specific sites. (A natural biological function of these enzymes is to protect bacteria by attacking viral and other foreign DNA.) Some restriction enzymes (rare-cutters) cut the DNA very infrequently, generating a small number of very large fragments (several thousand to a million bp). Most enzymes cut DNA more frequently, thus generating a large number of small fragments (less than a hundred to more than a thousand bp).

On average, restriction enzymes with

- 4-base recognition sites will yield pieces 256 bases long,
- 6-base recognition sites will yield pieces 4000 bases long, and
- 8-base recognition sites will yield pieces 64,000 bases long.

Since hundreds of different restriction enzymes have been characterized, DNA can be cut into many different small fragments.

Separating Chromosomes

Flow sorting

Pioneered at Los Alamos National Laboratory (LANL), flow sorting employs flow cytometry to separate, according to size, chromosomes isolated from cells during cell division when they are condensed and stable. As the chromosomes flow singly past a laser beam, they are differen-tiated by analyzing the amount of DNA present, and individual chromosomes are directed to specific collection tubes.

Somatic cell hybridization

In somatic cell hybridization, human cells and rodent tumor cells are fused (hybridized); over time, after the chromosomes mix, human chromosomes are preferentially lost from the hybrid cell until only one or a few remain. Those individual hybrid cells are then propagated and maintained as cell lines containing specific human chromosomes. Improvements to this technique have generated a number of hybrid cell lines, each with a specific single human chromosome.

Macrorestriction maps: Top-down mapping. In top-down mapping, a single chromosome is cut (with rare-cutter restriction enzymes) into large pieces, which are ordered and subdivided; the smaller pieces are then mapped further. The resulting macrorestriction maps depict the order of and distance between sites at which rare-cutter enzymes cleave (Fig. 9a). This approach yields maps with more continuity and fewer gaps between fragments than contig maps (see below), but map resolution is lower and may not be useful in finding particular genes; in addition, this strategy generally does not produce long stretches of mapped sites. Currently, this approach allows DNA pieces to be located in regions measuring about 100,000 bp to 1 Mb.

The development of pulsed-field gel (PFG) electrophoretic methods has improved the mapping and cloning of large DNA molecules. While conventional gel electrophoretic methods separate pieces less than 40 kb (1 kb = 1000 bases) in size, PFG separates molecules up to 10 Mb, allowing the application of both conventional and new mapping methods to larger genomic regions.

Contig maps: Bottom-up mapping. The bottom-up approach involves cutting the chromosome into small pieces, each of which is cloned and ordered. The ordered fragments form contiguous DNA blocks (contigs). Currently, the resulting "library" of clones

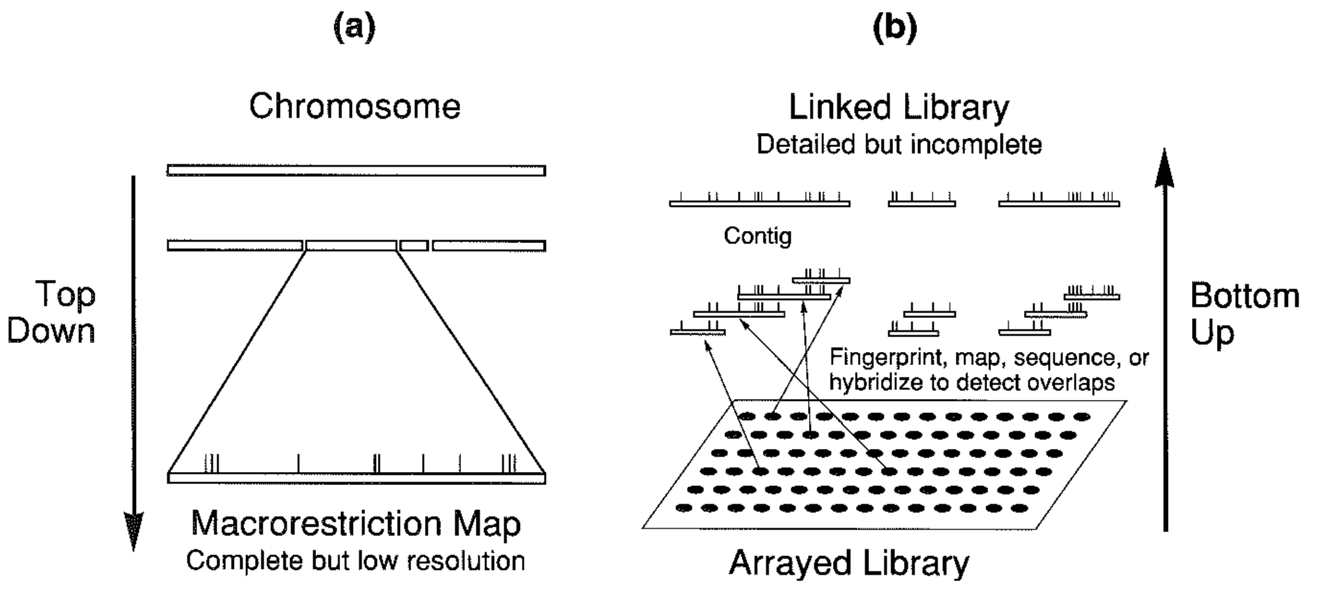


Fig. 9. Physical Mapping Strategies. *Top-down physical mapping* ***(a)*** *produces maps with few gaps, but map resolution may not allow location of specific genes. Bottom-up strategies* ***(b)*** *generate extremely detailed maps of small areas but leave many gaps. A combination of both approaches is being used. [Source: Adapted from P. R. Billings et al., "New Techniques for Physical Mapping of the Human Genome,"* The FASEB Journal ***5****(1), 29 (1991).]*

varies in size from 10,000 bp to 1 Mb (Fig. 9b, p. 204). An advantage of this approach is the accessibility of these stable clones to other researchers. Contig construction can be verified by FISH, which localizes cosmids to specific regions within chromosomal bands.

Contig maps thus consist of a linked library of small overlapping clones representing a complete chromosomal segment. While useful for finding genes localized to a small area (under 2 Mb), contig maps are difficult to extend over large stretches of a chromosome because all regions are not clonable. DNA probe techniques can be used to fill in the gaps, but they are time consuming. Figure 10 is a diagram relating the different types of maps.

Technological improvements now make possible the cloning of large DNA pieces, using artificially constructed chromosome vectors that carry human DNA fragments as large as 1 Mb. These vectors are maintained in yeast cells as artificial chromosomes (YACs). (For more explanation, see DNA Amplification, pp. 208–10.) Before YACs were developed, the largest cloning vectors (cosmids) carried inserts of only 20 to 40 kb. YAC methodology drastically reduces the number of clones to be ordered; many YACs span entire human genes. A more detailed map of a large YAC insert can be produced by subcloning, a process in which fragments of the original insert are cloned into smaller-insert vectors. Because some YAC regions are unstable, large-capacity bacterial vectors (i.e., those that can accommodate large inserts) are also being developed.

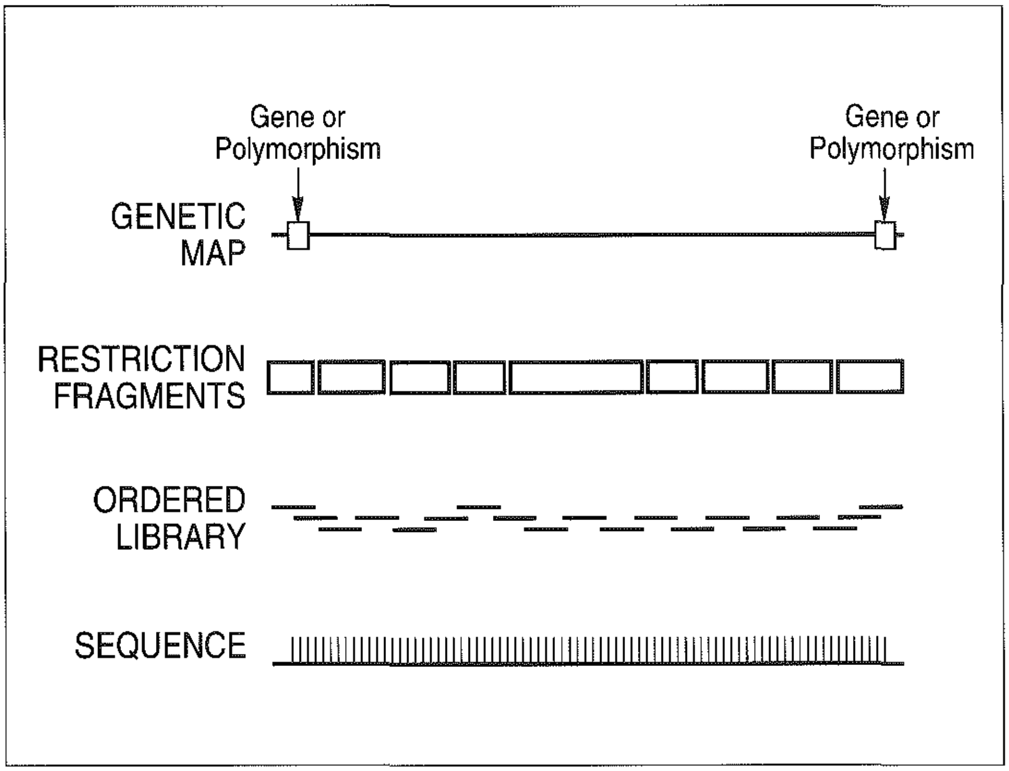


Fig. 10. Types of Genome Maps. *At the coarsest resolution, the genetic map measures recombination frequency between linked markers (genes or polymorphisms). At the next resolution level, restriction fragments of 1 to 2 Mb can be separated and mapped. Ordered libraries of cosmids and YACs have insert sizes from 40 to 400 kb. The base sequence is the ultimate physical map. Chromosomal mapping (not shown) locates genetic sites in relation to bands on chromosomes (estimated resolution of 5 Mb); new in situ hybridization techniques can place loci 100 kb apart. These direct strategies link the other four mapping approaches diagramed here. [Source: see Fig. 9.]*

Appendix A: Primer on Molecular Genetics

Sequencing Technologies

The ultimate physical map of the human genome is the complete DNA sequence—the determination of all base pairs on each chromosome. The completed map will provide biologists with a Rosetta stone for studying human biology and enable medical researchers to begin to unravel the mechanisms of inherited diseases. Much effort continues to be spent locating genes; if the full sequence were known, emphasis could shift to determining gene function. The Human Genome Project is creating research tools for 21st-century biology, when the goal will be to understand the sequence and functions of the genes residing therein.

Achieving the goals of the Human Genome Project will require substantial improvements in the rate, efficiency, and reliability of standard sequencing procedures. While technological advances are leading to the automation of standard DNA purification, separation, and detection steps, efforts are also focusing on the development of entirely new sequencing methods that may eliminate some of these steps. Sequencing procedures currently involve first subcloning DNA fragments from a cosmid or bacteriophage library into special sequencing vectors that carry shorter pieces of the original cosmid fragments (Fig. 11). The next step is to make the subcloned fragments into sets of nested fragments differing in length by one nucleotide, so that the specific base at the end of each successive fragment is detectable after the fragments have been separated by gel electrophoresis. Current sequencing technologies are discussed on p. 211.

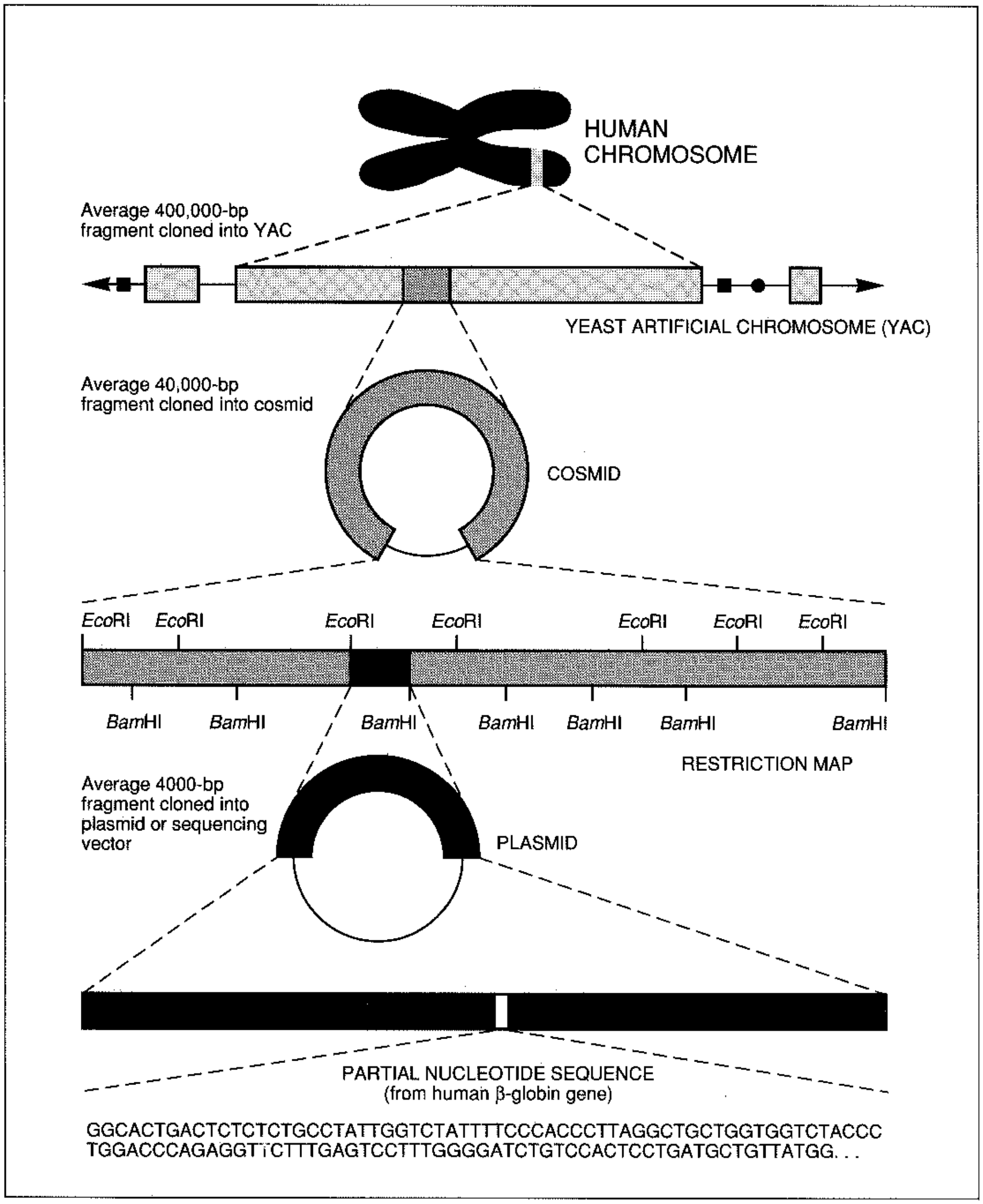

HUMAN CHROMOSOME
Average 400,000-bp fragment cloned into YAC
YEAST ARTIFICIAL CHROMOSOME (YAC)
Average 40,000-bp fragment cloned into cosmid
COSMID
EcoRI
EcoRI
EcoRI
EcoRI
EcoRI
EcoRI
EcoRI
BamHI
BamHI
BamHI
BamHI
BamHI
BamHI
BamHI
RESTRICTION MAP
Average 4000-bp fragment cloned into plasmid or sequencing vector
PLASMID
PARTIAL NUCLEOTIDE SEQUENCE
(from human β-globin gene)
GGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTGGTCTACCC
TGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGG. . .

Fig. 11. Constructing Clones for Sequencing. *Cloned DNA molecules must be made progressively smaller and the fragments subcloned into new vectors to obtain fragments small enough for use with current sequencing technology. Sequencing results are compiled to provide longer stretches of sequence across a chromosome. (Source: adapted from David A. Micklos and Greg A. Freyer,* DNA Science, A First Course in Recombinant DNA Technology, *Burlington, N.C.: Carolina Biological Supply Company, 1990.)*

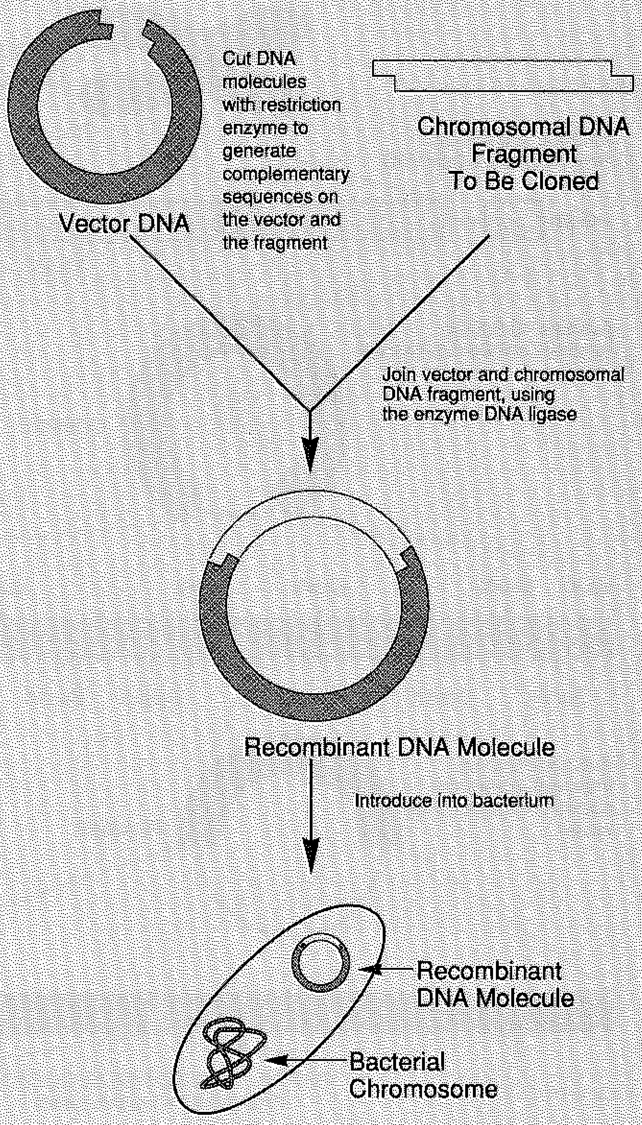
DNA Amplification: Cloning and Polymerase Chain Reaction (PCR)

Cloning (in vivo DNA amplification)

Cloning involves the use of recombinant DNA technology to propagate DNA fragments inside a foreign host. The fragments are usually isolated from chromosomes using restriction enzymes and then united with a carrier (a vector). Following introduction into suitable host cells, the DNA fragments can then be reproduced along with the host cell DNA. Vectors are DNA molecules originating from viruses, bacteria, and yeast cells. They accommodate various sizes of foreign DNA fragments ranging from 12,000 bp for bacterial vectors (plasmids and cosmids) to 1 Mb for yeast vectors (yeast artificial chromosomes). Bacteria are most often the hosts for these inserts, but yeast and mammalian cells are also used **(a)**.

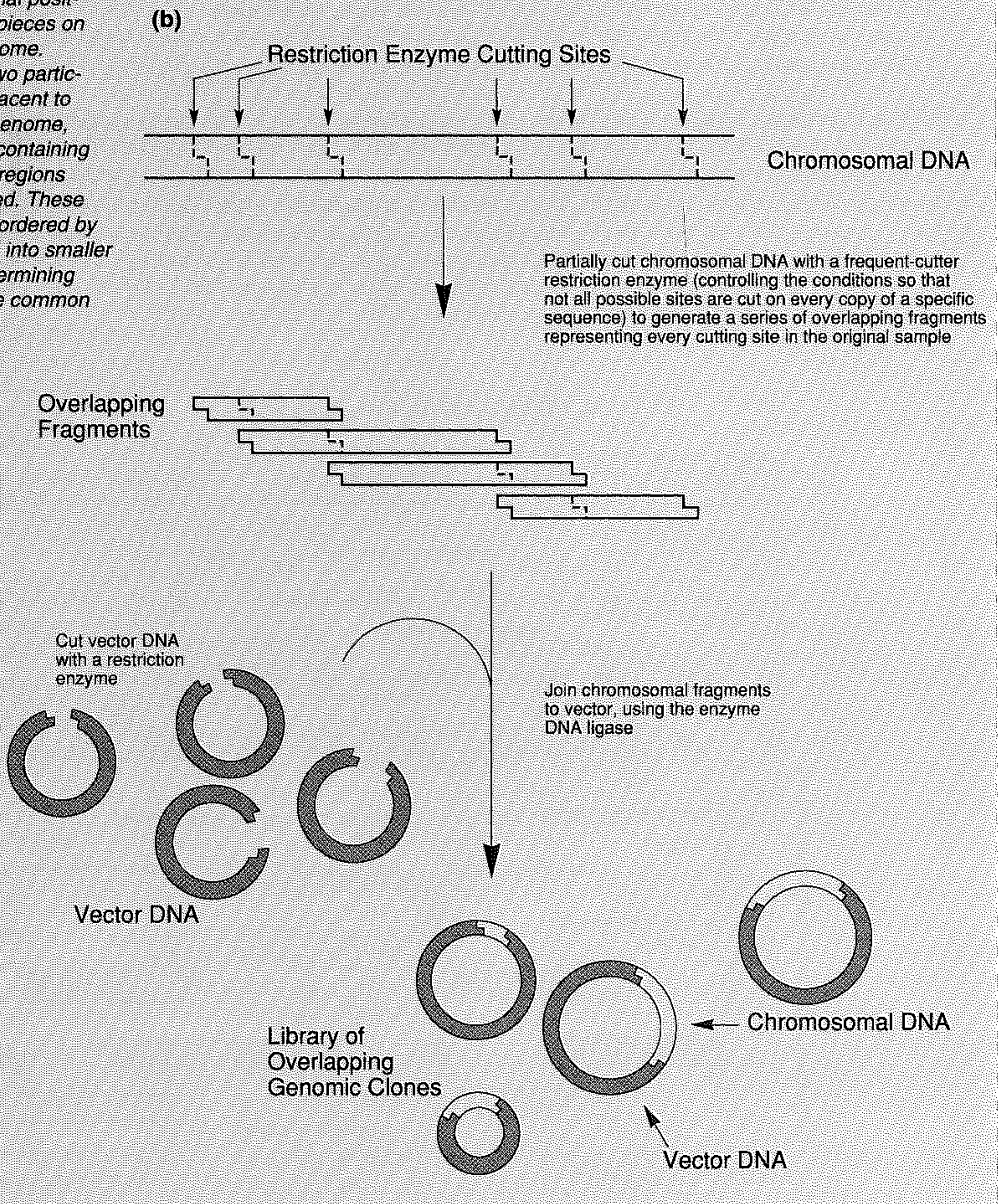
Cloning procedures provide unlimited material for experimental study. A random (unordered) set of cloned DNA fragments is called a library. Genomic libraries are sets of overlapping fragments encompassing an entire genome **(b)**. Also available are chromosome-specific libraries, which consist of fragments derived from source DNA enriched for a particular chromosome. (See Separating Chromosomes box, p. 203.)

(a)


Cut DNA molecules with restriction enzyme to generate complementary sequences on the vector and the fragment
Chromosomal DNA Fragment To Be Cloned
Vector DNA
Join vector and chromosomal DNA fragment, using the enzyme DNA ligase
Recombinant DNA Molecule
Introduce into bacterium
Recombinant DNA Molecule
Bacterial Chromosome

(a) ***Cloning DNA in Plasmids.*** *By fragmenting DNA of any origin (human, animal, or plant) and inserting it in the DNA of rapidly reproducing foreign cells, billions of copies of a single gene or DNA segment can be produced in a very short time. DNA to be cloned is inserted into a plasmid (a small, self-replicating circular molecule of DNA) that is separate from chromosomal DNA. When the recombinant plasmid is introduced into bacteria, the newly inserted segment will be replicated along with the rest of the plasmid.*

(b) *Constructing an Overlapping Clone Library.* *A collection of clones of chromosomal DNA, called a library, has no obvious order indicating the original positions of the cloned pieces on the uncut chromosome. To establish that two particular clones are adjacent to each other in the genome, libraries of clones containing partly overlapping regions must be constructed. These clone libraries are ordered by dividing the inserts into smaller fragments and determining which clones share common DNA sequences.*


(b)
Restriction Enzyme Cutting Sites
Chromosomal DNA
Overlapping Fragments
Cut vector DNA with a restriction enzyme
Vector DNA
Library of Overlapping Genomic Clones
Chromosomal DNA
Vector DNA

Partially cut chromosomal DNA with a frequent-cutter restriction enzyme (controlling the conditions so that not all possible sites are cut on every copy of a specific sequence) to generate a series of overlapping fragments representing every cutting site in the original sample

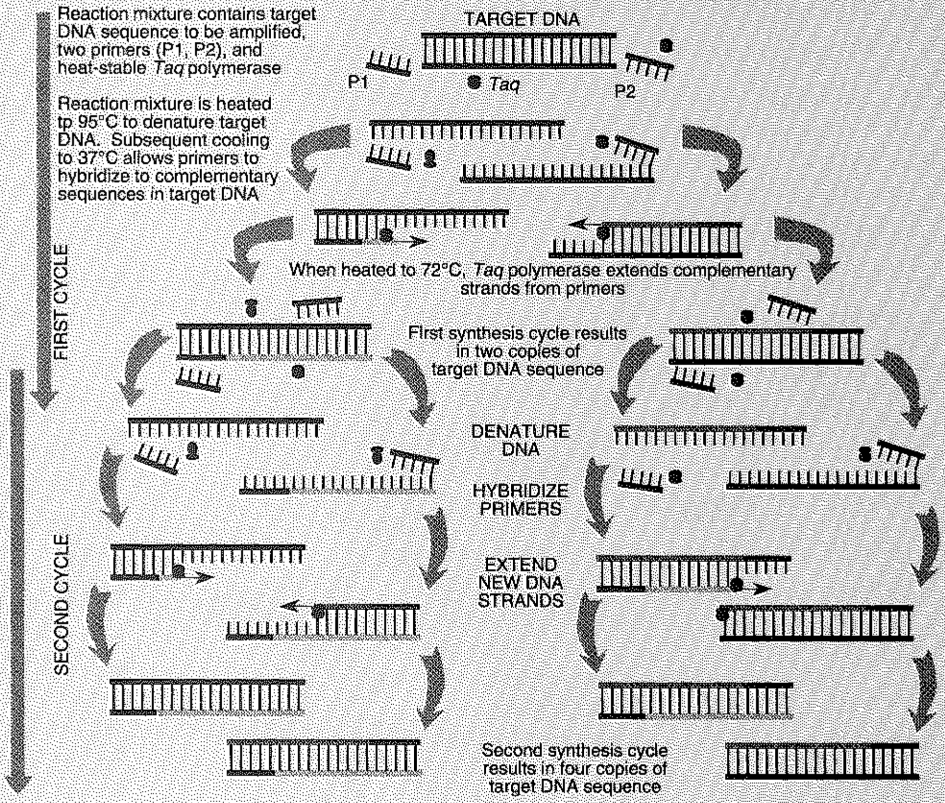
Join chromosomal fragments to vector, using the enzyme DNA ligase

PCR (in vitro DNA amplification)

Described as being to genes what Gutenberg's printing press was to the written word, PCR can amplify a desired DNA sequence of any origin (virus, bacteria, plant, or human) hundreds of millions of times in a matter of hours, a task that would have required several days with recombinant technology. PCR is especially valuable because the reaction is highly specific, easily automated, and capable of amplifying minute amounts of sample. For these reasons, PCR has also had a major impact on clinical medicine, genetic disease diagnostics, forensic science, and evolutionary biology.

PCR is a process based on a specialized polymerase enzyme, which can synthesize a complementary strand to a given DNA strand in a mixture containing the 4 DNA bases and 2 DNA fragments (primers, each about 20 bases long) flanking the target sequence. The mixture is heated to separate the strands of double-stranded DNA containing the target sequence and then cooled to allow (1) the primers to find and bind to their complementary sequences on the separated strands and (2) the polymerase to extend the primers into new complementary strands. Repeated heating and cooling cycles multiply the target DNA exponentially, since each new double strand separates to become two templates for further synthesis. In about 1 hour, 20 PCR cycles can amplify the target by a millionfold.

DNA Amplification Using PCR


Reaction mixture contains target DNA sequence to be amplified, two primers (P1, P2), and heat-stable *Taq* polymerase
TARGET DNA
P1
Taq
P2
Reaction mixture is heated tp 95°C to denature target DNA. Subsequent cooling to 37°C allows primers to hybridize to complementary sequences in target DNA
When heated to 72°C, *Taq* polymerase extends complementary strands from primers
First synthesis cycle results in two copies of target DNA sequence
FIRST CYCLE
DENATURE DNA
HYBRIDIZE PRIMERS
EXTEND NEW DNA STRANDS
SECOND CYCLE
Second synthesis cycle results in four copies of target DNA sequence

Source: *DNA Science*, see Fig. 11.

Current Sequencing Technologies

The two basic sequencing approaches, Maxam-Gilbert and Sanger, differ primarily in the way the nested DNA fragments are produced. Both methods work because gel electrophoresis produces very high resolution separations of DNA molecules; even fragments that differ in size by only a single nucleotide can be resolved. Almost all steps in these sequencing methods are now automated. Maxam-Gilbert sequencing (also called the chemical degradation method) uses chemicals to cleave DNA at specific bases, resulting in fragments of different lengths. A refinement to the Maxam-Gilbert method known as multiplex sequencing enables investigators to analyze about 40 clones on a single DNA sequencing gel. Sanger sequencing (also called the chain termination or dideoxy method) involves using an enzymatic procedure to synthesize DNA chains of varying length in four different reactions, stopping the DNA replication at positions occupied by one of the four bases, and then determining the resulting fragment lengths (Fig. 12).

These first-generation gel-based sequencing technologies are now being used to sequence small regions of interest in the human genome. Although investigators could use existing technology to sequence whole chromosomes, time and cost considerations make large-scale sequencing projects of this nature impractical. The smallest human chromosome (Y) contains 50 Mb; the largest (chromosome 1) has 250 Mb. The largest continuous DNA sequence obtained thus far, however, is approximately 350,000 bp, and the best available equipment can sequence only 50,000 to 100,000 bases per year at an approximate cost of $1 to $2 per base. At that rate, an unacceptable 30,000 work-years and at least $3 billion would be required for sequencing alone.

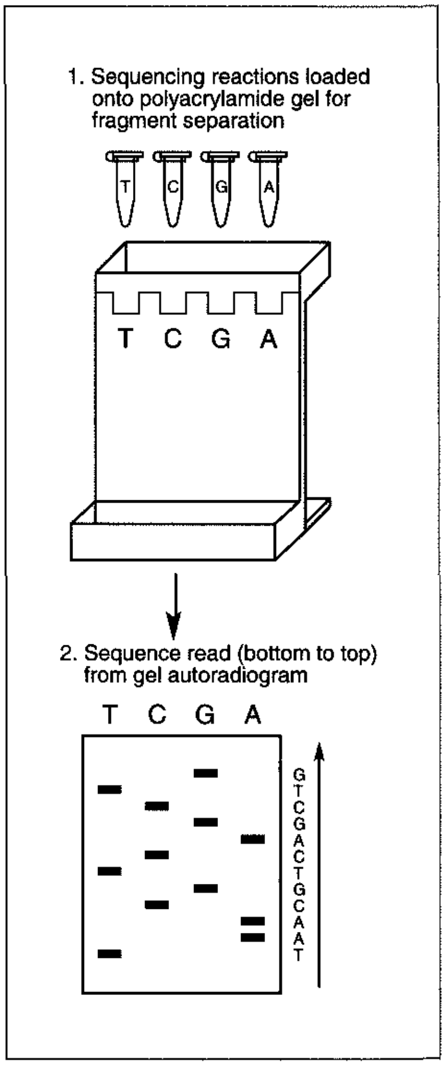

1. Sequencing reactions loaded onto polyacrylamide gel for fragment separation
T
C
G
A
T C G A
2. Sequence read (bottom to top) from gel autoradiogram
T C G A
GTCGACTGCAAT

Fig. 12. DNA Sequencing. *Dideoxy sequencing (also called chain-termination or Sanger method) uses an enzymatic procedure to synthesize DNA chains of varying lengths, stopping DNA replication at one of the four bases and then determining the resulting fragment lengths. Each sequencing reaction tube (T, C, G, and A) in the diagram contains*

- *a DNA template, a primer sequence, and a DNA polymerase to initiate synthesis of a new strand of DNA at the point where the primer is hybridized to the template;*
- *the four deoxynucleotide triphosphates (dATP, dTTP, dCTP, and dGTP) to extend the DNA strand;*
- *one labeled deoxynucleotide triphosphate (using a radioactive element or dye); and*
- *one dideoxynucleotide triphosphate, which terminates the growing chain wherever it is incorporated. Tube A has didATP, tube C has didCTP, etc.*

For example, in the A reaction tube the ratio of the dATP to didATP is adjusted so that each tube will have a collection of DNA fragments with a didATP incorporated for each adenine position on the template DNA fragments. The fragments of varying length are then separated by electrophoresis (1) and the positions of the nucleotides analyzed to determine sequence. The fragments are separated on the basis of size, with the shorter fragments moving faster and appearing at the bottom of the gel. Sequence is read from bottom to top (2). (Source: see Fig. 11.)

Sequencing Technologies Under Development

A major focus of the Human Genome Project is the development of automated sequencing technology that can accurately sequence 100,000 or more bases per day at a cost of less than $.50 per base. Specific goals include the development of sequencing and detection schemes that are faster and more sensitive, accurate, and economical. Many novel sequencing technologies are now being explored, and the most promising ones will eventually be optimized for widespread use.

Second-generation (interim) sequencing technologies will enable speed and accuracy to increase by an order of magnitude (i.e., 10 times greater) while lowering the cost per base. Some important disease genes will be sequenced with such technologies as (1) high-voltage capillary and ultrathin electrophoresis to increase fragment separation rate and (2) use of resonance ionization spectroscopy to detect stable isotope labels.

Third-generation gel-less sequencing technologies, which aim to increase efficiency by several orders of magnitude, are expected to be used for sequencing most of the human genome. These developing technologies include (1) enhanced fluorescence detection of individual labeled bases in flow cytometry, (2) direct reading of the base sequence on a DNA strand with the use of scanning tunneling or atomic force microscopies, (3) enhanced mass spectrometric analysis of DNA sequence, and (4) sequencing by hybridization to short panels of nucleotides of known sequence. Pilot large-scale sequencing projects will provide opportunities to improve current technologies and will reveal challenges investigators may encounter in larger-scale efforts.

Partial Sequencing To Facilitate Mapping, Gene Identification

Correlating mapping data from different laboratories has been a problem because of differences in generating, isolating, and mapping DNA fragments. A common reference system designed to meet these challenges uses partially sequenced unique regions (200 to 500 bp) to identify clones, contigs, and long stretches of sequence. Called sequence tagged sites (STSs), these short sequences have become standard markers for physical mapping.

Because coding sequences of genes represent most of the potentially useful information content of the genome (but are only a fraction of the total DNA), some investigators have begun partial sequencing of cDNAs instead of random genomic DNA. (cDNAs are derived from mRNA sequences, which are the transcription products of expressed genes.) In addition to providing unique markers, these partial sequences [termed expressed sequence tags (ESTs)] also identify expressed genes. This strategy can thus provide a means of rapidly identifying most human genes. Other applications of the EST approach include determining locations of genes along chromosomes and identifying coding regions in genomic sequences.

End Games: Completing Maps and Sequences; Finding Specific Genes

Starting maps and sequences is relatively simple; finishing them will require new strategies or a combination of existing methods. After a sequence is determined using the methods described above, the task remains to fill in the many large gaps left by current mapping methods. One approach is single-chromosome microdissection, in which a piece is physically cut from a chromosomal region of particular interest, broken up into smaller pieces, and amplified by PCR or cloning (see DNA Amplification, pp. 208–10). These fragments can then be mapped and sequenced by the methods previously described.

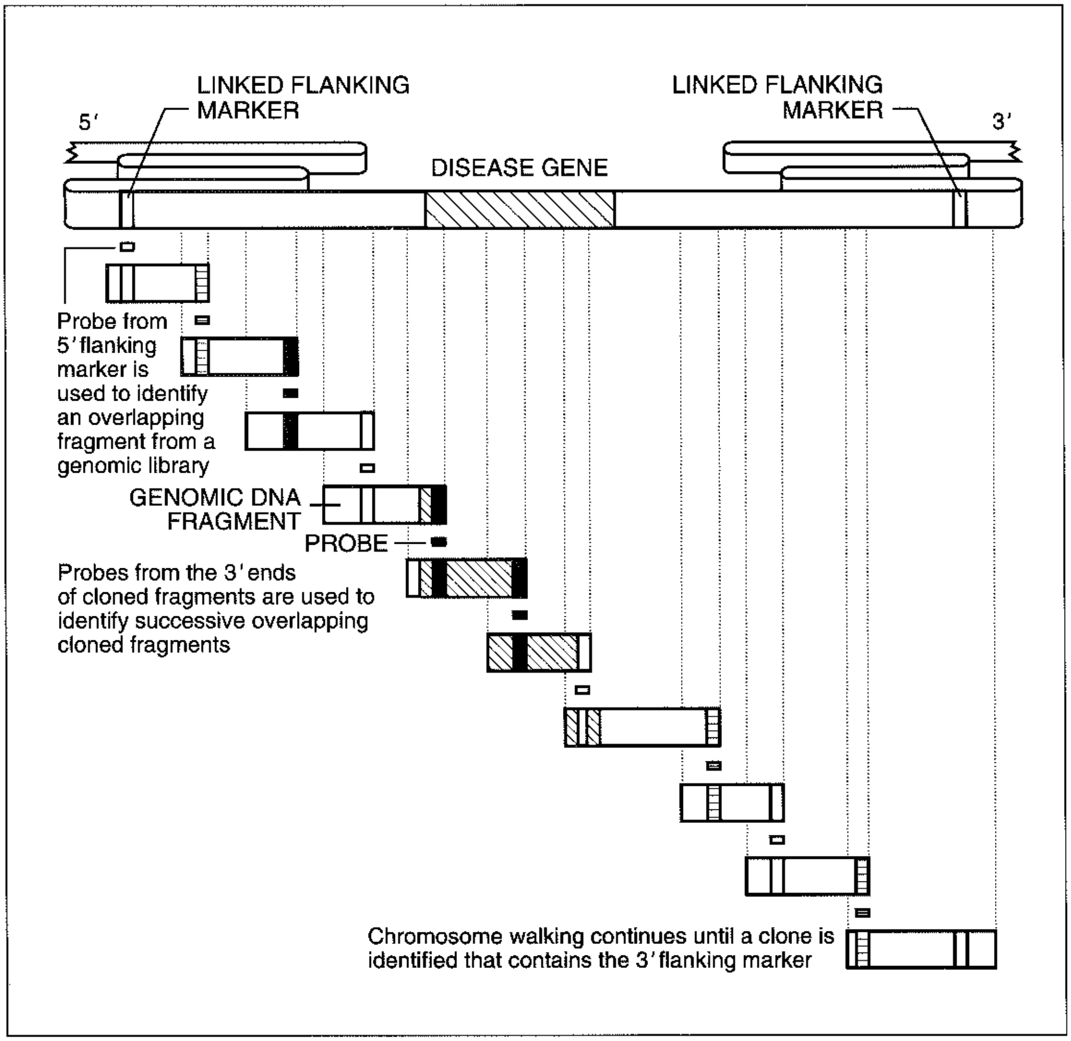
Chromosome walking, one strategy for filling in gaps, involves hybridizing a primer of known sequence to a clone from an unordered genomic library and synthesizing a short complementary strand (called "walking" along a chromosome). The complementary strand is then sequenced and its end used as the next primer for further walking; in this way the adjacent, previously unknown, region is identified and sequenced. The chromosome is thus systematically sequenced from one end to the other. Because primers must be synthesized chemically, a disadvantage of this technique is the large number of different primers needed to walk a long distance. Chromosome walking is also used to locate specific genes by sequencing the chromosomal segments between markers that flank the gene of interest (Fig. 13, p. 214).

The current human genetic map has about 1000 markers, or 1 marker spaced every 3 million bp; an estimated 100 genes lie between each pair of markers. Higher-resolution genetic maps have been made in regions of particular interest. New genes can be located by combining genetic and physical map information for a region. The genetic map basically describes gene order. Rough information about gene location is sometimes available also, but these data must be used with caution because recombination is not equally likely at all places on the chromosome. Thus the genetic map, compared to the physical map, stretches in some places and compresses in others, as though it were drawn on a rubber band.

The degree of difficulty in finding a disease gene of interest depends largely on what information is already known about the gene and, especially, on what kind of DNA alterations cause the disease. Spotting the disease gene is very difficult when disease results from a single altered DNA base; sickle cell anemia is an example of such a case, as are probably most major human inherited diseases. When disease results from a large DNA rearrangement, this anomaly can usually be detected as alterations in the physical map of the region or even by direct microscopic examination of the chromosome. The location of these alterations pinpoints the site of the gene.

Identifying the gene responsible for a specific disease without a map is analogous to finding a needle in a haystack. Actually, finding the gene is even more difficult, because even close up, the gene still looks like just another piece of hay. However, maps give clues on where to look; the finer the map's resolution, the fewer pieces of hay to be tested.

Once the neighborhood of a gene of interest has been identified, several strategies can be used to find the gene itself. An ordered library of the gene neighborhood can be constructed if one is not already available. This library provides DNA fragments that can be screened for additional polymorphisms, improving the genetic map of the region and further restricting the possible gene location. In addition, DNA fragments from the region can be used as probes to search for DNA sequences that are expressed (transcribed to RNA) or conserved among individuals. Most genes will have such sequences. Then individual gene candidates must be examined. For example, a gene responsible for liver disease is likely to be expressed in the liver and less likely in other tissues or organs. This type of evidence can further limit the search. Finally, a suspected gene may need to be sequenced in both healthy and affected individuals. A consistent pattern of DNA variation when these two samples are compared will show that the gene of interest has very likely been found. The ultimate proof is to correct the suspected DNA alteration in a cell and show that the cell's behavior reverts to normal.



Probe from 5' flanking marker is used to identify an overlapping fragment from a genomic library

Probes from the 3' ends of cloned fragments are used to identify successive overlapping cloned fragments

Chromosome walking continues until a clone is identified that contains the 3' flanking marker

Fig. 13. Cloning a Disease Gene by Chromosome Walking. *After a marker is linked to within 1 cM of a disease gene, chromosome walking can be used to clone the disease gene itself. A probe is first constructed from a genomic fragment identified from a library as being the closest linked marker to the gene. A restriction fragment isolated from the end of the clone near the disease locus is used to reprobe the genomic library for an overlapping clone. This process is repeated several times to walk across the chromosome and reach the flanking marker on the other side of the disease-gene locus. (Source: see Fig. 11.)*

Model Organism Research

Most mapping and sequencing technologies were developed from studies of nonhuman genomes, notably those of the bacterium *Escherichia coli*, the yeast *Saccharomyces cerevisiae*, the fruit fly *Drosophila melanogaster*, the roundworm *Caenorhabditis elegans*, and the laboratory mouse *Mus musculus*. These simpler systems provide excellent models for developing and testing the procedures needed for studying the much more complex human genome.

A large amount of genetic information has already been derived from these organisms, providing valuable data for the analysis of normal gene regulation, genetic diseases, and evolutionary processes. Physical maps have been completed for *E. coli*, and extensive overlapping clone sets are available for *S. cerevisiae* and *C. elegans*. In addition, sequencing projects have been initiated by the NIH genome program for *E. coli*, *S. cerevisiae*, and *C. elegans*.

Mouse genome research will provide much significant comparative information because of the many biological and genetic similarities between mouse and man. Comparisons of human and mouse DNA sequences will reveal areas that have been conserved during evolution and are therefore important. An extensive database of mouse DNA sequences will allow counterparts of particular human genes to be identified in the mouse and extensively studied. Conversely, information on genes first found to be important in the mouse will lead to associated human studies. The mouse genetic map, based on morphological markers, has already led to many insights into human biology. Mouse models are being developed to explore the effects of mutations causing human diseases, including diabetes, muscular dystrophy, and several cancers. A genetic map based on DNA markers is presently being constructed, and a physical map is planned to allow direct comparison with the human physical map.

Informatics: Data Collection and Interpretation

Collecting and Storing Data

The reference map and sequence generated by genome research will be used as a primary information source for human biology and medicine far into the future. The vast amount of data produced will first need to be collected, stored, and distributed. If compiled in books, the data would fill an estimated 200 volumes the size of a Manhattan telephone book (at 1000 pages each), and reading it would require 26 years working around the clock (Fig. 14, p. 216).

Because handling this amount of data will require extensive use of computers, database development will be a major focus of the Human Genome Project. The present challenge is to improve database design, software for

HUMAN GENETIC DIVERSITY:
The Ultimate Human Genetic Database

- Any two individuals differ in about 3 x 10^6 bases (0.1%).
- The population is now about 5 x 10^9.
- A catalog of all sequence differences would require 15 x 10^{15} entries.
- This catalog may be needed to find the rarest or most complex disease genes.

database access and manipulation, and data-entry procedures to compensate for the varied computer procedures and systems used in different laboratories. Databases need to be designed that will accurately represent map information (linkage, STSs, physical location, disease loci) and sequences (genomic, cDNAs, proteins) and link them to each other and to bibliographic text databases of the scientific and medical literature.

Interpreting Data

New tools will also be needed for analyzing the data from genome maps and sequences. Recognizing where genes begin and end and identifying their exons, introns, and regulatory sequences may require extensive comparisons with sequences from related species such as the mouse to search for conserved similarities (homologies). Searching a database for a particular DNA sequence may uncover these homologous sequences in a known gene from a model organism, revealing insights into the function of the corresponding human gene.

Correlating sequence information with genetic linkage data and disease gene research will reveal the molecular basis for human variation. If a newly identified gene is found to code for a flawed protein, the altered protein must be compared with the normal version to identify the specific abnormality that causes disease. Once the error is pinpointed, researchers must try to determine how to correct it in the human body, a task that will require knowledge about how the protein functions and in which cells it is active.

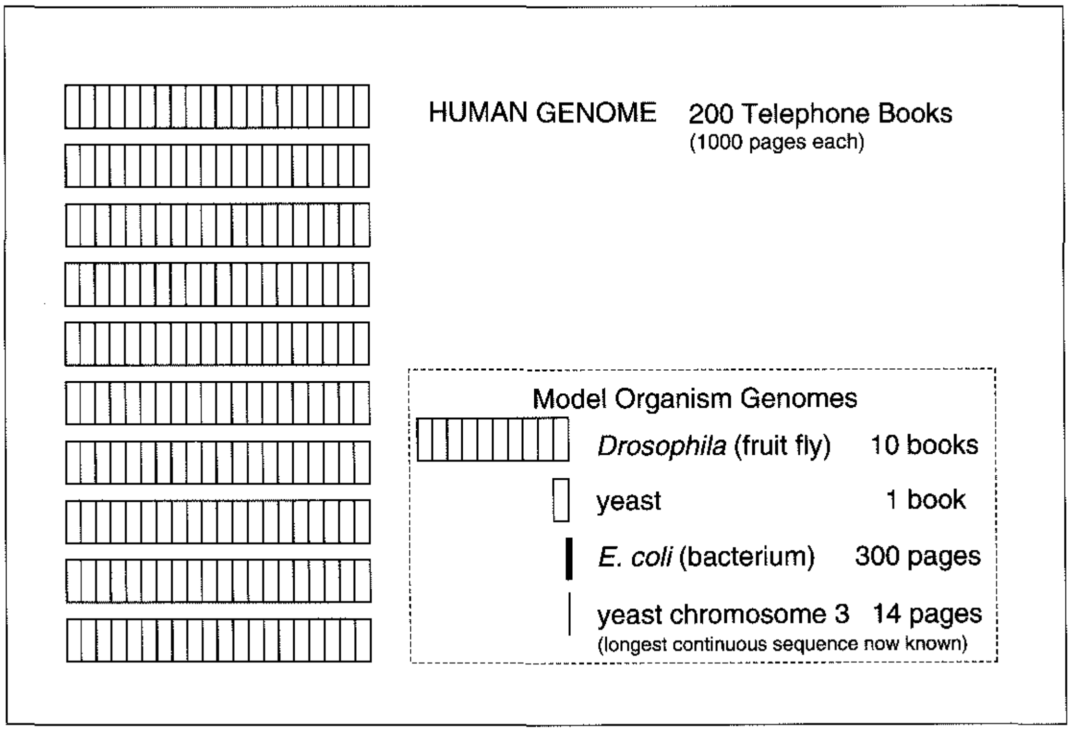


Fig. 14. Magnitude of Genome Data. *If the DNA sequence of the human genome were compiled in books, the equivalent of 200 volumes the size of a Manhattan telephone book (at 1000 pages each) would be needed to hold it all. New data-analysis tools will be needed for understanding the information from genome maps and sequences.*

Correct protein function depends on the three-dimensional (3D), or folded, structure the proteins assume in biological environments; thus, understanding protein structure will be essential in determining gene function. DNA sequences will be translated into amino acid sequences, and researchers will try to make inferences about functions either by comparing protein sequences with each other or by comparing their specific 3-D structures (Fig. 15).

Because the 3-D structure patterns (motifs) that protein molecules assume are much more evolutionarily conserved than amino acid sequences, this type of homology search could prove more fruitful. Particular motifs may serve similar functions in several different proteins, information that would be valuable in genome analyses. Currently, however, only a few protein motifs can be recognized at the sequence level. Continued development of analytic capabilities to facilitate grouping protein sequences into motif families will make homology searches more successful.

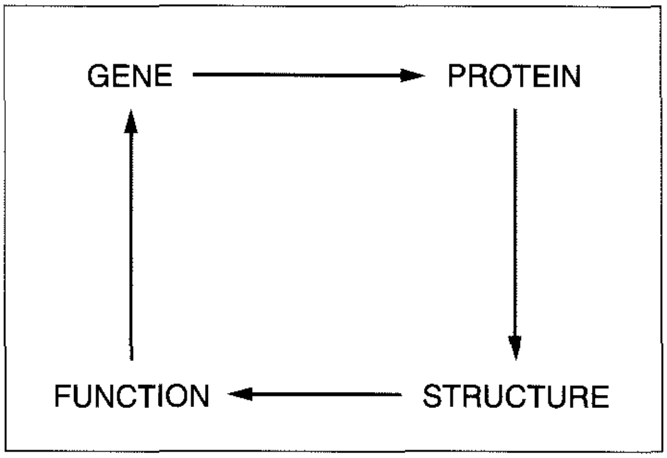


Fig. 15. Understanding Gene Function. *Understanding how genes function will require analyses of the 3-D structures of the proteins for which the genes code.*

Mapping Databases

The Genome Data Base (GDB), located at Johns Hopkins University (Baltimore, Maryland), provides location, ordering, and distance information for human genetic markers, probes, and contigs linked to known human genetic disease. GDB is presently working on incorporating physical mapping data. Also at Hopkins is the Online *Mendelian Inheritance in Man* database, a catalog of inherited human traits and diseases.

The Human and Mouse Probes and Libraries Database (located at the American *Type Culture* Collection in Rockville, Maryland) and the GBASE mouse database (located at Jackson Laboratory, Bar Harbor, Maine) include data on RFLPs, chromosomal assignments, and probes from the laboratory mouse.

Sequence Databases

Nucleic Acids (DNA and RNA)

GenBank®, the European Molecular Biology Laboratory (EMBL) sequence database, and the DNA Database of Japan (DDBJ) house over 70 Mb of sequence from more than 2500 different organisms. Compiled from both direct submissions and journal scans, GenBank is supported at IntelliGenetics (Mountain View, California) and LANL through a contract from the NIH National Institute of General Medical Sciences. Although responsibility for GenBank will move to the National Center for Biotechnology Information (NCBI) of the National Library of Medicine in September 1992, LANL will continue to handle direct data submissions from authors. International collaborations with EMBL and DDBJ will also continue. NCBI is also developing GenInfo, a data archive that will eventually offer integrated access to other databases.

Proteins

The major protein sequence databases are the Protein Identification Resource (National Biomedical Research Foundation), Swissprot, and GenPept (both distributed with GenBank). In addition to sequence information, they contain information on protein motifs and other features of protein structure.

Impact of the Human Genome Project

The atlas of the human genome will revolutionize medical practice and biological research into the 21st century and beyond. All human genes will eventually be found, and accurate diagnostics will be developed for most inherited diseases. In addition, animal models for human disease research will be more easily developed, facilitating the understanding of gene function in health and disease.

Researchers have already identified single genes associated with a number of diseases, such as cystic fibrosis, Duchenne muscular dystrophy, myotonic dystrophy, neurofibromatosis, and retinoblastoma. As research progresses, investigators will also uncover the mechanisms for diseases caused by several genes or by a gene interacting with environmental factors. Genetic susceptibilities have been implicated in many major disabling and fatal diseases including heart disease, stroke, diabetes, and several kinds of cancer. The identification of these genes and their proteins will pave the way to more-effective therapies and preventive measures. Investigators determining the underlying biology of genome organization and gene regulation will also begin to understand how humans develop from single cells to adults, why this process sometimes goes awry, and what changes take place as people age.

New technologies developed for genome research will also find myriad applications in industry, as well as in projects to map (and ultimately improve) the genomes of economically important farm animals and crops.

While human genome research itself does not pose any new ethical dilemmas, the use of data arising from these studies presents challenges that need to be addressed before the data accumulate significantly. To assist in policy development, the ethics component of the Human Genome Project is funding conferences and research projects to identify and consider relevant issues, as well as activities to promote public awareness of these topics.

Conferences, Meetings, and Workshops Sponsored by DOE

Appendix B

Appendix B: Conferences, Meetings, and Workshops Sponsored by DOE

4/89	Second Cold Spring Harbor Meeting on Genome Mapping and Sequencing: Cold Spring Harbor, NY
6/89	Chromosome 16 Workshop: New Haven, CT
10/89	First Annual Genome Sequencing Conference: Wolf Trap, Vienna, VA
12/89	Large Insert Cloning Workshop: Houston, TX
12/89	Human X Chromosome Workshop: Houston, TX
2/90	Chromosome 3 Workshop: San Antonio, TX
3/90	First Conference on Genetics, Religion, and Ethics: Houston, TX
4/90	Application of Mass Spectrometry to DNA Sequencing Workshop: Seattle, WA
4/90	Chromosome 21 Workshop: Bethesda, MD
4/90	Workshop on Mapping Human Chromosome 22: Paris
8/90	DOE-NIH Annual Planning and Evaluation Retreat: Hunt Valley, MD
8/90	Chromosome 19 Workshop: Charleston, SC
9/90	Genome Sequencing Conference II: Hilton Head, SC
9/90	First International Workshop on Human Chromosome 5: London
11/90	Fourth International Workshop on Mouse Genome Mapping: Annapolis, MD
1/91	Second X Chromosome Workshop: Oxford, England
2/91	Second DOE Contractor-Grantee Workshop: Santa Fe, NM
3/91	Chromosome 17 Workshop: Salt Lake City, UT
4/91	Workshop on Computational Molecular Biology: Seattle, WA
4/91	Chromosome 3 Workshop: Denver, CO
4/91	Chromosome 21 Workshop: Denver, CO
5/91	Sequencing by Hybridization Workshop: Gaithersburg, MD
5/91	Chromosome 11 Workshop: Paris
6/91	Workshop on Open Problems of Computational Molecular Biology: Telluride, CO
6/91	Chromosome 4 Workshop: Philadelphia, PA
9/91	DOE-NIH Annual Planning and Evaluation Retreat: Lafayette, CA
9/91	ELSI Working Group Meeting on Privacy: Bethesda, MD
9/91	First Panel Meeting "Predicting Future Diseases" at the National Academy of Sciences Institute of Medicine: Washington, DC
9/91	Genome Sequencing III: Hilton Head, SC
9/91	Workshop on Informatics Needs of Large-Scale Sequencing Projects: Hilton Head, SC

10/91	Conference on Identification of Transcribed Sequences in the Human Genome: Bethesda, MD
10/91	Workshop on DNA Sequence Acquisition and Interpretation: Cold Spring Harbor, NY
11/91	Conference on Justice and the Human Genome: Chicago, IL
11/91	Sequencing By Hybridization Workshop: Moscow
12/91	Human Genetics and Genome Analysis: A Practical Workshop for the Nonscientist: Cold Spring Harbor, NY
1/92	Chromosome 19 Workshop: Nijmegen, Netherlands
2/92	Chromosome 16 Workshop: Adelaide, Australia
3/92	Second Conference on Genetics, Religion, and Ethics: Houston, TX
3/92	Chromosome 17 Workshop: Salt Lake City, UT
3/92	Chromosome 3 Workshop: Tokyo, Japan
3/92	Chromosome 9 Workshop: Cambridge, England
5/92	Chromosome 5 Workshop: Chicago, IL
6/92	Chromosome 4 Workshop: Leiden, Netherlands
6/92	Chromosome 6 Workshop: Ann Arbor, MI
6/92	Chromosome 15 Workshop: Tucson, AZ
6/92	Chromosome 18 Workshop: Chicago, IL
6/92	DOE/NIH Annual Planning and Evaluation Retreat: Bethesda, MD

Partial Listing of Future DOE-Sponsored Workshops

9/92	Chromosome 11 Workshop; San Diego, CA
9/92	Chromosome 12 Workshop: Oxford, England
9/92	Chromosome 13 Workshop: New York, NY
11/92	Chromosome 2 Workshop: Lake Tahoe, CA
2/93	Third DOE Contractor-Grantee Workshop: Santa Fe, NM

Members of the DOE Health and Environmental Research Advisory Committee

Appendix C

Appendix C: Members of the DOE Health and Environmental Research Advisory Committee

Sheldon Wolff (Chair)	University of California, San Francisco
E. Morton Bradbury	Los Alamos National Laboratory
Eville Gorham	University of Minnesota
Jonathan Greer	Abbott Laboratories
Barbara Ann Hamkalo	University of California, Irvine
Sam Hurst	Atom Sciences, Inc.
Kenneth K. Kidd	Yale University
Leonard S. Lerman	Massachusetts Institute of Technology
Gordon J. MacDonald	University of California, San Diego
J. Justin McCormick	Michigan State University
Mortimer L. Mendelsohn	Lawrence Livermore National Laboratory
Mary Lou Pardue	Massachusetts Institute of Technology
Theodore L. Phillips	University of California, San Francisco
Richard C. Reba	University of Chicago
Melvin I. Simon	California Institute of Technology
Warren M. Washington	National Center for Atmospheric Research
Audrey Wegst	Diagnostic Technology Consultants, Inc.
Harel Weinstein	Mt. Sinai School of Medicine

Members of DOE-NIH Joint Working Groups

Appendix D

Appendix D: Members of NIH-DOE Joint Working Groups

Joint Working Group on Ethical, Legal, and Social Issues

(First met September 1989; first workshop held February 5–6, 1990)

Nancy Wexler (Chair)	Columbia University
Jonathan R. Beckwith	Harvard Medical School
Robert Cook-Deegan	National Academy of Sciences Institute of Medicine
Patricia King	Georgetown University Law Center
Victor A. McKusick	Johns Hopkins University Hospital
Robert F. Murray	Howard University
Thomas H. Murray	Case Western Reserve University

Joint Mapping Working Group

(First met December 1989)

David Botstein	Stanford University
Anthony V. Carrano	Lawrence Livermore National Laboratory
C. Thomas Caskey	Baylor College of Medicine
David R. Cox	University of California, San Francisco
Robert K. Moyzis	Los Alamos National Laboratory
Maynard V. Olson	Washington University

Joint Informatics Task Force (*ad hoc*)

(First met March 7–9, 1990; final meeting January 3, 1992)

Dieter Soll (Chair)	Yale University
George I. Bell	Los Alamos National Laboratory
David Botstein	Stanford University
Elbert Branscomb	Lawrence Livermore National Laboratory
John Devereux	Genetics Computer Group
Nathan Goodman	Whitehead Institute
Gregory Hamm	Rutgers University Waksman Institute
Eric Lander	Massachusetts Institute of Technology
Frank Olken	Lawrence Berkeley Laboratory
Mark L. Pearson	E. I. du Pont de Nemours & Company
Sylvia J. Spengler	Lawrence Berkeley Laboratory
Michael Waterman	University of Southern California

Joint Sequencing Working Group
(First met May 10, 1990)

Ellson Chen	Genentech, Inc.
Ronald Davis	Stanford University
John Devereux	Genetics Computer Group
Walter Gilbert	Harvard University
Leroy E. Hood	California Institute of Technology
Mark L. Pearson	E.I. du Pont de Nemours & Company
Joseph Sambrook	University of Texas
Phillip A. Sharp	Massachusetts Institute of Technology
William Studier	Brookhaven National Laboratory

Joint Working Group on the Mouse
(First met May 6, 1991)

Verne Chapman (Chair)	Roswell Park Memorial Institute
Frank Constantini	Columbia University
Neal Copeland	National Cancer Institute-Frederick Cancer Research and Development Center
William Dove	University of Wisconsin, Madison
Joseph Nadeau	Jackson Laboratory
Roger Reeves	Johns Hopkins University
Janet Rossant	Mt. Sinai Hospital
Oliver Smithies	University of North Carolina, Chapel Hill
Richard Woychik	Oak Ridge National Laboratory

Glossary

Appendix E

Appendix E: Glossary

Portions of the glossary text were taken directly or modified from definitions in the U.S. Congress Office of Technology Assessment document: *Mapping Our Genes—The Genome Projects: How Big, How Fast?* OTA-BA-373, Washington, D.C.: U.S. Government Printing Office, April 1988.

Adenine (A): A nitrogenous base, one member of the *base pair* A-T (adenine-*thymine*).

Alleles: Alternative forms of a genetic *locus*; a single allele for each locus is inherited separately from each parent (e.g., at a locus for eye color the allele might result in blue or brown eyes).

Amino acid: Any of a class of 20 molecules that are combined to form *proteins* in living things. The sequence of amino acids in a protein and hence protein function are determined by the *genetic code.*

Amplification: An increase in the number of copies of a specific DNA fragment; can be in vivo or in vitro. See *cloning, polymerase chain reaction.*

Arrayed library: Individual primary recombinant clones (hosted in *phage, cosmid, YAC,* or other *vector)* that are placed in two-dimensional arrays in microtiter dishes. Each primary clone can be identified by the identity of the plate and the clone location (row and column) on that plate. Arrayed libraries of clones can be used for many applications, including screening for a specific *gene* or genomic region of interest as well as for *physical mapping.* Information gathered on individual clones from various genetic *linkage* and *physical map* analyses is entered into a relational database and used to construct physical and genetic *linkage maps* simultaneously; clone identifiers serve to interrelate the multilevel maps. Compare *library, genomic library.*

Autoradiography: A technique that uses X-ray film to visualize radioactively labeled molecules or fragments of molecules; used in analyzing length and number of DNA fragments after they are separated by gel *electrophoresis.*

Autosome: A *chromosome* not involved in sex determination. The *diploid* human *genome* consists of 46 chromosomes, 22 pairs of autosomes, and 1 pair of *sex chromosomes* (the X and Y chromosomes).

Bacteriophage: See *phage.*

Base pair (bp): Two nitrogenous bases (*adenine* and *thymine* or *guanine* and *cytosine*) held together by weak bonds. Two strands of DNA are held together in the shape of a double helix by the bonds between base pairs.

Base sequence: The order of *nucleotide* bases in a DNA molecule.

Base sequence analysis: A method, sometimes automated, for determining the *base sequence.*

Biotechnology: A set of biological techniques developed through basic research and now applied to research and product development. In particular, the use by industry of *recombinant DNA,* cell fusion, and new bioprocessing techniques.

bp: See *base pair.*

cDNA: See *complementary DNA.*

Centimorgan (cM): A unit of measure of *recombination* frequency. One centimorgan is equal to a 1% chance that a marker at one genetic *locus* will be separated from a marker at a second locus due to *crossing over* in a single generation. In human beings, 1 centimorgan is equivalent, on average, to 1 million *base pairs*.

Centromere: A specialized *chromosome* region to which spindle fibers attach during cell division.

Chromosomes: The self-replicating genetic structures of cells containing the cellular DNA that bears in its *nucleotide* sequence the linear array of *genes*. In *prokaryotes*, chromosomal DNA is circular, and the entire genome is carried on one chromosome. *Eukaryotic* genomes consist of a number of chromosomes whose DNA is associated with different kinds of *proteins*.

Clone bank: See *genomic library*.

Clones: A group of cells derived from a single ancestor.

Cloning: The process of asexually producing a group of cells (clones), all genetically identical, from a single ancestor. In *recombinant DNA technology,* the use of DNA manipulation procedures to produce multiple copies of a single *gene* or segment of DNA is referred to as cloning DNA.

Cloning vector: DNA molecule originating from a *virus*, a *plasmid,* or the cell of a higher organism into which another DNA fragment of appropriate size can be integrated without loss of the vector's capacity for self-replication; vectors introduce foreign DNA into host cells, where it can be reproduced in large quantities. Examples are *plasmids*, *cosmids*, and *yeast artificial chromosomes*; vectors are often *recombinant* molecules containing DNA sequences from several sources.

cM: See *centimorgan*.

Code: See *genetic code*.

Codon: See *genetic code*.

Complementary DNA (cDNA): DNA that is synthesized from a *messenger RNA* template; the single-stranded form is often used as a *probe* in *physical mapping*.

Complementary sequences: *Nucleic acid base sequences* that can form a double-stranded structure by matching *base pairs*; the complementary sequence to G-T-A-C is C-A-T-G.

Conserved sequence: A *base sequence* in a DNA molecule (or an *amino acid* sequence in a *protein*) that has remained essentially unchanged throughout evolution.

Contig map: A map depicting the relative order of a linked *library* of small overlapping clones representing a complete chromosomal segment.

Appendix E: Glossary

Contigs: Groups of *clones* representing overlapping regions of a *genome.*

Cosmid: Artificially constructed *cloning vector* containing the *cos* gene of *phage* lambda. Cosmids can be packaged in lambda phage particles for infection into *E. coli*; this permits cloning of larger DNA fragments (up to 45 kb) than can be introduced into bacterial hosts in *plasmid* vectors.

Crossing over: The breaking during *meiosis* of one maternal and one paternal *chromosome*, the exchange of corresponding sections of DNA, and the rejoining of the chromosomes. This process can result in an exchange of *alleles* between chromosomes. Compare *recombination.*

Cytosine (C): A *nitrogenous base*, one member of the *base pair* G-C (*guanine* and cytosine).

Deoxyribonucleotide: See *nucleotide.*

Diploid: A full set of genetic material, consisting of paired *chromosomes*—one chromosome from each parental set. Most animal cells except the *gametes* have a diploid set of chromosomes. The diploid human *genome* has 46 chromosomes. Compare *haploid.*

DNA (deoxyribonucleic acid): The molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between *base pairs* of *nucleotides.* The four nucleotides in DNA contain the bases: *adenine* (A), *guanine* (G), *cytosine* (C), and *thymine* (T). In nature, *base pairs* form only between A and T and between G and C; thus the *base sequence* of each single strand can be deduced from that of its partner.

DNA probes: See *probe.*

DNA replication: The use of existing DNA as a template for the synthesis of new DNA strands. In humans and other *eukaryotes*, replication occurs in the cell *nucleus.*

DNA sequence: The relative order of *base pairs*, whether in a fragment of DNA, a *gene*, a *chromosome*, or an entire *genome*. See *base sequence analysis.*

Domain: A discrete portion of a *protein* with its own function. The combination of domains in a single protein determines its overall function.

Double helix: The shape that two linear strands of DNA assume when bonded together.

***E. coli*:** Common bacterium that has been studied intensively by geneticists because of its small genome size, normal lack of pathogenicity, and ease of growth in the laboratory.

Electrophoresis: A method of separating large molecules (such as DNA fragments or *proteins*) from a mixture of similar molecules. An electric current is passed through a medium containing the mixture, and each kind of molecule travels through the medium at a different rate, depending on its electrical charge and size. Separation is based on these differences. Agarose and acrylamide gels are the media commonly used for electrophoresis of proteins and nucleic acids.

Endonuclease: An *enzyme* that cleaves its nucleic acid substrate at internal sites in the *nucleotide* sequence.

Enzyme: A *protein* that acts as a catalyst, speeding the rate at which a biochemical reaction proceeds but not altering the direction or nature of the reaction.

EST: Expressed sequence tag. See *sequence tagged site.*

Eukaryote: Cell or organism with membrane-bound, structurally discrete *nucleus* and other well-developed subcellular compartments. Eukaryotes include all organisms except *viruses*, bacteria, and blue-green algae. Compare *prokaryote.* See *chromosomes.*

Evolutionarily conserved: See *conserved sequence.*

Exogenous DNA: DNA originating outside an organism.

Exons: The *protein*-coding DNA sequences of a *gene.* Compare *introns.*

Exonuclease: An *enzyme* that cleaves *nucleotides* sequentially from free ends of a linear nucleic acid substrate.

Expressed gene: See *gene expression.*

FISH (fluorescence in situ hybridization): A *physical mapping* approach that uses fluorescein tags to detect *hybridization* of *probes* with *metaphase chromosomes* and with the less-condensed *somatic interphase* chromatin.

Flow cytometry: Analysis of biological material by detection of the light-absorbing or fluorescing properties of cells or subcellular fractions (i.e., *chromosomes*) passing in a narrow stream through a laser beam. An absorbance or fluorescence profile of the sample is produced. Automated sorting devices, used to fractionate samples, sort successive droplets of the analyzed stream into different fractions depending on the fluorescence emitted by each droplet.

Flow karyotyping: Use of flow cytometry to analyze and/or separate *chromosomes* on the basis of their DNA content.

Gamete: Mature male or female reproductive cell (sperm or ovum) with a *haploid* set of *chromosomes* (23 for humans).

Gene: The fundamental physical and functional unit of heredity. A *gene* is an ordered sequence of *nucleotides* located in a particular position on a particular *chromosome* that encodes a specific functional product (i.e., a *protein* or *RNA* molecule). See *gene expression.*

Gene expression: The process by which a *gene's* coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into *mRNA* and then translated into *protein* and those that are transcribed into *RNA* but not translated into protein (e.g., *transfer* and *ribosomal RNAs*).

Appendix E: Glossary

Gene families: Groups of closely related *genes* that make similar products.

Gene library: See *genomic library.*

Gene mapping: Determination of the relative positions of *genes* on a DNA molecule (*chromosome* or *plasmid*) and of the distance, in *linkage* units or physical units, between them.

Gene product: The biochemical material, either *RNA* or *protein*, resulting from expression of a gene. The amount of gene product is used to measure how active a gene is; abnormal amounts can be correlated with disease-causing alleles.

Genetic code: The sequence of *nucleotides*, coded in triplets (*codons*) along the *mRNA*, that determines the sequence of *amino acids* in *protein* synthesis. The DNA sequence of a *gene* can be used to predict the mRNA sequence, and the genetic code can in turn be used to predict the *amino acid* sequence.

Genetic engineering technologies: See *recombinant DNA technologies.*

Genetic map: See *linkage map.*

Genetic material: See *genome.*

Genetics: The study of the patterns of inheritance of specific traits.

Genome: All the genetic material in the *chromosomes* of a particular organism; its size is generally given as its total number of *base pairs.*

Genome projects: Research and technology development efforts aimed at *mapping* and *sequencing* some or all of the *genome* of human beings and other organisms.

Genomic library: A collection of *clones* made from a set of randomly generated overlapping DNA fragments representing the entire *genome* of an organism. Compare *library, arrayed library.*

Guanine (G): A nitrogenous base, one member of the *base pair* G-C (guanine and *cytosine*).

Haploid: A single set of *chromosomes* (half the full set of genetic material), present in the egg and sperm cells of animals and in the egg and pollen cells of plants. Human beings have 23 chromosomes in their reproductive cells. Compare *diploid.*

Heterozygosity: The presence of different *alleles* at one or more *loci* on *homologous chromosomes.*

Homeobox: A short stretch of *nucleotides* whose *base sequence* is virtually identical in all the *genes* that contain it. It has been found in many organisms from fruit flies to human beings. In the fruit fly, a homeobox appears to determine when particular groups of genes are expressed during development.

Homologies: Similarities in DNA or *protein* sequences between individuals of the same species or among different species.

Homologous chromosomes: A pair of *chromosomes* containing the same linear *gene* sequences, each derived from one parent.

Human gene therapy: Insertion of normal DNA directly into cells to correct a genetic defect.

Human Genome Initiative: Collective name for several projects begun in 1986 by DOE to (1) create an ordered set of DNA segments from known chromosomal locations, (2) develop new computational methods for analyzing genetic map and DNA sequence data, and (3) develop new techniques and instruments for detecting and analyzing DNA. This DOE initiative is now known as the Human Genome Program. The national effort, led by DOE and NIH, is known as the Human Genome Project.

Hybridization: The process of joining two *complementary* strands of DNA or one each of DNA and RNA to form a double-stranded molecule.

Informatics: The study of the application of computer and statistical techniques to the management of information. In *genome* projects, informatics includes the development of methods to search databases quickly, to analyze DNA sequence information, and to predict *protein* sequence and structure from DNA sequence data.

In situ hybridization: Use of a DNA or RNA probe to detect the presence of the *complementary DNA* sequence in cloned bacterial or cultured *eukaryotic* cells.

Interphase: The period in the cell cycle when DNA is replicated in the nucleus; followed by *mitosis*.

Introns: The DNA *base sequences* interrupting the *protein*-coding sequences of a *gene*; these sequences are *transcribed* into *RNA* but are cut out of the message before it is *translated* into protein. Compare *exons*.

In vitro: Outside a living organism.

Karyotype: A photomicrograph of an individual's *chromosomes* arranged in a standard format showing the number, size, and shape of each chromosome type; used in low-resolution *physical mapping* to correlate gross chromosomal abnormalities with the characteristics of specific diseases.

kb: See *kilobase.*

Kilobase (kb): Unit of length for DNA fragments equal to 1000 *nucleotides*.

Library: An unordered collection of *clones* (i.e., cloned DNA from a particular organism), whose relationship to each other can be established by *physical mapping*. Compare *genomic library, arrayed library.*

Appendix E: Glossary

Linkage: The proximity of two or more *markers* (e.g., *genes*, *RFLP* markers) on a *chromosome*; the closer together the markers are, the lower the probability that they will be separated during DNA repair or replication processes (binary fission in *prokaryotes*, *mitosis* or *meiosis* in *eukaryotes*), and hence the greater the probability that they will be inherited together.

Linkage map: A map of the relative positions of genetic *loci* on a *chromosome*, determined on the basis of how often the loci are inherited together. Distance is measured in *centimorgans* (*cM*).

Localize: Determination of the original position (*locus)* of a *gene* or other *marker* on a chromosome.

Locus (pl. loci): The position on a *chromosome* of a *gene* or other chromosome *marker*; also, the DNA at that position. The use of *locus* is sometimes restricted to mean regions of DNA that are *expressed*. See *gene expression*.

Macrorestriction map: Map depicting the order of and distance between sites at which *restriction enzymes* cleave *chromosomes*.

Mapping: See *gene mapping, linkage map, physical map*.

Marker: An identifiable physical location on a *chromosome* (e.g., *restriction enzyme cutting site*, *gene*) whose inheritance can be monitored. Markers can be expressed regions of DNA (genes) or some segment of DNA with no known coding function but whose pattern of inheritance can be determined. See *RFLP, restriction fragment length polymorphism.*

Mb: See *megabase.*

Megabase (Mb): Unit of length for DNA fragments equal to 1 million *nucleotides* and roughly equal to 1 *cM.*

Meiosis: The process of two consecutive cell divisions in the *diploid* progenitors of sex cells. Meiosis results in four rather than two daughter cells, each with a *haploid* set of *chromosomes.*

Messenger RNA (mRNA): RNA that serves as a template for *protein* synthesis. See *genetic code.*

Metaphase: A stage in *mitosis* or *meiosis* during which the *chromosomes* are aligned along the equatorial plane of the cell.

Mitosis: The process of nuclear division in cells that produces daughter cells that are genetically identical to each other and to the parent cell.

mRNA: See *messenger RNA.*

Multifactorial or multigenic disorders: See *polygenic disorders.*

Multiplexing: A *sequencing* approach that uses several pooled samples simultaneously, greatly increasing sequencing speed.

Mutation: Any heritable change in DNA *sequence*. Compare *polymorphism*.

Nitrogenous base: A nitrogen-containing molecule having the chemical properties of a base.

Nucleic acid: A large molecule composed of *nucleotide* subunits.

Nucleotide: A subunit of DNA or *RNA* consisting of a nitrogenous base (*adenine*, *guanine*, *thymine*, or *cytosine* in DNA; adenine, guanine, *uracil*, or cytosine in RNA), a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA). Thousands of *nucleotides* are linked to form a DNA or RNA molecule. See *DNA, base pair, RNA.*

Nucleus: The cellular organelle in *eukaryotes* that contains the genetic material.

Oncogene: A *gene*, one or more forms of which is associated with cancer. Many oncogenes are involved, directly or indirectly, in controlling the rate of cell growth.

Overlapping clones: See *genomic library.*

PCR: See *polymerase chain reaction.*

Phage: A *virus* for which the natural host is a bacterial cell.

Physical map: A map of the locations of identifiable landmarks on DNA (e.g., *restriction enzyme cutting sites*, *genes*), regardless of inheritance. Distance is measured in *base pairs*. For the human *genome*, the lowest-resolution *physical map* is the banding patterns on the 24 different *chromosomes*; the highest-resolution map would be the complete *nucleotide* sequence of the chromosomes.

Plasmid: Autonomously replicating, extrachromosomal circular DNA molecules, distinct from the normal bacterial *genome* and nonessential for cell survival under nonselective conditions. Some plasmids are capable of integrating into the host genome. A number of artificially constructed plasmids are used as *cloning vectors.*

Polygenic disorders: Genetic disorders resulting from the combined action of *alleles* of more than one *gene* (e.g., heart disease, diabetes, and some cancers). Although such disorders are inherited, they depend on the simultaneous presence of several alleles; thus the hereditary patterns are usually more complex than those of single-gene disorders. Compare *single-gene disorders.*

Polymerase chain reaction (PCR): A method for amplifying a DNA *base sequence* using a heat-stable *polymerase* and two 20-base *primers*, one *complementary* to the (+)-strand at one end of the sequence to be amplified and the other complementary to the (–)-strand at the other end. Because the newly synthesized DNA strands can subsequently serve as additional templates for the same primer sequences, successive rounds of primer annealing, strand elongation, and dissociation produce rapid and highly specific amplification of the desired sequence. PCR also can be used to detect the existence of the defined sequence in a DNA sample.

Polymerase, DNA or RNA: *Enzymes* that catalyze the synthesis of *nucleic acids* on preexisting nucleic acid templates, assembling RNA from ribonucleotides or DNA from deoxyribonucleotides.

Polymorphism: Difference in DNA sequence among individuals. Genetic variations occurring in more than 1% of a population would be considered useful polymorphisms for genetic *linkage* analysis. Compare *mutation.*

Primer: Short preexisting polynucleotide chain to which new deoxyribonucleotides can be added by DNA *polymerase.*

Probe: Single-stranded DNA or *RNA* molecules of specific base *sequence,* labeled either radioactively or immunologically, that are used to detect the *complementary* base sequence by *hybridization.*

Prokaryote: Cell or organism lacking a membrane-bound, structurally discrete *nucleus* and other subcellular compartments. Bacteria are prokaryotes. Compare *eukaryote.* See *chromosomes.*

Promoter: A site on DNA to which *RNA polymerase* will bind and initiate *transcription.*

Protein: A large molecule composed of one or more chains of *amino acids* in a specific order; the order is determined by the *base sequence* of *nucleotides* in the *gene* coding for the protein. Proteins are required for the structure, function, and regulation of the body's cells, tissues, and organs, and each protein has unique functions. Examples are hormones, *enzymes,* and antibodies.

Purine: A nitrogen-containing, single-ring, basic compound that occurs in nucleic acids. The purines in DNA and RNA are adenine and guanine.

Pyrimidine: A nitrogen-containing, double-ring, basic compound that occurs in nucleic acids. The pyrimidines in DNA are cytosine and thymine; in RNA, cytosine and uracil.

Rare-cutter enzyme: See *restriction enzyme cutting site.*

Recombinant clones: *Clones* containing *recombinant DNA molecules.* See *recombinant DNA technologies.*

Recombinant DNA molecules: A combination of DNA molecules of different origin that are joined using *recombinant DNA technologies.*

Recombinant DNA technologies: Procedures used to join together DNA segments in a cell-free system (an environment outside a cell or organism). Under appropriate conditions, a recombinant DNA molecule can enter a cell and replicate there, either autonomously or after it has become integrated into a cellular *chromosome.*

Recombination: The process by which progeny derive a combination of *genes* different from that of either parent. In higher organisms, this can occur by *crossing over.*

Regulatory regions or sequences: A DNA *base sequence* that controls *gene expression.*

Resolution: Degree of molecular detail on a *physical map* of DNA, ranging from low to high.

Restriction enzyme, endonuclease: A *protein* that recognizes specific, short *nucleotide sequences* and cuts DNA at those sites. Bacteria contain over 400 such *enzymes* that recognize and cut over 100 different DNA sequences. See *restriction enzyme cutting site.*

Restriction enzyme cutting site: A specific *nucleotide sequence* of DNA at which a particular *restriction enzyme* cuts the DNA. Some sites occur frequently in DNA (e.g., every several hundred *base pairs*), others much less frequently (*rare-cutter*; e.g., every 10,000 base pairs).

Restriction fragment length polymorphism (RFLP): Variation between individuals in DNA fragment sizes cut by specific *restriction enzymes*; *polymorphic sequences* that result in RFLPs are used as *markers* on both *physical maps* and genetic *linkage maps.* RFLPs are usually caused by *mutation* at a cutting site. See *marker.*

RFLP: See *restriction fragment length polymorphism.*

Ribonucleic acid (RNA): A chemical found in the *nucleus* and cytoplasm of cells; it plays an important role in *protein* synthesis and other chemical activities of the cell. The structure of RNA is similar to that of DNA. There are several classes of RNA molecules, including *messenger RNA*, *transfer RNA*, *ribosomal RNA*, and other small RNAs, each serving a different purpose.

Ribonucleotides: See *nucleotide.*

Ribosomal RNA (rRNA): A class of RNA found in the ribosomes of cells.

Ribosomes: Small cellular components composed of specialized ribosomal RNA and protein; site of protein synthesis. See *ribonucleic acid* (*RNA*).

RNA: See *ribonucleic acid.*

Sequence*:* See *base sequence.*

Sequence tagged site (STS): Short (200 to 500 *base pairs*) DNA sequence that has a single occurrence in the human *genome* and whose location and base sequence are known. Detectable by *polymerase chain reaction*, STSs are useful for localizing and orienting the mapping and sequence data reported from many different laboratories and serve as landmarks on the developing *physical map* of the human genome. Expressed sequence tags (ESTs) are STSs derived from cDNAs.

Sequencing: Determination of the order of *nucleotides* (*base sequences*) in a DNA or *RNA* molecule or the order of *amino acids* in a *protein.*

Sex chromosomes: The X and Y *chromosomes* in human beings that determine the sex of an individual. Females have two X chromosomes in diploid cells; males have an X and a Y chromosome. The sex chromosomes comprise the 23rd chromosome pair in a *karyotype.* Compare *autosome.*

Shotgun method: *Cloning* of DNA fragments randomly generated from a *genome.* See *library, genomic library.*

Single-gene disorder: Hereditary disorder caused by a *mutant* allele of a single *gene* (e.g., Duchenne muscular dystrophy, retinoblastoma, sickle cell disease). Compare *polygenic disorders.*

Somatic cells: Any cell in the body except *gametes* and their precursors.

Southern blotting: Transfer by absorption of DNA fragments separated in electrophoretic gels to membrane filters for detection of specific *base sequences* by radiolabeled complementary probes.

STS: See *sequence tagged site.*

Tandem repeat sequences: Multiple copies of the same *base sequence* on a *chromosome*; used as a marker in *physical mapping.*

Technology transfer: The process of converting scientific findings from research laboratories into useful products by the commercial sector.

Telomere: The ends of *chromosomes.* These specialized structures are involved in the replication and stability of linear DNA molecules. See *DNA replication.*

Thymine (T): A nitrogenous base, one member of the *base pair* A-T (*adenine*-thymine).

Transcription: The synthesis of an *RNA* copy from a *sequence* of DNA (a *gene*); the first step in *gene expression.* Compare *translation.*

Transfer RNA (tRNA): A class of *RNA* having structures with triplet *nucleotide* sequences that are *complementary* to the triplet nucleotide coding sequences of *mRNA.* The role of tRNAs in protein synthesis is to bond with *amino acids* and transfer them to the ribosomes, where proteins are assembled according to the genetic code carried by mRNA.

Transformation: A process by which the genetic material carried by an individual cell is altered by incorporation of exogenous DNA into its *genome.*

Translation: The process in which the genetic code carried by mRNA directs the synthesis of *proteins* from *amino acids.* Compare *transcription.*

tRNA: See *transfer RNA.*

Uracil: A nitrogenous base normally found in RNA but not DNA; uracil is capable of forming a *base pair* with *adenine.*

Vector: See *cloning vector.*

Virus: A noncellular biological entity that can reproduce only within a host cell. Viruses consist of *nucleic acid* covered by *protein*; some animal viruses are also surrounded by membrane. Inside the infected cell, the virus uses the synthetic capability of the host to produce progeny virus.

VLSI: Very large-scale integration allowing over 100,000 transistors on a chip.

YAC: See *yeast artificial chromosome.*

Yeast artificial chromosome (YAC): A vector used to clone DNA fragments (up to 400 kb); it is constructed from the telomeric, centromeric, and replication origin sequences needed for replication in yeast cells. Compare *cloning vector, cosmid.*

Index to Principal and Coinvestigators Listed in Abstracts

Index to Principal and Coinvestigators Listed in Abstracts

Adams, Mark 97
Adamson, Anne 164
Alexander, Peter 182
Allen, Michael 177
Allison, David 125
Amemiya, Chris 100, 104
Anderson, N. Leigh 168
Anderson, Norman 168
Anderson, W. Holt 167
Andreason, Grai 174
Antonarakis, Stylianos 172
Apostolou, Sinoula 108
Apsell, Paula 156
Arenstorf, Hartwig 171
Arlinghaus, Heinrich 127, 165, 168, 177
Ashworth, Linda 104
Aslandidis, Charalampos 100
Athwal, Raghbir 82
Bacha, Hamid 185
Baker, Elizabeth 108
Baker, Mark 123
Balding, David 152
Balhorn, Rodney 177, 181
Balooch, Mehdi 177
Barker, David 172
Beckwith, Jonathon 186
Beeson, Diane 158
Benner, W. Henry 128
Berg, Claire 137, 178
Berg, Douglas 178
Beugelsdijk, Tony 111, 116
Birdsall, David 181
Birren, Bruce 105
Black, Lindsay 93
Blackwell, Tom 152
Bonaldo, Maria 96
Bouma, Hessel III 157
Boyartchuk, Victor 83
Bradbury, E. Morton 82
Brandriff, Brigitte 109
Branscomb, Elbert 103, 139
Brase, James 181
Bremer, Meire 105
Brennan, Thomas 119
Bridgers, Michael 141
Brody, Linnea 98
Bronstein, Irena 167, 170
Brown, Gilbert 119, 125, 127
Brown, Henry 141
Brown, Stephen 96
Brule, James 185
Bulger, Ruth 156
Burks, Christian 141
Callen, David 106, 108
Campbell, Evelyn 83, 89
Campbell, Mary 83, 89
Cantor, Charles 111, 163
Carrano, Anthony 84, 88, 94, 100, 103, 104, 109, 139
Casey, Denise 164
Caskey, C. Thomas 99, 157
Cassatt, James 140
Chait, Brian 136
Chedd, Graham 156
Chen, Chira 104
Chen, C. H. Winston 120
Chen, Ed 143
Chen, Jiun 132
Chen, Liang 108
Chen, Shizhong 174
Cheng, Jan-Fang 83, 100
Cherkauer, Kevin 151
Church, George 121
Cinkosky, Michael 141, 141
Clancy, Suzanne 101
Clark, Steven 101
Collins, Debra 157
Combs, Jesse 104
Copeland, Alex 104
Corona, Angela 184
Crandall, Lee 161
Craven, Mark 151
Crkvenjakov, Radomir 121
Davidson, Jack 122
Davidson, K. Alicia 164
Deaven, Larry 83, 84, 89, 106
de Jong, Pieter 84, 100, 103, 104
Denton, M. Bonner 123
Djbali, Malek 174
Doggett, Norman 106, 108
Dougherty, Randall 141
Douthart, Richard 140
Dovichi, Norman 124
Drmanac, Radoje 121
Dubnick, Mark 97
Dunn, John 138

Durkin, Scott 92
Duster, Troy 158
Earle, Colin 123
Edmonds, Charles 136
Edwards, Brooks 170
Efstratiadis, Agiris 96
Einstein, Ralph 154
Entine, Gerald 173
Epling, Gary 96
Eubanks, James 174
Evans, Glen 101, 174
Faber, Vance 141
Fader, Betsy 186
Fain, Pamela 172
Fairfield, Frederic 152
Fawcett, John 89
Feitshans, Ilise 159
Ferrell, Thomas 125
Fickett, James 141, 152
Fields, Christopher 97, 142
Fischer, Peggy 186
Flatley, Jay 167
Fockler, Carita 105
Foote, Robert 119, 122, 125, 127
Francomano, Clare 148
Fullarton, Jane 156
Furuya, Frederic 166
Gabra, Nashua 175
Gatewood, Joe 82
Generoso, Estela 107
Gesteland, Raymond 126
Gibson, William 165
Giddings, J. Calvin 112
Gingrich, Jeffrey 102
Giovannini, Marco 174
Glazer, Alexander 132
Goldberg, Mark 141
Gong, Kevin 145
Grad, Frank 159
Grady, Deborah 84
Gray, Joe 85, 181
Greener, Phillip 98
Gusfield, Daniel 144, 148
Hahn, Peter 88
Hainfeld, James 112, 166
Hansen, Tony 113
Hart, Reece 174
Hartman, John 167, 169
Haugland, Richard 179
Hempfner, Philip 141
Henderson, Margaret 161
Hermanson, Gary 101, 174
Hewitt, Peter 179
Hieter, Philip 172
Hildebrand, C. Edgar 106, 108
Himawan, Jeff 133
Hoekstra, Merl 101
Hoffman, Lance 186
Hofmann, Gunter 174
Hollen, Robert 111, 116
Holmes, Linda 163
Holtzman, Neil 159
Honda, Sandra 144
Hood, Leroy 126, 143
Hopkins, Janet 95
Hozier, John 88
Huang, Henry 178
Huang, Xiaohua 132
Huber, Hans 180
Huhn, Greg 174
Hunkapiller, Tim 143, 147, 148
Hurst, Gerald 166
Hutchinson, Marge 145, 155
Imara, Mwalimu 161
Jackson, Cynthia 86
Jacobson, K. Bruce 119, 120, 125, 125, 127
Jaklevic, Joseph 113, 115, 117, 128, 128
Jelenc, Pierre 96
Jett, James 114, 129
Johnson, Lori 104
Juo, Rouh-Rong 170
Jurka, Jerzy 144
Kandpal, Rajendra 171
Kang, Hee-Chol 179
Kao, Fa-Ten 86
Karger, Barry 114
Katz, Joseph 115, 128
Kaufman, Daniel 174
Keller, Richard 123, 129
Kelley, Jenny 97
Kerlavage, Anthony 97
Khan, Akbar 95
Knoche, Kimberly 166
Kolbe, William 113, 115, 128

Index to Principal and Coinvestigators Listed in Abstracts

Kopelman, Raoul 130
Korenberg, Julie 87
Kozman, Helen 108
Kozubel, Mark 116
Kuo, Wen-Lin 85
Lai, Tran 141
Lane, Michael 88
Lane, Sharon 108
Langmore, John 130
Lappé, Marc 186
Larimer, Frank 119, 127
Lawler, Eugene 144, 148
Lawrence, Charles 144
Lee, Bill 94
Lennon, Gregory 88
Leonard, Lisa 174
Lerman, Leonard 115, 175
Lewis, Suzanna 145, 145, 147
Longmire, Jon 84
Loo, Joseph 136
Lowery, Robert 166
Lowry, Steven 102
Lumley, Amanda 163
Macken, Catherine 152
Maglott, Donna 92, 171
Makarov, Vladimir 130
Mann, Reinhold 154
Mansfield, Betty 164
Mark, Hon Fong 86
Markowitz, Victor 145, 146, 147
Marr, Thomas 148
Martin, Christopher 170
Martin, Christopher H. 89, 103, 131
Martin, John 114, 129
Martin, Sheryl 164
Matheson, Nina 148
Mathies, Richard 132
Maurer, Susanne 174
Mayeda, Carol 89, 103, 131
McAllister, Douglass 167
McCarthy, John 145
McCormick, MaryKay 84, 89, 106
McElligott, David 101, 174
McInerney, Joseph 159
McKean, Ronald 175
McKusick, Victor 148
Mead, David 132
Medvick, Patricia 111, 116
Meincke, Linda 89
Merrill, Carl 92
Meyne, Julie 91
Michael, Sharon 97
Micklos, David 161
Milosavljevic, Aleksandar 144
Mohrenweiser, Harvey 88, 103
Moir, Donald 90
Moore, Stefan 160
Moreno, Ruben 97
Mosley, Ray 161
Moyzis, Robert 84, 89, 91, 106
Mucenski, Mike 94
Mulley, John 108
Mundt, Mark 141
Mural, Richard 153, 154
Murphy, Timothy 186
Myers, Eugene 144
Nagle, James 97
Nancarrow, Julie 108
Natowicz, Marvin 160
Nelson, David 139
Nelson, David L. 91, 99
Nelson, Debra 141
Nelson, J. Robert 157
Nierman, William 92, 171
Nikolic, Julia 100
Noordewier, Michiel 151
Oehler, Chuck 166
Okumura, K. 106
Olken, Frank 145, 145, 148
Olsen, Anne 104
Orpana, Arto 95
Orr, Bradford 130
Overbeek, Ross 182
Page, George 160
Palazzolo, Michael 89, 103, 131
Parimoo, Satish 171
Patanjali, Sankhavaram 171
Payne, Marvin 120
Pearson, Peter 148
Pecherer, Robert 141, 141
Pelkey, Joanne 140
Peters, Don 85
Pfeifer, Gerd 135
Phillips, Hilary 108
Phoenix, David 161
Pinkel, Dan 85

Pirrung, Michael 133
Polymeropoulos, Mihael 92
Powell, Richard 166
Pratt, Lorien 151
Quesada, Mark 132
Radspinner, David 123
Ramsey, Roswitha 119
Rao, Venigalla 93
Ratliff, Robert 91
Reilly, Phillip 161
Reiner, Andrew 148
Richards, Robert 108
Richardson, Charles 133, 180
Riggs, Arthur 135
Rinchik, Eugene 93, 94, 107, 153
Ringold, Gordon 167
Robbins, Robert 148
Roberts, Randy 111, 116
Roman, Maria 174
Romo, Anthony 174
Roszak-MacDonell, Darlene 167
Rush, John 180
Rye, Hays 132
Sachleben, Richard 119, 125, 127
Sainz, Jesus 105
Saleh, Mary 174
Schenk, Karen 152
Schimke, R. Neil 157
Schmitt, Eric 175
Schwartz, Stanley 183
Searls, David 150
Segebrecht, Linda 157
Selleri, Licia 101, 174
Sgro, Peichen 141
Shavlik, Jude 151
Shen, Yang 108
Shera, E. Brooks 129
Shi, Zhong You 130
Shizuya, Hiroaki 105
Shoshani, Arie 146
Siciliano, Michael 94
Siekhaus, Wigbert 177
Sikela, James 95
Simon, Melvin 105
Sindelar, Linda 113
Slezak, Tom 139
Smith, Cassandra 105
Smith, Lloyd 132, 135, 136
Smith, Michael 101
Smith, Richard 136
Smith, Steven 130
Snider, Ken 174
Soares, Marcelo 96
Soderlund, Carol 142
Solomon, David 169
Sorenson, Doug 141, 141
Speed, Terence 148
Spengler, Sylvia 164
Stallings, Raymond 106, 108
Stevens, Tamara 95
Stiegman, Jeffrey 183, 184
Stinnett, Donna 164
Stormon, Charles 185
Storti, George 176
Stovall, Leonard 116
Strathmann, Michael 131
Strausbaugh, Linda 137
Stricker, Jenny 159
Stubbs, Lisa 107
Studier, F. William 138
Sudar, Damir 85
Sun, Tian-Qiang 97
Sutherland, Betsy 96
Sutherland, Grant 106, 108
Sutherland, Robert 141
Swaroop, Anand 171
Szeto, Ernest 146
Tabor, Stanley 133, 180
Tan, Weihong 130
Tang, Jane 100
Thakhar, Vishakha 93
Theil, Edward 113, 117
Thompson, Andrew 108
Thonnard, Norbert 127, 165
Thundat, Thomas 125
Thurman, David 140
Toliver, Greg 174
Torney, David 141, 152
Towell, Geoffrey 151
Trask, Barbara 84, 103, 109
Trebes, James 181
Trimmer, David 116
Trottier, Ralph 161
Troup, Charles 141
Tynan, Katherine 103, 109
Uber, Donald 113, 117

Appendix E: Index to Principal and Coinvestigators Listed in Abstracts

Uberbacher, Edward 153, 154
van den Engh, Ger 84, 88, 109
Varghese, Alison 170
Vaux, Kenneth 186
Venter, J. Craig 97
Vos, Jean-Michel 97
Voyta, John 170
Wagner, Caryn 174
Wahl, Geoffrey 98
Walichiewicz, Jolanta 144
Wang, Denan 105
Warburton, Dorothy 159
Ward, David 106, 171
Warmack, Robert 125
Wassom, John 164
Waterman, Michael 143
Weiss, Robert 126
Weissman, Sherman 171
Wendroff, Burton 152
West, John 185
Whitaker, James 179
Whitmore, Scott 108
Whitsitt, Andrew 151
Whittaker, Clive 152
Wilcox, Andrea 95
Wilder, Mark 114
Williams, Peter 138
Winternitz, Katherine 159
Witkowski, Jan 161
Wohlpart, Alfred 163
Woodbury, Neal 138
Woychik, Richard 93, 94, 119, 127, 153
Wright, James 163
Wyrick, Judy 164
Xiao, Hong 92
Yang, Sherman 144
Yantis, Bonnie 141
Yesley, Michael 164
Yeung, Edward 117
Yokobata, Kathy 100
Yorkey, Thomas 181
Yoshida, Kaoru 105
Youderian, Philip 98
Yu, Jing-Wei 86
Yust, Laura 164
Zhao, Jun 174
Zorn, Manfred 145, 155

Acronym List

AEC	Atomic Energy Commission
ANL*	Argonne National Laboratory, Argonne, IL
ATCC	American *Type Culture* Collection, Rockville, MD
BNL*	Brookhaven National Laboratory, Upton, NY
CEPH	Centre d'Etude du Polymorphisme Humain
CRADA	Cooperative Research and Development Agreement
DKFZ	German Cancer Research Center
DOE	Department of Energy
ERDA	Energy Research and Development Administration
FCCSET	Federal Coordinating Council on Science, Engineering, and Technology
GDB*†	Genome Data Base
HERAC*	Health and Environmental Research Advisory Committee
HGCC*	Human Genome Coordinating Committee
HGMIS*	Human Genome Management Information System (ORNL)
HUGO	Human Genome Organization (international)
JHU	Johns Hopkins University
JITF*†	Joint Informatics Task Force
LANL*	Los Alamos National Laboratory, Los Alamos, NM
LBL*	Lawrence Berkeley Laboratory, Berkeley, CA
LLNL*	Lawrence Livermore National Laboratory, Livermore, CA
MRC	Medical Research Council (U.K.)
NAS	National Academy of Sciences (U.S.)
NCHGR†	National Center for Human Genome Research
NIH†	National Institutes of Health, Bethesda, MD
NLGLP*	National Laboratory Gene Library Project (LANL, LLNL)
NRC	National Research Council (NAS)
NSF	National Science Foundation
OHER*	Office of Health and Environmental Research
ORNL*	Oak Ridge National Laboratory, Oak Ridge, TN
OSTP	Office of Scientific and Technology Policy (White House)
OTA	Office of Technology Assessment (U.S. Congress)
PACHG†	Program Advisory Committee on the Human Genome
PNL*	Pacific Northwest Laboratory, Richland, WA
SBIR	Small Business Innovation Research
SCC	Scientific Coordinating Committee
TWAS	Third World Academy of Sciences
UNESCO	United Nations Educational, Scientific, and Cultural Organization
USDA	U.S. Department of Agriculture

*Denotes U.S. Department of Energy organizations.

†Denotes U.S. Department of Health and Human Services organizations.