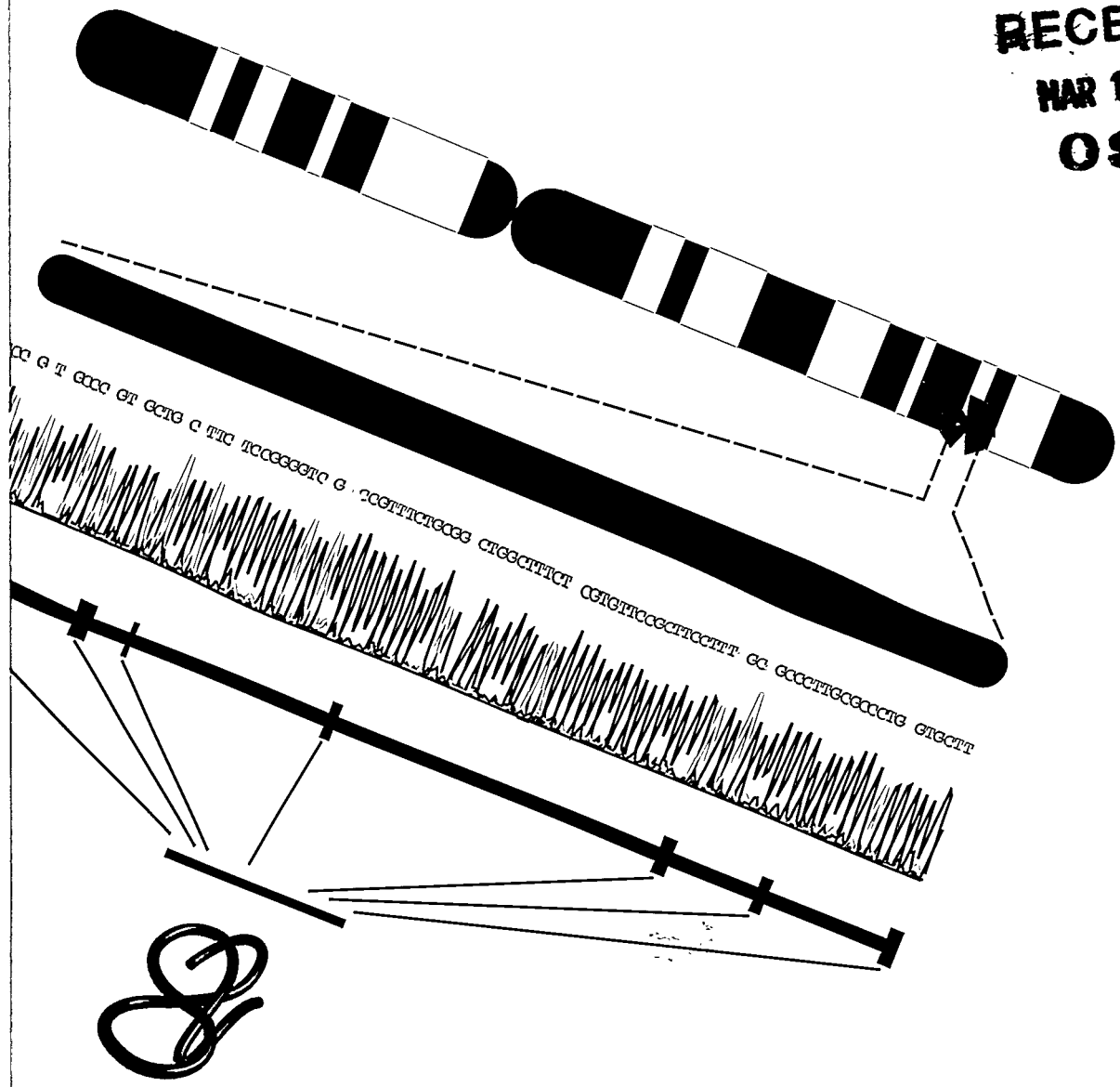
ORNL/M-6225

DOE/ER-0713 (Part 1)

HUMAN GENOME PROGRAM REPORT

Part 1, Overview and Progress



UNITED STATES DEPARTMENT OF ENERGY
OFFICE OF ENERGY RESEARCH
OFFICE OF BIOLOGICAL AND ENVIRONMENTAL RESEARCH

DOE HUMAN GENOME PROGRAM 1997

This is Part 1 of a two-part report published in 1997 to reflect research and progress in the U.S. Department of Energy (DOE) Human Genome Program from 1994 through 1996, with specified updates made just before publication. Part 1 is the program overview and report on progress, and Part 2 consists of 1996 research abstracts.

Print copies of Parts 1 and 2 and subsequent reports on DOE genome research are available upon request from the Human Genome Management Information System (HGMIS); Oak Ridge National Laboratory; 1060 Commerce Park; Oak Ridge, TN 37830 (423/576-6669, Fax: /574-9888, *bkq@ornl.gov*). Electronic versions are accessible via the DOE and HGMIS Web sites below.

- *http://www.er.doe.gov/production/ober/hug_top.html*
- *http://www.ornl.gov/hgmis/research.html*

This publication is available to DOE and DOE contractors from the Office of Scientific and Technical Information; P.O. Box 62; Oak Ridge, TN 37831 (423/576-8401). It is available to the public from the National Technical Information Service; U.S. Department of Commerce; 5285 Port Royal Road; Springfield, VA 22161.

COVER

Detailed chromosome descriptions, together with other biological resources, software, and instrumentation generated in the first 7 years of the DOE Human Genome Program (HGP), are enabling researchers to begin focusing on their most challenging goal: Determining the sequence of DNA subunits (the bases A, T, C, and G) found in the 24 different human chromosomes. Differences in DNA sequence underlie much of life's diversity.

The cover depicts the progress of human genome research, beginning with a microscopic view of a duplicated chromosome (top). Genome researchers begin with a very small chromosomal fragment (asterisk), using enzymes to cut it into the smaller pieces (red bars) required for DNA sequencing. Automated technology determines the DNA sequence of all or part of each fragment (graph with color-coded peaks).

Another HGP goal is to identify the estimated 70,000 to 100,000 genes, which account for only about 5% of human DNA. Computer analysis of DNA sequences is one way investigators identify gene features in DNA sequences (solid line with tick marks). In a living cell, individual gene segments from DNA molecules are assembled into short-lived intermediary molecules (short red line), and the information is translated by the cell's machinery into three-dimensional proteins (black globular structure at bottom). All organisms are made up largely of proteins that provide the structural components and specialized enzymes required by cells and tissues.

Public resources and technologies arising from the HGP and other genome efforts worldwide are laying the foundation for future explorations into the functions of each protein encoded by the genes. This research, which also will investigate how proteins work together in systems and pathways and react to external cues, will extend far into the future.

DOE/ER-0713 (Part 1)

ORNL/M--6225

HUMAN GENOME PROGRAM REPORT

Part 1, Overview and Progress

Date Published: November 1997

Prepared for the
U.S. Department of Energy
Office of Energy Research
Office of Biological and Environmental Research
Germantown, MD 20874-1290

Prepared by the
Human Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830-6480
managed by
Lockheed Martin Energy Research Corporation
for the
U.S. Department of Energy
Under Contract DE-AC05-96OR22464

Major Events in the U.S. Human Genome Project and Related Programs

1983

LANL and LLNL begin production of DNA clone (cosmid) libraries representing single chromosomes.

1984

DOE OHER and ICPEMC cosponsor Alta, Utah, conference highlighting the growing role of recombinant DNA technologies. OTA incorporates Alta proceedings into a 1986 report acknowledging value of human genome reference sequence.

1985

* Robert Sinsheimer holds meeting on human genome sequencing at University of California, Santa Cruz.

At OHER, Charles DeLisi and David A. Smith commission the first Santa Fe conference to assess the feasibility of a Human Genome Initiative.

1986

Following the Santa Fe conference, DOE OHER announces Human Genome Initiative. With $5.3 million, pilot projects begin at DOE national laboratories to develop critical resources and technologies.

1987

DOE advisory committee, HERAC, recommends a 15-year, multidisciplinary, scientific, and technological undertaking to map and sequence the human genome. DOE designates multidisciplinary human genome centers.

* NIH NIGMS begins funding of genome projects.

1988

* Reports by OTA and NAS NRC recommend concerted genome research program.

HUGO founded by scientists to coordinate efforts internationally.

* First annual Cold Spring Harbor Laboratory meeting held on human genome mapping and sequencing.

DOE and NIH sign MOU outlining plans for cooperation on genome research.

Telomere (chromosome end) sequence having implications for aging and cancer research is identified at LANL.

1989

DNA STSs recommended to correlate diverse types of DNA clones.

DOE and NIH establish Joint ELSI Working Group.

1990

DOE and NIH present joint 5-year U.S. HGP plan to Congress. The 15-year project formally begins.

Projects begun to mark genes on chromosome maps as sites of mRNA expression.

R&D begun for efficient production of more stable, large-insert BACs.

1991

Human chromosome mapping data repository, GDB, established.

1992

* Low-resolution genetic linkage map of entire human genome published.

Guidelines for data release and resource sharing announced by DOE and NIH.

1993

International IMAGE Consortium established to coordinate efficient mapping and sequencing of gene-representing cDNAs.

DOE-NIH Joint ELSI Working Group's Task Force on Genetic Information and Insurance releases recommendations.

DOE and NIH revise 5-year goals [*Science* **262**, 43–46 (Oct. 1,1993)].

* French Généthon provides mega-YACs to the genome community.

IOM releases U.S. HGP-funded report, "Assessing Genetic Risks."

GRAIL sequence interpretation service with Internet access initiated at ORNL.

ADA	Americans with Disabilities Act
ANL	Argonne National Laboratory
BAC	bacterial artificial chromosome
cDNA	complementary deoxyribonucleic acid
CGAP	Cancer Genome Anatomy Project
DNA	deoxyribonucleic acid
DHHS	Department of Health and Human Services (NIH)
DOE	Department of Energy
EEOC	Equal Employment Opportunity Commission
ELSI	ethical, legal, and social issues
GDB	Genome Database
GRAIL	Gene Recognition and Analysis Internet Link
HERAC	Health and Environmental Research Advisory Committee
HGP	Human Genome Project, Human Genome Program
HUGO	Human Genome Organisation
ICPEMC	International Commission for Protection Against Environmental Mutagens and Carcinogens
IMAGE	Integrated Molecular Analysis of Gene Expression
IOM	Institute of Medicine (NAS)

DISCLAIMER

Portions of this document may be illegible electronic image products. Images are produced from the best available original document.

1994

* Genetic-mapping 5-year goal achieved 1 year ahead of schedule.

Completion of second-generation DNA clone libraries representing each human chromosome by LLNL and LBNL.

Genetic Privacy Act, first U.S. HGP legislative product, proposed to regulate collection, analysis, storage, and use of DNA samples and genetic information obtained from them; endorsed by DOE-NIH Joint ELSI Working Group.

DOE Microbial Genome Program launched; spin-off of HGP.

LLNL chromosome paints commercialized.

SBH technologies from ANL commercialized.

DOE HGP Information Web site activated for public and researchers.

1995

LANL and LLNL announce high-resolution physical maps of chromosome 16 and chromosome 19, respectively.

* Moderate-resolution maps of chromosomes 3, 11, 12, and 22 maps published.

* First (nonviral) whole genome sequenced (for the bacterium *Haemophilus influenzae*).

Sequence of smallest bacterium, *Mycoplasma genitalium*, completed, displaying the minimum number of genes needed for independent existence.

* EEOC guidelines extend ADA employment protection to cover discrimination based on genetic information related to illness, disease, or other conditions.

1996

Methanococcus jannaschii genome sequenced; confirms existence of third major branch of life, the Archaea.

DOE-NIH Task Force on Genetic Testing releases interim principles.

* Integrated STS-based detailed human physical map with 30,000 STSs achieves an HGP goal.

* Health Care Portability and Accountability Act prohibits use of genetic information in certain health-insurance eligibility decisions, requires DHHS to enforce health-information privacy provisions.

DOE-NIH Joint ELSI Working Group releases guidelines on informed consent for large-scale sequencing projects.

DOE and NCHGR issue guidelines on use of human subjects for large-scale sequencing projects.

* *Saccharomyces cerevisiae* (yeast) genome sequence completed by international consortium.

Sequence of the human T-cell receptor region completed.

Wellcome Trust sponsors large-scale sequencing strategy meeting in Bermuda for international coordination of human genome sequencing.

1997

DOE forms Joint Genome Institute for implementing high-throughput sequencing at DOE HGP centers.

* NIH NCHGR becomes NHGRI.

* *Escherichia coli* genome sequence completed.

Second large-scale sequencing strategy meeting held in Bermuda.

* High-resolution physical maps of chromosomes X and 7 completed.

Methanobacterium thermoautotrophicum genome sequence completed.

Archaeoglobus fulgidus genome sequence completed.

* NCI CGAP begins.

---

* DOE had limited or no involvement in this event.

LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LLNL	Lawrence Livermore National Laboratory
MGP	Microbial Genome Project
MOU	Memorandum of Understanding
mRNA	messenger ribonucleic acid
NAS	National Academy of Sciences
NCHGR	National Center for Human Genome Research (NIH)
NCI	National Cancer Institute (NIH)
NHGRI	National Human Genome Research Institute (NIH)
NIGMS	National Institute of General Medical Sciences (NIH)
NIH	National Institutes of Health
NRC	National Research Council
OHER	Office of Health and Environmental Research
ORNL	Oak Ridge National Laboratory
OTA	Office of Technology Assessment
R&D	Research and Development
SBH	sequencing by hybridization
STS	sequence tagged site
YAC	yeast artificial chromosome

Preface

More than a decade ago, the Office of Health and Environmental Research (OHER) of the U.S. Department of Energy (DOE) struck a bold course in launching its Human Genome Initiative, convinced that its mission would be well served by a comprehensive picture of the human genome. Organizers recognized that the information the project would generate—both technological and genetic—would contribute not only to a new understanding of human biology and the effects of energy technologies but also to a host of practical applications in the biotechnology industry and in the arenas of agriculture and environmental protection.

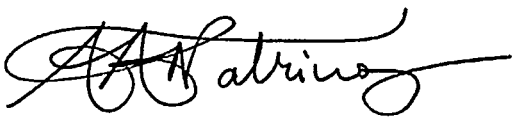
Today, the project's value appears beyond doubt as worldwide participation contributes toward the goals of determining the human genome's complete sequence by 2005 and elucidating the genome structure of several model organisms as well. This report summarizes the content and progress of the DOE Human Genome Program (HGP). Descriptive research summaries, along with information on program history, goals, management, and current research highlights, provide a comprehensive view of the DOE program.

Last year marked an early transition to the third and final phase of the U.S. Human Genome Project as pilot programs to refine large-scale sequencing strategies and resources were funded by DOE and the National Institutes of Health, the two sponsoring U.S. agencies. The human genome centers at Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory had been serving as the core of DOE multidisciplinary HGP research, which requires extensive contributions from biologists, engineers, chemists, computer scientists, and mathematicians. These team efforts were complemented by those at other DOE-supported laboratories and about 60 universities, research organizations, companies, and foreign institutions. Now, to focus DOE's considerable resources on meeting the challenges of large-scale sequencing, the sequencing efforts of the three genome centers have been integrated into the Joint Genome Institute. The institute will continue to bring together research from other DOE-supported laboratories. Work in other critical areas continues to develop the resources and technologies needed for production sequencing; computational approaches to data management and interpretation (called informatics); and an exploration of the important ethical, legal, and social issues arising from use of the generated data, particularly regarding the privacy and confidentiality of genetic information.

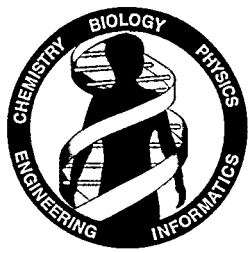
Insights, technologies, and infrastructure emerging from the Human Genome Project are catalyzing a biological revolution. Health-related biotechnology is already a success story—and is still far from reaching its potential. Other applications are likely to beget similar successes in coming decades; among these are several of great importance to DOE. We can look to improvements in waste control and an exciting era of environmental bioremediation, we will see new approaches to improving energy efficiency, and we can hope for dramatic strides toward meeting the fuel demands of the future.

In 1997 OHER, renamed the Office of Biological and Environmental Research (OBER), is celebrating 50 years of conducting research to exploit the boundless promise of energy technologies while exploring their consequences to the public's health and the environment. The DOE Human Genome Program and a related spin-off project, the Microbial Genome Program, are major components of the Biological and Environmental Research Program of OBER.

DOE OBER is proud of its contributions to the Human Genome Project and welcomes general or scientific inquiries concerning its genome programs. Announcements soliciting research applications appear in *Federal Register*, *Science*, *Human Genome News*, and other publications. The deadline for formal applications is generally midsummer for awards to be made the next year, and submission of preproposals in areas of potential interest is strongly encouraged. Further information may be obtained by contacting the program office or visiting the DOE home page (301/903-6488, Fax: -8521, *genome@oer.doe.gov*, URL: *http://www.er.doe.gov/production/ober/hug_top.html*).



Aristides Patrinos, Associate Director
Office of Biological and Environmental Research
U.S. Department of Energy
November 3, 1997

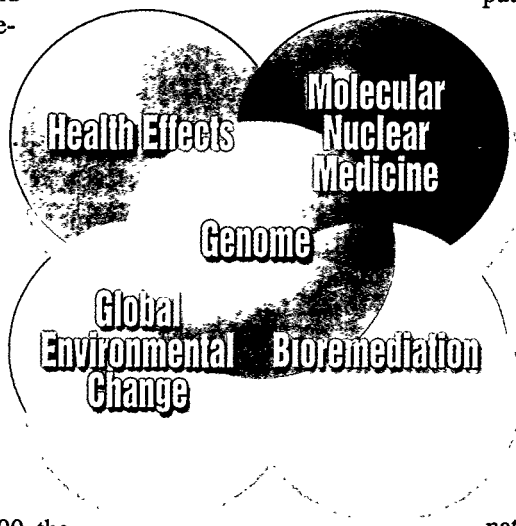


Contents

genome (jē′nōm), n.
all the genetic material in the chromosomes of an organism.

Scientific and technical terms are defined in the Glossary, p. 101. More historical details and other information appear in the Appendices beginning on p. 71.

Now completing its first decade, the Human Genome Program of the U.S. Department of Energy (DOE) is the longest-running federally funded program to analyze the genetic material—the genome— that determines an individual's characteristics at the most fundamental level. Part of the Biological and Environmental Research (BER) Program sponsored by the DOE Office of Biological and Environmental Research (OBER*), the genome program is a major component of the larger U.S. Human Genome Project.



Since October 1990, the project has been supported jointly by DOE and the National Institutes of Health (NIH) National Human Genome Research Institute (formerly National Center for Human Genome Research). Together, the DOE and NIH components make up the world's largest centrally coordinated biology research project ever undertaken.

The U.S. Human Genome Project is a 15-year endeavor to characterize the human genome by improving existing human genetic maps, constructing physical maps of entire chromosomes, and ultimately determining a complete sequence of the deoxyribonucleic acid (DNA) subunits. Parallel studies are being carried out on selected model organisms to facilitate interpretation of human gene function.

*In 1997 the Office of Health and Environmental Research (OHER) was renamed Office of Biological and Environmental Research (OBER).

The ultimate goal of the U.S. project is to identify the estimated 70,000 to 100,000 human genes and render them accessible for future biological study. A complete human DNA sequence will provide physicians and researchers in many biological disciplines with an extraordinary resource: an "encyclopedia" of human biology obtainable by computer and available to all.

For 50 years, programs in the DOE Office of Biological and Environmental Research have crossed traditional research boundaries in seeking new solutions to energy-related biological and environmental challenges (see Appendix F, p. 95, and **http://www.er.doe.gov/production/ober/ober.html).**

Obtaining the complete sequence by 2005 will require a highly coordinated and focused international effort generating advances in biological methodology; instrumentation (particularly automation); and computer-based methods for collecting, storing, managing, and analyzing the rapidly growing body of data.

Project Origins

The potential value of detailed genetic information was recognized early; until recently, however, obtaining this information was far beyond the capabilities of biomedical research. DOE OBER and its two predecessor agencies—the Atomic Energy Commission and the Energy Research and Development Administration—had long sponsored genetic research in both microbial and higher systems. These studies included explorations into population genetics; genome structure, maintenance, replication, damage, and repair; and the consequences of genetic mutations. These traditional DOE activities evolved naturally into the Human Genome Program.

OBER's mission is described more fully in the Program Management section (p. 59) of this report.

By 1985, progress in genetic and DNA technologies led to serious discussions in the scientific community about initiating a major project to analyze the structure of the human genome. After concluding that a DNA sequence would offer the most useful approach for detecting inherited mutations, DOE in 1986 announced its Human Genome Initiative. The initiative emphasized development of resources and technologies for genome mapping, sequencing, computation, and infrastructure support that would culminate in a complete sequence of the human genome.

The National Research Council issued a report in 1988 recommending a dedicated research budget of $200 million annually for 15 years to determine the sequence of the 3 billion chemical subunits (base pairs) in the human genome and to map and identify all human genes.

To launch the nation's Human Genome Project, Congress appropriated funds to DOE and also to NIH, which had long supported research in genetics and molecular biology as an integral part of its mission to improve the health of all Americans. Other federal agencies and foundations outside the Human Genome Project also contribute to genome research, and many other countries are making important contributions through their own genome research projects.

Coordinated Efforts

In 1988 DOE and NIH signed a Memorandum of Understanding in which the agencies agreed to work together, coordinate technical research and activities, and share results. The two agencies assumed a joint systematic approach toward establishing goals to satisfy both short- and long-term project needs.

Early guidelines projected three 5-year phases, for which the first plan was presented to Congress in 1990. The 1990

Anticipated Benefits of Genome Research

Predictions of biology as "the science of the 21st century" have been made by observers as diverse as Microsoft's Bill Gates and U.S. President Bill Clinton. Already revolutionizing biology, genome research has spawned a burgeoning biotechnology industry and is providing a vital thrust to the increasing productivity and pervasiveness of the life sciences.

Technology and resources promoted by the Human Genome Project already have had profound impacts on biomedical research and promise to revolutionize biological research and clinical medicine. Increasingly detailed genome maps have aided researchers seeking genes associated with dozens of genetic conditions, including myotonic dystrophy, fragile X syndrome, neurofibromatosis types 1 and 2, a kind of inherited colon cancer, Alzheimer's disease, and familial breast cancer.

Current and potential applications of genome research will address national needs in molecular medicine, waste control and environmental cleanup, biotechnology, energy sources, and risk assessment.

Molecular Medicine

On the horizon is a new era of molecular medicine characterized less by treating symptoms and more by looking to the most fundamental causes of disease. Rapid and more specific diagnostic tests will make possible earlier treatment of countless maladies. Medical researchers also will be able to devise novel therapeutic regimens based on new classes of drugs, immunotherapy techniques, avoidance of environmental conditions that may trigger disease, and possible augmentation or even replacement of defective genes through gene therapy.

Microbial Genomes

In 1994, taking advantage of new capabilities developed by the genome project, DOE formulated the Microbial Genome Initiative to sequence the genomes of bacteria useful in the areas of energy production, environmental remediation, toxic waste reduction, and industrial processing. In the resulting Microbial Genome Project, six microbes that live under extreme conditions of temperature and pressure have been sequenced completely as

plan emphasized the creation of chromosome maps, software, and automated technologies to enable sequencing.

By 1993, unexpectedly rapid progress in chromosome mapping required updating the goals [*Science* **262**, 43–46 (October 1, 1993)], which now project through 1998 (see p. 5). This plan is being revised again in anticipation of the approaching high-throughput sequencing phase of the project. Last year marked an early transition to this phase as many more genome sequencing projects were funded. The second and third phases of the project will optimize resources, refine sequencing strategies, and, finally, completely determine the sequence of all base pairs in the genome.

Another area of DOE and NIH cooperation is in exploring the ethical, legal, and social issues (ELSI) arising from increased availability of genetic data and growing genetic-testing capabilities. The two agencies established a joint working group to confront these ELSI challenges and have cosponsored joint projects and workshops.

DOE Genome Program

A general overview follows of recent progress made in the DOE Human Genome Program. Refer to the timeline (pp. ii–iii) for other achievements toward U.S. goals, including contributions made outside DOE.

Physical maps

For DOE, an early goal was to develop chromosome physical maps, which involves reconstructing the order of cloned DNA fragments to represent their specific originating chromosomes. (A set of such cloned fragments is called a library.) Critical to this effort were the libraries of individual human chromosomes

of August 1997. Structural studies are under way to learn what is unique about the proteins of these organisms—the ultimate aim being to use the microbes and their enzymes for such practical purposes as waste control and environmental cleanup.

Biotechnology

The potential for commercial development presents U.S. industry with a wealth of opportunities. Sales of biotechnology products are projected to exceed $20 billion by the year 2000. The genome project already has stimulated significant investment by large corporations and prompted the creation of new biotechnology companies hoping to capitalize on the far-reaching implications of its research.

Energy Sources

Biotechnology, fueled by insights reaped from the genome project, will play a significant role in improving the use of fossil-based resources. Increased energy demands, projected over the next 50 years, require strategies to circumvent the many problems associated with today's dominant energy technologies. Biotechnology promises to help address these needs by providing cleaner means for the bioconversion of raw materials to refined products. In addition, there is the possibility of developing entirely new biomass-based energy sources. Having the genomic sequence of the methane-producing microorganism *Methanococcus jannaschii*, for example, will enable researchers to explore the process of methanogenesis in more detail and could lead to cheaper production of fuel-grade methane.

Risk Assessment

Understanding the human genome will have an enormous impact on the ability to assess risks posed to individuals by environmental exposure to toxic agents. Scientists know that genetic differences make some people more susceptible—and others more resistant—to such agents. Far more work must be done to determine the genetic basis of such variability. This knowledge will directly address DOE's long-term mission to understand the effects of low-level exposures to radiation and other energy-related agents, especially in terms of cancer risk.

produced at Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL). These libraries allowed the huge task of mapping and sequencing the entire 3 billion bases in the human genome to be broken down into 24 much smaller single-chromosome units. Availability of the libraries has enabled the participation of many laboratories worldwide. Some three generations of clone libraries with improving characteristics have been produced and widely distributed. In the DOE-supported projects, DNA clones representing chromosomes 16, 19, and 22 have been ordered (mapped) and are now providing material needed for large-scale sequencing.

Sequencing

Toward the goal of greatly increasing the speed and decreasing the cost of DNA sequencing, DOE has supported improvements in standard technologies and has pioneered support for revolutionary sequencing systems. Marked improvements have been made in reagents, enzymes, and raw data quality. Such novel approaches as sequencing by hybridization (using DNA "chips") and mass spectrometry have already found important, previously unanticipated applications outside the Human Genome Project.

Joint Genome Institute

In early 1997, the human genome centers at Lawrence Berkeley National Laboratory, LANL, and LLNL began collaborating in the Joint Genome Institute (JGI), within which high-throughput sequencing will be implemented [see p. 26 and *Human Genome News* **8**(2), 1–2]. The initial JGI focus will be on sequencing areas of high biological interest on several chromosomes, including human chromosomes 5, 16, and 19. Establishment of JGI represents a major transition in the DOE Human Genome Program.

Previously, most goals were pursued by small- to medium-sized teams, with modest multisite collaborations. The JGI will house high-throughput implementations of successful technologies that will be run with increasingly stringent process- and quality-control systems.

In addition, a small component aimed at understanding how genes function in the body—a field known as functional genomics—has been established and will grow as sequencing targets are met. High-throughput functional genomics represents a new era in human biology, one which will have profound implications for solving biological problems.

Informatics

In preparation for the production-sequencing phase, many algorithms for interpreting DNA sequence have been developed, and an increasing number have become available as services over the Internet. Last year, the GRAIL (for Gene Recognition and Analysis Internet Link) and GenQuest servers, developed and maintained at Oak Ridge National Laboratory, processed an average of almost 40 million bases of sequence each month.

As technology improves and data accumulates exponentially, continued progress in the Human Genome Project will depend increasingly on the development of sophisticated computational tools and resources to manage and interpret the information. The ease with which researchers can access and use the data will provide a measure of the project's success. Critical to this success is the creation of interoperable databases and other computing and informatics tools to collect, organize, and interpret thousands of DNA clones.

For additional information on the DOE genome programs, refer to Research Highlights, p. 9; Research Narratives, p. 25; this report's *Part 2, 1996 Research Abstracts*; and the Web site (*http://www.ornl.gov/hgmis*).

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

October 1, 1993, to September 30, 1998 (FY 1994 through FY 1998)*

Major events in the U.S. Human Genome Project, including progress made toward these goals, are charted in a timeline on pp. ii–iii.

Genetic Mapping

- Complete the 2- to 5-cM map by 1995.
- Develop technology for rapid genotyping.
- Develop markers that are easier to use.
- Develop new mapping technologies.

Physical Mapping

- Complete a sequence tagged site (STS) map of the human genome at a resolution of 100 kb.

DNA Sequencing

- Develop efficient approaches to sequencing one- to several-megabase regions of DNA of high biological interest.
- Develop technology for high-throughput sequencing, focusing on systems integration of all steps from template preparation to data analysis.
- Build up a sequencing capacity to allow sequencing at a collective rate of 50 Mb per year by the end of the period. This rate should result in an aggregate of 80 Mb of DNA sequence completed by the end of FY 1998.

Gene Identification

- Develop efficient methods for identifying genes and for placement of known genes on physical maps or sequenced DNA.

Technology Development

- Substantially expand support of innovative technological developments as well as improvements in current technology for DNA sequencing and for meeting the needs of the Human Genome Project as a whole.

Model Organisms

- Finish an STS map of the mouse genome at a 300-kb resolution.
- Finish the sequence of the *Escherichia coli* and *Saccharomyces cerevisiae* genomes by 1998 or earlier.
- Continue sequencing *Caenorhabditis elegans* and *Drosophila melanogaster* genomes with the aim of bringing *C. elegans* to near completion by 1998.
- Sequence selected segments of mouse DNA side by side with corresponding human DNA in areas of high biological interest.

Informatics

- Continue to create, develop, and operate databases and database tools for easy access to data, including effective tools and standards for data exchange and links among databases.
- Consolidate, distribute, and continue to develop effective software for large-scale genome projects.
- Continue to develop tools for comparing and interpreting genome information.

Ethical, Legal, and Social Implications

- Continue to identify and define issues and develop policy options to address them.
- Develop and disseminate policy options regarding genetic testing services with potential widespread use.
- Foster greater acceptance of human genetic variation.
- Enhance and expand public and professional education that is sensitive to sociocultural and psychological issues.

Training

- Continue to encourage training of scientists in interdisciplinary sciences related to genome research.

Technology Transfer

- Encourage and enhance technology transfer both into and out of centers of genome research.

Outreach

- Cooperate with those who would establish distribution centers for genome materials.
- Share all information and materials within 6 months of their development. This should be accomplished by submission of information to public databases or repositories, or both, where appropriate.

---

*Original 1990 goals were revised in 1993 due to rapid progress. A second revision was being developed at press time.

Evolution of a Vision: Genome Project Origins,

In an interview at a DNA sequencing conference in Hilton Head, South Carolina, David Smith, a founder and former Director of the DOE Human Genome Program, recalled the establishment of this country's first human genome project. The impressive early achievements and spin-off benefits, he noted, offer more than mere vindication for project founders. They also provide a tantalizing glimpse into the future where, he observed, "scientists will be empowered to study biology and make connections in ways undreamt of before."*

The DOE Human Genome Program began as a natural outgrowth of the agency's long-term mission to develop better technologies for measuring health effects, particularly induced mutations. As Smith explained it, "DOE had been supporting mutation studies in Japan, where no heritable mutations could be detected in the offspring of populations exposed to the atomic blasts at Hiroshima and Nagasaki. The program really grew out of a need to characterize DNA differences between parents and children more efficiently. DOE led the development of many mutation tests, and we were interested in developing even more sensitive detection methods. Mortimer Mendelsohn of Lawrence Livermore National Laboratory, a member of the International Commission for Protection Against Environmental Mutagens and Carcinogens, and I decided to hold a workshop to discuss DNA-based methods (see Human Genome Project chronology, p. ii).

"Ray White (University of Utah) organized the meeting, which took place in Alta, Utah, in December 1984. It was a small meeting but very stimulating intellectually. We concluded the obvious—that if you really wanted to use DNA-based technologies, you had to come up with more efficient ways to characterize the DNA of much larger regions of the genome. And the ultimate sensitivity would be the capability to compare the complete DNA sequences of parents and their offspring."

Project Begins

Smith recalled reaction to the first public statement that DOE was starting a program with the aim of sequencing the human genome. "I announced it at the Cold Spring Harbor meeting in May 1986, and there was a big hullabaloo." After a year-long review, a National Academy of Sciences National Research Council panel endorsed the project and the basic strategy proposed. Smith pointed out that NIH and others were also having discussions on the feasibility of sequencing the human genome. "Once NIH got interested, many more people became involved. DOE and NIH signed a Memorandum of Understanding in October 1988 to coordinate our activities aimed at characterizing the human genome." But, he observed, it wasn't all smooth sailing. The nascent project had many detractors.

Responding to Critics

Many scientists, prominent biologists among them, thought having the sequence would be a misuse of scarce resources. Smith, laughing now, recalls one scientist complaining, "Even if I had the sequence, I wouldn't know what to do with it." Other critics worried that the genome project would siphon shrinking research funds away from individual investigator-initiated research projects. Smith takes the opposite view. "In fact, individual investigators can do things they would never be able to do otherwise. We're beginning to see that demonstrated at this meeting. For the first time, we're finding people exploring systematic ways of looking at gene function in organisms. The genome project opens up enormous new research fields to be mined. Cottage-industry biologists won't need a lot of robots, but they will have to be computer literate to put the information all together."

The genome project also is providing enabling technologies essential to the future of the emerging biotechnology industry, catalyzing its tremendous growth. According to Smith, the technologies are capable of more than elucidating the human genome. "We're developing an infrastructure for future research. These technologies will allow us to efficiently characterize any of the organisms out there that pertain to various DOE missions, with such applications as better fuels from biomass, bioremediation, and waste control. They also will lead to a greater understanding of global cycles, such as the carbon cycle, and the identification of potential biological interventions. Look at the ocean; an amazing number of microbes are in there, but we don't know how to use them to influence cycles to control some of the harmful things that might be happening. Up to now, biotechnology has been nearly all health oriented, but applications of genome research to modern biology really go beyond health. That's one of the things motivating our program to try to develop some of these other biotechnological applications."

> ***"Genomics has come of age, and it is opening the door to entirely new approaches to biology."***

Responding to criticism about not researching gene function early in the project, Smith reasserted that the purpose of the Human Genome Project is to build technologies and resources that will enable researchers to learn about biology in a much

*The Seventh International Genome Sequencing and Analysis Conference, September 1995.

Present and Future Challenges, Far-Reaching Benefits

more efficient way. "The genome budget is devoted to very specific goals, and we make sure that projects contribute toward reaching them."

International Scope

Smith credited the international community with contributing to many project successes. "The initial planning was for a U.S. project, but the outcome, of course, is that it is truly international, and we would not be nearly as far as we are today without those contributions. Also, there's been a fair amount of money from private companies, and support from the Muscular Dystrophy Association in France and The Wellcome Trust in the United Kingdom has been extremely important."

Technology Advances

While noting enormous advances across the board, Smith cited automation progress and observed that tremendously powerful robots and automated processes are changing the way molecular biology is done. "A lot of novel technologies probably won't be useful for initial sequencing but will be very valuable for comparing sequences of different people and for polymorphism studies. One of the most gratifying recent successes is the DNA polymerase engineering project. Researchers made a fairly simple change, but it resulted in a thermosequenase that may answer a lot of problems, reduce the cost of sequencing, and give us better data."

Progress in genome research requires the use of maturing technologies in other fields. "The combination of technologies that are coming together has been fortuitous; for example, advances in informatics and data-handling technologies have had a tremendous impact on the genome project. We would be in deep trouble if they were at a less-mature stage of development. They have been an important DOE focus."

ELSI

Smith described tangible progress toward goals associated with programs on the ethical, legal, and social issues (ELSI) related to data produced by the genome project. "ELSI programs have done a lot to educate the thinkers, and this has produced a higher level of discourse in the country about these issues. DOE is spending a large fraction of its ELSI money on informing special populations who can reach others. Educating judges has been especially well received because they realize the potential impact of DNA technology on the courts."

According to Smith, more people and groups need to be involved in ELSI matters. "We have some ELSI products: the DOE-NIH Joint ELSI Working Group has an insurance task force report, and a DOE ELSI grantee has produced draft privacy legislation. Now it's time for others to come and translate ELSI efforts into policy. Perhaps the new National Bioethics Advisory Commission can do some of this."

New Model for Biological Research

Smith spoke of a changing paradigm guiding DOE-supported biology. "Some years ago, the central idea or dogma in molecular biology research was that information in DNA directs RNA, and RNA directs proteins. Today, I think there is a new paradigm to guide us: Sequence implies structure, and structure implies function. The word 'implies' in our new paradigm means there are rules," continued Smith, "but these are rules we don't understand today. With the aid of structural information, algorithms, and computers, we will be able to relate sequence to structure and eventually relate structure to function. Our effort focuses on developing the technologies and tools that will allow us to do this efficiently."

"That's how I think about what we do at DOE," he said. "We're working a lot on technology and projects aimed at human and microbial genome sequencing. For understanding sequence implications, we are making major, increasing investments in synchrotrons, synchrotron user facilities, neutron user facilities, and big nuclear magnetic resonance machines. These are all aimed at rapid structure determination." Smith explained that now we are seeing the beginnings of the biotechnology revolution implied by the sequence-to-structure-to-function paradigm. "If you really understand the relationship between sequence and function, you can begin to design sequences for particular purposes. We don't yet know that much about the world around us, but there are capabilities out there in the biological world, and if we can understand them, we can put those capabilities to use."

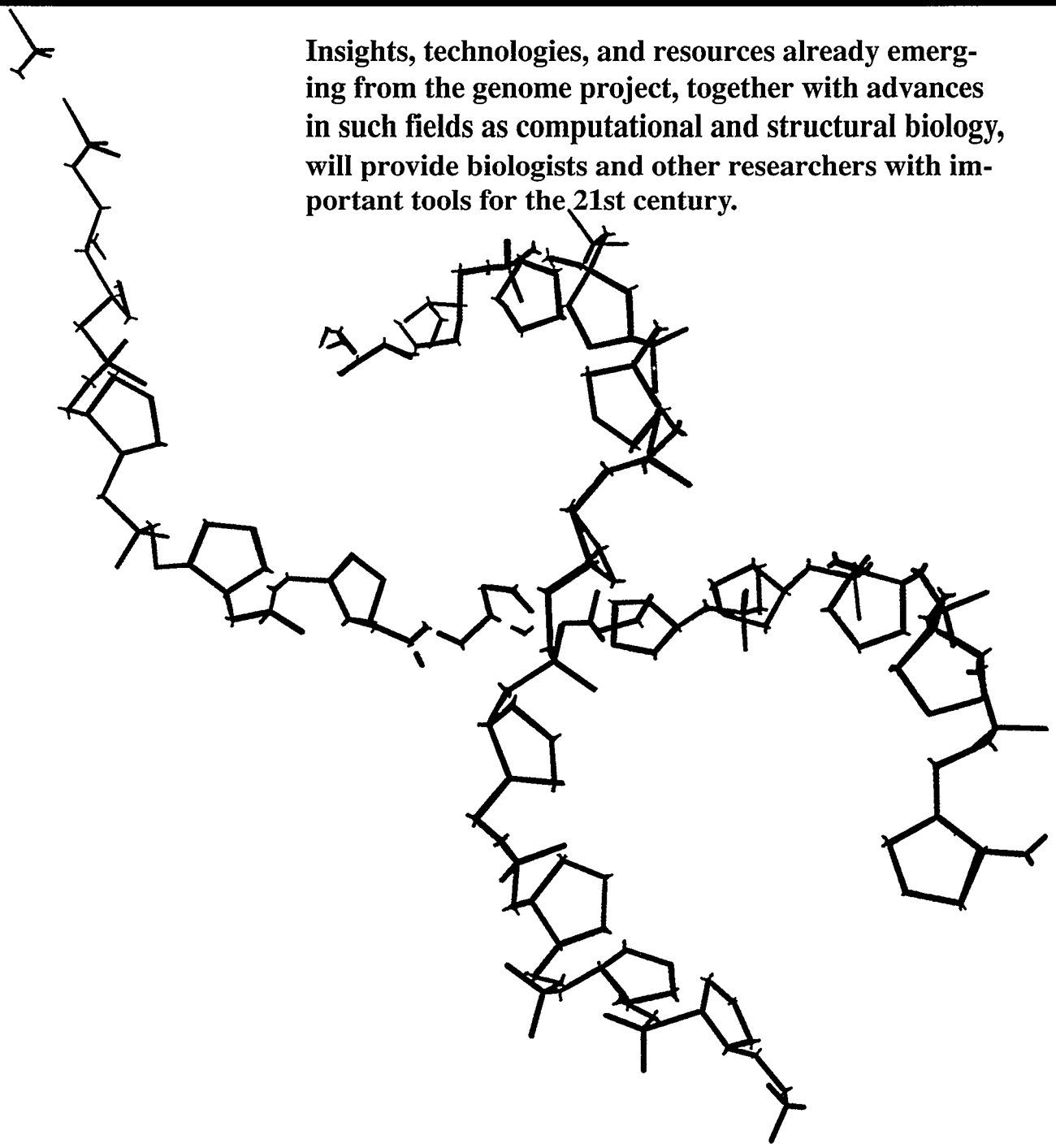
"Comparative genomics," he continued, "will teach us a tremendous amount about human evolution. The current phylogenetic tree is based on ribosomal RNA sequences, but when we have determined whole genomic sequences of different microbes, they will probably give us different ideas about relationships among archaebacteria, eukaryotes, and prokaryotes."

Feeling good about progress over the previous 5 years, Smith summed it up succinctly: "Genomics has come of age, and it is opening the door to entirely new approaches to biology."

David Smith retired at the end of January 1996. Taking responsibility for the DOE Human Genome Program is Aristides Patrinos, who is also Associate Director of the DOE Office of Biological and Environmental Research. Marvin Frazier is Director of the Health Effects and Life Sciences Research Division, which manages the Human Genome Program.

Looking to the Future

Insights, technologies, and resources already emerging from the genome project, together with advances in such fields as computational and structural biology, will provide biologists and other researchers with important tools for the 21st century.



Transitioning to large-scale sequencing

The early years of the Human Genome Program have been remarkably successful. Critical resources and infrastructures have been established, and technologies have been developed for producing several useful types of chromosomal maps. These gains are supporting the project's transition to the large-scale sequencing phase. Some highlights and trends in the U.S. Department of Energy's (DOE) Human Genome Program after FY 1993 are presented in this section.

Clone Resources for Mapping, Sequencing, and Gene Hunting

The demands of large chromosomal mapping and sequencing efforts have necessitated the development of several different types of clone collections (called libraries) carrying human DNA. Three generations of DOE-developed libraries are being distributed to research teams in the United States and abroad. In these libraries, human DNA segments of various lengths are maintained in bacterial cells.

NLGLP Libraries

The first two generations are chromosome-specific libraries carrying small inserts of human DNA (15,000 to 40,000 base pairs). As part of the National Laboratory Gene Library Project (NLGLP) begun in 1983, these libraries were prepared at Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL) using DOE flow-sorting technology to separate individual chromosomes. Library availability has allowed the very difficult whole-genome tasks to be divided into 24 more manageable single-chromosome projects that could be pursued at separate research centers. Completed in 1994, NLGLP libraries have provided critical resources to genome researchers worldwide (*http://www-bio.llnl.gov/genome/html/cosmid.html*). Very high resolution chromosome maps based principally on NLGLP libraries were published in 1995 for chromosomes 16 and 19. These are described in detail in the Research Narratives section of this report (see LLNL, p. 27, and LANL, p. 35).

PACs and BACs

The third generation of clone resources supporting chromosome mapping is composed of P1 artificial chromosome (PAC) and bacterial artificial chromosome (BAC) libraries. A prototype PAC library was produced by the team of Leon Rosner (then at DuPont) many years ago, but more efficient production began with improvements introduced by the DOE-supported teams headed by Melvin Simon at Caltech (BACs) and Pieter de Jong at Roswell Park (PACs).

In contrast to cosmids, BACs and PACs provide a more uniform representation of the human genome, and the greater length of their inserts (90,000 to

DOE Genome Research Web Site
http://www.ornl.gov/hgmis/research.html

Research Narratives

Separate narratives, beginning on p. 25, contain detailed descriptions of research programs and accomplishments at these major DOE genome research facilities.

- Lawrence Livermore National Laboratory
- Los Alamos National Laboratory
- Lawrence Berkeley National Laboratory
- University of Washington Genome Sequencing Laboratory
- Genome Database
- National Center for Genome Resources

Research Abstracts

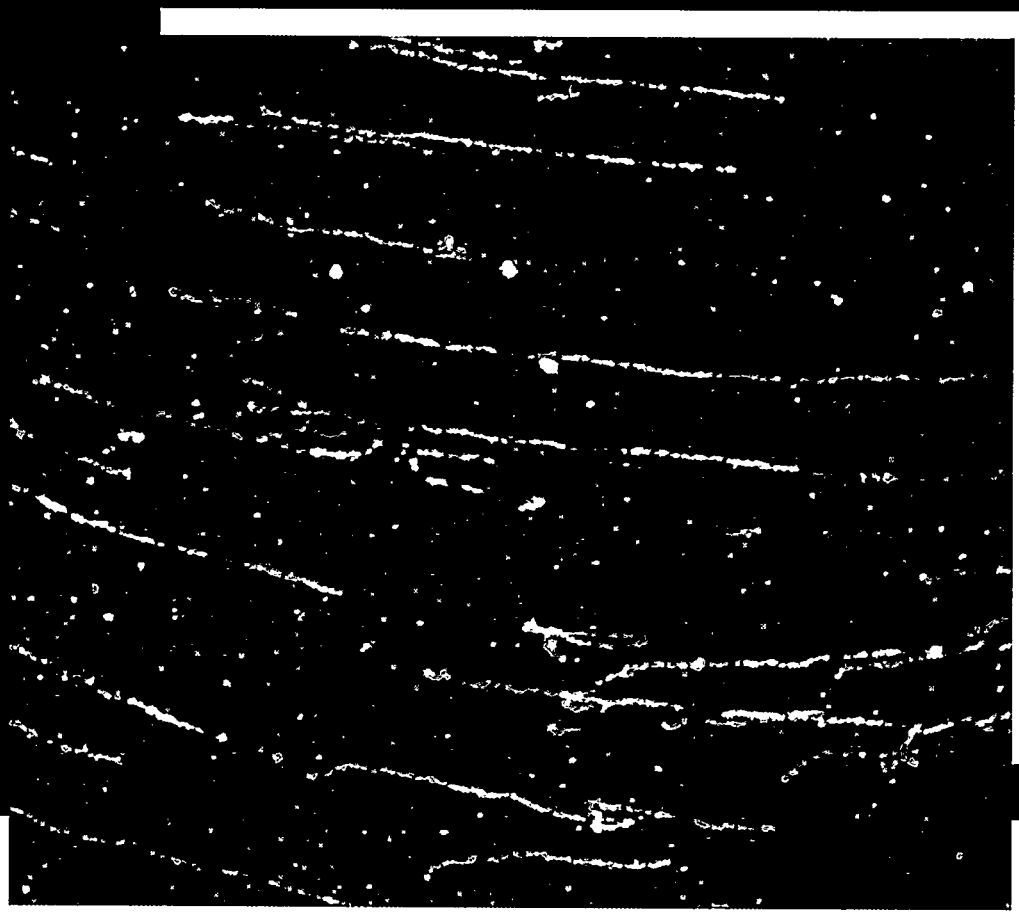
Descriptions of individual research projects at other institutions are given in *Part 2, 1996 Research Abstracts.*

300,000 base pairs) facilitates both mapping and sequencing. Their usefulness was illustrated dramatically in 1993 when the first breast cancer–susceptibility gene (*BRCA*1) was found in a BAC clone after other types of resources had failed. The next year, with major support from NIH, de Jong's PACs contributed to the isolation of the second human breast cancer–susceptibility gene (*BRCA*2).

Mapping

The assembly of ordered, overlapping sets (contigs) of high-quality clones has long been considered an essential step toward human genome sequencing. Because the clones have been mapped to precise genomic locations, DNA sequences obtained from them can be located on the chromosomes with minimal uncertainty.

The large insert size of BACs and PACs allows researchers to visually map them on chromosomes by using fluorescence in situ hybridization (FISH) technology (see photomicrograph below). These mapped BACs and PACs represent very valuable resources for the cytogeneticist exploring chromosomal abnormalities. Two major medical genetics resources have been developed: (1) The Resource for Molecular Cytogenetics at the University of California, San Francisco, in collaboration with the Lawrence Berkeley National Laboratory (LBNL) team led by Joe Gray (*http://rmc-www.lbl.gov*) and (2) The Total Human Genome BAC-PAC Resource at Cedars-Sinai Medical Center, Los Angeles, developed by Julie Korenberg's laboratory (see map, p. 12, and Web site, *http://www.csmc.edu/genetics/korenberg/korenberg.html*).



FISH Mapping on DNA Fibers. The fluorescence microscope reveals several individual cloned DNA fibers from yeast artificial chromosomes (YACs, in blue) after molecular combing to attach and stretch the DNA molecules across a glass microscope slide. Also shown are the locations of two P1 clones, labeled green and red, mapped onto the YAC fibers using FISH. Digital imaging technology can be used to assemble physical maps of chromosomes with a resolution of about 3 to 5 kilobases. [Source: Joe Gray, University of California, San Francisco]

Coordinated Mapping and Sequencing

A simple strategy was proposed in 1996 for choosing BACs or PACs to elongate sequenced regions most efficiently [*Nature* **381**, 364–66 (1996)]. The first step is to develop a BAC end sequence database, with each entry having the BAC clone name and the sequences of its human insert ends. In toto, the source BACs should represent a 15- to 20-fold coverage of the human genome. Then for any BAC or chromosomal region sequenced, a comparison against the database will return a list of BACs (or PACs) that overlap it. Optimal choices for the next BACs (or PACs) to be sequenced can then be made, entailing minimal everlap (and therefore minimal redundancy of sequencing).

Two pilot BAC-PAC end-sequencing projects were initiated in September of 1996 to explore feasibility, optimize technologies, establish quality controls, and design the necessary informatics infrastructure. Particular benefits are anticipated for small laboratories that will not have to maintain large libraries of clones and can avoid preliminary contig mapping (see abstracts of Glen Evans; Julie Korenberg; Mark Adams, Leroy Hood, and Melvin Simon; and Pieter de Jong in Part 2 of this report).

Updated information on BAC-PAC resources can be found on the Web (*http://www.ornl.gov/meetings/bacpac/95bac.html*). [See Appendix C: Human Subjects Guidelines, p. 77 or *http://www.ornl.gov/hgmis/archive/nchgrdoe.html* for DOE-NIH guidelines on using DNA from human subjects for large-scale sequencing.]

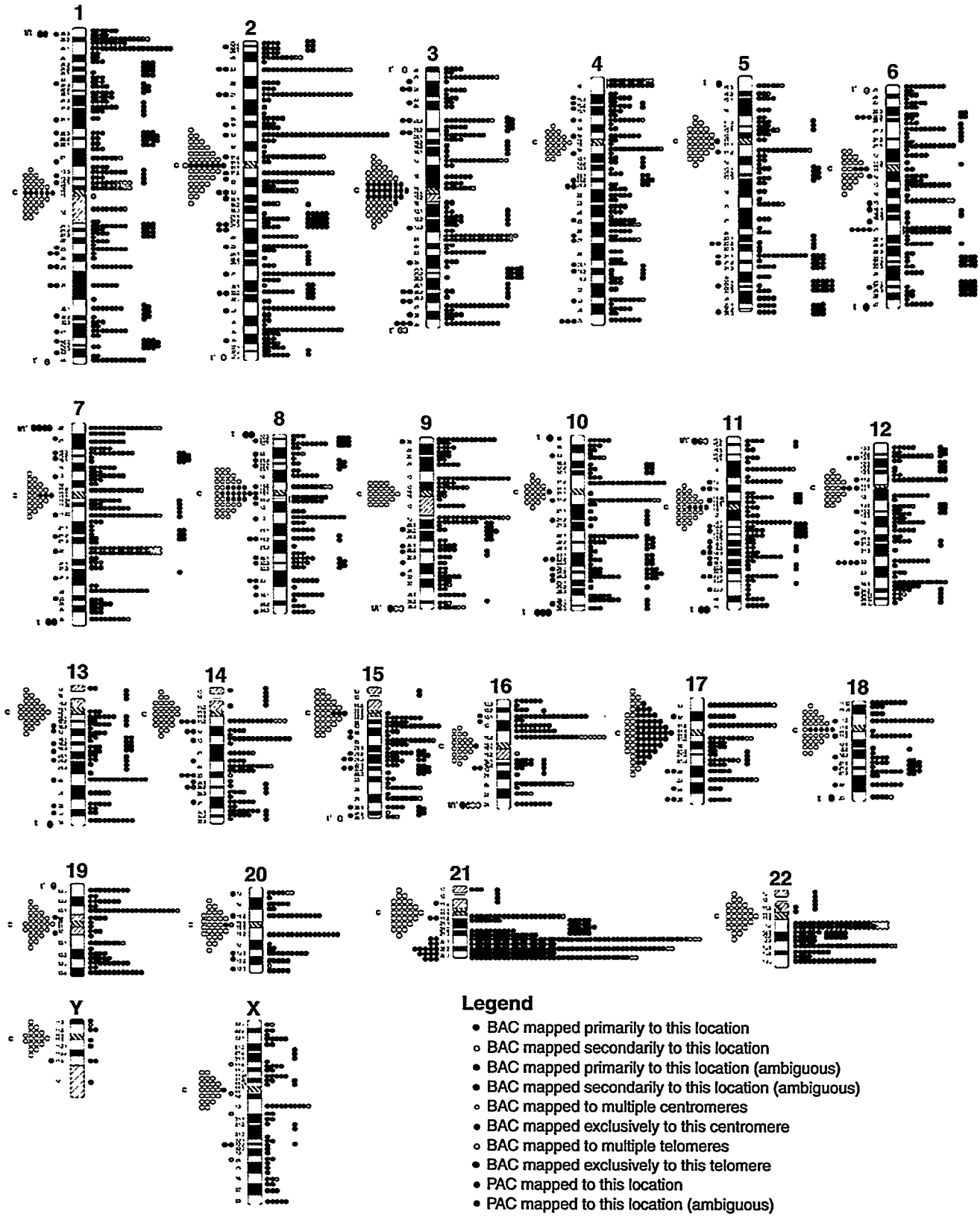
cDNA Libraries

In 1990, DOE initiated projects to enrich the developing chromosome contig maps with markers for genes. Although the protein-encoding messenger RNAs are good representatives of their source genes, they are unstable and must be converted to complementary DNAs (cDNAs) for practical applications. These conversions are tricky, and artifacts are introduced easily. The team led by Bento Soares (University of Iowa) has optimized the steps and continues to produce cDNA libraries of the highest quality. At LLNL, individual cDNA clones are put into standard arrays and then distributed worldwide for characterization by the international IMAGE (for Integrated Molecular Analysis of Gene Expression) Consortium (see box, p. 13).

Initially supported under a DOE cDNA initiative, Craig Venter's team (now at The Institute for Genomic Research) greatly improved technologies for reading sequences from cDNA ends (expressed sequence tags, called ESTs). Together with complementary analysis software, ESTs were shown to be a valuable resource for categorizing cDNAs and providing the first clues to the functions of the genes from which they are derived. This fast EST approach has attracted millions of dollars in commercial investment. Mapping the cDNA onto a chromosome can identify the location of its corresponding gene. Many laboratories worldwide are contributing to the continuing task of mapping the estimated 70,000 to 100,000 human genes.

HAECs

All the previously described DNA clones are maintained in bacterial host cells. However, for unknown reasons, some regions of the human genome appear to be unclonable or unstable in bacteria. The team led by Jean-Michel Vos (University of North Carolina, Chapel Hill) has developed a human artificial episomal chromosome (HAEC) system based on the Epstein-Barr virus that may be useful for coverage of these especially difficult regions. In the broader biomedical community, HAECs also show promise for use in gene therapy.



BAC-PAC Map. *The Total Human Genome BAC-PAC Resource represents an important tool for understanding the genes responsible for human development and disease (http://www.csmc.edu/genetics/korenberg/korenberg.html). The Resource, consisting of more than 5000 BAC and PAC clones, covers every human chromosome band and 25% of the entire human genome. Each color dot represents a single BAC or PAC clone mapped by FISH to a specific chromosome band represented in black and white. The clones, which are stable and useful for sequencing, have been integrated with the genetic and physical chromosome maps.* *[Source: Julie Korenberg, Cedars-Sinai Medical Center]*

Resources for Gene Discovery

Hunting for disease genes is not a specific goal of the DOE Human Genome Program. However, DOE-supported libraries sent to researchers worldwide have facilitated gene hunts by many research teams. DOE libraries have played a role in the discovery of genes for cystic fibrosis, the most common lethal inherited disease in Caucasians; Huntington's disease, a progressive lethal neurological disorder; Batten's disease, the most prevalent neurodegenerative childhood disease; two forms of dwarfism; Fanconi anemia, a rare disease characterized by skeletal abnormalities and a predisposition to cancer; myotonic dystrophy, the most common adult form of muscular dystrophy; a rare inherited form of breast cancer; and polycystic kidney disease, which affects an estimated 500,000 people in the United States at a healthcare cost of over $1 billion per year.

The team led by Fa-Ten Kao (Eleanor Roosevelt Institute) has microdissected several chromosomes and made derivative clone libraries broadly available to disease-gene hunters. This resource played a critical role in isolating the gene responsible for some 15% of colon cancers.

Of Mice and Humans: The Value of Comparative Analyses

A remaining challenge is to recognize and discriminate all the functional constituents of a gene, particularly regulatory components not represented within cDNAs, and to predict what each gene may actually do in human biology. Comparing human and mouse sequences is an exceptionally powerful way to identify homologous genes and regulatory elements that have been substantially conserved during evolution.

Researchers led by Leroy Hood (University of Washington, Seattle) have analyzed more than 1 million bases of sequence from T-cell receptor (TCR)

To IMAGE the Human Gene Map

Since 1993, the Integrated Molecular Analysis of Gene Expression (IMAGE) Consortium has played a major role in the development of a human gene map. Founding members of the IMAGE Consortium are Bento Soares (Columbia University, now at University of Iowa), Gregory Lennon (LLNL), Mihael Polymeropoulos (National Institutes of Health's National Institute of Mental Health), and Charles Auffrey (Généthon, in France). Because cDNA molecules represent coding (expressed-gene) areas of the genome, sets of cloned cDNAs are a valuable resource to the gene-mapping community. The cDNA libraries representing different tissues have many members in common. Thus, good coordination among participating laboratories can minimize redundant work. The international IMAGE Consortium laboratories fulfill this role by developing and arraying cDNA clones for worldwide use. *[http://www-bio.llnl.gov/bbrp/image/image.html]*

From the IMAGE cDNA clones, researchers at the Washington University (St. Louis) Sequencing Center determine ESTs with support from Merck, Inc. The data, which are used in gene localization, are then entered into public databases. More than 10,000 chromosomal assignments have been entered into Genome Database (*http://www.gdb.org*). Including replica copies, over 3 million clones have been distributed, probably representing about 50,000 distinct human genes.

The IMAGE infrastructure is being used in two additional programs. At LLNL, the IMAGE laboratory arrays mouse cDNA libraries produced by Soares for the Washington University Mouse EST project (*http://genome.wustl.edu/est/mouse_esthmpg.html*) with sequencing sponsored by the Howard Hughes Medical Institute. Additional clone libraries are being used in a collaborative sequencing project sponsored by the NIH National Cancer Institute as part of the Cancer Genome Anatomy Project to identify and fully sequence genes implicated in major cancers (*http://www.ncbi.nlm.nih.gov/ncicgap*).

chromosome regions of both human and mouse genomes. Many subtle functional elements can be recognized only by comparing human and mouse sequences. TCRs play a major role in immunity and autoimmune disease, and insights into their mechanisms may one day help treat or even prevent such diseases as arthritis, diabetes, and multiple sclerosis (possibly even AIDS).

Comparative analysis is also used to model human genetic diseases. Given sequence information, researchers can produce targeted mutations in the mouse as a rapid and economical route to elucidating gene function. Such studies continue to be used effectively at Oak Ridge National Laboratory (ORNL).

DNA Sequencing

From the beginning of the genome project, DOE's DNA sequencing-technology program has supported both improvements to established methodologies and innovative higher-risk strategies. The first major sequencing project, a test bed for incremental improvements, culminated with elucidation of the highly complex TCR region (described above) by a team led by Hood.

A novel "directed" sequencing strategy initiated at LBNL in 1993 provides a potential alternative approach that can include automation as a core design feature. In this approach, every sequencing template is first mapped to its original position on a chromosome (resolution, 30 bases). The advantages of this method include a large reduction in the number of sequencing reactions needed and in the sequence-assembly steps that follow. To date, this directed strategy has achieved significant results with simpler, less repetitive nonhuman sequences, particularly in the NIH-funded *Drosophila* genome program. The system also is in use at the Stanford Human Genome Center and Mercator Genetics, Inc.

The preparation of DNA clones for sequencing involves several biochemical processing steps that require different solution environments. At the Whitehead Institute, Trevor Hawkins has improved systems for reversible binding of DNA molecules to magnetic beads that are compatible with complete robotic management. The second-generation Sequatron fits on a tabletop with a single robotic arm moving sample trays between servicing stations. This very compact system, supported by sophisticated software, may be ideal for laboratories with limited or costly floor space.

Fluorescent tags are critical components of conventional automated sequencing approaches. The team of Richard Mathies and Alexander Glazer (University of California, Berkeley) has made a series of improvements in fluorescence systems that have decreased DNA input needs and markedly increased the quality of raw data, thereby supporting longer useful reads of DNA sequence.

Complementary improvements in enzymology have been achieved by the team of Charles Richardson and Stanley Tabor (Harvard Medical School). Current widely used procedures for automated DNA sequencing involve cycling between high and low temperatures. The Harvard researchers used information about the three-dimensional structure of polymerases (enzymes needed for DNA replication) and how they function to engineer an improved Taq polymerase. ThermoSequenase, which is now produced commercially as part of the ThermoSequenase kit, reduces the amount of expensive sequencing reagents required and supports popular cycle-sequencing protocols.

The application of higher electrical fields in gel electrophoresis separation of DNA fragments can increase sequencing speed and efficiency. Conventional thick gels cannot adequately dissipate the additional heat produced, however. Two promising routes to "thinness" are ultrathin slab gels and

capillary systems. An ultrathin gel system was developed by Lloyd Smith (University of Wisconsin, Madison) and licensed for commercial development.

The replacement of gels by pumpable solutions of long polymers is making capillary array electrophoresis (CAE) potentially practical for DNA sequencing. The first CAE system for DNA was demonstrated by the team of Barry Karger (Northeastern University). In 1995, Karger and Norman Dovichi (University of Alberta, Canada) separately identified CAE conditions under which DNA sequencing reads could be extended usefully up to the 1000-base range. Another CAE system, developed by Edward Yeung (Iowa State University), has been licensed for commercial production (see box, p. 23). Mathies has developed a system in which a confocal microscope displays DNA bands. Application of this system to the sizing of larger DNA fragments binding multiple fluors allows single-molecule detection.

Replacing the gel-separation step with mass spectroscopy (MS) is another promising approach for rapid DNA sequencing. MS uses differences in mass-to-charge ratios to separate ionized atoms or molecules. Early efforts at MS sequencing were plagued by chemical reactivity during the "launching" phase of matrix-assisted laser desorption ionization (MALDI). MALDI badly degraded the DNA sample input. However, the degradation chemistry was elucidated in Smith's laboratory, leading to improvements. At ORNL, the team of Chung-Hsuan Chen has performed extensive trials of alternative matrices and has achieved significant improvements that now support sequence reads up to 100 DNA bases. The system is undergoing trials for DNA diagnostic applications.

The most revolutionary sequencing technology is being pursued by the team of Richard Keller and James Jett at LANL. Their goal is to read out sequence from single DNA molecules, work that builds on LANL's expertise in flow cytometry. The strand to be sequenced is labeled first with fluors that distinguish the four DNA subunits and is then suspended in a flow stream. An exonuclease cleaves the subunits, which flow past an interrogating laser system that reports the subunits' identities. All system constituents are operational but limited by the low subunit release rates of commercially available exonucleases. A current developmental focus is on identifying more active exonucleases.

Synthetic DNA strands in the 15- to 30-base range (oligomers) play essential roles in DNA sequencing; in sample-preparation steps for the polymerase chain reaction, which copies DNA strands millions of times; and in DNA-based diagnostics. The cost of custom oligomer synthesis once was a limiting factor in many research projects. A more economical, highly parallel oligomer synthesis technology was developed by Thomas Brennan at Stanford University (see last bullet, p. 22, for further details).

The sequencing by hybridization (SBH) technology provides information only on short stretches of DNA in a single trial (interrogation), but thousands of low-cost interrogations can be performed in parallel. SBH is very useful for rapid classification of short DNAs such as cDNAs, very low cost DNA resequencing, and detection of DNA sequence differences (polymorphisms) over short regions. The team of Radomir Crkvenjakov and Radoje Drmanac invented one format of SBH while in Yugoslavia, made substantial improvements at Argonne National Laboratory (ANL), and later started Hyseq Inc. to commercialize these technologies. At ANL, another implementation, SBH on matrices (SHOM) of gels, holds promise for high-accuracy sequence proofreading and diverse DNA diagnostics. The ANL team, led by Andrei Mirzabekov, collaborates

with the Englehardt Institute in Moscow, where SHOM was demonstrated initially.

Informatics: Data Collection and Analysis

Explosive growth of information and the challenges of acquiring, representing, and providing access to data pose continuing monumental tasks for the large public databases. Over the last 3 years, the Genome Database (GDB), the major international repository of human genome mapping data, has made extensive changes culminating in the enhanced representation of genomic maps and gene information in GDB V6.0. Major issues for the Genome Sequence DataBase (GSDB), established in 1994, are to capture and annotate the sequence data and to represent it in a form capable of supporting complex, ad hoc queries. Both GDB and GSDB have been restructured recently to handle the increasing flood of data and make it more useful for downstream biology (see Research Narratives, GDB, p. 49, and GSDB, p. 55. *[http://www.gdb.org* and *http://www.ncgr.org/gsdb]*

Victor Markowitz, formerly of LBNL, has developed a suite of database tools allowing substantial modifications of underlying data structures while the biologists' query tools remain stable. *[http://gizmo.lbl.gov/DM_TOOLS/DMTools.html]*

The Genome Annotation Consortium (based at ORNL) was initiated in 1997 to be a modular, distributed informatics facility for analyzing and processing (e.g., annotating) genome-scale sequence data.

The many improvements in World Wide Web software now enable maps to be downloaded simply by using a browser with accessory software provided by GDB. Computers sift stretches of DNA sequence for patterns that identify such biologically important features as protein-coding regions (exons), regulatory areas, and RNA splice sites. Other computer tools are used to compare a new sequence (i.e., a putative gene) against all other database entries, retrieve any homologous sequences that already have been entered, and indicate the degree of similarity.

The Gene Recognition and Analysis Internet Link (GRAIL) at ORNL localizes genes and other biologically important sequence features (see box, p. 17).

Another analytical service that returns informative, annotated data is MAGPIE, provided through ANL by Terry Gaasterland. MAGPIE is designed to reside locally at the site of a genome project and actively carry out analysis of genome sequence data as it is generated, with automated continued reevaluation as search databases grow (*http://www.mcs.anl.gov/home/gaasterl/magpie.html*). Once an automated functional overview has been established, it remains to pinpoint the organisms' exact metabolic pathways and establish how they interact. To this end, the WIT (What is There) system, which succeeds PUMA, supports the construction of metabolic pathways. Such constructions or models are based on sequence data, the clearly established biochemistry of specific organisms, and an understanding of the interdependencies of biochemical mechanisms. WIT, which was developed by Evgenij Selkov and Ross Overbeek at ANL, offers a particularly valuable tool for testing current hypotheses about microbial biology. *[http://www.cme.msu.edu/WIT]*

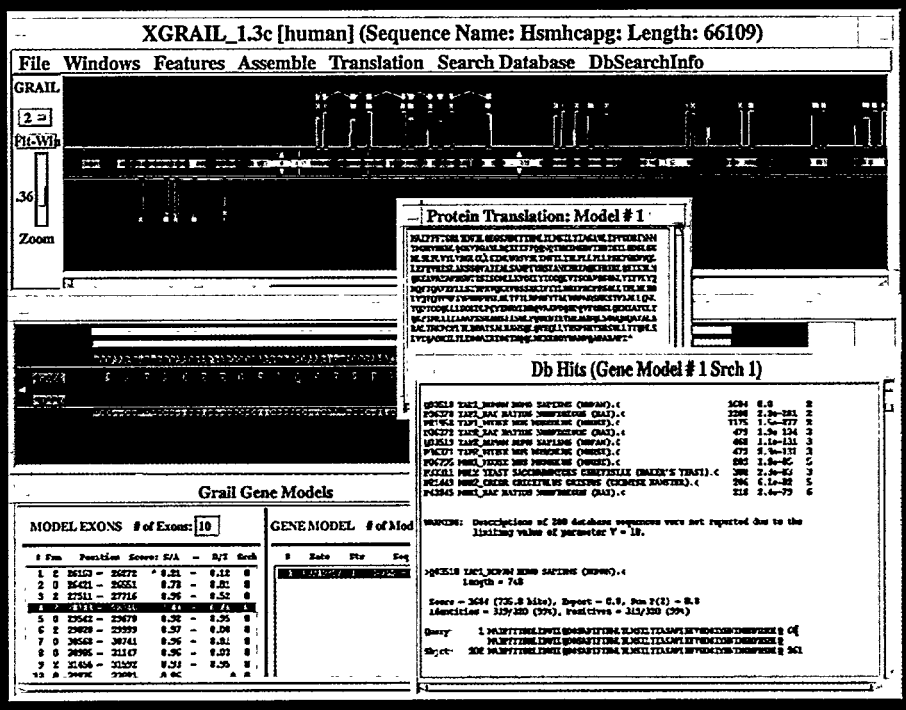
Researchers at the University of Colorado have developed another approach for predicting coding regions in genomic DNA, combining multiple types of evidence into a single scoring function, and returning both optimal and ranked suboptimal solutions. The approach is robust to substitution errors but sensitive to frameshift errors. The group is now exploring methods for predicting other classes of sequence regions, especially promoters. *[*software

GRAIL and GenQuest

In 1996 the Gene Recognition and Analysis Internet Link (GRAIL) processed nearly 40 million bases of sequence per month, making it the most widely used "gene-finding" system available. Developed at Oak Ridge National Laboratory (ORNL) by a team led by Ed Uberbacher, GRAIL uses artificial intelligence and machine learning to discover complex relationships in sequence data. The genQuest server, also at ORNL, compares information generated by GRAIL with data in protein, DNA, and motif databases to add further value to annotation of DNA sequences.

GRAIL's latest version (1.3) combines a Motif Graphical Client with improved sensitivity and splice-site recognition, better performance in AT-rich regions, new analysis systems for model organisms, and frameshift detection. This system can be used on a wide variety of UNIX platforms, including Sun, DEC, and SGI. The many ways to access GRAIL include a command line sockets client that permits remote program calls to all basic GRAIL-genQuest analysis services, thus allowing convenient integration of GRAIL results into automated analysis pipelines.

Contact GRAIL staff through the Web site at *http://compbio.ornl.gov* or at *GRAILMAIL@ornl.gov* for e-mail and ftp access.



The figure above shows the GRAIL analysis of part of the human major histocompatibility locus, which carries genes responsible for cellular immunity. Included in this analysis are potential exons (gene-coding regions), gene models, CpG islands (areas rich in bases C and G found in most mammalian genes), and repetitive DNA elements. [Source: Richard Mural, ORNL]

and information: *http://beagle.colorado.edu/~eesnyder/GeneParser.html]*

The Baylor College of Medicine (BCM) Search Launcher improves user access to the wide variety of database-search tools available on the Web. Search Launcher features a single point of entry for related searches, the addition of hypertext links to results returned by remote servers, and a batch client. *[http://gc.bcm.tmc.edu:8088/search-launcher/launcher.html]*

FASTA-SWAP, also from the BCM group, is a new pattern-search tool for databases that improves sensitivity and specificity to help detect related sequences. BEAUTY, an enhanced version of the BLAST database-search program, improves access to information about the functions of matched sequences and incorporates additional hypertext links. Graphical displays allow correlation of hit positions with annotated domain positions. Future plans include providing access to information from and direct links to other databases, including organism-specific databases.

PROCRUSTES uses comparisons of the same gene of different species to delimit gene structure much more accurately. The product of a collaboration between Pavel Pevzner (University of Southern California) and two Russian researchers, PROCRUSTES is based on the spliced-alignment algorithm, which explores all possible exon assemblies and finds the multiexon structure that best fits a related protein. *[http://www-hto.usc.edu/software/procrustes]*

The Ethical, Legal, and Social Issues component of the DOE Human Genome Program supports projects to help judges understand the scientific validity of the genetics-based claims that are poised to flood the nation's courtrooms. Robert F. Orr (left) of the North Carolina Supreme Court and Francis X. Spina of the Massachusetts Appeals Court at the New England Regional Conference on the Courts and Genetics (July 1997) participate in a hands-on laboratory session. As a prelude to learning the fundamentals of DNA science and genetic testing, the judges are precipitating DNA (seen as streaks on the glass rod in the tube) from a solution containing the bacterium Escherichia coli. *[Courts and Science On-Line Magazine:* http://www.ornl.gov/courts*]*

Ethical, Legal, and Social Issues (ELSI)

From the outset of the Human Genome Project, researchers recognized that the resulting increase in knowledge about human biology and personal genetic information would raise complex ethical and policy issues for individuals and society. Rapid worldwide progress in the project has heightened the urgency of this challenge.

Most observers agree that personal knowledge of genetic susceptibility can be expected to serve humankind well, opening the door to more accurate diagnoses, preventive intervention, intensified screening, lifestyle changes, and early and effective treatment. But such knowledge has another side, too: risk of anxiety, unwelcome changes in personal relationships, and the danger of stigmatization. Often, genetic tests can indicate possible future medical conditions far in advance of any symptoms or available therapies or treatments. If handled carelessly, genetic information could threaten an individual with discrimination by potential employers and insurers.

Other issues are perhaps less immediate than these personal concerns but no less challenging. How, for example, are products of the Human Genome Project to be patented and commercialized? How are the judicial, medical, and educational communities—not to mention the public at large—to be educated effectively about genetic research and its implications?

To confront these issues, the DOE and NIH ELSI programs jointly established an ELSI working group to coordinate policy and research between the two agencies. *[*An FY 1997 report evaluating the joint ELSI group is available on the Web (*http://www.ornl.gov/hgmis/archive/elsirept.html*).*]*

The DOE Human Genome Program has focused its ELSI efforts on education, privacy, and the fair use of genetic information (including ownership and commercialization); workplace issues, especially screening for susceptibilities to environmental agents; and implications of research findings regarding interactions among multiple genes and environmental influences.

A few highlights from the DOE ELSI portfolio for FY 1994 through FY 1997 are outlined below.

- Three high school curriculum modules developed by the Biological Sciences Curriculum Study (BSCS). *[http://www.bscs.org]*
- An educational program in Los Angeles to develop a culturally and linguistically appropriate genetics curriculum based on a BSCS module (see above) for Hispanic students and their families. *[http://vflylab.calstatela.edu/hgp]*
- A series of workshops to educate a core group of 1000 judges around the nation and a handbook with companion videotape to assist federal and state judges in understanding and assessing genetic evidence in an increasing number of civil and criminal cases (see photo above).

- Educational materials developed by the Science+Literacy for Health Project of the American Association for the Advancement of Science (AAAS) and targeted at or above the 6th- to 8th-grade reading levels. [AAAS: 202/326-6453; *Your Genes, Your Choices* booklet: *http://www.nextwave.org/ehr/books/index.html]*
- A program at the University of Chicago aimed at developing a knowledge base for physicians and nurses who will train other practitioners to introduce new genetic services.
- A series of radio programs (see photo at right) on the science and ethical issues of the genome project and a TV documentary program on ELSI issues. *[http://www.pbs.org]*
- *The Gene Letter*, a monthly online newsletter on ELSI issues for healthcare professionals and consumers. *[http://www.geneletter.org]*
- A congressional fellowship program in human genetics, administered through AAAS, for one annual fellowship for a mid-career geneticist. *[society@genetics.faseb.org]*
- The draft Genetic Privacy Act, prepared as a model for privacy legislation and covering the collection, analysis, storage, and use of DNA samples and the genetic information derived from them. *[http://www.ornl.gov/hgmis/resource/privacy/privacy1.html]*
- Privacy studies at the Center for Social and Legal Research, including an analysis of the effects of new genetic technologies on individuals and institutions.

For details on these and other projects, see ELSI Abstracts, p. 45, in Part 2 of this report. In addition to the specific projects listed in Part 2, the DOE program sponsors a number of conferences and workshops on ELSI topics.

Leroy Hood (left) of the University of Washington, Seattle, talks with Bari Scott at the 1996 DOE Human Genome Program Contractor-Grantee Workshop. Scott represented the Genome Radio Project (see text at left), which is supported by the Ethical, Legal, and Social Issues Program of the DOE Human Genome Program. (See the project's abstract in Part 2 of this report for more information.)

DOE ELSI Web Site
http://www.ornl.gov/hgmis/resource/elsi.html

Protection of Human Research Subjects

In 1996, President Clinton appointed the National Bioethics Advisory Commission to provide guidance on the ethical conduct of current and future biological and behavioral research, especially that related to genetics and the rights and welfare of human research subjects (*http://www.nih.gov/nbac/nbac.htm*).

Also in 1996, DOE and NIH issued a document providing investigators with guidance in the use of DNA from human subjects for large-scale sequencing projects (see Appendix C: Human Subjects Guidelines, p. 77). *[http://www.ornl.gov/hgmis/archive/nchgrdoe.html]*

Lawrence Livermore National Laboratory researcher Maria de Jesus, who designed software to automate DNA isolation. [Source: Linda Ashworth, LLNL]

Converting scientific knowledge into commercially useful products

Transferring technology to the private sector, a primary mission of DOE, is strongly encouraged in the Human Genome Program to enhance the nation's investment in research and technological competitiveness. Human genome centers at Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (LANL) provide opportunities for private companies to collaborate on joint projects or use laboratory resources. These opportunities include access to information (including databases), personnel, and special facilities; informal research collaborations; Cooperative Research and Development Agreements (CRADAs); and patent and software licensing. For information on recently developed resources, contact individual genome research centers or see Research Highlights, beginning on p. 9. Many universities have their own licensing and technology transfer offices.

Some collaborations and technology-transfer highlights from FY 1994 through FY 1996 are described below.

Collaborations

Involvement of the private sector in research and development can facilitate successful transfer of technology to the marketplace, and collaborations can speed production of essential tools for genome research. A number of interactive projects are now under way, and others are in preliminary stages.

CRADAs

One technology-transfer mechanism used by DOE national laboratories is the CRADA, a legal agreement with a nongovernmental organization to collaborate on a defined research project. Under a CRADA, the two entities share scientific and technological expertise, with the governmental organization providing personnel, services, facilities, equipment, or other resources. Funds must come from the nongovernmental partner. A benefit to participating companies is the opportunity to negotiate exclusive licenses for inventions arising from these collaborations. For periods through 1996, the CRADAs in place in the DOE Human Genome Program included the following:

- LLNL with Applied Biosystems Division of Perkin-Elmer Corporation to develop analytical instrumentation for faster DNA sequencing instrumentation;
- LANL with Amgen, Inc., to develop bioassays for cell growth factors;
- Oak Ridge National Laboratory (ORNL) with Darwin Molecular, Inc., for mouse models of human immunologic disease;
- ORNL with Proctor & Gamble, Inc., for analyses of liver regeneration in a mouse model; and
- Brookhaven National Laboratory with U.S. Biochemical Corporation to identify proteins useful for primer-walking methods and large-scale sequencing.

Work for Others

In other collaborations, the LBNL genome center is participating in a Work for Others agreement with Amgen to automate the isolation and characterization of large numbers of mouse cDNAs. The center group is focusing on adapting LBNL's automated colony-picking system to cDNA protocols and applying methods to generate large numbers of filter replicas for colony

Technology Transfer Legislation

Technology transfer involves converting scientific knowledge into commercially useful products. Through the 1980s, a series of laws was enacted to encourage the development of commercial applications of federally funded research at universities and federal laboratories. Such laws [chiefly the Bayh-Dole Act of 1980, Stevenson-Wydler Act of 1980, and Federal Technology Transfer Act of 1986 (Public Laws 96-517, 96-480, and 99-502, respectively)] were not aimed specifically at genome or even biomedical research. However, such research and the surrounding commercial biotechnology enterprises clearly have benefited from them. The biotechnology sector's success owes much to federal policies on technology transfer and intellectual property. [Source: U.S. Congress, Office of Technology Assessment, *Federal Technology Transfer and the Human Genome Project*, OTA-BP-EHR-162 (Washington, D.C.: U.S. Government Printing Office, September 1995)]

filter hybridization and subsequent analysis. ["Work for Others" projects supported by an agency or organization other than DOE (e.g., NIH, National Cancer Institute, or a private company) can be conducted at a DOE installation because this work is complementary to DOE research missions and usually requires multidisciplinary DOE facilities and technologies.]

The Resource for Molecular Cytogenetics was established at LBNL and the University of California (UC), San Francisco, with the support of the Office of Biological and Environmental Research and Vysis, Inc. (formerly Imagenetics). The Resource aims to apply fluorescent in situ hybridization (FISH) techniques to genetic analysis of human tissue samples; produce probe reagents; design and develop digital-imaging microscopy; distribute probes, analysis technology, and educational materials in the molecular cytogenetic community; and transfer useful reagents, processes, and instruments to the private sector for commercialization.

NIST Advanced Technology Program

Several commercial applications of research sponsored by the U.S. Human Genome Project have been furthered by the Advanced Technology Program (ATP) of the U.S. National Institute of Standards and Technology. ATP's mission is to stimulate economic growth and industrial competitiveness by encouraging high-risk but powerful new technologies. Its Tools for DNA Diagnostics program uses collaborations among researchers and industry to develop (1) cost-effective methods for determining, analyzing, and storing DNA sequences for a wide variety of diagnostic applications ranging from healthcare to agriculture to the environment and (2) a new and potentially very large market for DNA diagnostic systems.

Awardees have included companies developing DNA diagnostic chips, more powerful cytogenetic diagnostic techniques based on comparative genomic hybridization, DNA sequencing instrumentation, and DNA analysis technology. Eventually, commercialization of these underlying technologies is expected to generate hundreds of thousands of jobs. *[800/287-3863, Fax: 301/926-9524, atp@micf.nist.gov, http://www.atp.nist.gov]*

Patenting and Licensing Highlights, FY 1994–96

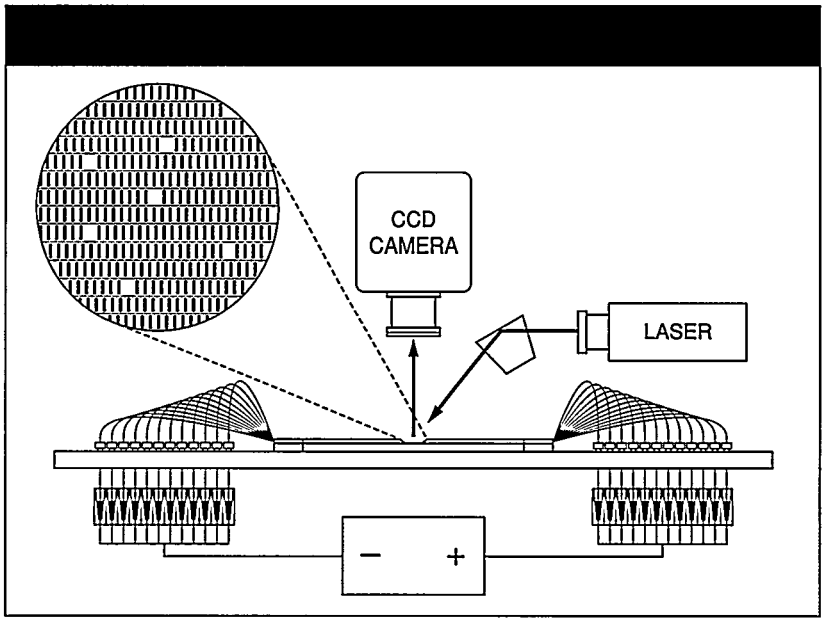
- A development license for single-molecule DNA sequencing replaced the 1991–94 CRADA (the first CRADA to be established in the U.S. Human Genome Project) between LANL and Life Technologies, Inc. (LTI).
- In 1995, a broad patent was awarded to UC for chromosome painting. This technology uses FISH to stain specific locations in cells and chromosomes for diagnosing, imaging, and studying chromosomal abnormalities and cancer. Resulting from a 1989 CRADA between LLNL and UC, FISH was licensed exclusively to Vysis.
- Hyseq, Inc., was founded in 1993 by former Argonne National Laboratory researchers Radoje Drmanac and Radomir Crkvenjakov to commercialize the sequencing by hybridization (SBH) technology. Hyseq has exclusive patent rights to a variation known as format 3 of SBH or the "super chip." Hyseq later won an Advanced Technology Program award from the U.S. National Institute of Standards and Technology to develop the technology further.
- Oligomers—short, single-stranded DNAs—are crucial reagents for genome research and biomedical diagnostics. ProtoGene Laboratories, Inc., was founded to commercialize new DNA synthesis technology (developed initially at LBNL with completed prototypes at Stanford University) and to offer the first lower-cost custom oligomer synthesis. The Parallel Array Synthesis system, which independently synthesizes 96 oligomers per run in a standard 96-well microtiter plate format, shows great promise for significant cost reductions. ProtoGene first

licensed sales and distribution to LTI and, later, production rights as well. LTI operates production centers in the United States, Europe, and Japan.

- The GRAIL-genQuest sequence-analysis software developed at ORNL was licensed by Martin Marietta Energy Systems (now Lockheed Martin Energy Research) to ApoCom, Inc., for pharmaceutical and biotechnology company researchers who cannot use the Internet because of data-security concerns. The public GRAIL-genQuest service remains freely available on the Internet (see box, p. 17).
- In 1995, an exclusive license was granted to U.S. Biochemical Corporation for a genetically engineered, heat-stable, DNA-replicating enzyme with much-improved sequencing properties. The enzyme was developed by Stanley Tabor at Harvard University Medical School.
- In 1995, an advanced capillary array electrophoresis system for sequencing DNA was patented by Iowa State University. The system was licensed to Premier American Technologies Corporation for commercialization (see graphic at right and R&D 100 Awards, next page).
- In 1996, a patent was granted to LANL researchers for DNA fragment sizing and sorting by laser-induced fluorescence. An exclusive license was awarded to Molecular Technology, Inc., for commercialization of the single-molecule detection capability related to DNA sizing (see R&D 100 Awards, next page).

SBIR and STTR

Small Business Innovation Research (SBIR) Program awards are designed to stimulate commercialization of new technology for the benefit of both the private and public sectors. The highly competitive program emphasizes cutting-edge, high-risk research with potential for high payoff in different areas, including human genome research. Small business firms with fewer than 500 employees are invited to submit applications. SBIR human genome topics concentrate on innovative and experimental approaches for carrying out the goals of the Human Genome Project (see SBIR, p. 63, in Part 2 of this report). The Small Business Technology Transfer (STTR) Program fosters transfers between research institutions and small businesses. *[*DOE SBIR and STTR contact: Kay Etzler (301/903-5867, Fax: -5488, *Kay.Etzler@oer.doe.gov*), *http://sbir.er.doe.gov/sbir, http://sttr.er.doe.gov/sttr]*



Capillary Array Electrophoresis (CAE). *CAE systems promise dramatically faster and higher-resolution fragment separation for DNA sequencing. A multiplexed CAE system designed by Edward Yeung (Iowa State University) has been developed for commercial production by Premier American Technologies Corporation (PATCO). In the PATCO ESY9600 model, DNA samples are introduced into the 96-capillary array; as the separated fragments pass through the capillaries, they are irradiated all at once with laser light. Fluorescence is measured by a charged coupled device that acts as a simultaneous multichannel detector. (Inset circle at upper left: Closeup view of individual capillary lanes with separated samples.) Because every fragment length exists in the sample, bases are identified in order according to the time required for them to reach the laser-detector region.* *[Source: Thomas Kane, PATCO]*

Technology Transfer Award

A Federal Laboratory Consortium Award for Excellence in Technology Transfer was presented to Edward Yeung and a research team at Iowa State University's Ames Laboratory in 1993. Their laser-based method for indirect fluorescence of biological samples may have applications for routine high-speed DNA sequencing (see graphic, p. 23). Yeung also won the 1994 American Chemical Society Award for Analytical Chemistry.

1997 R&D 100 Awards

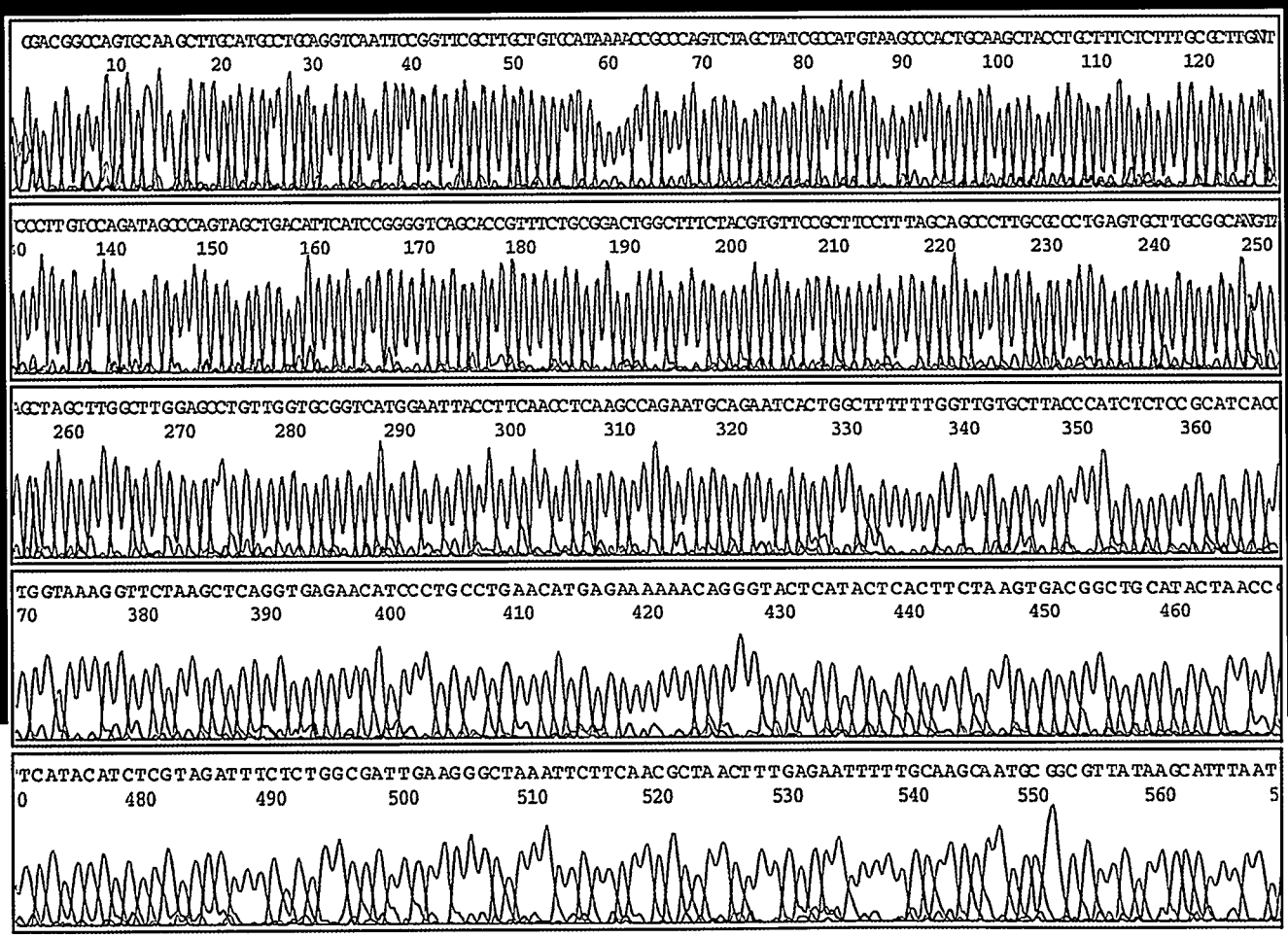
DOE researchers in 12 facilities across the country won 36 of the R&D 100 Awards given by *Research and Development Magazine* for 1996 work. DOE award-winning research ranged from advances in supercomputing to the biological recycling of tires. Announced in July 1997, these awards bring DOE's R&D 100 total to 453, the most of any single organization and twice as many as all other government agencies combined.

Two DOE genome-related research projects received 1997 R&D 100 Awards. One was to Yeung (see text at left and graphic, p. 23) for "ESY9600 Multiplexed Capillary Electrophoresis DNA Sequencer."

The other award was to Richard Keller and James Jett (LANL) with Amy Gardner (Molecular Technologies, Inc.) for "Rapid-Size Analysis of Individual DNA Fragments." This technology speeds determination of DNA fragment sizes, making DNA fingerprinting applications in biotechnology and other fields more reliable and practical.

R&D Magazine began making annual awards in 1963 to recognize the 100 most significant new technologies, products, processes, and materials developed throughout the world during the previous year (*http://www.rdmag.com/rd100/100award.htm*). Winners are chosen by the magazine's editors and a panel of 75 respected scientific experts in a variety of disciplines. Previous winners of R&D 100 Awards include such well-known products as the flashcube (1965), antilock brakes (1969), automated teller machine (1973), fax machine (1975), digital compact cassette (1993), and Taxol anticancer drug (1993).

Readout from an automated DNA sequencing machine depicts the order of the four DNA bases (A, T, C, and G) in a DNA fragment of more than 500 bases. [Source: Linda Ashworth, LLNL]



Joint Genome Institute

DOE Merges Sequencing Efforts of Genome Centers

http://www.jgi.doe.gov

Elbert Branscomb
JGI Scientific Director
Lawrence Livermore
National Laboratory
7000 East Avenue, L-452
Livermore, CA 94551
510/422-5681
***elbert@alu.llnl.gov* or**
elbert@shotgun.llnl.gov

In a major restructuring of its Human Genome Program, on October 23, 1996, the DOE Office of Biological and Environmental Research established the Joint Genome Institute (JGI) to integrate work based at its three major human genome centers.

The JGI merger represents a shift toward large-scale sequencing via intensified collaborations for more effective use of the unique expertise and resources at Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (see Research Narratives, beginning on p. 27 in this report). Elbert Branscomb (LLNL) serves as JGI's Scientific Director. Capital equipment has been ordered, and operational support of about $30 million is projected for the 1998 fiscal year.

With easy access to both LBNL and LLNL, a building in Walnut Creek, California, is being modified. Here, starting in late FY 1998, production DNA sequencing will be carried out for JGI. Until that time, large-scale sequencing will continue at LANL, LBNL, and LLNL. Expectations are that within 3 to 4 years the Production Sequencing Facility will house some 200 researchers and technicians working on high-throughput DNA sequencing using state-of-the-art robotics.

Initial plans are to target gene-rich regions of around 1 to 10 megabases for sequencing. Considerations include gene density, gene families (especially clustered families), correlations to model organism results, technical capabilities, and relevance to the DOE mission (e.g., DNA repair, cancer susceptibility, and impact of genotoxins). The JGI program is subject to regular peer review.

Sequence data will be posted daily on the Web; as the information progresses to finished quality, it will be submitted to public databases.

As JGI and other investigators involved in the Human Genome Project are beginning to reveal the DNA sequence of the 3 billion base pairs in a reference human genome, the data already are becoming valuable reagents for explorations of DNA sequence function in the body, sometimes called "functional genomics." Although large-scale sequencing is JGI's major focus, another important goal will be to enrich the sequence data with information about its biological function. One measure of JGI's progress will be its success at working with other DOE laboratories, genome centers, and non-DOE academic and industrial collaborators. In this way, JGI's evolving capabilities can both serve and benefit from the widest array of partners.

Production DNA Sequencing Begun Worldwide

The year 1996 marked a transition to the final and most challenging phase of the U.S. Human Genome Project, as pilot programs aimed at refining large-scale sequencing strategies and resources were funded by DOE and NIH (see Research Highlights, DNA Sequencing, p. 14). Internationally, large-scale human genome sequencing was kicked off in late 1995 when The Wellcome Trust announced a 7-year, $75-million grant to the private Sanger Centre to scale up its sequencing capabilities. French investigators also have announced intentions to begin production sequencing.

Funding agencies worldwide agree that rapid and free release of data is critical. Other issues include sequence accuracy, types of annotation that will be most useful to biologists, and how to sustain the reference sequence.

The international Human Genome Organisation maintains a Web page to provide information on current and future sequencing projects and links to sites of participating groups (*http://hugo.gdb.org*). The site also links to reports and resources developed at the February 1996 and 1997 Bermuda meetings on large-scale human genome sequencing, which were sponsored by The Wellcome Trust.

Lawrence Livermore National Laboratory Human Genome Center

http://www-bio.llnl.gov/bbrp/genome/genome.html

The Human Genome Center at Lawrence Livermore National Laboratory (LLNL) was established by DOE in 1991. The center operates as a multidisciplinary team whose broad goal is understanding human genetic material. It brings together chemists, biologists, molecular biologists, physicists, mathematicians, computer scientists, and engineers in an interactive research environment focused on mapping, DNA sequencing, and characterizing the human genome.

Goals and Priorities

In the past 2 years, the center's goals have undergone an exciting evolution. This change is the result of several factors, both intrinsic and extrinsic to the Human Genome Project. They include: (1) successful completion of the center's first-phase goal, namely a high-resolution, sequence-ready map of human chromosome 19; (2) advances in DNA sequencing that allow accelerated scaleup of this operation; and (3) development of a strategic plan for LLNL's Biology and Biotechnology Research Program that will integrate the center's resources and strengths in genomics with programs in structural biology, individual susceptibility, medical biotechnology, and microbial biotechnology.

The primary goal of LLNL's Human Genome Center is to characterize the mammalian genome at optimal resolution and to provide information and material resources to other in-house or collaborative projects that allow exploitation of genomic biology in a synergistic manner. DNA sequence information provides the biological driver for the center's priorities:

- Generation of highly accurate sequence for chromosome 19.
- Generation of highly accurate sequence for genomic regions of high biological interest to the mission of the DOE Office of Biological and Environmental Research (e.g., genes involved in DNA repair, replication, recombination, xenobiotic metabolism, and cell-cycle control).
- Isolation and sequence of the full insert of cDNA clones associated with genomic regions being sequenced.
- Sequence of selected corresponding regions of the mouse genome in parallel with the human.
- Annotation and position of the sequenced clones with physical landmarks such as linkage markers and sequence tagged sites (STSs).
- Generation of mapped chromosome 19 and other genomic clones [cosmids, bacterial artificial chromosomes (BACs), and P1 artificial chromosomes (PACs)] for collaborating groups.
- Sharing of technology with other groups to minimize duplication of effort.
- Support of downstream biology projects, for example, structural biology, functional studies, human variation, transgenics, medical biotechnology, and microbial biotechnology with know-how, technology, and material resources.

Center Organization and Activities

Completion and publication of the metric physical map of human chromosome 19 (see p. 28) in 1995 has led to consolidation of many functions associated with physical mapping, with increased emphasis on DNA sequencing. The center is organized into five broad areas of research and support: sequencing, resources, functional genomics, informatics and analytical genomics, and instrumentation. Each area consists of multiple projects, and extensive interaction occurs both within and among projects.

Human Genome Center
Lawrence Livermore National Laboratory
Biology and Biotechnology Research Program
7000 East Avenue, L-452
Livermore, CA 94551

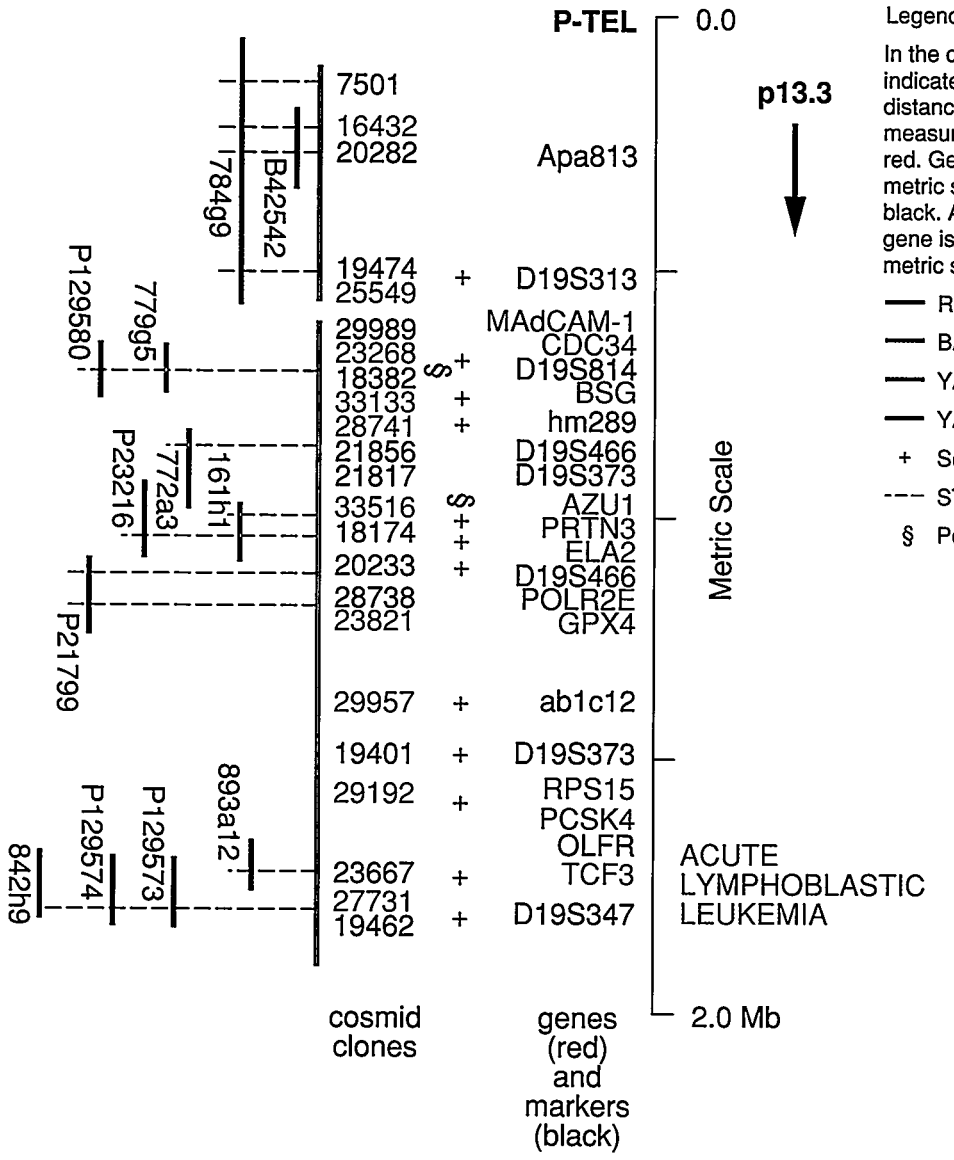
Anthony V. Carrano
Director
510/422-5698, Fax: /423-3110
carrano1@llnl.gov

Linda Ashworth
Assistant to Center Director
510/422-5665, Fax: -2282
ashworth1@llnl.gov

In lieu of individual abstracts, research projects and investigators at the LLNL Human Genome Center are represented in this narrative. More information can be found on the center's Web site (see URL above).

Update

In 1997 Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory began collaborating in a Joint Genome Institute to implement high-throughput sequencing [see p. 26 and *Human Genome News* 8(2), 1–2].



Legend

In the column labeled cosmid clones, black indicates a FISH-ordered clone where distance between clones has been measured. Other cosmids are shown in red. Genes are in red to the left of the metric scale. Other markers are labeled in black. A disease associated with a specific gene is shown in blue to the right of the metric scale.

- Restriction-mapped contig
- BAC, PAC, or P1 clone
- YAC with known and concordant size
- YAC with unknown or discordant size
- \+ Sequence tagged site (STS)
- --- STS and/or hybridization results
- § Polymorphic marker

Chromosome 19 Map. *In the current map (at left) of the first 2 million bases at the p-telomere end of chromosome 19, the* EcoR I *restriction-mapped contigs (represented by red lines) provide the starting material for genomic sequencing across a region.*

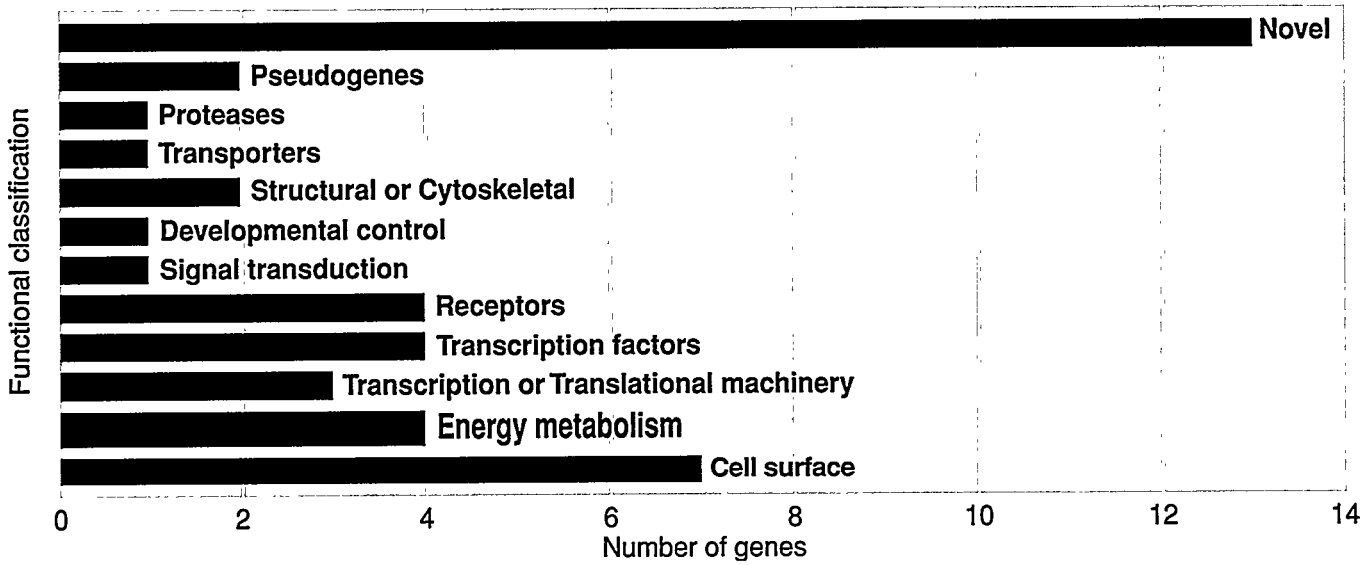
Construction of the human chromosome 19 physical map was based on a similar strategy for mapping the roundworm Caenorhabditis elegans. *View the complete map on the World Wide Web (*http://www-bio.llnl.gov/genome/html/chrom_map.html). *[Source: Adapted from figure provided by Linda Ashworth, LLNL]*

Sequencing

The sequencing group is divided into several subprojects. The core team is responsible for the construction of sequence libraries, sequencing reactions, and data collection for all templates in the random phase of sequencing. The finishing team works with data produced by the core team to produce highly redundant, highly accurate "finish" sequence on targets of interest. Finally, a team of researchers focuses specifically on development, testing, and implementation of new protocols for the entire group, with an emphasis on improving the efficiency and cost basis of the sequencing operation.

Resources

The resources group provides mapped clonal resources to the sequencing teams. This group performs physical mapping as needed for the DNA sequencing group by using fingerprinting, restriction mapping, fluorescence in situ hybridization, and other techniques. A small mapping effort is under way to identify, isolate, and characterize BAC



Putative-Gene Classification. *The figure depicts the functional classification of putative genes identified in a 1.02-Mb region on the long arm of human chromosome 19. Analysis of the completed sequence between markers D19S208 and COX7A1 revealed 43 open reading frames (ORFs) or putative genes. (An ORF is a DNA region containing specific sequences that signal the beginning and ending of a gene.)*

Thirty of these putative genes were found to have sequence similarities to a wide variety of known genes or proteins, including some involved in transcription, cell adhesion and signaling, and metabolism. Many appear to be related functionally to such known proteins as the GTP-ase activating proteins or the ETS family of transcription factors. Others seem to be new members of existing gene families, for example, the mRNA splicing factor, or of such pseudogenes as the elongation factor Tu.

In addition to those that could be classified, 13 novel genes were identified, including one with high similarity to a predicted ORF of unknown function in the roundworm Caenorhabditis elegans. *[Source: Adapted from graph provided by Linda Ashworth, LLNL]*

clones (from anywhere in the human genome) that relate to susceptibility genes, for example, DNA repair. These clones will be characterized and provided for sequencing and at the same time contribute to understanding the biology of the chromosome, the genome, and susceptibility factors. The mapping team also collaborates with others using the chromosome 19 map as a resource for gene hunting.

Functional Genomics

The functional genomics team is responsible for assembling and characterizing clones for the Integrated Molecular Analysis of Gene Expression (called IMAGE) Consortium and cDNA sequencing, as well as for work on gene expression and comparative mouse genomics. The effort emphasizes genes involved in DNA repair and links strongly to LLNL's gene-expression and structural biology efforts. In addition, this team is working closely with Oak Ridge National Laboratory (ORNL) to develop a comparative map and the sequence data for mouse regions syntenic to human chromosome 19 (see p. 32).

Informatics and Analytical Genomics

The informatics and analytical genomics group provides computer science support to biologists. The sequencing informatics team works directly with the DNA sequencing group to facilitate and automate sample handing, data acquisition and storage, and DNA sequence analysis and annotation. The

analytical genomics team provides statistical and advanced algorithmic expertise. Tasks include development of model-based methods for data capture, signal processing, and feature extraction for DNA sequence and fingerprinting data and analysis of the effectiveness of newly proposed methods for sequencing and mapping.

Instrumentation

The instrumentation group also has multiple components. Group members provide expertise in instrumentation and automation in high-throughput electrophoresis, preparation of high-density replicate DNA and colony filters, fluorescence labeling technologies, and automated sample handling for DNA sequencing. To facilitate seamless integration of new technologies into production use, this group is coupled tightly to the biologist user groups and the informatics group.

Collaborations

The center interacts extensively with other efforts within the LLNL Biology and Biotechnology Research Program and with other programs at LLNL, the academic community, other research institutes, and industry. More than 250 collaborations range from simple probe and clone sharing to detailed gene family studies. The following list reflects some major collaborations.

- Integration of the genetic map of human chromosome 19 with corresponding mouse chromosomes (ORNL).
- Miniaturized polymerase chain reaction instrumentation (LLNL).
- Sequencing of IMAGE Consortium cDNA clones (Washington University, St. Louis).
- Mapping and sequencing of a gene associated with Finnish congenital nephrotic syndrome (University of Oulu, Finland).

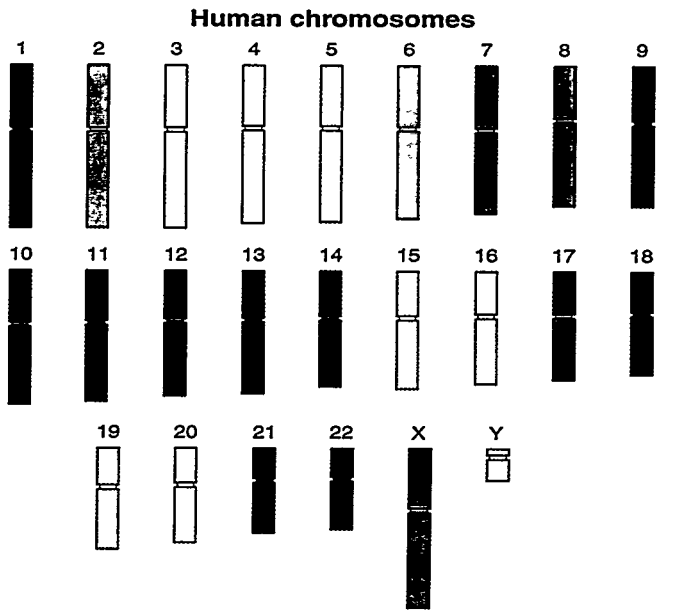
Accomplishments

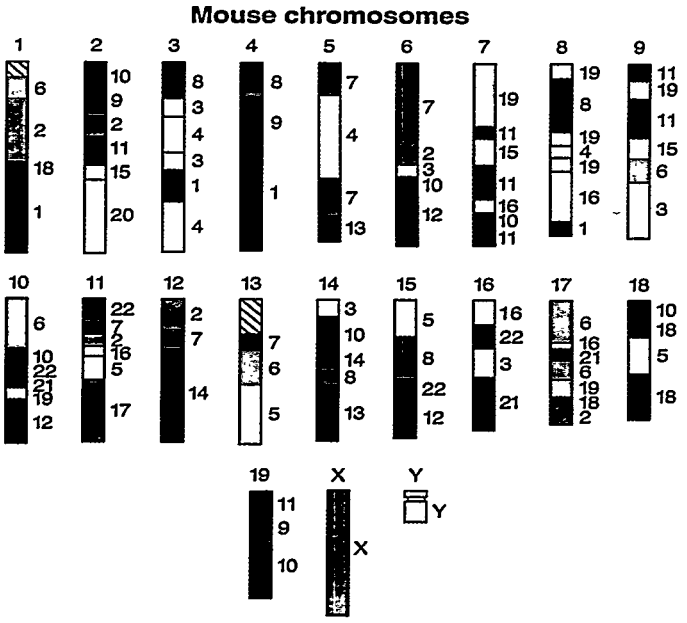
The LLNL Human Genome Center has excelled in several areas, including comparative genomic sequencing of DNA repair genes in human and rodent species, construction of a metric physical map of human chromosome 19, and development and application of new biochemical and mathematical approaches for constructing ordered clone maps. These and other major accomplishments are highlighted below.

- Completion of highly accurate sequencing totaling 1.6 million bases of DNA, including regions spanning human DNA repair genes, the candidate region for a congenital kidney disease gene, and other regions of biological interest on chromosome 19.
- Completion of comparative sequence analysis of 107,500 bases of genomic DNA encompassing the human DNA repair gene *ERCC*2 and the corresponding regions in mouse and hamster (p. 32). In addition to *ERCC*2, analysis revealed the presence of two previously undescribed genes in all three species. One of these genes is a new member of the kinesin motor protein family. These proteins play a wide variety of roles in the cell, including movement of chromosomes before cell division.
- Complete sequencing of human genomic regions containing two additional DNA repair genes. One of these, *XRCC*3, maps to human chromosome 14 and encodes a protein that may be required for chromosome stability. Analysis of the genomic sequence identified another kinesin motor protein gene physically linked to *XRCC*3. The second human repair gene, *HHR*23A, maps to 19p13.2. Sequence analysis of 110,000 bases containing *HHR*23A identified six other genes, five of which are new genes with similarity

to proteins from mouse, human, yeast, and *Caenorhabditis elegans*.

- Complete sequencing of full-length cDNAs for three new DNA repair genes (*XRCC*2, *XRCC*3, and *XRCC*9) in collaboration with the LLNL DNA repair group.
- Generation of a metric physical map of chromosome 19 spanning at least 95% of the chromosome. This unique map incorporates a metric scale to estimate the distance between genes or other markers of interest to the genetics community.
- Assembly of nearly 45 million bases of *Eco*R I restriction-mapped cosmid contigs for human chromosome 19 using a combination of fingerprinting and cosmid walking. Small gaps in cosmid continuity have been spanned by BAC, PAC, and P1 clones, which are then integrated into the restriction maps. The high depth of coverage of these maps (average redundancy, 4.3-fold) permits selection of a minimum overlapping set of clones for DNA sequencing.
- Placement of more than 400 genes, genetic markers, and other loci on the chromosome 19 cosmid map. Also, 165 new STSs associated with premapped cosmid contigs were generated and added to the physical map.
- Collaborations to identify the gene (*COMP*) responsible for two allelic genetic diseases, pseudoachondroplasia and multiple epiphyseal dysplasia, and the identification of specific mutations causing each condition.
- Through sequence analysis of the 2A subfamily of the human cytochrome P450 enzymes, identification of a new variant that exists in 10% to 20% of individuals and results in reduced ability to metabolize nicotine and the antiblood-clotting drug Coumadin.
- Location of a zinc finger gene that encodes a transcription factor regulating blood-cell development adjacent to telomere repeat sequences, possibly the gene nearest one end of chromosome 19.
- Completion of the genomic and cDNA sequence of the gene for the human Rieske Fe-S protein involved in mitochondrial respiration.
- Expansion of the mouse-human comparative genomics collaboration with ORNL to include study of new groups of clustered transcription factors found on human chromosome 19q and as syntenic homologs on mouse chromosome 7 (p. 32).
- Numerous collaborations (in particular, with Washington University and Merck) continuing to expand the LLNL-based IMAGE Consortium, an effort to characterize the transcribed human genome. The IMAGE clone collection is now the largest public collection of sequenced cDNA clones, with more than one million arrayed clones, 800,000 sequences in public databases, and 10,000 mapped cDNAs.
- Development and deployment of a comprehensive system to handle sample tracking needs of production DNA sequencing. The system combines databases and graphical interfaces running on both Mac and Sun platforms and scales easily to handle large-scale production sequencing.
- Expansion of the LLNL genome center's World Wide Web site to include tables that link to each gene being sequenced, to the quality scores and assembled bases collected each night during the sequencing process, and to the submitted GenBank sequence when a clone is completed. *[http://bbrp.llnl.gov/test-bin/projqcsummary]*



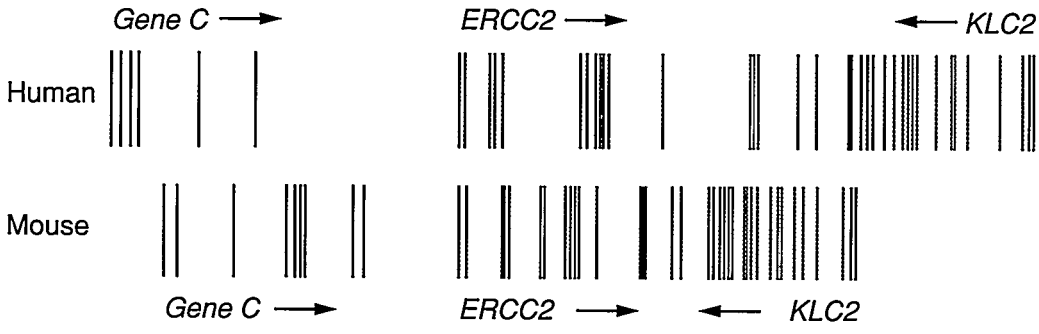


Human-Mouse Homologies. *LLNL researcher Lisa Stubbs (above) is shown in the Mouse Genetics Research Facility at ORNL.* [ORNL photo]

The figure at left demonstrates the genetic similarity (homology) of the superficially dissimilar mouse and human species. The similarity is such that human chromosomes can be cut (schematically at least) into about 150 pieces (only about 100 are large enough to appear here), then reassembled into a reasonable approximation of the mouse genome. The colors and corresponding numbers on the mouse chromosomes indicate the human chromosomes containing homologous segments. [Source: Lisa Stubbs, LLNL]

Comparative sequencing of homologous regions in human and mouse at LLNL has enhanced the ability to identify protein-coding (exon) and noncoding DNA regions that have remained unchanged over the course of evolution. Colors in the figure below depict similarities in mouse and human genes involved in DNA repair, a research interest rooted in DOE's mission to develop better technologies for measuring health effects, particularly mutations. [Source: Linda Ashworth, LLNL]

ERCC2 Region



5 kb

Legend

— Exons from "Gene C"
— Exons of ERCC2 gene
— Exons of KLC2 gene
— Non-coding conserved element

- Implementation of a new database to support sequencing and mapping work on multiple chromosomes and species. Web-based automated tools were developed to facilitate construction of this database, the loading of over 100 million bytes of chromosome 19 data from the existing LLNL database, and automated generation of Web-based input interfaces.
- Significant enhancement of the LLNL Genome Graphical Database Browser software to display and link information obtained at a subcosmid resolution from both restriction map hybridization and sequence feature data. Features, such as genes linked to diseases, allow tracking to fragments as small as 500 base pairs of DNA.
- Development of advanced microfabrication technologies to produce electrophoresis microchannels in large glass substrates for use in DNA sequencing.
- Installation of a new filter-spotting robot that routinely produces $6 \times 6 \times 384$ filters. A $16 \times 16 \times 384$ pattern has been achieved.
- Upgrade of the Lawrence Berkeley National Laboratory colony picker using a second computer so that imaging and picking can occur simultaneously.

Future Plans

Genomic sequencing currently is the dominant function of Livermore's Human Genome Center. The physical mapping effort will ensure an ample supply of sequence-ready clones. For sequencing targets on chromosome 19, this includes ensuring that the most stable clones (cosmids, BACs, and PACs) are available for sequencing and that regions with such known physical landmarks as STSs and expressed sequenced tags (ESTs) are annotated to facilitate sequence assembly and analysis. The following targets are emphasized for DNA sequencing:

- Regions of high gene density, including regions containing gene families.
- Chromosome 19, of which at least 42 million bases are sequence ready.
- Selected BAC and PAC clones representing regions of about 0.2 million to 1 million bases throughout the human genome; clones would be selected based on such high-priority biological targets as genes involved in DNA repair, replication, recombination, xenobiotic metabolism, cell-cycle checkpoints, or other specific targets of interest.
- Selected BAC and PAC clones from mouse regions syntenic with the genes indicated above.
- Full-insert cDNAs corresponding to the genomic DNA being sequenced.

The informatics team is continuing to deploy broader-based supporting databases for both mapping and sequencing. Where appropriate, Web- and Java-based tools are being developed to enable biologists to interact with data. Recent reorganization within this group enables better direct support to the sequencing group, including evaluating and interfacing sequence-assembly algorithms and analysis tools, data and process tracking, and other informatics functions that will streamline the sequencing process.

The instrumentation effort has three major thrusts: (1) continued development or implementation of laboratory automation to support high-throughput sequencing; (2) development of the next-generation DNA sequencer; and (3) development of robotics to support high-density BAC clone screening. The last two goals warrant further explanation.

The new DNA sequencer being developed under a grant from the National Institutes of Health, with minor support

through the DOE genome center, is designed to run 384 lanes simultaneously with a low-viscosity sieving medium. The entire system would be loaded automatically, run, and set up for the next run at 3-hour intervals. If successful, it should provide a 20- to 40-fold increase in throughput over existing machines.

An LLNL-designed high-precision spotting robot, which should allow a density of 98,304 spots in 96 cm^2, is now operating. The goal of this effort is to create high-density filters representing a 10× BAC coverage of both human and mouse genomes (30,000 clones = 1× coverage). Thus each filter would provide ~3× coverage, and eight such filters would provide the desired coverage for both genomes. The filters would be hybridized with amplicons from individual or region-specific cDNAs and ESTs; given the density of the BAC libraries, clones that hybridize should represent a binned set of BACs for a region of interest. These BACs could be the initial substrate for a BAC sequencing strategy. Performing hybridizations in parallel in mouse and human DNA facilitates the development of the mouse map (with ORNL involvement), and sequencing BACs from both species identifies evolutionarily conserved and, perhaps, regulatory regions.

Information generated by sequencing human and mouse DNA in parallel is expected to expand LLNL efforts in functional genomics. Comparative sequence data will be used to develop a high-resolution synteny map of conserved mouse-human domains and incorporate automated northern expression analysis of newly identified genes. Long range, the center hopes to take advantage of a variety of forms of expression analysis, including site-directed mutation analysis in the mouse.

Summary

The Livermore Human Genome Center has undergone a dramatic shift in emphasis toward commitment to large-scale, high-accuracy sequencing of chromosome 19, other chromosomes, and targeted genomic regions in the human and mouse. The center also is committed to exploiting sequence information for functional genomics studies and for other programs, both in house and collaboratively.

Los Alamos National Laboratory Center for Human Genome Studies

http://www-ls.lanl.gov/masterhgp.html

Biological research was initiated at Los Alamos National Laboratory (LANL) in the 1940s, when the laboratory began to investigate the physiological and genetic consequences of radiation exposure. Eventual establishment of the national genetic sequence databank called GenBank, the National Flow Cytometry Resource, numerous related individual research projects, and fulfillment of a key role in the National Laboratory Gene Library Project all contributed to LANL's selection as the site for the Center for Human Genome Studies in 1988.

Center Organization and Activities

The LANL genome center is organized into four broad areas of research and support: Physical Mapping, DNA Sequencing, Technology Development, and Biological Interfaces. Each area consists of a variety of projects, and work is distributed among five LANL Divisions (Life Sciences; Theoretical; Computing, Information, and Communications; Chemical Science and Technology; and Engineering Sciences and Applications). Extensive interdisciplinary interactions are encouraged.

Physical Mapping

The construction of chromosome- and region-specific cosmid, bacterial artificial chromosome (BAC), and yeast artificial chromosome (YAC) recombinant DNA libraries is a primary focus of physical mapping activities at LANL. Specific work includes the construction of high-resolution maps of human chromosomes 5 and 16 and associated informatics and gene discovery tasks.

Accomplishments

- Completion of an integrated physical map of human chromosome 16 consisting of both a low-resolution YAC contig map and a high-resolution cosmid contig map (pp. 37–39). With sequence tagged site (STS) markers provided on average every 125,000 bases, the YAC-STS map provides almost-complete coverage of the chromosome's euchromatic arms. All available loci continue to be incorporated into the map.
- Construction of a low-resolution STS map of human chromosome 5 consisting of 517 STS markers regionally assigned by somatic-cell hybrid approaches. Around 95% mega-YAC–STS coverage (50 million bases) of 5p has been achieved. Additionally, about 40 million bases of 5q mega-YAC–STS coverage have been obtained collaboratively.
- Refinement of BAC cloning procedures for future production of chromosome-specific libraries. Successful partial digestion and cloning of microgram quantities of chromosomal DNA embedded in agarose plugs. Efforts continue to increase the average insert size to about 100,000 bases.

DNA Sequencing

DNA sequencing at the LANL center focuses on low-pass sample sequencing (SASE) of large genomic regions. SASE data is deposited in publicly available databases to allow for wide distribution. Finished sequencing is prioritized from initial SASE analysis and pursued by parallel primer walking. Informatics development includes data tracking, gene-discovery integration with the Sequence Comparison ANalysis (SCAN) program, and functional genomics interaction.

Accomplishments

- SASE sequencing of 1.5 million bases from the p13 region of human chromosome 16.
- Discovery of more than 100 genes in SASE sequences.

Center for Human Genome Studies
Los Alamos National Laboratory
P.O. Box 1663
Los Alamos, NM 87545

Larry L. Deaven
Acting Director
505/667-3912, Fax: -2891
ldeaven@telomere.lanl.gov

Lynn Clark
Technical Coordinator
505/667-9376, Fax: -2891
clark@telomere.lanl.gov

Robert K. Moyzis
Director, 1989–97*

In lieu of individual abstracts, research projects and investigators at the LANL Center for Human Genome Studies are represented in this narrative. More information can be found on the center's Web site (see URL above).

Update

In 1997 Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory began collaborating in a Joint Genome Institute to implement high-throughput sequencing [see p. 26 and *Human Genome News* 8(2), 1–2].

*Now at University of California, Irvine

- Generation of finished sequence for a 240,000-base telomeric region of human chromosome 7q. From initial sequences generated by SASE, oligonucleotides were synthesized and used for primer walking directly from cosmids comprising the contig map. Complete sequencing was performed to determine what genes, if any, are near the 7q terminus. This intriguing region lacks significant blocks of subtelomeric repeat DNA typically present near eukaryotic telomeres.
- Complete single-pass sequencing of 2018 exon clones generated from LANL's flow-sorted human chromosome 16 cosmid library. About 950 discrete sequences were identified by sequence analysis. Nearly 800 appear to represent expressed sequences from chromosome 16.
- Development of Sequence Viewer to display ABI sequences with trace data on any computer having an Internet connection and a Netscape World Wide Web browser.
- Sequencing and analysis of a novel pericentromeric duplication of a gene-rich cluster between 16p11.1 and Xq28 (in collaboration with Baylor College of Medicine).

Technology Development

Technology development encompasses a variety of activities, both short and long term, including novel vectors for library construction and physical mapping; automation and robotics tools for physical mapping and sequencing; novel approaches to DNA sequencing involving single-molecule detection; and novel approaches to informatics tools for gene identification.

Accomplishments

- Development of SCAN program for large-scale sequence analysis and annotation, including a translator converting SCAN data to GIO format for submission to Genome Sequence DataBase.
- Application of flow-cytometric approach to DNA sizing of P1 artificial chromosome (PAC) clones. Less than one picogram of linear or supercoiled DNA is analyzed in under 3 minutes. Sizing range has been extended down to 287 base pairs. Efforts continue to extend the upper limit beyond 167,000 bases.
- Characterization of the detection of single, fluorescently tagged nucleotides cleaved from multiple DNA fragments suspended in the flow stream of a flow cytometer (see picture, p. 70). The cleavage rate for Exo III at 37°C was measured to be about 5 base pairs per second per M13 DNA fragment. To achieve a single-color sequencing demonstration, either the background burst rate (currently about 5 bursts per second) must be reduced or the exonuclease cleavage rate must be increased significantly. Techniques to achieve both are being explored.
- Construction of a simple and compact apparatus, based on a diode-pumped Nd:YAG laser, for routine DNA fragment sizing.
- Development of a new approach to detect coding sequences in DNA. This complete spectral analysis of coding and noncoding sequences is as sensitive in its first implementations as the best existing techniques.
- Use of phylogenetic relationships to generate new profiles of amino acid usage in conserved domains. The profiles are particularly useful for classification of distantly related sequences.

Biological Interfaces

The Biological Interfaces effort targets genes and chromosome regions associated with DNA damage and repair, mitotic stability, and chromosome structure and function as primary subjects for physical mapping and sequencing. Specific disease-associated genes on human chromosome 5 (e.g., Cri-du-Chat syndrome) and on 16 (e.g., Batten's disease and Fanconi anemia) are the subjects of collaborative biological projects.

Accomplishments

- Identification of two human 7q exons having 99% homology to the cDNA of a known human gene, vasoactive intestinal peptide receptor 2A. Preliminary data suggests that the *VIPR*2A gene is expressed.
- Identification of numerous expressed sequence tags (ESTs) localized to the 7q region. Since three of the ESTs contain at least two regions with high confidence of homology (~90%), genes in addition to *VIPR*2A may exist in the terminal region of 7q.
- Generation of high-resolution cosmid coverage on human chromosome 5p for the larynx and critical regions identified with Cri-du-Chat syndrome, the most common human terminal-deletion syndrome (in collaboration with Thomas Jefferson University).
- Refinement of the Wolf-Hirschhorn syndrome (WHS) critical region on human chromosome 4p. Using the SCAN program to identify genes likely to contribute to WHS, the project serves as a model for defining the interaction between genomic sequencing and clinical research.
- Collaborative construction of contigs for human chromosome 16, including 1.05 million bases in cosmids through the familial Mediterranean fever (FMF) gene region (with members of the FMF Consortium) and 700,000 bases in P1 clones encompassing the polycystic kidney disease gene (with Integrated Genetics, Inc.).
- Collaborative identification and determination of the complete genomic structure of the Batten's disease gene (with members of the BDG Consortium), the gamma subunit of the human amiloride-sensitive epithelial channel (Liddle's syndrome, with University of Iowa), and the polycystic kidney disease gene (with Integrated Genetics).
- Participation in an international collaborative research consortium that successfully identified the gene responsible for Fanconi anemia type A.

Chromosome 16 Physical Map (pp. 38–39). *A condensed chromosome 16 physical map constructed at Los Alamos National Laboratory (LANL) is shown in two parts on the following pages. Besides facilitating the isolation and characterization of disease genes, the map provides the framework for a large-scale sequencing effort by LANL, The Institute for Genomic Research, and the Sanger Centre.*

Distinct types of maps and data are shown as levels or tiers on the integrated map. At the top of each page is a view of the banded human chromosome to which the map is aligned. A somatic-cell hybrid breakpoint map, which divides the chromosome into 90 intervals, was used as a backbone for much of the map integration.

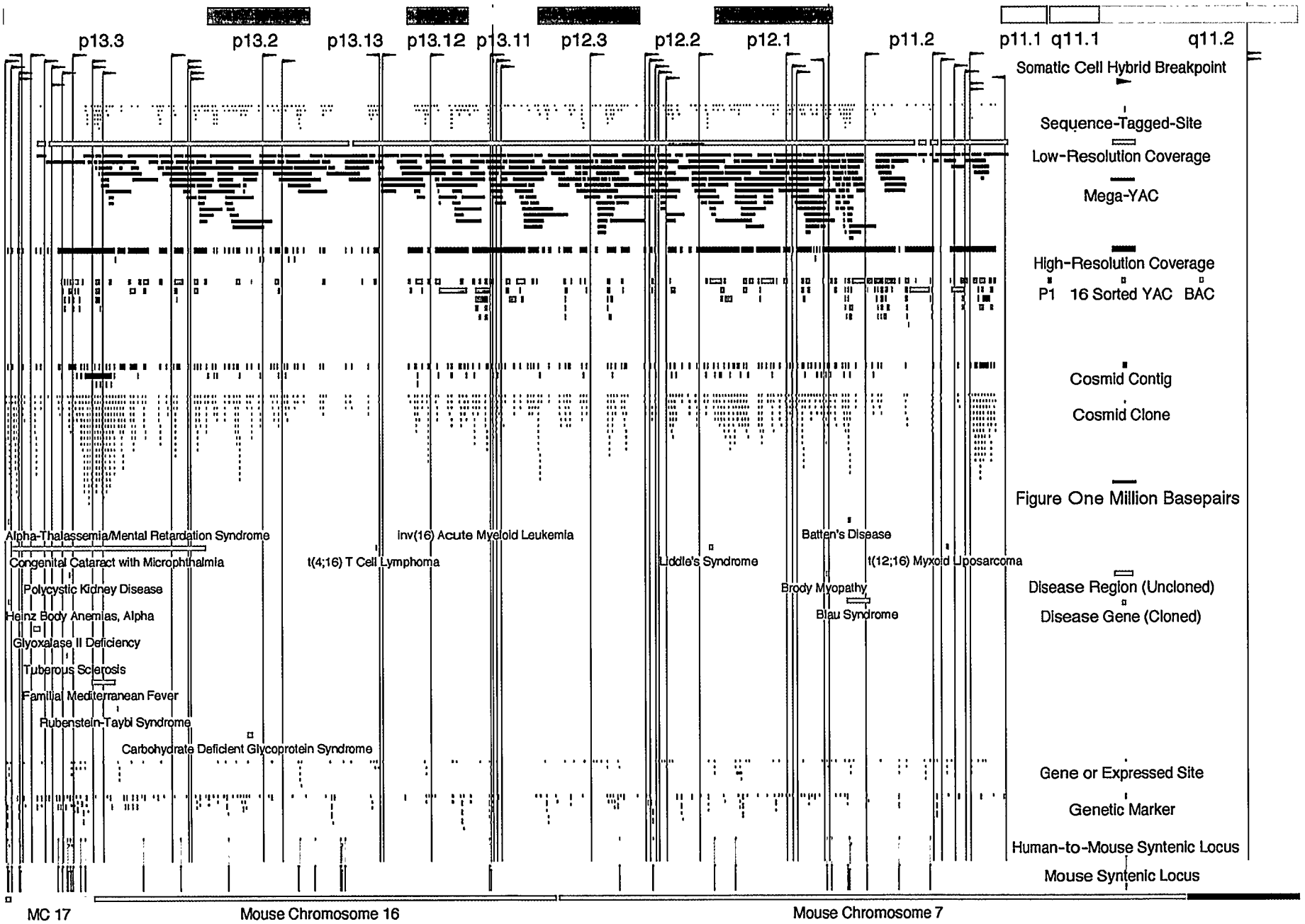
The physical map consists of both a low-resolution yeast artificial chromosome (YAC) contig map localized to and ordered within the breakpoint intervals with sequence tagged sites (STSs) and a high-resolution bacteria-based clone map. The YAC-STS map provides almost complete coverage of the chromosome's euchromatic arm, with STS markers on average every 100,000 bases.

A high-resolution, sequence-ready cosmid contig map is anchored to the YAC and breakpoint maps via STSs developed from cosmid contigs and by hybridizations between YACs and cosmids.

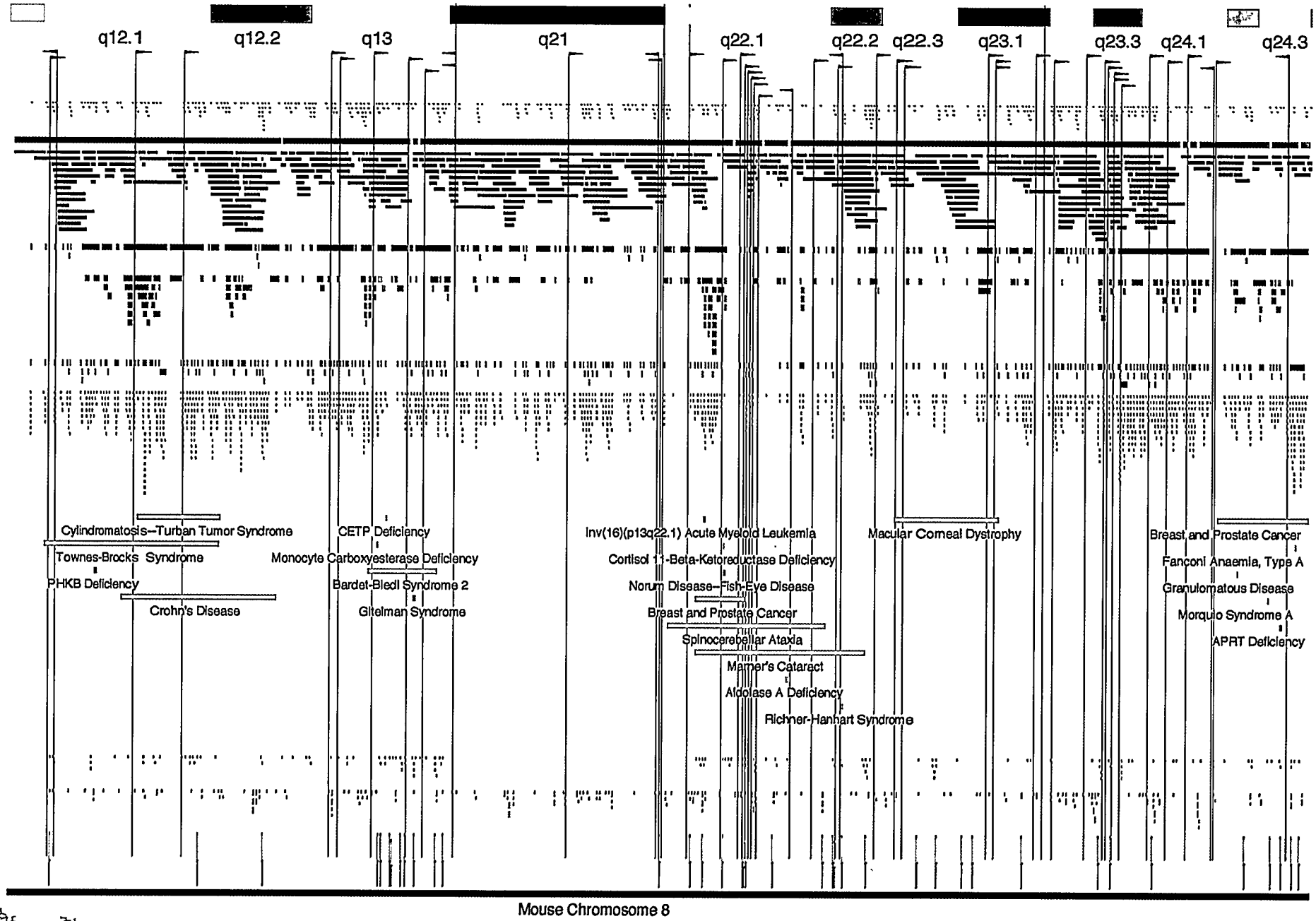
As part of the ongoing effort to incorporate all available loci onto a single map of this chromosome, the integrated map also features genes, expressed sequence tags, exons (gene-coding regions), and genetic markers.

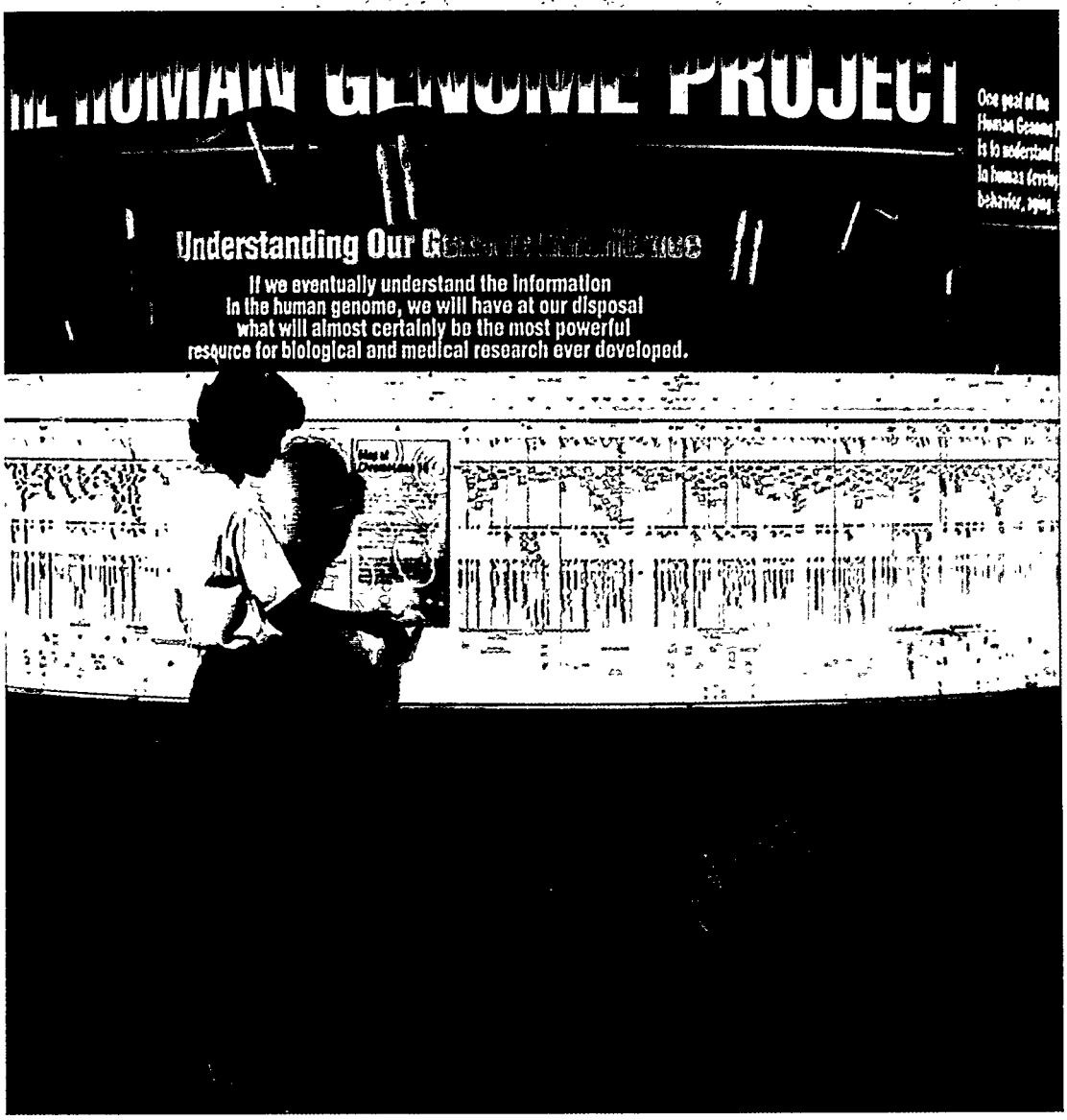
The mouse chromosome segments at the bottom of the map contain groups that correspond to human genes mapped to the regions shown above them. *[Source: Norman Doggett, LANL]*

Human Chromosome 16



Human Chromosome 16





The exhibit "Understanding Our Genetic Inheritance" at the Bradbury Science Museum in Los Alamos, New Mexico, describes the LANL Center for Human Genome Studies' contributions to the Human Genome Project. The exhibit's centerpiece is a 16-foot-long version of LANL's map of human chromosome 15. [Source: LANL Center for Human Genome Studies]

Patents, Licenses, and CRADAs

- Rhett L. Affleck, James N. Demas, Peter M. Goodwin, Jay A. Schecker, Ming Wu, and Richard A. Keller, "Reduction of Diffusional Defocusing in Hydrodynamically Focused Flows by Complexing with a High Molecular Weight Adduct," United States Patent, filed December 1996.
- R.L. Affleck, W.P. Ambrose, J.D. Demas, P.M. Goodwin, M.E. Johnson, R.A. Keller, J.T. Petty, J.A. Schecker, and M. Wu, "Photobleaching to Reduce or Eliminate Luminescent Impurities for Ultrasensitive Luminescence Analysis," United States Patent, S-87, 208, accepted September 1997.
- J.H. Jett, M.L. Hammond, R.A. Keller, B.L. Marrone, and J.C. Martin, "DNA Fragment Sizing and Sorting by Laser-Induced Fluorescence," United States Patent, S.N. 75,001, allowed May 1996.
- James H. Jett, "Method for Rapid Base Sequencing in DNA and RNA with Three Base Labeling," in preparation.
- Development license and exclusive license to LANL's DNA sizing patent obtained by Molecular Technology, Inc., for commercialization of single-molecule detection capability to DNA sizing.

Future Plans

LANL has joined a collaboration with California Institute of Technology and The Institute for Genomic Research to construct a BAC map of the *p* arm of human chromosome 16 and to complete the sequence of a 20-million–base region of this map.

In its evolving role as part of the new DOE Joint Genome Institute, LANL will continue scaleup activities focused on high-throughput DNA sequencing. Initial targets include genes and DNA regions associated with chromosome structure and function, syntenic breakpoints, and relevant disease-gene loci.

A joint DNA sequencing center was established recently by LANL at the University of New Mexico. This facility is responsible for determining the DNA sequence of clones constructed at LANL, then returning the data to LANL for analysis and archiving.

Lawrence Berkeley National Laboratory Human Genome Center

http://www-hgc.lbl.gov/GenomeHome.html

Since 1937 the Ernest Orlando Lawrence Berkeley National Laboratory (LBNL) has been a major contributor to knowledge about human health effects resulting from energy production and use. That was the year John Lawrence went to Berkeley to use his brother Ernest's cyclotrons to launch the application of radioactive isotopes in biological and medical research. Fifty years later, Berkeley Lab's Human Genome Center was established.

Now, after another decade, an expansion of biological research relevant to Human Genome Project goals is being carried out within the Life Sciences Division, with support from the Information and Computing Sciences and Engineering divisions. Individuals in these research projects are making important new contributions to the key fields of molecular, cellular, and structural biology; physical chemistry; data management; and scientific instrumentation. Additionally, industry involvement in this growing venture is stimulated by Berkeley Lab's location in the San Francisco Bay area, home to the largest congregation of biotechnology research facilities in the world.

In July 1997 the Berkeley genome center became part of the Joint Genome Institute (see p. 26).

Sequencing

Large-scale genomic sequencing has been a central, ongoing activity at Berkeley Lab since 1991. It has been funded jointly by DOE (for human genome production sequencing and technology development) and the NIH National Human Genome Research Institute [for sequencing the *Drosophila melanogaster* model system, which is carried out in partnership with the University of California, Berkeley (UCB)]. The human genome sequencing area at Berkeley Lab consists of five groups: Bioinstrumentation, Automation, Informatics, Biology, and Development. Complementing these activities is a group in Life Sciences Division devoted to functional genomics, including the transgenics program.

The directed DNA sequencing strategy at Berkeley Lab was designed and implemented to increase the efficiency of genomic sequencing (see figure, p. 45). A key element of the directed approach is maintaining information about the relative positions of potential sequencing templates throughout the entire sequencing process. Thus, intelligent choices can be made about which templates to sequence, and the number of selected templates can be kept to a minimum. More important, knowledge of the interrelationship of sequencing runs guides the assembly process, making it more resistant to difficulties imposed by repeated sequences. As of July 3, 1997, Berkeley Lab had generated 4.4 megabases of human sequence and, in collaboration with UCB, had tallied 7.6 megabases of *Drosophila* sequence.

Instrumentation and Automation

The instrumentation and automation program encompasses the design and fabrication of custom apparatus to facilitate experiments, the programming of laboratory robots to automate repetitive procedures, and the development of (1) improved hardware to extend the applicability range of existing commercial robots and (2) an integrated operating system to control and monitor experiments. Although some discrete instrumentation modules used in the integrated protocols are obtained commercially, LBNL designs its own custom instruments when existing capabilities are inadequate. The instrumentation modules are then integrated into a large system to facilitate large-scale production sequencing. In addition, a significant effort is devoted to improving

Human Genome Center
Lawrence Berkeley National Laboratory
1 Cyclotron Road
Berkeley, CA 94720

Contact:
Mohandas Narla
510/486-7029, Fax: -6746
mohandas_narla@macmail.lbl.gov

Joyce Pfeiffer
Administrative Assistant

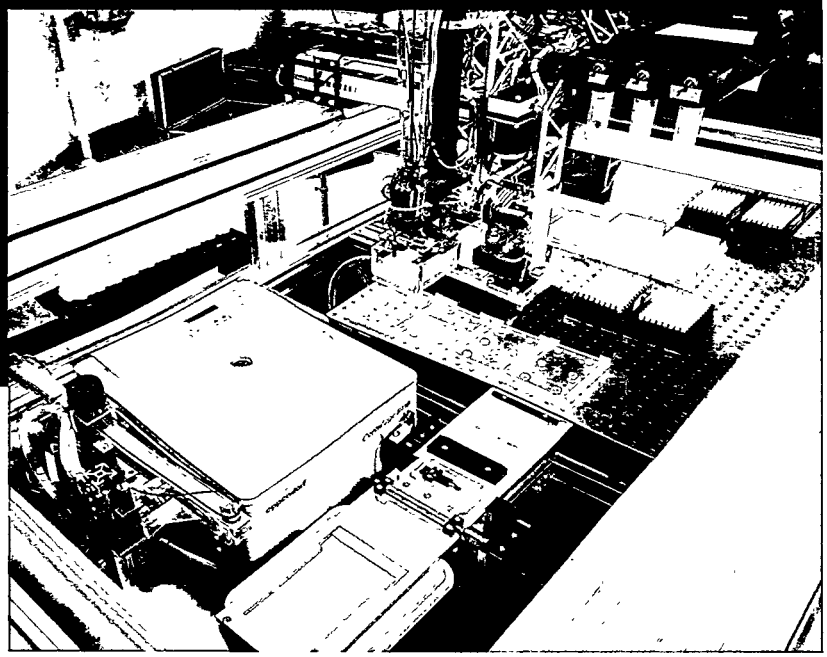
Michael Palazzolo*
Director, 1996–97

In lieu of individual abstracts, research projects and investigators at the LBNL Human Genome Center are represented in this narrative. More information can be found on the center's Web site (see URL above).

Update

In 1997 Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, and Los Alamos National Laboratory began collaborating in a Joint Genome Institute to implement high-throughput sequencing [see p. 26 and *Human Genome News* 8(2), 1–2].

*Now at Amgen, Inc.



DNA Prep Machine. *The DNA Prep machine (above) was designed by Berkeley Lab's Martin Pollard to perform plasmid preparation on 192 samples (2 microtiter plates) in about 2.5 to 4 hours, depending on the protocol. Controlled by a personal computer running a Visual Basic Control program, the instrument includes a gantry robot equipped with pipettors, reagent dispensers, hot and cold temperature stations, and a pneumatic gripper. [Source: LBNL]*

fluorescence-assay methods, including DNA sequence analysis and mass spectrometry for molecular sizing.

Recent advances in the instrumentation group include DNA Prep machine and Prep Track. These instruments are designed to automate completely the highly repetitive and labor-intensive DNA-preparation procedure to provide higher daily throughput and DNA of consistent quality for sequencing (see photos, p. 43, and Web pages: *http://hgighub.lbl.gov/esd/DNAPrep/TitlePage.html* and *http://hgighub.lbl.gov/esd/prepTrackWebpage/preptrack.htm*).

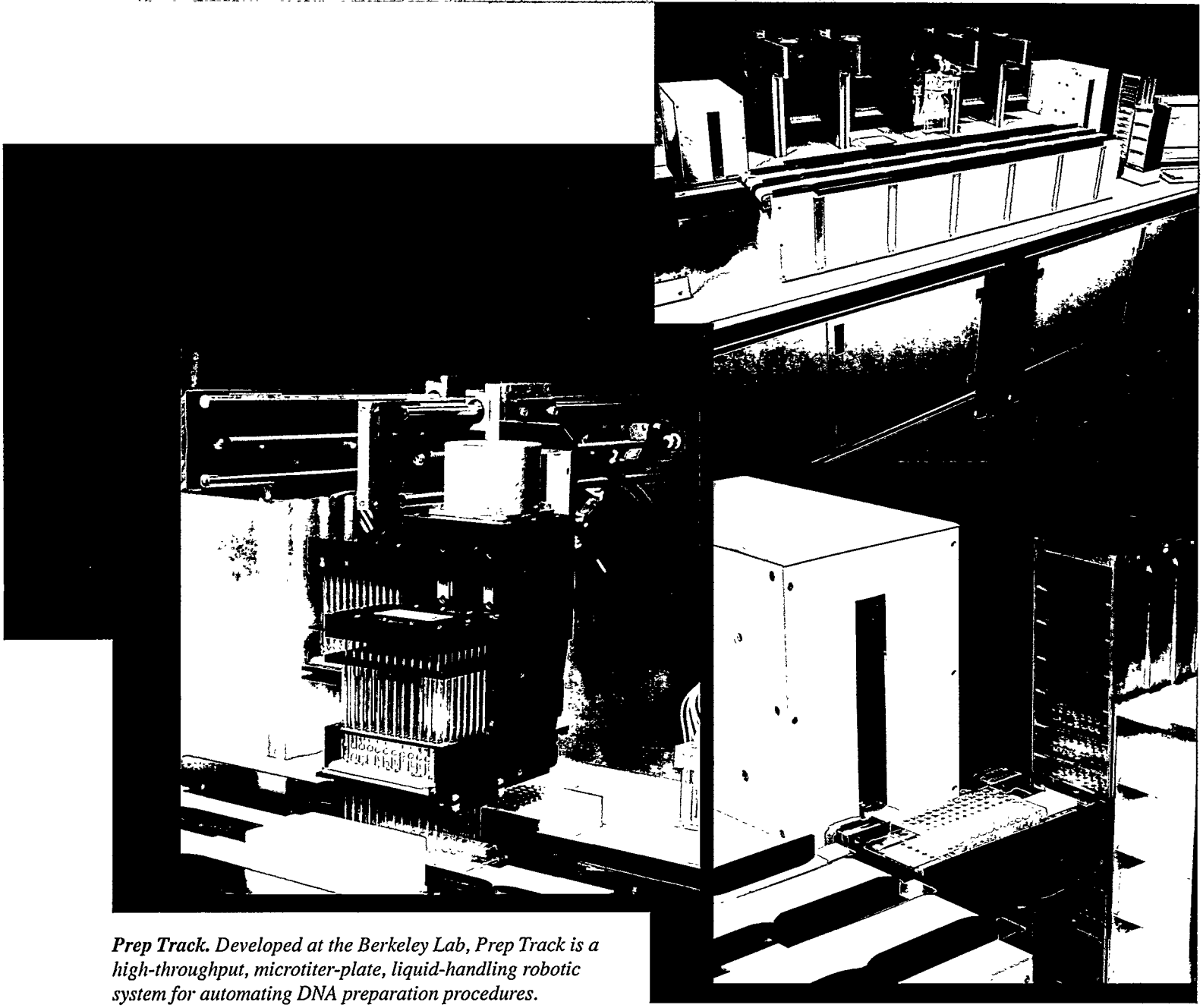
Berkeley Lab's near-term needs are for 960 samples per day of DNA extracted from overnight bacteria growths. The DNA protocol is a modified boil prep prepared in a 96-well format. Overnight bacteria growths are lysed, and samples are separated from cell debris by centrifugation. The DNA is recovered by ethanol precipitation.

Informatics

The informatics group is focused on hardware and software support and system administration, software development for end sequencing, transposon mapping and sequence template selection, data-flow automation, gene finding, and sequence analysis. Data-flow automation is the main emphasis. Six key steps have been identified in this process, and software is being written and tested to automate all six. The first step involves controlling gel quality, trimming vector sequence, and storing the sequences in a database. A program module called Move-Track-Trim, which is now used in production, was written to handle these steps. The second through fourth steps in this process involve assembling, editing, and reconstructing P1 clones of 80,000 base pairs from 400-base traces. The fifth step is sequence annotation, and the sixth is data submission.

Annotation can greatly enhance the biological value of these sequences. Useful annotations include homologies to known genes, possible gene locations, and gene signals such as promoters. LBNL is developing a workbench for automatic sequence annotation and annotation viewing and editing. The goal is to run a series of sequence-analysis tools and display the results to compare the various predictions. Researchers then will be able to examine all the annotations (for example, genes predicted by various gene-finding methods) and select the ones that look best.

Nomi Harris developed Genotator, an annotation workbench consisting of a stand-alone annotation browser and several sequence-analysis functions. The back end runs several gene finders, homology searches (using BLAST), and signal searches and saves the results in ".ace" format. Genotator thus automates the tedious process of operating a dozen different sequence-analysis programs with many different input and output formats. Genotator can function via command-line arguments or with the graphical user interface (*http://www-hgc.lbl.gov/inf/annotation.html*).



Prep Track. *Developed at the Berkeley Lab, Prep Track is a high-throughput, microtiter-plate, liquid-handling robotic system for automating DNA preparation procedures. Microtiter plates are fetched from cassettes, moved to one of two conveyor belts, and transported to protocol-defined modules. Plates are moved continuously and automatically through the system as each module simultaneously processes plates in the module lift stations. The plates exit the system and are stored in microtiter-plate cassettes.*

Modules include a station capable of dispensing liquids in volumes from as low as 5 microliters to several milliliters, four 96-channel pipettors, and the plate-fetching module. Each module is controlled independently by programmable logic controllers (PLCs). The overall system is controlled by a personal computer and a Visual Basic Control master that determines the order in which plates are processed. The actions of each lift station and dispenser or pipettor are determined locally by programs resident in each module's PLC. The Visual Basic Control program moves the plates through the system based on the predefined protocol and on module status reports as monitored by PLCs.

The current belt length on the Prep Track supports eight standard modules, which can be reconfigured to any order. Standardization of mechanical, electrical, and communication components allows new modules to be designed and manufactured easily. The current standard module footprint is 250 mm wide, 600 mm deep, and 250 mm to the conveyor belt deck. The first protocol to be implemented on Prep Track will be polymerase chain reaction setups, with sequence-reaction setups to follow. [Source: LBNL]

Progress to Date

Chromosome 5

Over the last year, the center has focused its production genomic sequencing on the distal 40 megabases of the human chromosome 5 long arm. This region was chosen because it contains a cluster of growth factor and receptor genes and is likely to yield new and functionally related genes through long-range sequence analysis. Results to date include:

- 40-megabase nonchimeric map containing 82 yeast artificial chromosomes (YACs) in the chromosome 5 distal long arm.
- 20-megabase contig map in the region of 5q23-q33 that contains 198 P1s, 60 P1 artificial chromosomes, and 495 bacterial artificial chromosomes (BACs) linked by 563 sequenced tagged sites (STSs) to form contigs.
- 20-megabase bins containing 370 BACs in 74 bins in the region of 5q33-q35.

Chromosome 21

An early project in the study of Down syndrome (DS), which is characterized by chromosome 21 trisomy, constructed a high-resolution clone map in the chromosome 21 DS region to be used as a pilot study in generating a contiguous gene map for all of chromosome 21. This project has integrated P1 mapping efforts with transgenic studies in the Life Sciences Division. P1 maps provide a suitable form of genomic DNA for isolating and mapping cDNA.

- 186 clones isolated in the major DS region of chromosome 21 comprising about 3 megabases of genomic DNA extending from D21S17 to ETS2. Through cross-hybridization, overlapping P1s were identified, as well as gaps between two P1 contigs, and transgenic mice were created from P1 clones in the DS region for use in phenotypic studies.

Transgenic Mice

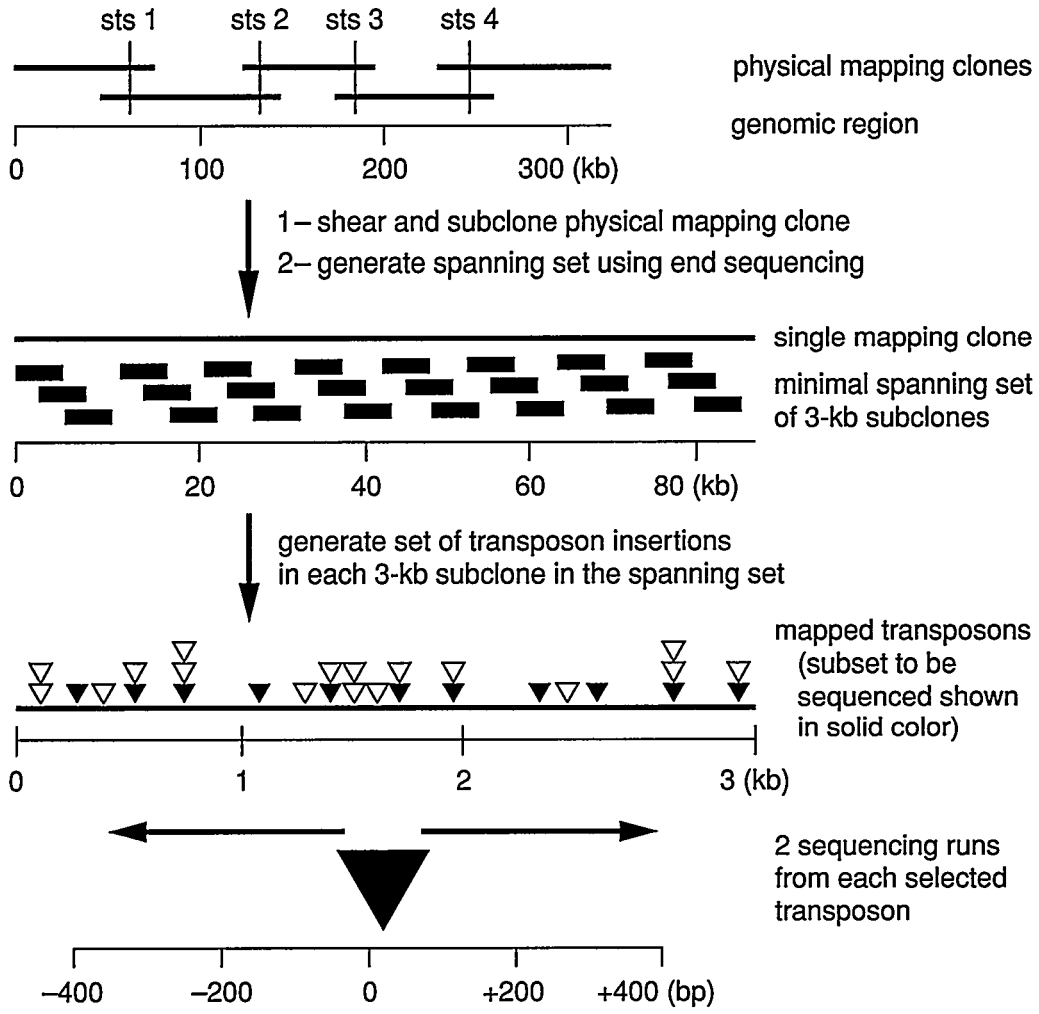
One of the approaches for determining the biological function of newly identified genes uses YAC transgenic mice. Human sequence harbored by YACs in transgenic mice has been shown to be correctly regulated both temporally and spatially. A set of nonchimeric overlapping YACs identified from the 5q31 region has been used to create transgenic mice. This set of transgenic mice, which together harbor 1.5 megabases of human sequence, will be used to assess the expression pattern and potential function of putative genes discovered in the 5q31 region. Additional mapping and sequencing are under way in a region of human chromosome 20 amplified in certain breast tumor cell lines.

Resource for Molecular Cytogenetics

Divining landmarks for human disease amid the enormous plain of the human genetic map is the mission of an ambitious partnership among the Berkeley Lab; University of California, San Francisco; and a diagnostics company. The collaborative Resource for Molecular Cytogenetics is charting a course toward important sites of biological interest on the 23 pairs of human chromosomes (*http://rmc-www.lbl.gov*).

The Resource employs the many tools of molecular cytogenetics. The most basic of these tools, and the cornerstone of the Resource's portfolio of proprietary technology, is a method generally known as "chromosome painting," which uses a technique referred to as fluorescence in situ hybridization or FISH. This technology was invented by LBNL Resource leaders Joe Gray and Dan Pinkel.

A technology to emerge recently from the Resource is known as "Quantitative DNA Fiber Mapping (QDFM)." High-resolution human genome maps in a form suitable for DNA sequencing traditionally have been constructed by



Sequencing Strategy. *The directed sequencing strategy used at LBNL involves four steps: (1) generate a P1-based physical map (using STS-content mapping) to provide a set of minimally overlapping clones, (2) shear and subclone each P1 clone into 3-kilobase fragments and identify a minimally overlapping subclone set, (3) generate and map transposon inserts in each subclone, and (4) sequence using commercial primer-binding sites engineered into the transposon. Subclone sequences are then assembled and edited, and the gaps are identified. P1 clones are reconstructed, and the resulting composite data is analyzed, annotated, and finally submitted to the databases. The production sequencing effort has generated 12 megabases of finished, double-stranded genomic DNA sequence from both* Drosophila *and human templates.* *[Source: Adapted from figure provided by LBNL]*

various methods of fingerprinting, hybridization, and identification of overlapping STSs. However, these techniques do not readily yield information about sequence orientation, the extent of overlap of these elements, or the size of gaps in the map. Ulli Weier of the Resource developed the QDFM method of physical map assembly that enables the mapping of cloned DNA directly onto linear, fully extended DNA molecules. QDFM allows unambiguous assembly of critical elements leading to high-resolution physical maps. This task now can be accomplished in less than 2 days, as compared with weeks by conventional methods. QDFM also enables detection and characterization of gaps in existing physical maps—a crucial step toward completing a definitive human genome map.

Lawrence Livermore National Laboratory scientist Stephanie Stilwagen loads a sample into an automated DNA sequencing system. [Source: Linda Ashworth, LLNL]

The Human Genome Project soon will need to increase rapidly the scale at which human DNA is analyzed. The ultimate goal is to determine the order of the 3 billion bases that encode all heritable information. During the 20 years since effective methods were introduced to carry out DNA sequencing by biochemical analysis of recombinant-DNA molecules, these techniques have improved dramatically. In the late 1970s, segments of DNA spanning a few thousand bases challenged the capacity of world-class sequencing laboratories. Now, a few million base pairs per year represent state-of-the-art output for a single sequencing center.

However, the Human Genome Project is directed toward completing the human sequence in 5 to 10 years, so the data must be acquired with technology available now. This goal, while clearly feasible, poses substantial organizational and technical challenges. Organizationally, genome centers must begin building data-production units capable of sustained, cost-effective operation. Technically, many incremental refinements of current technology must be introduced, particularly those that remove impediments to increasing the scale of DNA sequencing. The University of Washington (UW) Genome Center is active in both areas.

Production Sequencing

Both to gain experience in the production of high-quality, low-cost DNA sequence and to generate data of immediate biological interest, the center is sequencing several regions of human and mouse DNA at a current throughput of 2 million bases per year. This "production sequencing" has three major targets: the human leukocyte antigen (HLA) locus on human chromosome 6, the mouse locus encoding the alpha subunit of T-cell receptors, and an "anonymous" region of human chromosome 7.

The HLA locus encodes genes that must be closely matched between organ donors and organ recipients. This sequence data is expected to lead to long-term improvements in the ability to achieve good matches between unrelated organ donors and recipients.

The mouse locus that encodes components of the T-cell–receptor family is of interest for several reasons. The locus specifies a set of proteins that play a critical role in cell-mediated immune responses. It provides sequence data that will help in the design of new experimental approaches to the study of immunity in mice—one of the most important experimental animals for immunological research. In addition, the locus will provide one of the first large blocks of DNA sequence for which both human and mouse versions are known.

Human-mouse sequence comparisons provide a powerful means of identifying the most important biological features of DNA sequence because these features are often highly conserved, even between such biologically different organisms as human and mouse. Finally, sequencing an "anonymous" region of human chromosome 7, a region about which little was known previously, provides experience in carrying out large-scale sequencing under the conditions that will prevail throughout most of the Human Genome Project.

Technology for Large-Scale Sequencing

In addition to these pilot projects, the UW Genome Center is developing incremental improvements in current sequencing technology. A particular focus is on enhanced computer software to process raw data acquired with automated laboratory instruments that are used in DNA mapping and sequencing. Advanced instrumentation is commercially available for determining DNA sequence via the "four-color–fluorescence method," and this instrumentation is expected to carry

University of Washington Genome Center
Department of Medicine
Box 352145
Seattle, WA 98195

Maynard Olson
Director
206/685-7366, Fax: -7344
mvo@u.washington.edu

For more information on research projects and investigators at the University of Washington Genome Center, see abstracts in Part 2 of this report and the center's Web site (see URL above).

the main experimental load of the Human Genome Project. Raw data produced by these instruments, however, require extensive processing before they are ready for biological analysis.

Large-scale sequencing involves a "divide-and-conquer" strategy in which the huge DNA molecules present in human cells are broken into smaller pieces that can be propagated by recombinant-DNA methods. Individual analyses ultimately are carried out on segments of less than 1000 bases. Many such analyses, each of which still contains numerous errors, must be melded together to obtain finished sequence. During the melding, errors in individual analyses must be recognized and corrected. In typical large-scale sequencing projects, the results of thousands of analyses are melded to produce highly accurate sequence (less than one error in 10,000 bases) that is continuous in blocks of 100,000 or more bases. The UW Genome Center is playing a major role in developing software that allows this process to be carried out automatically with little need for expert intervention. Software developed in the UW center is used in more than 50 sequencing laboratories around the world, including most of the large-scale sequencing centers producing data for the Human Genome Project.

High-Resolution Physical Mapping

The UW Genome Center also is developing improved software that addresses a higher-level problem in large-scale sequencing. The starting point for large-scale sequencing typically is a recombinant-DNA molecule that allows propagation of a particular human genomic segment spanning 50,000 to 200,000 bases. Much effort during the last decade has gone into the physical mapping of such molecules, a process that allows huge regions of chromosomes to be defined in terms of sets of overlapping recombinant-DNA molecules whose precise positions along the chromosome are known. However, the precision required for knowing relationships of recombinant-DNA molecules derived from neighboring chromosomal portions increases as the Human Genome Project shifts its emphasis from mapping to sequencing.

High-resolution maps both guide the orderly sequencing of chromosomes and play a critical role in quality control. Only by mapping recombinant-DNA molecules at high resolution can subtle defects in particular molecules be recognized. Such defective human DNA sources, which are not faithful replicas of the human genome, must be weeded out before sequencing can begin. The UW Genome Center has a major program in high-resolution physical mapping which, like the work on sequencing itself, uses advanced computing tools. The center is producing maps of regions targeted for sequencing on a just-in-time basis. These highly detailed maps are proving extremely valuable in facilitating the production of high-quality sequence.

Ultimate Goal

Although many challenges currently posed by the Human Genome Project are highly technical, the ultimate goal is biological. The project will deliver immense amounts of high-quality, continuous DNA sequence into publicly accessible databases. These data will be annotated so that biologists who use them will know the most likely positions of genes and have convenient access to the best available clues about the probable function of these genes. The better the technical solutions to current challenges, the better the center will be able to serve future users of the human genome sequence.

Genome Database

http://www.gdb.org

Genome Database
Johns Hopkins University
2024 E. Monument Street
Baltimore, MD 21205-2236

Stanley Letovsky
Informatics Director

Robert Cottingham
Operations Director

Telephone for both: 410/955-9705
Fax for both: 410/614-0434

David Kingsbury
Director, 1993–97*

In lieu of individual abstracts, research projects and investigators at GDB are represented in this narrative. More information can be found on GDB's Web site (see URL above).

The release of Version 6 of the Genome Database (GDB) in January 1996 signaled a major change for both the scientific community and GDB staff. GDB 6.0 introduced a number of significant improvements over previous versions of GDB, most notably a revised data representation for genes and genomic maps and a new curatorial model for the database. These new features, along with a remodeled database structure and new schema and user interface, provide a resource with the potential to integrate all scientific information currently available on human genomics. GDB rapidly is becoming the international biomedical research community's central source for information about genomic structure, content, diversity, and evolution.

A New Data Model

Inherent in the underlying organization of information in GDB is an improved model for genes, maps, and other classes of data. In particular, genomic segments (any named region of the genome) and maps are being expanded regularly. New segment types have been added to support the integration of mapping and sequencing data (for example, gene elements and repeats) and the construction of comparative maps (syntenic regions). New map types include comparative maps for representing conserved syntenies between species and comprehensive maps that combine data from all the various submitted maps within GDB to provide a single integrated view of the genome. Experimental observations such as order, size, distance, and chimerism are also available.

Through the World Wide Web, GDB links its stored data with many other biological resources on the Internet. GDB's External Link category is a growing collection of cross-references established between GDB entities and related information in other databases. By providing a place for these cross-references, GDB can serve as a central point of inquiry into technical data regarding human genomics.

Direct Community Data Submission and Curation

Two methods for data submission are in use. For individuals submitting small amounts of data, interactive editing of the database through the Web became available in April 1996, and the process has undergone several simplifications since that time. This continues to be an area of development for GDB because all editing must take place at the Baltimore site, and Internet connections from outside North America may be too slow for interactive editing to be practical. Until these difficulties are resolved, GDB encourages scientists with limited connectivity to Baltimore to submit their data via more traditional means (e-mail, fax, mail, phone) or to prepare electronic submissions for entry by the data group on site.

For centers submitting large quantities of data, GDB developed an electronic data submission (EDS) tool, which provides the means to specify login password validation and commands for inserting and updating data in GDB. The EDS syntax includes a mechanism for relating a center's local naming conventions to GDB objects. Data submitted to GDB may be stored privately for up to 6 months before it automatically becomes public. The database is programmed to enforce this Human Genome Project policy. Detailed specifications of GDB's EDS syntax and other submission instructions are available (EDS prototype, *http://www.gdb.org/eds*).

Since the EDS system was implemented, GDB has put forth an aggressive effort to increase the amount of data stored in the database. Consequently, the database has grown tremendously. During 1996 it grew from 1.8 to 6.7 gigabytes.

To provide accountability regarding data quality, the shift to community curation introduced the idea that individuals and

*Now at Chiron Pharmaceuticals, Emeryville, California

laboratories own the data they submit to GDB and that other researchers cannot modify it. However, others should be able to add information and comments, so an additional feature is the community's ability to conduct electronic online public discussions by annotating the database submissions of fellow researchers. GDB is the first database of its kind to offer this feature, and the number of third-party annotations is increasing in the form of editorial commentary, links to literature citations, and links to other databases external to GDB. These links are an important part of the curatorial process because they make other data collections available to GDB users in an appropriate context.

Improved Map Representation and Querying

Accompanying the release of GDB 6.0, the program Mapview creates graphical displays of maps. Mapview was developed at GDB to display a number of map types (cytogenetic, radiation hybrid, contig, and linkage) using common graphical conventions found in the literature. Mapview is designed to stand alone or to be used in conjunction with a Web browser such as Netscape, thereby creating an interactive graphical display system. When used with Netscape, Mapview allows the user to retrieve details about any displayed map object.

Maps are accessed through the query form for genomic segment and its subclasses via a special program that allows the user to select whole maps or slices of maps from specific regions of interest and to query by map type. The ability to browse maps stored in GDB or download them in the background was also incorporated into GDB 6.0.

GDB stores many maps of each chromosome, generated by a variety of mapping methods. Users who are interested in a region, such as the neighborhood of a gene or marker, will be able to see all maps that have data in that region, whether or not they contain the desired marker. To support database querying by region of interest, integrated maps have been developed that combine data from all the maps for each chromosome. These are called *Comprehensive Maps*.

Queries for all loci in a region of interest are processed against the comprehensive maps, thereby searching all relevant maps. Comprehensive maps are also useful for display purposes because they organize the content of a region by class of locus (e.g., gene, marker, clone) rather than by data source. This approach yields a much less complex presentation than an alignment of numerous primary maps. Because such information as detailed orders, order discrepancies between maps, and nonlinear metric relations between maps is not always captured in the comprehensive maps, GDB continues to provide access to aligned displays of primary maps.

A Variety of Searching Strategies

Recognizing the eclectic user community's need to search data and formulate queries, GDB offers a spectrum of simple to complex search strategies. In addition, direct programming access is available using either GDB's object query language to the Object Broker software layer or standard query language to the underlying Sybase relational database.

Querying by Object Directly from GDB's Home Page

The simplest methods search for objects according to known GDB accession numbers; sequence database–accession numbers; specified names, including wildcard symbols that will automatically match synonyms and primary names; and keywords contained anywhere in the text.

Querying by Region of Interest

A region of interest can be specified using a pair of flanking markers, which can be cytogenetic bands, genes, amplimers (sequence tagged sites), or any other mapped objects. Given a region of interest, the comprehensive maps are searched to find all loci that fall within them. These loci can be displayed in a table, graphically as a slice through a comprehensive map, or as slices through a chosen set of primary maps. A comprehensive map slice shows all loci in the region, including genes, expressed sequence tags (ESTs), amplimers, and clones. A region also can be specified as a neighborhood around a single marker of interest.

Results of queries for genes, amplimers, ESTs, or clones can be displayed on a GDB comprehensive map. Results are spread across several chromosomes displayed in Mapview (see figure, p. 52). A query for all the PAX genes (specified as symbol = PAX* on the gene query form) retrieves genes on multiple chromosomes. Double-clicking on one of these genes brings up detailed gene information via the Web browser.

Querying by Polymorphism

GDB contains a large number of polymorphisms associated with genes and other markers. Queries can be constructed for a particular type of marker (e.g., gene, amplimer, clone), polymorphism (i.e., dinucleotide repeat), or level of heterozygosity. These queries can be combined with positional queries to find, for example, polymorphic amplimers in a region bounded by flanking markers or in a particular chromosomal band. If desired, the retrieved markers can be viewed on a comprehensive map.

Work in Progress

Mapview 2.3

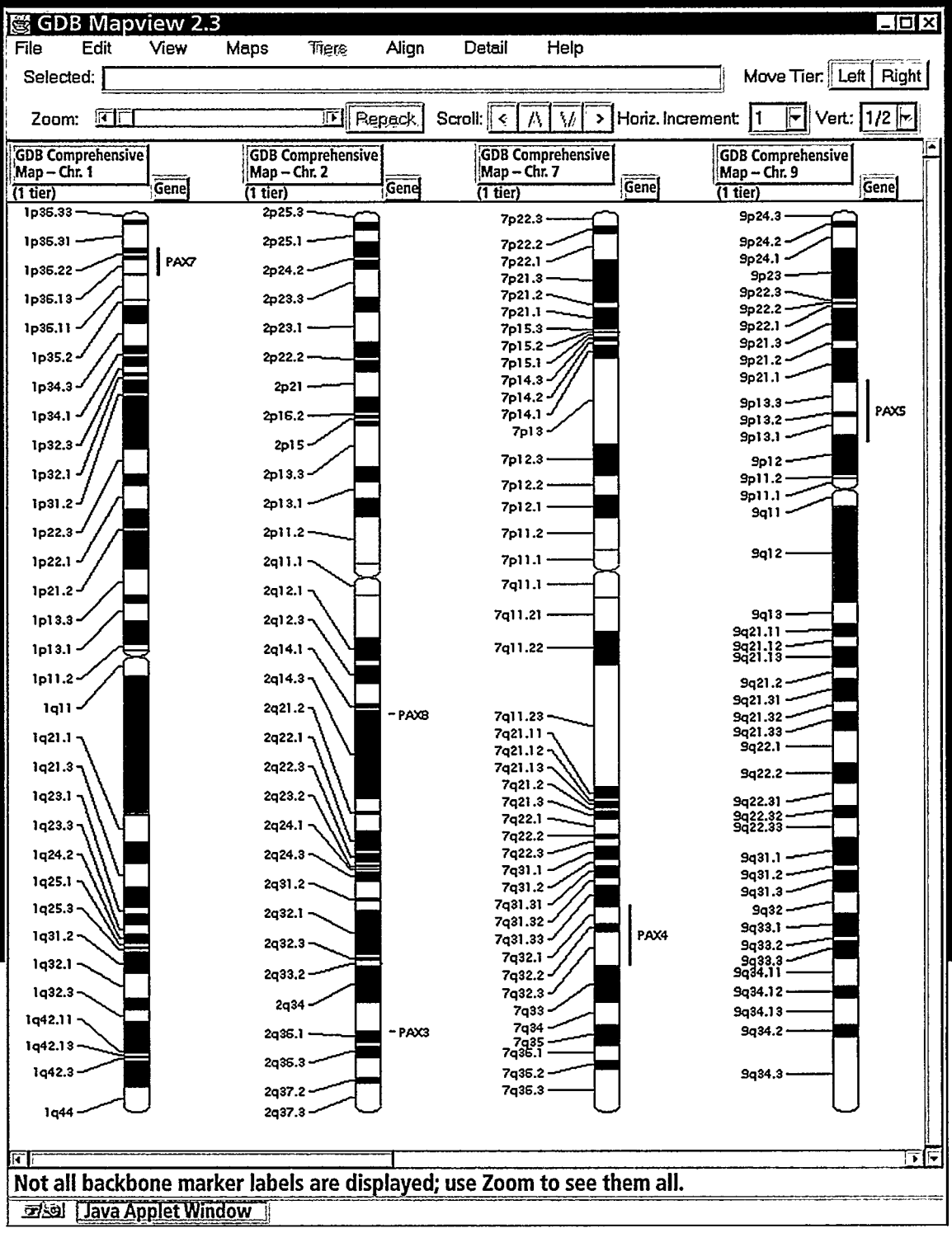
Mapview 2.1, the next generation of the GDB map viewer, was released in March 1997. The latest version, Mapview 2.3, is available in all common computing environments because it is written in the Java programming language. Most important, the new viewer can display multiple aligned maps side by side in the window, with alignment lines indicating common markers in neighboring maps. As before, users can select individual markers to retrieve more information about them from the database.

GDB developers have entered into a collaborative relationship with other members of the bioWidget Consortium so the Java-based alignment viewer will become part of a collection of freely available software tools for displaying biological data (*http://goodman.jax.org/projects/biowidgets/consortium*).

Future plans for Mapview include providing or enhancing the ability to generate manuscript-ready Postscript map images, highlight or modify the display of particular classes of map objects based on attribute values, and requery for additional information.

Variation

Since its inception, GDB has been a repository for polymorphism data, with more than 18,000 polymorphisms now in GDB. A collaboration has been initiated with the Human Gene Mutation Database (HGMD) based in Cardiff, Wales, and headed by David Cooper and Michael Krawczak. HGMD's extensive collection of human mutation data, covering many disease-causing loci, includes sequence-level mutation characterizations. This data set will be included in GDB and updated from HGMD on an ongoing basis. The HGMD team also will provide advice

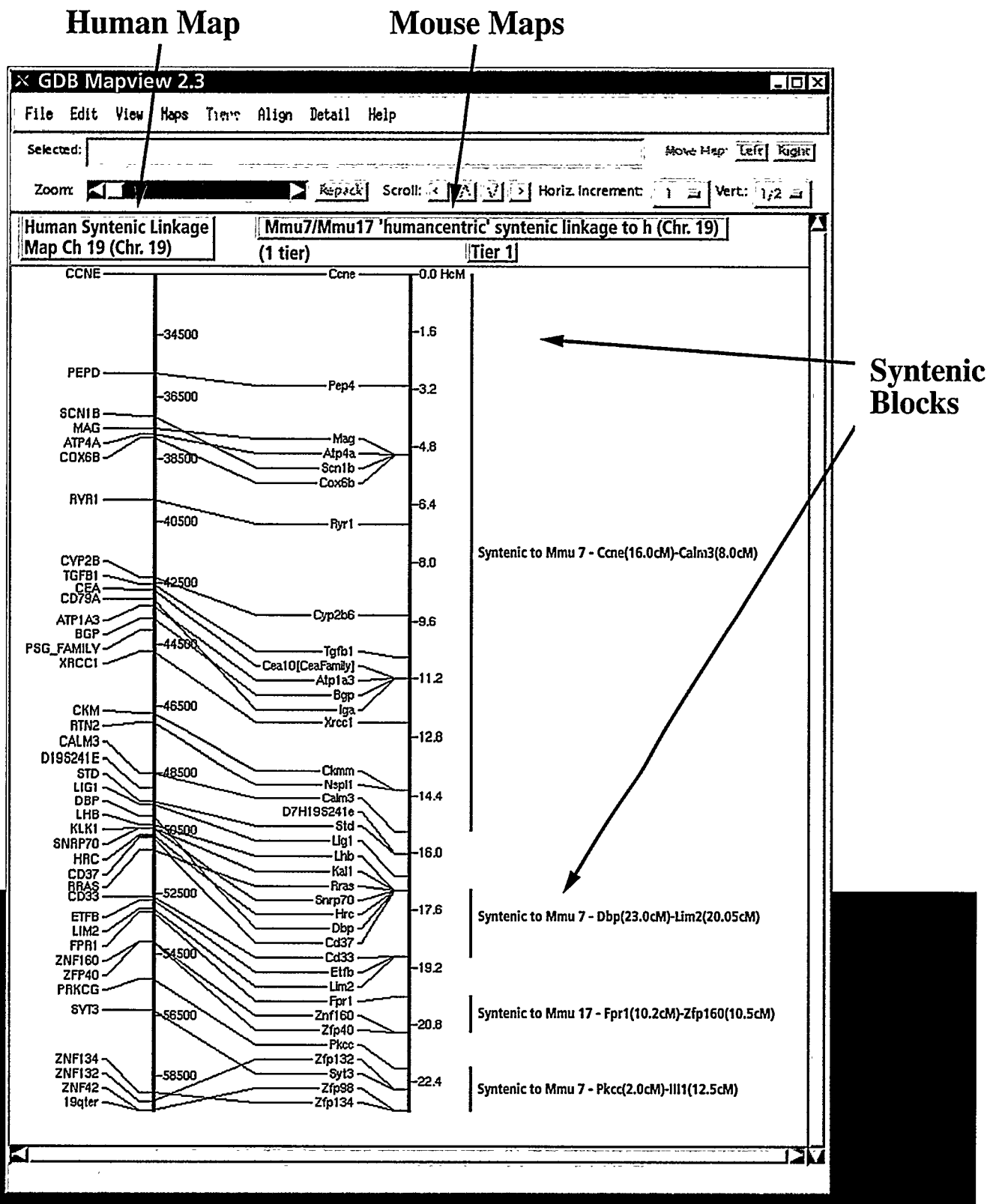


Graphical Display of Results of Query for Genes with Names matching "PAX." [Source: Robert Cottingham, GDB]*

on GDB's representation of genetic variation, which is being enhanced to model mutations and polymorphisms at the sequence level. These modifications will allow GDB to act as a repository for single-nucleotide polymorphisms, which are expected to be a major source of information on human genetic variation in the near future.

Mouse Synteny

Genomic relationships between mouse and man provide important clues regarding gene location, phenotype, and function (see figure, p. 53). One of GDB's goals is to enable direct comparisons between these two organisms, in collaboration with the Mouse Genome Database



Rearranged Mouse Map Aligned Against Human Chromosome. [Source: Robert Cottingham]

at Jackson Laboratory. GDB is making additions to its schema to represent this information so that it can be displayed graphically with Mapview. In addition, algorithmic work is under way to use mapping data to automatically identify regions of conserved synteny between mouse and man. These algorithms will allow the synteny maps to be updated regularly. An important application of comparative mapping is the ability to predict the existence and location of unknown human homologs of known, mapped mouse genes. A set of such predictions is available in a report at the GDB Web site, and similar data will be available in the database itself in the spring of 1998.

Collaborations

GDB is a participant in the Genome Annotation Consortium (GAC) project, whose goal is to produce high-quality, automatic annotation of genomic sequences (*http://compbio.ornl.gov/CoLab*). Currently, GDB is developing a prototype mechanism to transition from GDB's Mapview display to the GAC sequence-level browser over common genome regions. GAC also will establish a human genome reference sequence that will be the base against which GDB will refer all polymorphisms and mutations. Ultimately, every genomic object in GDB should be related to an appropriate region of the reference sequence.

Sequencing Progress

The sequencing status of genomic regions now can be recorded in GDB. Based on submissions to sequence databases, GAC will determine genomic regions that have been completed. GDB also will be collaborating with the European Bioinformatics Institute, in conjunction with the international Human Genome Organisation (HUGO), to maintain a single shared Human Sequence Index that will record commitments and status for sequencing clones or regions. As a result, the sequencing status of any region can be displayed alongside other GDB mapping data.

Outreach

The Genome Database continues to seek direct community feedback and interact with the broader science community via various sources:

- International Scientific Advisory Committee meets annually to offer input and advice.
- Quarterly Review Committee confers frequently with the staff to track GDB progress and suggest change.
- HUGO nomenclature, chromosome, and other editorial committees have specialized functions within GDB, providing official names and consensus maps and ensuring the high quality of GDB's content.

Copies of GDB are available worldwide from ten mirror sites (nodes) that make the data more easily accessible to the international research community. GDB staff meet annually with node managers to facilitate interaction and to benefit from other user perspectives.

Genome Sequence DataBase
1800 Old Pecos Trail, Suite A
Santa Fe, NM 87505

Peter Schad
Vice-President, Bioinformatics and Biotechnology
505/995-4447, Fax: -4432
cnc@ncgr.org

Carol Harger
GSDB Manager
505/982-7840, Fax: -7690
cah@ncgr.org

The National Center for Genome Resources (NCGR) is a not-for-profit organization created to design, develop, support, and deliver resources in support of public and private genome and genetic research. To accomplish these goals, NCGR is developing and publishing the Genome Sequence DataBase (GSDB) and the Genetics and Public Issues (GPI) program.

NCGR is a center to facilitate the flow of information and resources from genome projects into both public and private sectors. A broadly based board of governors provides direction and strategy for the center's development.

NCGR opened in Santa Fe in July 1994, with its initial bioinformatics work being developed through a cooperative 5-year agreement with the Department of Energy funded in July 1995. Committed to serving as a resource for all genomic research, the center works collaboratively with researchers and seeks input from users to ensure that tools and projects under development meet their needs.

Genome Sequence DataBase

GSDB is a relational database that contains nucleotide sequence data (see pie chart) and its associated annotation from all known organisms (*http://www.ncgr.org/gsdb*). All data are freely available to the public. The major goals of GSDB are to provide the support structure for storing sequence data and to furnish useful data-retrieval services.

GSDB adheres to the philosophy that the database is a "community-owned" resource that should be simple to update to reflect new discoveries about sequences. A corollary to this is GSDB's conviction that researchers know their areas of expertise much better than a database curator and, therefore, they should be given ownership and control over the data they submit to the database. The true role of the GSDB staff is to help researchers submit data to and retrieve data from the database.

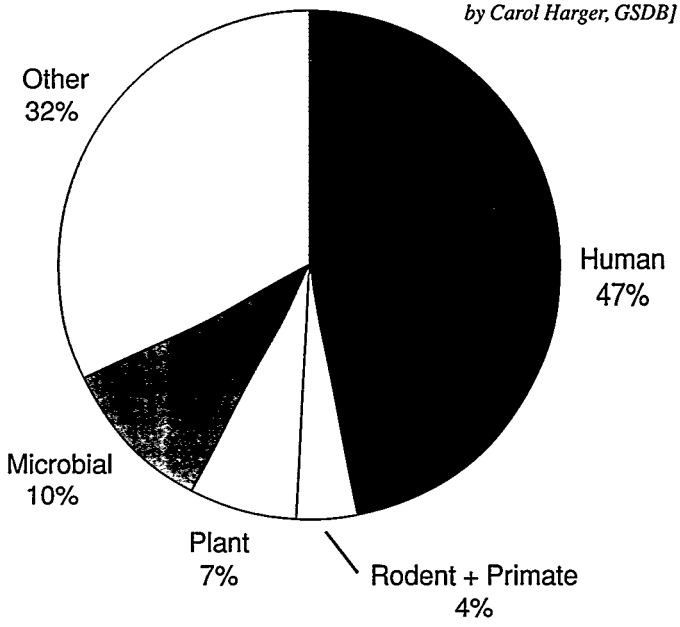
GSDB Enhancements

During 1996, GSDB underwent a major renovation to support new data types and concepts that are important to genomic research. Tables within the database were restructured, and new tables and data fields were added. Some key additions to GSDB include the support of data ownership, sequence alignments, and discontiguous sequences.

The concept of data ownership is a cornerstone to the functioning of the new GSDB. Every piece of data (e.g., sequence or feature) within the database is owned by the submitting researcher, and changes can be made only by the data owner or GSDB staff. This implementation of data ownership provides GSDB with the ability to support community (third-party) annotation—the addition of annotation to a sequence by other community researchers.

In lieu of individual abstracts, research projects and investigators at NCGR are represented in this narrative. More information can be found on the center's Web site (see URL above).

This chart illustrates the taxonomic distribution of the 1,076,481,102 base pairs in the Genome Sequence DataBase. About 47% of the base pairs and 58% of the total database records represent human sequences (August 1997).
[Source: Adapted from chart provided by Carol Harger, GSDB]



A second enhancement of GSDB is the ability to store and represent sequence alignments. GSDB staff has been constructing alignments to several key sequences including the env and pol (reverse transcriptase) genes of the HIV genome, the complete chromosome VIII of Saccharomyces cerevisiae, and the complete genome of *Haemophilus influenzae*. These alignments are useful as possible sites of biological interest and for rapidly identifying differences between sequences.

A third key GSDB enhancement is the ability to represent known relationships of order and distance between separate individual pieces of sequence. These sets of sequences and their relative positions are grouped together as a single discontiguous sequence. Such a sequence may be as simple as two primers that define the ends of a sequence tagged site (STS), it may comprise all exons that are part of a single gene, or it may be as complex as the STS map for an entire chromosome.

GSDB staff has constructed discontiguous sequences for human chromosomes 1 through 22 and X that include markers from Massachusetts Institute of Technology–Whitehead Institute STS maps and from the Stanford Human Genome Center. The set of 2000 STS markers for chromosome X, which were mapped recently by Washington University at St. Louis, also have been added to chromosome X. About 50 genomic sequences have been added to the chromosome 22 map by determining their overlap with STS markers. Genomic sequences are being added to all the chromosomes as their overlap with the STS markers is determined. These discontiguous sequences can be retrieved easily and viewed via their sequence names using the GSDB Annotator. Sequence names follow the format of HUMCHR#MP, where # equals 1 through 22 or X.

GSDB staff also has utilized discontiguous sequences to construct maps for maize and rice. The maize discontiguous sequences were constructed using markers from the University of Missouri, Columbia. Markers for the rice discontiguous sequence were obtained from the Rice Genome Database at Cornell University and the Rice Genome Research Project in Japan.

New Tools

As a result of the major GSDB renovation, new tools were needed for submitting and accessing database data. Annotator was developed as a graphical interface that can be used to view, update, and submit sequence data *(http://www.ncgr.org/gsdb/beta.html)*. Maestro, a Web-based interface, was developed to assist researchers in data retrieval (http://www.ncgr.org/gsdb/maestrobeta.html). Although both these tools currently are available to researchers, GSDB is continuing development to add increased capabilities.

Annotator displays a sequence and its associated biological information as an image, with the scale of the image adjustable by the user. Additional information about the sequence or an associate biological feature can be obtained in a pop-up window. Annotator also allows a user to retrieve a sequence for review, edit existing data, or add annotation to the record. Sequences can be created using Annotator, and any sequences created or edited can be saved either to a local file for later review and further editing or saved directly to the database.

Correct database structures are important for storing data and providing the research community with tools for searching and retrieving data. GSDB is making a concerted effort to expand and improve these services. The first generation of the Maestro query tool is available from the GSDB Web pages. Maestro allows researchers to perform queries on 18 different fields, some of which are queryable only through GSDB, for example, D segment numbers from the Genome Database at Johns Hopkins University in Baltimore.

Additionally, Maestro allows queries with mixed Boolean operators for a more refined search. For example, a user may wish to compare relatively long mouse and human sequences that do not contain identified coding regions. To obtain all sequences meeting these criteria, the scientific name field would be searched first for "Mus musculus" and then for "Homo sapiens" using the Boolean term "OR." Then the sequence-length filter could be used to refine the search to sequences longer than 10,000 base pairs. To exclude sequences containing identified coding-region features, the "BUT NOT" term can be used with the Feature query field set equal to "coding region."

With Maestro, users can view the list of search matches a few at a time and retrieve more of the list as needed. From the list, users can select one or several sequences according to their short descriptions and review or download the sequence information in GIO, FASTA, or GSDB flatfile format.

Future Plans

Although most pieces necessary for operation are now in place, GSDB is still improving functionality and adding enhancements. During the next year GSDB, in collaboration with other researchers, anticipates creating more discontiguous sequence maps for several model organisms, adding more functionality to and providing a Web-based submission tool and tool kit for creating GIO files.

Microbial Genome Web Pages

NCGR also maintains informational Web pages on microbial genomes. These pages, created as a community reference, contain a list of current or completed eubacterial, Archaeal, and eukaryotic genome sequencing projects. Each main page includes the name of the organism being sequenced, sequencing groups involved, background information on the organism, and its current location on the Carl Woese Tree of Life. As the Microbial Genome Project progresses, the pages will be updated as appropriate.

Genetics and Public Issues Program

GPI serves as a crucial resource for people seeking information and making decisions about genetics or genomics (http://www.ncgr.org/gpi). GPI develops and provides information that explains the ethical, legal, policy, and social relevance of genetic discoveries and applications.

To achieve its mission, GPI has set forth three goals: (1) preparation and development of resources, including careful delineation of ethical, legal, policy, and social issues in genetics and genomics; (2) dissemination of genetic information targeted to the public, legal and health professionals, policymakers, and decision makers; and (3) creation of an information network to facilitate interaction among groups.

GPI delivers information through four primary vehicles: online resources, conferences, publications, and educational programs. The GPI program maintains a continually evolving World Wide Web site containing a range of material freely accessible over the Internet.

Los Alamos National Laboratory researcher David Bruce uses an automated system for gridding chromosome library clones in preparation of very dense filter arrays for hybridization experiments. [Source: Lynn Clark, LANL]

http://www.er.doe.gov/production/ober/hug_top.html

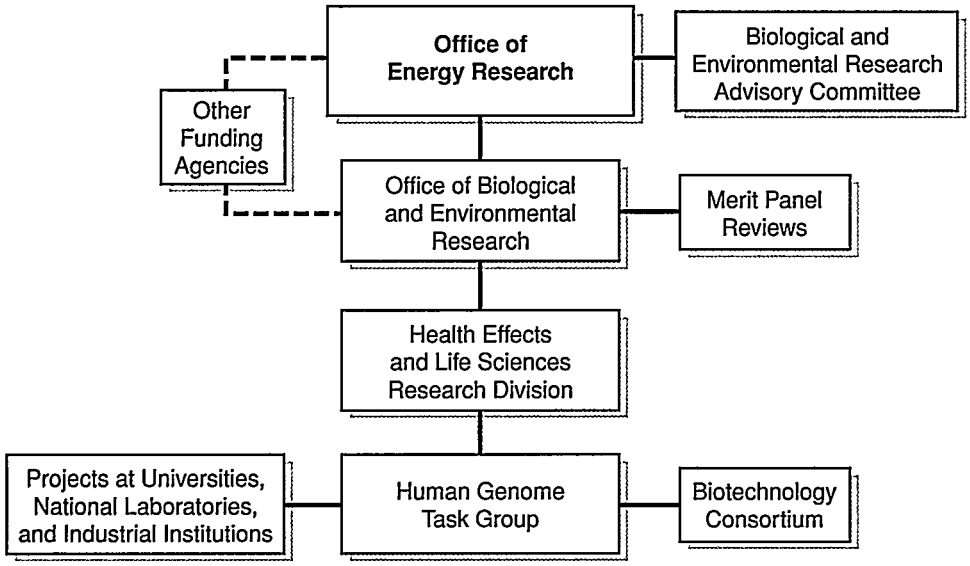
The Human Genome Program was conceived in 1986 as an initiative within the DOE Office of Health and Environmental Research, which has been renamed Office of Biological and Environmental Research (OBER) (see chart below). The program is administered primarily through the OBER Health Effects and Life Sciences Research Division (HELSRD), both directed by David A. Smith until his retirement in January 1996. Marvin Frazier is now Director of HELSRD, and OBER is led by Associate Director Aristides Patrinos, who also serves as Human Genome Program manager. Previous directors and managers are listed in the table below. OBER is within the Office of Energy Research, directed by Martha Krebs.

DOE OBER Mission

Based on mandates from Congress, DOE OBER's principal missions are to (1) develop the knowledge necessary to identify, understand, and anticipate long-term health and environmental consequences of energy use and development and (2) employ DOE's unique scientific and technological capabilities in solving major scientific problems in medicine, biology, and the environment.

Genome integrity and radiation biology have been a long-term concern of OBER at DOE and its predecessors—the Atomic Energy Commission (AEC) and the Energy Research and Development Administration (ERDA). In the United States, the first federal support

See Appendix A, p. 73, for information on Human Genome Project history, including enabling legislation.



OHER Associate or Acting Directors		Human Genome Program Managers	
Charles De Lisi	1985	Benjamin J. Barnhart	1988
Robert W. Wood	1987	David A. Smith	1991
David J. Galas	1990	Aristides Patrinos	1996
Aristides Patrinos	1993		

Institutions Conducting DOE-Sponsored Genome Research	
DOE national laboratories	7
Academic institutions	28
Private-sector institutions	10
Companies, including Small Business Innovation Research	11
Foreign institutions (Russia, Canada, Israel)	7

DOE Human Genome Task Group

Member	Specialty
Chair: Aristides Patrinos	Physical sciences
Benjamin J. Barnhart	Genetics, Radiation biology
Elbert Branscomb	Scientific Director, Joint Genome Institute
Daniel W. Drell	Biology, ELSI, Informatics, Microbial genome
Ludwig Feinendegen	Medicine, Radiation biology
Marvin Frazier	Molecular and cellular biology
Gerald Goldstein†	Physical science, Instrumentation
D. Jay Grimes†	Microbiology
Roland Hirsch	Structural biology, Instrumentation
Arthur Katz*	Physical sciences
Anna Palmisano*†	Microbiology, Microbial genome
Michael Riches	Physical sciences
Jay Snoddy†	Molecular biology, Informatics
Marvin Stodolsky	Molecular biology, Biophysics
David G. Thomassen	Cell and molecular biology
John C. Wooley	Computational biology

*Joined, 1997.
†Left OBER, 1997.

Biotechnology Consortium

Chair: Aristides Patrinos	DOE Office of Biological and Environmental Research
Charles Arntzen*	Cornell University
Elbert Branscomb	Lawrence Livermore National Laboratory
Charles Cantor	Boston University
Anthony Carrano	Lawrence Livermore National Laboratory
Thomas Caskey	Merck Research Laboratories
David Eisenberg	University of California, Los Angeles
Chris Fields†	National Center for Genome Resources
David Galas	Darwin Molecular, Inc.
Raymond Gesteland	University of Utah
Keith Hodgson	Stanford University
Leroy Hood	University of Washington, Seattle
David Kingsbury†	Chiron Pharmaceuticals
Robert Moyzis†	University of California, Irvine
Mohandas Narla*	Lawrence Berkeley National Laboratory
Michael Palazzolo	Amgen, Inc.
Melvin Simon*	California Institute of Technology
Hamilton Smith*	Johns Hopkins University School of Medicine
Lloyd Smith	University of Wisconsin, Madison
Lisa Stubbs	Lawrence Livermore National Laboratory
Edward Uberbacher*	Oak Ridge National Laboratory
Marc Van Montagu*	Ghent University, Belgium
Executive Officer: Sylvia Spengler	Lawrence Berkeley National Laboratory

*Appointed after October 1996.
†Resigned, 1997.
Note: All members of the DOE Human Genome Task Group are ex-officio members of the Biotechnology Consortium.

for genetic research was through AEC. In the early days of nuclear energy development, the focus was on radiation effects and broadened later under ERDA and DOE to include health implications of all energy technologies and their by-products.

Today, extensive OBER-sponsored research programs on genomic structure, maintenance, damage, and repair continue at the national laboratories and universities. These and other OBER efforts support a DOE shift toward a preventive approach to health, environment, and safety concerns. World-class scientists in top facilities working on leading-edge problems spawn the knowledge to revolutionize the technology, drive the future, and add value to the U.S. economy. Major OBER research includes characterization of DNA repair genes and improvement of methodologies and resources for quantifying and characterizing genetic polymorphisms and their relationship to genetic susceptibilities.

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

- Sponsors peer-reviewed research and development projects at universities, in the private sector, and at DOE national laboratories (see box, p. 59).
- Considers novel, beneficial initiatives with input from the scientific community and governmental sectors.
- Provides expertise to various governmental working groups.
- Supports the capabilities of multidisciplinary DOE national laboratories and their unique user facilities for the nation's benefit (p. 61).

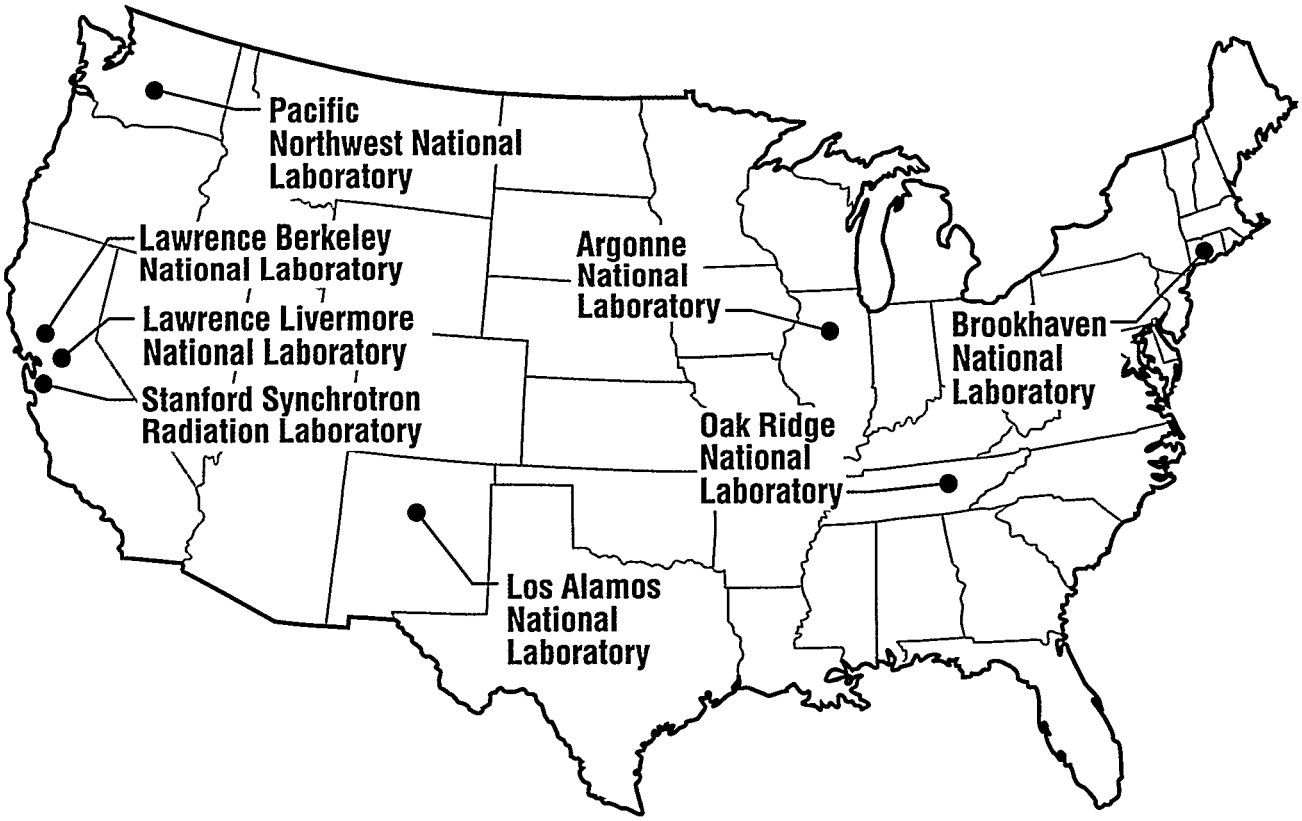
Human Genome Program resources and technologies are focused on sequencing the human genome and related informatics and supportive infrastructure (see chart and tables, p. 62). The genomes of selected microorganisms are analyzed under the separate Microbial Genome Program.

Major DOE User Facilities and Resources Relevant to Molecular Biology Research

Although the genome program is contributing fundamental information about the structure of chromosomes and genes, other types of knowledge are required to understand how genes and their products function. Three-dimensional protein structure studies are still essential because structure cannot be predicted fully from its encoded DNA sequence.

To enhance these and other studies, DOE builds and maintains structural biology user facilities that enable scientists to gain an understanding of relationships between biological structures and their functions, study disease processes, develop new pharmaceuticals, and conduct basic research in molecular biology and environmental processes. These resources are used heavily by both academic and private-sector scientists.

Other important resources available to the research community include the clone libraries developed in the National Laboratory Gene Library Project and distributed worldwide, the GRAIL Online Sequence Interpretation Service, and the Mouse Genetics Research Facility.



Argonne National Laboratory
- Advanced Photon Source

Brookhaven National Laboratory
- High-Flux Beam Reactor
- National Synchrotron Light Source
- Protein Structure Data Bank
- Scanning Transmission Electron Microscope

Lawrence Berkeley National Laboratory
- Advanced Light Source
- Center for X-Ray Optics
- National Energy Research Scientific Computing Center

Lawrence Livermore National Laboratory
- National Laboratory Gene Library Project

Los Alamos National Laboratory
- National Flow-Cytometry Resource
- National Laboratory Gene Library Project
- Neutron-Scattering Center

Oak Ridge National Laboratory
- GRAIL, Online Sequence Interpretation Service
- Mouse Genetics Research Facility

Pacific Northwest National Laboratory
- Environmental Molecular Sciences Laboratory

Stanford University
- Synchrotron Radiation Laboratory

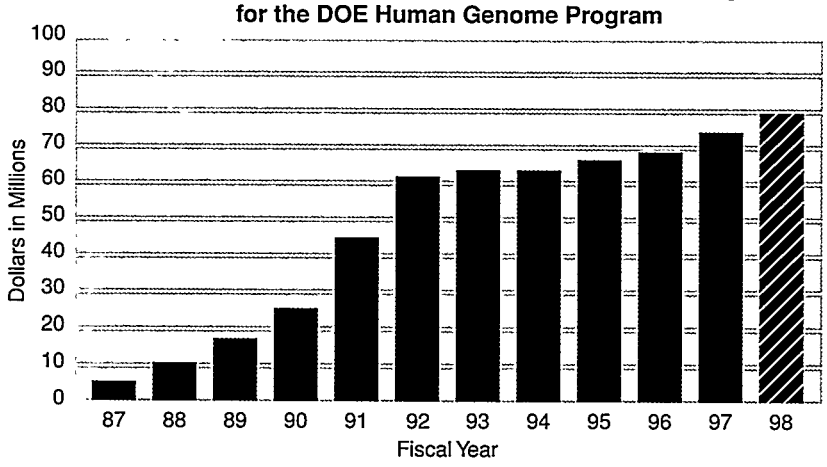
Human Genome Program

Coordination and Resources

Program coordination is the responsibility of the Human Genome Task Group (see box, p. 60), which, beginning in 1997, includes Elbert Branscomb, the Joint Genome Institute's Scientific Director. The task group is aided by the Biotechnology Consortium (which succeeded the former Human Genome Coordination Committee; see box, p. 60) to foster information exchange and dissemination. The task group administers the DOE Human Genome Program and its evolving needs and reports to the Associate Director for Biological and Environmental Research (currently Aristides Patrinos). The task group arranges periodic workshops and coordinates site reviews for genome centers, the Joint Genome Institute, databases, and other large projects. It also coordinates peer review of research proposals, administration of awards, and collaboration with all concerned agencies and organizations.

The Biotechnology Consortium provides the OBER Associate Director with external expertise in all aspects of genomics and informatics and a mechanism by which OBER can keep track of the latest developments in the field. It facilitates development and dissemination of novel genome technologies throughout the DOE system, ensures appropriate management and sharing of data and resources by all DOE contractors and grantees, and promotes interactions with other national and international genomic entities.

Operating Expenditures and FY 1998 Projected Budget for the DOE Human Genome Program



Human Genome Program Fiscal Year Expenditures ($M)

Year	Operating	Capital Equipment	Construction	Total
1996	68.3	5.6	5.7	79.6
1997	73.9	6.0	1.0	80.9
1998*	79.9	5.2	0.0	85.1

*Projected expenses.

Human Genome Program Operating Funds Distribution in FY 1996 ($K)

FY 1996	Mapping	Sequencing	Sequencing Technology	Informatics	ELSI	Administration	Totals	%
DOE Laboratories	8,980	11,015	11,128	6,840	313	2,783 *	41,059	60.1
Academic	6,671	4,368	3,257	6,178	642	4	21,120	30.9
Nonprofit	563	0	467	2,783	1,311	38	5,162	7.5
Federal	0	0	0	0	0	1,000 **	1,000	1.5
Total	16,214	15,383	14,852	15,801	2,266	3,825	68,341	
% of Total	23.8	22.5	21.7	23.1	3.3	5.6	100	

*Includes DOE laboratories' nonresearch costs but not U.S. government administration or SBIR.
**DOE contribution to the international Human Frontiers Neurosciences Program.

Communication

The DOE Human Genome Program communicates information in a variety of ways. These communication systems include the Human Genome Management Information System (HGMIS), projects in the Ethical, Legal, and Social Issues (ELSI) Program, electronic resources, meetings, and fellowships. Some of these mechanisms are described below. For more details, see Research Highlights, ELSI projects, p. 18.

HGMIS

HGMIS provides technical communication and information services for the DOE OBER Human Genome Program Task Group. HGMIS is charged with (1) helping to communicate genome-related matters and research to contractors, grantees, other (nongenome project) researchers, and other multipliers of information pertaining to genetic research; (2) serving as a clearinghouse for inquiries about the U.S. genome project; and (3) reducing research duplication by providing a forum for interdisciplinary information exchange (including resources developed) among genetic investigators worldwide.

HGMIS publishes the newsletter *Human Genome News*, sponsored by OBER. Over 14,000 *HGN* subscribers include genome and basic researchers at national laboratories, universities, and other research institutions; professors and teachers; industry representatives; legal personnel; ethicists; students; genetic counselors; physicians; science writers; and other interested individuals.

HGMIS also produces the DOE *Primer on Molecular Genetics*; a compilation of ELSI abstracts; and reports on the DOE Human Genome and Microbial Genome Programs, contractor-grantee workshops, and other related subjects.

Electronic versions of the primer and other HGMIS publications are available via the World Wide Web. HGMIS also initiates and maintains other related Web sites (see DOE Electronic Genome Resources section below and DOE Web Sites at right).

In addition to their print and online publishing efforts, HGMIS staff members answer questions generated via Web sites, telephone, fax, and e-mail. They also furnish customized information about the genome project for multipliers of information (contact: Betty Mansfield at 423/576-6669, Fax: /574-9888, *mansfieldbk@ornl.gov*).

DOE Electronic Genome Resources

Web Sites. The DOE Human Genome Program Home Page displays pointers to other programs within OBER and the Office of Energy Research. Links are made to additional biological and environmental information and to HGMIS, Genome Database, and other sites.

HGMIS initiates and maintains the searchable Human Genome Project Information Web site. This site contains more than 1700 text files of information for multidisciplinary technical audiences as well as for lay persons interested in learning about the science, goals, progress, and history of the project. Users include almost all levels of students; education, medical, and legal professionals; genetic society and support group members; biotechnology and pharmaceutical industry personnel; administrators; policymakers; and the press.

The site also houses a section of frequently asked questions, a quick fact finder, *Primer on Molecular Genetics*, all issues of *Human Genome News*, DOE Human Genome Program and contractor-grantee workshop reports, *To Know Ourselves*, historical documents, research abstracts, calendars of genome events, and hundreds of links to genome research and educational sites. More than 1000 other Web pages link to this site, resulting in more than 100,000 text file transfers each month. This

DOE Web Sites

DOE Human Genome Program
http://www.er.doe.gov/production/ober/hug_top.html

OBER
http://www.er.doe.gov/production/ober/ober_top.html

Office of Energy Research
http://www.er.doe.gov

Human Genome Project Information
http://www.ornl.gov/hgmis

HGP and Related Meetings
http://www.ornl.gov/meetings

Courts
http://www.ornl.gov/courts

The DOE Human Genome Program and Human Genome Project Information Web sites offer both general and scientific audiences thousands of text files and links for comprehensive coverage of all aspects of genome research worldwide. See text (pp. 63 and 65) for further details.

HGMIS site has received a Four-Star designation from the Magellan Group and the Editor's Choice Award from LookSmart.

Genome-project and related meetings are listed at a Web site (see box, p. 63), through which users can register and submit research abstracts. Another listed related site discusses issues at the critical intersection of genetics and the court system. This Web page is part of a project to educate and prepare the judiciary for the coming onslaught of cases involving genetic issues and data.

Newsgroup. The Human Genome Program Newsgroup operates through the BIOSCI electronic bulletin board network to allow researchers worldwide to communicate, share ideas, and find solutions to problems. Genome-related information is distributed through the newsgroup, including requests for grant applications, reports from recent scientific and advisory meetings, announcements of future events, and listings of free software and services (*gnome-pr@net.bio.net* or *http://www.bio.net*).

Postdoctoral Fellowships

OBER established the Human Genome Distinguished Postdoctoral Research Program in 1990 to support research on projects related to the DOE Human Genome Program. Beginning in FY 1996, the Human Genome Distinguished Postdoctoral Fellowships were merged with the Alexander Hollaender Distinguished Postdoctoral Fellowships, which provide support in all areas of OBER-sponsored research. Postdoctoral programs are administered by the Oak Ridge Institute for Science and Education, a university consortium and DOE contractor. For additional information, contact Linda Holmes (423/576-3192, *holmesl@orau.gov*) or see the Web site *(http://www.orau.gov/ober/hollaend.htm)*.

Human Genome Distinguished Postdoctoral Fellows

Names of past and current fellows in genome topics are given below with their research institutions and titles of proposed research. For 1996 research abstracts, refer to Index of Principal and Coinvestigators on p. 71 in Part 2 of this report.

1994 **Mark Graves** (Baylor College of Medicine): Graph Data Models for Genome Mapping

William Hawe (Duke University): Synthesis of Peptide Nucleic Acids for DNA Sequencing by Hybridization

Jingyue Ju (University of California, Berkeley): Design, Synthesis, and Use of Oligonucleotide Primers Labeled with Energy Transfer–Coupled Dyes

Mark Shannon (Oak Ridge National Laboratory): Comparative Study of a Conserved Zinc Finger Gene Region

1995 **Evan Eichler** (Lawrence Livermore National Laboratory): Identification, Organization, and Characterization of Zinc Finger Genes in a 2-Mb Cluster on 19p12

Kelly Ann Frazer (Lawrence Berkeley National Laboratory): In Vivo Complementation of the Murine Mutations Grizzled, Mocha, and Jitteri

Soo-in Hwang (Lawrence Berkeley National Laboratory): Positional Cloning of Oncogenes on 20q13.2

James Labrenz (University of Washington, Seattle): Error Analysis of Principal Sequencing Data and Its Role in Process Optimization for Genome-Scale Sequencing Projects

Marie Ruiz-Martinez (Northeastern University): Multiplex Purification Schemes for DNA Sequencing–Reaction Products: Application to Gel-Filled Capillary Electrophoresis

Todd Smith (University of Washington, Seattle): Managing the Flow of Large-Scale DNA Sequence Information

Alexander Hollaender Distinguished Postdoctoral Fellows in Genome Research

1996 **Cymbeline Culiat** (Oak Ridge National Laboratory): Cloning of a Mouse Gene Causing Severe Deafness and Balance Defects

Tau-Mu Yi (Laboratory of Structural Biology and Molecular Medicine, Los Angeles): Structure-Function Analysis of Alpha-Factor Receptor

1997 **Jeffrey Koshi** (Los Alamos National Laboratory): Construction, Analysis, and Use of Optimal DNA Mutation Matrices

Sandra McCutchen-Maloney (Lawrence Livermore National Laboratory): Structure and Function of a Damage-Specific Endonuclease Complex

The laser-based flow cytometer developed at DOE national laboratories enables researchers to separate human chromosomes for analysis. [Source: Los Alamos National Laboratory]

Enhancing genome research capabilities

The U.S. Human Genome Project is supported jointly by the Department of Energy (DOE) and the National Institutes of Health (NIH), each of which emphasizes different facets. The two agencies coordinate their efforts through development of common project goals and joint support of some programs addressing ethical, legal, and social issues (ELSI) arising from new genome tools, technology, and data.

Extraordinary advances in genome research are due to contributions by many investigators in this country and abroad. In the United States, such research (including nonhuman) also is funded by other federal agencies and private foundations and groups. Many countries are major contributors to the project through international collaborations and their own focused programs. Coordinating and facilitating these diverse research efforts around the world is the aim of the nongovernmental international Human Genome Organisation.

Some details of U.S. and worldwide coordination are provided below.

U.S. Human Genome Project: DOE and NIH

In 1988 DOE and NIH developed a Memorandum of Understanding that formalized the coordination of their efforts to decipher the human genome and thus "enhance the human genome research capabilities of both agencies." In early 1990 they presented Congress with a joint plan, *Understanding Our Genetic Inheritance, The U.S. Human Genome Project: The First Five Years (1991–1995)*. Referred to as the Five-Year Plan, it contained short-term scientific goals for the coordinated, multiyear research project and a comprehensive spending plan. Unexpectedly rapid progress in mapping prompted early revision of the original 5-year goals in the fall of 1993 [*Science* **262,** 43–46 (October 1, 1993)]. Current goals, which run through September 30, 1998, are listed on page 5; text of both 5-year plans is accessible via the Web (*http://www.ornl.gov/hgmis/project/hgp.html*).

DOE and NIH have adopted a joint policy to promote sharing of genome data and resources for facilitating progress and reducing duplicated work. (See Appendix B: DOE-NIH Sharing Guidelines, p. 75.)

ELSI Considerations

NIH and DOE devote at least 3% of their respective genome program budgets to identifying, analyzing, and addressing the ELSI considerations surrounding genome technology and the data it produces. The DOE ELSI component focuses on research into the privacy and confidentiality of personal genetic information, genetics relevant to the workplace, commercialization (including patenting) of genome research data, and genetic education for the general public and targeted communities. The NIH ELSI component supports studies on a range of ethical issues surrounding the conduct of genetic research and responsible clinical integration of new genetic technologies, especially in testing for mutations associated with cystic fibrosis and heritable breast, ovarian, and colon cancers.

In 1990, the DOE-NIH Joint ELSI Working Group was established to identify, address, and develop policy options; stimulate bioethics research; promote education of professional and lay groups; and collaborate with such international groups as the Human Genome Organisation (HUGO); United Nations Educational, Scientific, and Cultural Organization; and the European Community. Research funded by the U.S. Human Genome Project through the joint working group has produced policy recommendations in various areas. In May 1993, for

example, the DOE-NIH Joint ELSI Working Group Task Force on Genetic Information and Insurance issued a report with recommendations for managing the impact of advances in human genetics on the current system of healthcare coverage. In 1996, the working group released guidelines for investigators on using DNA from human subjects for large-scale sequencing projects. The guidance emphasizes numerous ways to preserve donor anonymity [see Appendix C, p. 77, and the World Wide Web (*http://www.ornl.gov/hgmis/archive/nchgrdoe.html*)].

In 1997, following an evaluation, the two agencies modified the ELSI working group into the ELSI Research and Program Evaluation Group (ERPEG). ERPEG will focus more specifically on research activities supported by DOE and NIH ELSI programs.

Other U.S. Programs

The potential impact of genome research on society and the rapid growth of the biotechnology industry have spurred the initiation of other genome research projects in this country and worldwide. These projects aim to create maps of the human genome and the genomes of model organisms and several economically important microbes, plants, and animals.

- The DOE Microbial Genome Program, begun in 1994, is producing complete genome sequence data on industrially important microorganisms, including those that live under extreme environmental conditions. The sequences of several microbial genomes have been completed. *[http://www.er.doe.gov/production/ober/EPR/mig_top.html]*
- In 1990, the National Science Foundation, DOE, and the U.S. Department of Agriculture (USDA) initiated a project to map and sequence the genome of the model plant *Arabidopsis thaliana*. The goal of this project is to enhance fundamental understanding of plant processes. In 1996, the three agencies began funding systematic, large-scale genomic sequencing of the 120-megabase *Arabidopsis* genome, with the goal of completing it by 2004, with DOE support through the Office of Basic Energy Sciences. *[http://pgec-genome.pw.usda.gov/agi.html]*
- USDA also funds animal genome research projects designed to obtain genome maps for economically important species (e.g., corn, soybeans, poultry, cattle, swine, and sheep) to enable genetic modifications that will increase resistance to diseases and pests, improve nutrient value, and increase productivity.
- The Advanced Technology Program (ATP) of the U.S. National Institute of Standards and Technology promotes industry-government partnerships in DNA sequencing and biotechnology through the Tools for DNA Diagnostics component. DOE staff participates in the ATP review process (see box, p. 22). *[http://www.atp.nist.gov]*
- In 1997 the NIH National Cancer Institute established the Cancer Genome Anatomy Project (CGAP) to develop new diagnostic tools for understanding molecular changes that underlie all cancers (*http://www.ncbi.nlm.nih.gov/ncicgap*). DOE researchers are generating clone libraries to support this effort.

International Collaborations

The current DOE-NIH Five-Year Plan commends the "spirit of international cooperation and sharing" that has characterized the Human Genome Project and played a major role in its success. Cooperation includes collaborations among laboratories in the United States

and abroad as well as extensive sharing of materials and information among genome researchers around the world. The DOE Human Genome Program supports many international collaborations as well as grantees in several foreign institutions.

Collaborations involving the DOE human genome centers include mapping chromosomes 16 and 19, developing resources, and constructing the human gene map from shared cDNA libraries. These libraries were generated by the Integrated Molecular Analysis of Gene Expression (called IMAGE) Consortium initiated by groups at Lawrence Livermore National Laboratory, Columbia University, NIH National Institute of Mental Health, and Généthon (France).

Investigators from almost every major sequencing center in the world met in Bermuda in February 1996 and again in 1997 to discuss issues related to large-scale sequencing. These meetings were designed to help researchers coordinate, compare, and evaluate human genome mapping and sequencing strategies; consider new sequencing and informatics technologies; and discuss release of data.

Human Genome Organisation

Founded by scientists in 1989, HUGO is a nongovernmental international organization providing coordination functions for worldwide genome efforts. HUGO activities range from support of data collation for constructing genome maps to organizing workshops. HUGO also fosters exchange of data and biomaterials, encourages technology sharing, and serves as a coordinating agency for building relationships among various government funding agencies and the genome community.

HUGO offers short-term (2- to 10-week) travel awards up to $1500 for investigators under age 40 to visit another country to learn new methods or techniques and to facilitate collaborative research between the laboratories.

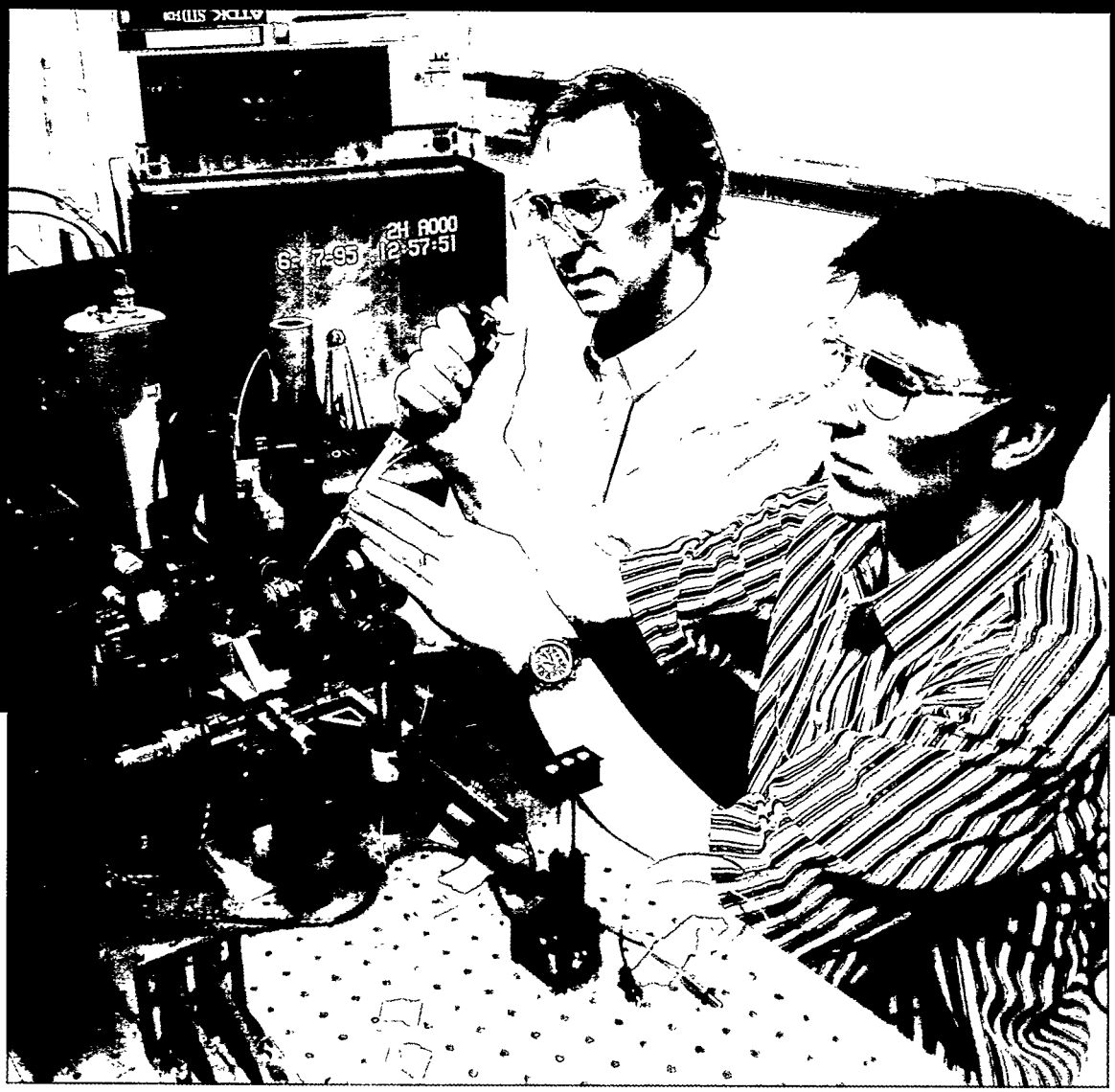
HUGO has worked closely with international funding agencies to sponsor single-chromosome workshops (SCWs) and other genome meetings. Due to the success of these workshops as well as the shift in emphasis from mapping to sequencing, DOE and NIH began to phase out their funding for international SCWs in FY 1996 but encouraged applications for individual SCWs as needed. In 1996, HUGO partially funded an international strategy meeting in Bermuda on large-scale sequencing. Principles regarding data release and a resources list developed at the meeting are available on the HUGO Web site (*http://hugo.gdb.org/hugo.html*).

Membership in HUGO (over 1000 people in more than 50 countries) is extended to persons concerned with human genome research and related scientific subjects. Its current president is Grant R. Sutherland (Adelaide Women and Children's Hospital, Australia). Directed by an 18-member international council, HUGO is supported by grants from the Howard Hughes Medical Institute and The Wellcome Trust.

Countries with Genome Programs

Countries with genome programs or strong programs in human genetics include Australia, Brazil, Canada, China, Denmark, European Union, France, Germany, Israel, Italy, Japan, Korea, Mexico, Netherlands, Russia, Sweden, United Kingdom, and United States.

Los Alamos National Laboratory researchers Peter Goodwin and Rhett Affleck load a sample of fluorescently labeled DNA into an ultrasensitive flow cytometer used to detect single cleaved nucleotides. [Source: Lynn Clark, LANL]



Appendix A
DOE Human Genome Program: Early History, Enabling Legislation

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 DOE Office of Biological and Environmental Research (OBER) 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. (See Appendix E for a discussion of the DOE Biological and Environmental Research Program.)

In 1984, OBER [then named Office of Health and Environmental Research (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. 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 (called 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

Enabling Legislation

In the United States, the first federal support for genetics research was through the Atomic Energy Commission. In the early days of nuclear energy development, the focus was on radiation effects and later broadened under the Energy Research and Development Administration (ERDA) and the Department of Energy to include the health implications of all energy technologies and their by-products. Major enabling legislation follows.

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.

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

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

Federal Non-Nuclear Energy Research and Development Act of 1974 (P.L. 93-577): Authorized ERDA to conduct a comprehensive non-nuclear energy research, development, and demonstration program to include the environmental and social consequences of the various technologies.

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 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).

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 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 and 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 1988, DOE and NIH signed a Memorandum of Understanding in which the agencies agreed to work together, coordinate technical research and activities, and share results. In 1990, DOE and NIH submitted a joint research plan outlining short- and long-term goals of the project.

Appendix B
DOE-NIH Guidelines for Sharing Data and Resources

*At its December 7, 1992, meeting, the DOE-NIH Joint Subcommittee on the Human Genome approved the following sharing guidelines, developed from the DOE draft of September 1991.**

The information and resources generated by the Human Genome Project have become substantial, and the interest in having access to them is widespread. It is therefore desirable to have a statement of philosophy concerning the sharing of these resources that can guide investigators who generate the resources as well as those who wish to use them.

A key issue for the Human Genome Project is how to promote and encourage the rapid sharing of materials and data that are produced, especially information that has not yet been published or may never be published in its entirety. Such sharing is essential for progress toward the goals of the program and to avoid unnecessary duplication. It is also desirable to make the fruits of genome research available to the scientific community as a whole as soon as possible to expedite research in other areas.

Although it is the policy of the Human Genome Project to maximize outreach to the scientific community, it is also necessary to give investigators time to verify the accuracy of their data and to gain some scientific advantage from the effort they have invested. Furthermore, in order to assure that novel ideas and inventions are rapidly developed to the benefit of the public, intellectual property protection may be needed for some of the data and materials.

After extensive discussion with the community of genome researchers, the advisors of the NIH and DOE genome programs have determined that consensus is developing around the concept that a 6-month period from the time the data or materials are generated to the time they are made available publicly is a reasonable maximum in almost all cases. More rapid sharing is encouraged.

Whenever possible, data should be deposited in public databases and materials in public repositories. Where appropriate repositories do not exist or are unable to accept the data or materials, investigators should accommodate requests to the extent possible.

The NIH and DOE genome programs have decided to require all applicants expecting to generate significant amounts of genome data or materials to describe in their application how and when they plan to make such data and materials available to the community. Grant solicitations will specify this requirement. These plans in each application will be reviewed in the course of peer review and by staff to assure they are reasonable and in conformity with program philosophy. If a grant is made, the applicant's sharing plans will become a condition of the award and compliance will be reviewed before continuation funding is provided. Progress reports will be asked to address the issue.

---

*Reprinted from *Human Genome News* 4(5), 4 (1993).

Appendix C
NIH-DOE Guidance on Human Subjects Issues in Large-Scale DNA Sequencing

Date issued: August 9, 1996

Introduction

The Human Genome Project (HGP) is now entering into large-scale DNA sequencing. To meet its complete sequencing goal, it will be necessary to recruit volunteers willing to contribute their DNA for this purpose. The guidance provided in this document is intended to address ethical issues that must be considered in designing strategies for recruitment and protection of DNA donors for large-scale sequencing.

Nothing in this document should be construed to differ from, or substitute for, the policies described in the Federal Regulations for the Protection of Human Subjects [45CFR46 (NIH) and 10CFR745 (DOE)]. Rather, it is intended to supplement those policies by focusing on the particular issues raised by large-scale human DNA sequencing. This statement addresses six topics: (1) benefits and risks of genomic DNA sequencing; (2) privacy and confidentiality; (3) recruitment of DNA donors as sources for library construction; (4) informed consent; (5) IRB approval; and (6) use of existing libraries.

The guidance provided in this statement is intended to afford maximum protection to DNA donors and is based on the belief that protection can best be achieved by a combination of approaches including:

- ensuring that the initial version of the complete human DNA sequence is derived from multiple donors;
- providing donors with the opportunity to make an informed decision about whether to contribute their DNA to this project; and
- taking effective steps to ensure the privacy and confidentiality of donors.

1. Benefits and Risks of Genomic DNA Sequencing

The HGP offers great promise for the improvement of human health. As a consequence of the HGP, there will be a more thorough understanding of the genetic bases of human biology and of many diseases. This, in turn, will lead to better therapies and, perhaps more importantly, prevention strategies for many of those diseases. Similarly, as the technology developed by the HGP is applied to understanding the biology of other organisms, many other human activities will be affected including agriculture, environmental management, and biologically based industrial processes.

While the HGP offers great promise to humanity, there will be no direct benefit, in either clinical or financial terms, to any of the individuals who choose to donate DNA for large-scale sequencing. Rather, the motivation for donation is likely to be an altruistic willingness to contribute to this historic research effort.

However, individuals who donate DNA to this effort may face certain risks. Information derived from the donors will become available in public databases. Such information may reveal, for example, DNA sequence-based information about disease susceptibility. If the donor becomes aware of such information, it could lead to emotional distress on her/his part. If such health-related information becomes known to others, discrimination against the donor (e.g., in insurance or in employment) could result. Unwanted notoriety is another potential risk to donors. Therefore, those engaged in large-scale sequencing must be sensitive to the unique features of this type of research and ensure that both the protections normally afforded research subjects and the special issues associated with human genomic DNA sequencing are thoroughly addressed.

While some risks to donors can already be identified, the probability of adverse events materializing appears to be low. However, the risks of harm to individuals will increase if confidentiality is not maintained and/or the number of donors is limited to a very few individuals. Either, or both, of these situations would increase the possibility of a donor's identity being revealed without his/her knowledge or permission.

A final issue to consider is characterized in a statement taken from the OPRR Guidebook[1] which points out that "some areas [of genetic research] present issues for which no clear guidance can be given at this point, either because enough is not known about the risks presented by the research, or because no consensus on the appropriate resolution of the problem exists." It is anticipated that the DNA sequence information produced by the Human Genome Project will be used in the future for types of research which cannot now be predicted and the risks of which cannot be assessed or disclosed.

2. Privacy and Confidentiality

In general, one of the most effective ways of protecting volunteers from the unexpected, unwelcome or unauthorized use of information about them is to ensure that there are no opportunities for linking an individual donor with information about him/her that is revealed by the research. By not collecting information about the identity of a research subject and any biological material or records developed in the course of the research, or by subsequently removing all

identifiers ("anonymizing" the sample), the possibility of risk to the subject stemming from the results of the research is greatly reduced. Large-scale DNA sequence determination represents an exception because each person's DNA sequence is unique and, ultimately, there is enough information in any individual's DNA sequence to absolutely identify her/him. However, the technology that would allow the unambiguous identification of an individual from his/her DNA sequence is not yet mature. Thus, for the foreseeable future, establishing effective confidentiality, rather than relying on anonymity, will be a very useful approach to protecting donors.

Investigators should introduce as many disconnects between the identity of donors and the publicly available information and materials as possible. There should not be any way for anyone to establish that a specific DNA sequence came from a particular individual, other than resampling an individual's DNA and comparing it to the sequence information in the public database. In particular, no phenotypic or demographic information about donors should be linked to the DNA to be sequenced.[2] For the purposes of the HGP such information will rarely be useful, and recording such information could result in possible misuse and compromise donor confidentiality.

Confidentiality should be "two way." Not only should others be unable to link a DNA sequence to a particular individual, but no individual who donates DNA should be able to confirm directly that a particular DNA sequence was obtained from their DNA sample.[3] This degree of confidentiality will preclude the possibility of re-contacting DNA donors, providing another degree of protection for them. It should be clear to both investigators and to donors that the contact involved in obtaining the initial specimen will be the only contact.[4]

Another approach for protecting all DNA donors is to reduce the incentive for wanting to know the identities of particular donors. If the initial human sequence is a "mosaic" or "patchwork" of sequenced regions derived from a number of different individuals, rather than that of a single individual, there would be considerably less interest in who the specific donors were. Although there may be scientific justification that each clone library used for sequencing should be derived from one person, there is no scientific reason that the entire initial human DNA sequence should be that of a single individual. As approximately 99.9% of the human DNA sequence is common between any two individuals, most of the fundamental biological information contained in the human DNA sequence is common to all people.

To increase the likelihood that the first human DNA sequence will be an amalgam of regions sequenced from different sources, a number of clone libraries must be made available. Although a number of large insert libraries have been made, most do not meet all of the standards set in this document; therefore, these libraries should be used as substrates for large-scale sequencing only under circumscribed conditions (see section 6, p. 79). Starting immediately, new libraries will be developed that have the advantage of being constructed in accordance with the ethical principles discussed in this document; they may also confer some additional scientific benefit. Such libraries are critical for the long-range needs of the HGP.

3. Source/Recruitment of DNA Donors for Library Construction

Another implication of the fact that 99.9% of the human DNA sequence is shared by any two individuals is that the backgrounds of the individuals who donate DNA for the first human sequence will make no scientific difference in terms of the usefulness and applicability of the information that results from sequencing the human genome. At the same time, there will undoubtedly be some sensitivity about the choice of DNA sources. There are no scientific reasons why DNA donors should not be selected from diverse pools of potential donors.[5]

There are two additional issues that have arisen in considering donor selection. These warrant particular discussion:

- It is recognized that women have historically been underrepresented in research, so it can be anticipated that concerns might arise if males (sperm DNA) were used exclusively as the source of DNA for large-scale sequencing. Although there would be no scientific basis for concern, because even in the case of a male source, half of the donor's DNA would have come from his mother and half from his father, nevertheless perceptions are not to be dismissed. While the choice of donors will not be dictated to investigators, it is expected that, because multiple libraries will be produced, a number of them will be made from female sources while others will be made from male sources.
- Staff of laboratories involved in library construction and DNA sequencing may be eager to volunteer to be donors because of their interest and belief in the HGP. However, proximity to the research may create some special vulnerabilities for laboratory staff members. It is also possible that they will feel pressure to donate and there may be an increased likelihood that confidentiality would be breached. Finally, there is a potential that the choice of persons so closely involved in the research may be interpreted as elitist. For all of these reasons, it is recommended that donors should not be recruited from laboratory staff, including the principal investigator.

4. Informed Consent

Obtaining informed consent specifically for the purpose of donating DNA for large-scale sequencing raises some unique concerns. Because anonymity cannot be guaranteed and confidentiality protections are not absolute, the disclosure process to potential donors must clearly specify what the process of DNA donation involves, what may make it different from other types of research, and what the implications are of one's DNA sequence information being a public scientific resource.

Federal regulations (45CFR46 and 10CFR745) require the disclosure of a number of issues in any informed consent document. They include such issues as potential benefits of the research, potential risks to the donor, control and ownership of donated material, long-term retention of donated material for future use, and the procedures that will be followed. In addition, there are several other disclosures that are of special importance for donors of DNA for large-scale sequencing. These include:

- the meaning of confidentiality and privacy of information in the context of large-scale DNA sequencing, and how these issues will be addressed;
- the lack of opportunity for the donor to later withdraw the libraries made from his/her DNA or his/her DNA sequence information from public use;
- the absence of opportunity for information of clinical relevance to be provided to the donor or her/his family;
- the possibility of unforeseen risks; and
- the possible extension of risk to family members of the donor or to any group or community of interest (e.g., gender, race, ethnicity) to which a donor might belong.

Many academic human genetics units have considerable experience in dealing with research subjects and obtaining informed consent, while the laboratories that are likely to be involved in making the libraries for sequencing have, in general, much less experience of this type. Therefore, library makers are encouraged to establish a collaboration with one or more human genetics units, with the latter being responsible for recruiting donors, obtaining informed consent, obtaining the necessary biological samples, and providing a blinded sample to the library maker. Collaboration with tissue banks may be considered as long as these banks are collecting tissues in accordance with this guidance. The library maker should have no contact with the donor and no opportunity to obtain any information about the donor's identity.

5. IRB Approval

Effective immediately, projects to construct libraries for large-scale DNA sequencing must obtain Institutional Review Board (IRB) approval before work is initiated. IRBs should carefully consider the unique aspects of large-scale sequencing projects. Some of the informed consent provisions outlined may be somewhat at odds with the usual and customary disclosures found in most protocols involving human subjects and which IRBs usually consider. For example, research subjects usually are given the opportunity to withdraw from a research project if they change their minds about participating. In the case of donors for large-scale sequencing, it will not be possible to withdraw either the libraries made from their DNA or the DNA sequence information obtained using those libraries once the information is in the public domain. By the time a significant amount of DNA sequence data has been collected, the libraries, as well as individual clones from them, will have been widely distributed and the sequence information will have been deposited in and distributed from public databases. In addition, there will be no possibility of returning information of clinical relevance to the donor or his/her family.

6. Use of Existing Libraries for Large-Scale Sequencing

Many of the existing libraries (including those derived from anonymous donors) were not made in complete conformity with the principles elaborated above. The potential risks that may result from their use will be minimized by the rapid introduction of several new libraries constructed in accordance with this guidance, which NCHGR and DOE are taking steps to initiate. This will ensure that the existing libraries will only contribute small amounts to the first complete human DNA sequence. In the interim, existing libraries can continue to be used for large-scale sequencing, only if IRB approval and consent for "continued use" are obtained[6] and approval by the funding agency is granted.

It is important that in obtaining consent for contined use of existing libraries, no coercion of the DNA donor occur. It is therefore recommended that consideration be given to whether it is appropriate for the individual who previously recruited the donor to recontact him/her to obtain this consent. In some cases an IRB may determine that the recontact should be made by a third party to assure that the donors are fully informed and allowed to choose freely whether their DNA can continue to be used for this purpose.

Conclusion

This document is intended to provide guidance to investigators and IRBs who are involved in large-scale sequencing efforts. It is designed to alert them to special ethical concerns that may arise in such projects. In particular, it provides guidance for the use of existing and the construction of new DNA libraries. Adhering to this guidance will ensure that the initial version of the complete human sequence is derived from multiple, diverse donors; that donors will have the opportunity to make an informed decision about whether to contribute their DNA to this project; and that effective steps will be taken by investigators to ensure the privacy and confidentiality of donors.

Investigators funded by NCHGR and DOE to develop new libraries for large-scale human DNA sequencing will be required to have their plans for the recruitment of DNA donors, including the informed consent documents, reviewed and approved by the funding agency before donors are recruited. Investigators involved in large-scale human sequencing will also be asked to observe those aspects of this guidance that pertain to them.

Approved August 17, 1996, by:

Francis S. Collins, M.D., Ph.D., Director, National Center for Human Genome Research, National Institutes of Health

Aristides N. Patrinos, Ph.D., Associate Director, Office of Health and Environmental Research, U.S. Department of Energy

Footnotes

1. Office of Protection from Research Risks, Protecting Human Research Subjects: Institutional Review Board Guidebook (OPRR: U.S. Government Printing Office, 1993).

2. It is recognized that it will be trivially easy to determine the sex of the donor of the library, by assaying for the presence or absence of Y chromosome in the library.

3. There are a number of approaches to preventing a DNA donor from knowing that his/her DNA was actually sequenced as part of the HGP. For example, each time a clone library is to be made, an appropriately diverse pool of between five and ten volunteers can be chosen in such a way that none of them knows the identity of anyone else in the pool. Samples for DNA preparation and for preparation of a cell line can be collected from all of the volunteers (who have been told that their specimen may or may not eventually be used for DNA sequencing) and one of those samples is randomly and blindly selected as the source actually used for library construction. In this way, not only will the identity of the individual whose DNA is chosen not be known to the investigators, but that individual will also not be sure that s/he is the actual source.

4. Although recontacting donors should not be possible, investigators will potentially want to be able to resample a donor's genome. Thus, at the time the initial specimen is obtained, in addition to making a clone library representing the donor's genome, it should also be used to prepare an additional aliquot of high molecular weight DNA for storage and a permanent cell line. Either resource could then be used as a source of the donor's genome in case additional DNA were needed or comparison with the results of the analysis of the cloned DNA were desired.

5. There has been discussion in the scientific community about the sex of DNA donors. A library prepared from a female donor will contain DNA from the X chromosome in an amount equivalent to the autosomes, but will completely lack Y chromosomal DNA. Conversely, a library prepared from a male donor will contain Y DNA, but both X and Y DNA will only be present at half the frequency of the DNA from the other chromosomes. Scientifically, then, there are both advantages and disadvantages inherent in the use of either a male or a female donor. The question of the sex of the donor also involves the question of the use of somatic or germ line DNA to make libraries. For making libraries, useful amounts of germ line DNA can only be obtained from a male source (i.e., from sperm); it is not possible to obtain enough ova from a female donor to isolate germ line DNA for this purpose. Opinion is divided in the scientific community about whether germ line or somatic DNA should be used for large-scale sequencing. Somatic DNA is known to be rearranged, relative to germ line DNA, in certain regions (e.g., the immunoglobulin genes) and the possibility has been raised that other developmentally based rearrangements may occur, although no example of the latter has been offered. While some believe that the sequence product should not contain any rearrangements of this sort, others consider this potential advantage of germ line DNA to be relatively minor in comparison to the need to have the X chromosome fully represented in sequencing efforts and prefer the use of somatic DNA.

6. Individuals whose DNA was used for library construction (with the exception of those created from deceased or anonymous individuals) should be fully informed about the risks and benefits described above, should freely choose whether they would like their DNA to continue to be used for this purpose, and their decision should be documented.

Executive Summary of Joint NIH-DOE Human Subjects Guidance

1. Those engaged in large-scale sequencing must be sensitive to the unique features of this type of research and ensure that both the protections normally afforded research subjects and the special issues associated with human genomic DNA sequencing are thoroughly addressed.

2. For the foreseeable future, establishing effective confidentiality, rather than relying on anonymity, will be a very useful approach to protecting donors.

3. Investigators should introduce as many disconnects between the identity of donors and the publicly available information and materials as possible.

4. No phenotypic or demographic information about donors should be linked to the DNA to be sequenced.

5. There are no scientific reasons why DNA donors should not be selected from diverse pools of potential donors.

6. While the choice of donors will not be dictated to investigators, it is expected that, because multiple libraries will be produced, a number of them will be made from female sources while others will be made from male sources.

7. It is recommended that donors should not be recruited from laboratory staff, including the principal investigator.

8. The disclosure process to potential donors must clearly specify what the process of DNA donation involves, what may make it different from other types of research, and what the implications are of one's DNA sequence information being a public scientific resource.

9. Library makers are encouraged to establish a collaboration with one or more human genetics units [or tissue banks].

10. The library maker should have no contact with the donor and no opportunity to obtain any information about the donor's identity.

11. Effective immediately, projects to construct libraries for large-scale DNA sequencing must obtain Institutional Review Board (IRB) approval before work is initiated.

12. Existing libraries can continue to be used for large-scale sequencing, only if IRB approval and consent for continued use are obtained and approval by the funding agency is granted.

13. It is important that in obtaining informed consent for continued use of existing libraries, no coercion of the DNA donor occur.

Appendix D
Human Genome Project and Genetics on the World Wide Web

August 1997

The World Wide Web offers the easiest path to information about the Human Genome Project and related genetics topics. Some useful sites to visit are included in the list below.

Human Genome Project

DOE Human Genome Program
http://www.er.doe.gov/production/ober/hug_top.html

Devoted to the DOE component of the U.S. Human Genome Project and to the DOE Microbial Genome Program. Links to many other sites.

Human Genome Project Information
http://www.ornl.gov/hgmis

Comprehensive site covering topics related to the U.S. and worldwide Human Genome Projects. Useful for updating scientists and providing educational material for nonscientists, in support of DOE's commitment to public education. Developed and maintained for DOE by the Human Genome Management Information System (HGMIS) at Oak Ridge National Laboratory.

NIH National Human Genome Research Institute
http://www.nhgri.nih.gov

Site of the NIH sector of the U.S. Human Genome Project.

DOE Human Genome Program Publications

***Human Genome News**
http://www.ornl.gov/hgmis/publicat/publications.html

Quarterly newsletter reporting on the worldwide Human Genome Project.

Biological Sciences Curriculum Study (BSCS) Teaching Modules

Online versions in preparation; hardcopies available from 719/531-5550

- "Genes, Environment, and Human Behavior," tentative title, in preparation
- "Mapping and Sequencing the Human Genome: Science, Ethics, and Public Policy" (1992)
- "The Human Genome Project: Biology, Computers, and Privacy" (1996)
- "The Puzzle of Inheritance: Genetics and the Methods of Science" (1997)

***Primer on Molecular Genetics, 1992**
http://www.ornl.gov/hgmis/publicat/publications.html#primer

Explains the science behind the genome research.

***To Know Ourselves, 1996**
http://www.ornl.gov/hgmis/tko

Booklet reviewing DOE's role, history, and achievements in the Human Genome Project and introducing the science and other aspects of the project.

Ethical, Legal, and Social Issues Related to Genetics Research

HGMIS Gateways Web page
http://www.ornl.gov/hgmis/links.html

Choose "Ethical, Legal, and Social Issues."

Center for Bioethics, University of Pennsylvania
http://www.med.upenn.edu/~bioethic

Full-text articles about such ethical issues as human cloning; includes a primer on bioethics.

Courts and Science On-Line Magazine (CASOLM)
http://www.ornl.gov/courts

Coverage of genetic issues affecting the courts.

ELSI in Science
http://www.lbl.gov/Education/ELSI/ELSI.html

Teaching modules designed to stimulate discussion on implications of scientific research.

Eubios Ethics Institute
http://www.biol.tsukuba.ac.jp/~macer/index.html

Site includes newsletter summarizing literature in bioethics and biotechnology.

Genetic Privacy Act
http://www.ornl.gov/hgmis/resource/elsi.html

Model legislation written with support of the DOE Human Genome Program.

MCET—The Human Genome Project
http://phoenix.mcet.edu/humangenome/index.html

ELSI issues for high school students.

*Print copy available from HGMIS (see p. 87 or inside front cover for contact information).

National Bioethics Advisory Committee
http://www.nih.gov/nbac/nbac.htm

The bioethics committee offers advice to the National Science and Technology Council and others on bioethical issues arising from research related to human biology and behavior.

National Center for Genomic Resources
http://www.ncgr.org

Comprehensive Genetics and Public Issues page; includes congressional bills related to genetic privacy.

The Gene Letter
http://www.geneletter.org/genetalk.html

Bimonthly newsletter to inform consumers and professionals about advances in genetics and encourage discussion about emerging policy dilemmas.

Your Genes, Your Choices
http://www.nextwave.org/ehr/books/index.html

Booklet written in simple English, describing the Human Genome Project; the science behind it; and how ethical, legal, and social issues raised by the project may affect people's everyday lives.

General Genetics and Biotechnology

Many of the following sites contain links to both educational and technical material.

HGMIS Community Education and Outreach Gateways Web Page
http://www.ornl.gov/hgmis/links.html

Access Excellence
http://outcast.gene.com/ae/index.html

Extensive genetic and biotechnology resources for teachers and nonscientists.

BIO Online (Biotechnology Industry Organization)
http://www.bio.com

Comprehensive directory of biotechnology sites on the Internet.

Biospace
http://www.biospace.com

Biotech industry site; profiles biotech companies by region.

BioTech
http://biotech.chem.indiana.edu

An interactive educational resource and biotech reference tool; includes a dictionary of 6000 life science terms.

Biotechnology Information Center, USDA National Agricultural Library
http://www.nal.usda.gov/bic

Comprehensive agricultural biotechnology resource; includes a bibliography on patenting biotechnology products and processes (*http://www.nal.usda.gov/bic/Biblios/patentag.htm*).

Bugs 'N Stuff
http://www.ncgr.org/microbe

List of microbial genomes being sequenced, research groups, genome sizes, and facts about selected organisms. Links to related sites.

Careers in Genetics
http://www.faseb.org/genetics/gsa/careers/bro-menu.htm

Online booklet from the Genetics Society of America, including several profiles of geneticists. See also career sections of sites specified above, such as Access Excellence.

Carolina Biological Supply Company
http://www.carosci.com/Tips.htm

Teaching materials for all levels. Includes mini-lessons on selected scientific topics, two online magazines, What's New, software, catalogs, and publications.

Cell & Molecular Biology Online
http://www.tiac.net/users/pmgannon

Links to electronic publications, current research, educational and career resources, and more.

CERN Virtual Library, Genetics section, Biosciences Division
http://www.ornl.gov/TechResources/Human_Genome/genetics.html

Includes an organism index linking to other pertinent databases, information on the U.S. and international Human Genome Projects, and links to research sites.

Classic Papers in Genetics
http://www.esp.org

Covers the early years, with introductory notes. See also Access Excellence site above for genetics history.

Community of Science Web Server
http://cos.gdb.org/best.html

Links to Medline, U.S. Patent Citation Database, Commerce Business Daily, The Federal Register, and other resources.

Database of Genome Sizes
http://www.cbs.dtu.dk/databases/DOGS/index.html

Lists numerous organisms with genome sizes, scientific and common names, classifications, and references.

Genetic and biological resources links
http://www.er.doe.gov/production/ober/bioinfo_center.html

Genetics Education Center, University of Kansas Medical Center
http://www.kumc.edu/instruction/medicine/genetics/homepage.html

Educational information on human genetics, career resources.

Genetics Glossary
http://www.ornl.gov/hgmis/publicat/glossary.html

Glossary of terms related to genetics.

Genetics Webliography
http://www.dml.georgetown.edu/%7Edavidsol/len.html

Extensive links for researchers and nonscientists from Georgetown University Library.

Genomics: A Global Resource
http://www.phrma.org/genomics/index.html

Many links. Website a joint project of the Pharmaceutical Research and Manufacturers of America and the American Institute of Biological Sciences; includes Genomics Today, a daily update on the latest news in the field.

Hispanic Educational Genome Project
http://vflylab.calstatela.edu/hgp

Designed to educate high school students and their families about genetics and the Human Genome Project. Links to other projects.

Howard Hughes Medical Institute
http://www.hhmi.org

Home page of major U.S. philanthropic organization that supports research in genetics, cell biology, immunology, structural biology, and neuroscience. Excellent introductory information on these topics.

Library of Congress
http://lcweb.loc.gov/homepage/lchp.html

Microbial Database
http://www.tigr.org/tdb/mdb/mdb.html

Lists completed and in-progress microbial genomes, with funding sources.

MIT Biology Hypertextbook
http://esg-www.mit.edu:8001/esgbio/7001main.html

All the basics.

Science and Mathematics Resources
http://www-sci.lib.uci.edu

More than 2000 Web references, including Frank Potter's Science Gems and Martindale's Health Science Guide. For teachers at all levels.

Virtual Courses on the Web
http://lenti.med.umn.edu/~mwd/courses.html

Links to Web tutorials in biology, genetics, and more.

Welch Web
http://www.welch.jhu.edu

Links to many Internet biomedical resources, dictionaries, encyclopedias, government sites, libraries, and more, from the Johns Hopkins University Welch Library.

Why Files
http://whyfiles.news.wisc.edu

Illustrated explanations of the science behind the news.

Images on the Web

Biochemistry Online
http://biochem.arach-net.com

Essays, courses, 3-D images of biomolecules, modeling, software.

Bugs in the News!
http://falcon.cc.ukans.edu/~jbrown/bugs.html

Microbiology information and a nice collection of images of biological molecules.

Cells Alive!
http://www.cellsalive.com

Images (some moving) of different types of cells.

Cn3D (See in 3-D)
http://www3.ncbi.nlm.nih.gov/Entrez/Structure/cn3d.html

3-D molecular structure viewer allowing the user to visualize and rotate structure data entries from Entrez. Highly technical, for researchers.

Cytogenetics Gallery
http://www.pathology.washington.edu:80/Cytogallery

Photos (karyotypes) of normal and abnormal chromosomes.

DNA Learning Center, Cold Spring Harbor Laboratory
http://darwin.cshl.org/index.html

Animated images of PCR and Southern Blotting techniques.

Gene Map from the 1996 Genome Issue of Science
http://www.ncbi.nlm.nih.gov/SCIENCE96

Click on particular areas of chromosomes and find genes.

Images of Biological Molecules
http://www.cc.ukans.edu/~micro/picts.html

3-D structures of proteins and nucleic acids obtained from Brookhaven National Laboratory Protein Database and others.

Lawrence Livermore National Laboratory Chromosome 19 Physical Map
http://www-bio.llnl.gov/bbrp/genome/genome.html

Los Alamos National Laboratory Chromosome 16 Physical Map
http://www-ls.lanl.gov/DBqueries/QueryPage.html

Journals and Magazines

HGMIS Journals Gateways Web page
http://www.ornl.gov/hgmis/links.html

Choose "Journals, Books, Periodicals."

Biochemistry and Molecular Biology Journals
http://biochem.arach-net.com/beasley/journals.html

Comprehensive list.

Nature, Nature Genetics, and Nature Biotechnology
http://www.nature.com

Abstracts of articles, full text of letters and editorials.

Science Magazine
http://www.sciencemag.org

Abstracts and some full-text articles.

Science Magazine Genome Issue (10/96)
http://www.sciencemag.org/science/content/vol274/issue5287

Full text includes a "clickable" gene map.

Science News
http://www.sciencenews.org

Online version of weekly popular science magazine with full text of selected articles.

Medical Genetics

Blazing a Genetic Trail
http://www.hhmi.org/GeneticTrail

Illustrated booklet from the Howard Hughes Medical Institute on hunting for disease genes.

Directory of National Genetic Voluntary Organizations and Related Resources
http://medhlp.netusa.net/agsg/agsgsup.htm

Support groups for people with genetic diseases and their families.

GeneCards
http://bioinformatics.weizmann.ac.il/cards

A database of more than 6000 genes; describes their functions, products, and biomedical applications.

Gene Therapy
http://www.mc.vanderbilt.edu/gcrc/gene/index.html

Web course covering the basics, with links to other sites.

Inherited-Disease Genes Found by Positional Cloning
http://www.ncbi.nlm.nih.gov/Baxevani/CLONE/index.html

Links to OMIM.

NIH Office of Recombinant DNA Activities
http://www.nih.gov/od/orda

Includes a database of human gene therapy protocols.

Online Mendelian Inheritance in Man (OMIM)
http://www.ncbi.nlm.nih.gov/Omim

A comprehensive, authoritative, and up-to-date human gene and genetic disorder catalog that supports medical genetics and the Human Genome Project.

Promoting Safe and Effective Genetic Testing in the United States (1997)
http://www.med.jhu.edu/tfgtelsi

Principles and recommendations by a joint NIH-DOE Human Genome Project group that examined the development and provision of gene tests in the United States.

Understanding Gene Testing
http://www.gene.com/ae/AE/AEPC/NIH/index.html

Illustrated brochure from the National Cancer Institute.

Science in the News

EurekAlert! *http://www.eurekalert.org*

InScight: *http://www.apnet.com/inscight*

SciWeb: *http://www.sciweb.com/news.html*

Short summaries of major stories, some with links to related articles in other sources.

HMS Beagle
http://biomednet.com/hmsbeagle

Biweekly electronic journal featuring major science stories, profiles, book reviews, and other items of interest.

Science Daily
http://www.sciencedaily.com

Headline stories, articles, and links to news services, newspapers, magazines, broadcast sources, journals, and organizations. Also offers weekly bulletins for updates by e-mail.

Science Guide
http://www.scienceguide.com

Daily news and information service and free science news e-mailer. Also contains directories of newsgroups, grant and funding resources, employment, and online journals.

ScienceNow
http://www.sciencenow.org

Daily online news service from Science magazine offers articles on major science news.

Web Search Tools

Biosciences Index to WWW Virtual Library
http://golgi.harvard.edu/htbin/biopages

Metacrawler
http://www.metacrawler.com

"Search the Net"
http://metro.turnpike.net/adorn/search.html

Comprehensive list of search tools, libraries, world fact books, and other useful information.

Search.com
http://www.search.com

Yahoo!
http://www.yahoo.com

Prepared August 1997 by
Human Genome Management Information System
Oak Ridge National Laboratory
1060 Commerce Park, MS 6480
Oak Ridge, TN 37830
423/576-6669, *caseydk@ornl.gov*
http://www.ornl.gov/hgmis

Appendix E
1996 Human Genome Research Projects

Research abstracts of these projects appear in Part 2 of this report.

Sequencing

Advanced Detectors for Mass Spectrometry
W.H. Benner and **J.M. Jaklevic**
Lawrence Berkeley National Laboratory, Berkeley, California

Mass Spectrometer for Human Genome Sequencing
Chung-Hsuan Chen
Oak Ridge National Laboratory, Oak Ridge, Tennessee

Genomic Sequence Comparisons
George Church
Harvard Medical School, Boston, Massachusetts

A PAC/BAC End-Sequence Data Resource for Sequencing the Human Genome: A 2-Year Pilot Study
Pieter de Jong
Roswell Park Cancer Institute, Buffalo, New York

Multiple-Column Capillary Gel Electrophoresis
Norman Dovichi
University of Alberta, Edmonton, Canada

DNA Sequencing with Primer Libraries
John J. Dunn and **F. William Studier**
Brookhaven National Laboratory, Upton, New York

Rapid Preparation of DNA for Automated Sequencing
John J. Dunn and **F. William Studier**
Brookhaven National Laboratory, Upton, New York

A PAC/BAC End-Sequence Database for Human Genomic Sequencing
Glen A. Evans
University of Texas Southwestern Medical Center, Dallas, Texas

Automated DNA Sequencing by Parallel Primer Walking
Glen A. Evans
University of Texas Southwestern Medical Center, Dallas, Texas

*Parallel Triplex Formation as Possible Approach for Suppression of DNA-Viruses Reproduction
V.L. Florentiev
Russian Academy of Sciences, Moscow, Russia

Advanced Automated Sequencing Technology: Fluorescent Detection for Multiplex DNA Sequencing
Raymond F. Gesteland
University of Utah, Salt Lake City, Utah

Resource for Molecular Cytogenetics
Joe Gray and **Daniel Pinkel**
University of California, San Francisco

DNA Sample Manipulation and Automation
Trevor Hawkins
Whitehead Institute and Massachusetts Institute of Technology, Cambridge, Massachusetts

Construction of a Genome-Wide Characterized Clone Resource for Genome Sequencing
Leroy Hood, Mark D. Adams,[1] and Melvin Simon[2]
University of Washington, Seattle
[1]The Institute for Genomic Research, Rockville, Maryland
[2]California Institute of Technology, Pasadena, California

DNA Sequencing Using Capillary Electrophoresis
Barry L. Karger
Northeastern University, Boston, Massachusetts

Ultrasensitive Fluorescence Detection of DNA
Richard A. Mathies and **Alexander N. Glazer**
University of California, Berkeley

Joint Human Genome Program Between Argonne National Laboratory and the Engelhardt Institute of Molecular Biology
Andrei Mirzabekov
Argonne National Laboratory, Argonne, Illinois, and
Engelhardt Institute of Molecular Biology, Moscow, Russia

High-Throughput DNA Sequencing: SAmple SEquencing (SASE) Analysis as a Framework for Identifying Genes and Complete Large-Scale Genomic Sequencing
Robert K. Moyzis
Los Alamos National Laboratory, Los Alamos, New Mexico

One-Step PCR Sequencing
Barbara Ramsay Shaw
Duke University, Durham, North Carolina

---

*Projects designated by an asterisk were funded through small emergency grants to Russian scientists following December 1992 site reviews by David Galas (formerly of OHER, renamed OBER in 1997), Raymond Gesteland (University of Utah), and Elbert Branscomb (LLNL).

Automation of the Front End of DNA Sequencing

Lloyd M. Smith and **Richard A. Guilfoyle**
University of Wisconsin, Madison

High-Speed DNA Sequence Analysis by Matrix-Assisted Laser Desorption Mass Spectrometry

Lloyd M. Smith
University of Wisconsin, Madison

Analysis of Oligonucleotide Mixtures by Electrospray Ionization-Mass Spectrometry

Richard D. Smith
Pacific Northwest National Laboratory, Richland, Washington

High-Speed Sequencing of Single DNA Molecules in the Gas Phase by FTICR-MS

Richard D. Smith
Pacific Northwest National Laboratory, Richland, Washington

Characterization and Modification of DNA Polymerases for Use in DNA Sequencing

Stanley Tabor
Harvard University, Boston, Massachusetts

Modular Primers for DNA Sequencing

Levy Ulanovsky[1,2]
[1]Argonne National Laboratory, Argonne, Illinois
[2]Weizmann Institute of Science, Rehovot, Israel

Time-of-Flight Mass Spectroscopy of DNA for Rapid Sequence

Peter Williams
Arizona State University, Tempe, Arizona

Development of Instrumentation for DNA Sequencing at a Rate of 40 Million Bases Per Day

Edward S. Yeung
Iowa State University, Ames, Iowa

Mapping

Resolving Proteins Bound to Individual DNA Molecules

David Allison and **Bruce Warmack**
Oak Ridge National Laboratory, Oak Ridge, Tennessee

*Improved Cell Electrotransformation by Macromolecules

Alexandre S. Boitsov
St. Petersburg State Technical University, St. Petersburg, Russia

Overcoming Genome Mapping Bottlenecks

Charles R. Cantor
Boston University, Boston, Massachusetts

Preparation of PAC Libraries

Pieter J. de Jong
Roswell Park Cancer Institute, Buffalo, New York

Chromosomes by Third-Strand Binding

Jacques R. Fresco
Princeton University, Princeton, New Jersey

Chromosome Region-Specific Libraries for Human Genome Analysis

Fa-Ten Kao
Eleanor Roosevelt Institute for Cancer Research, Denver, Colorado

*Identification and Mapping of DNA-Binding Proteins Along Genomic DNA by DNA-Protein Crosslinking

V.L. Karpov
Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

A PAC/BAC Data Resource for Sequencing Complex Regions of the Human Genome: A 2-Year Pilot Study

Julie R. Korenberg
Cedars Sinai Medical Center, Los Angeles, California

Mapping and Sequencing of the Human X Chromosome

D. L. Nelson
Baylor College of Medicine, Houston, Texas

*Sequence-Specific Proteins Binding to the Repetitive Sequences of High Eukaryotic Genome

Olga Podgornaya
Institute of Cytology, Russian Academy of Sciences, St. Petersburg, Russia

*Protein-Binding DNA Sequences

O.L. Polanovsky
Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

*Development of Intracellular Flow Karyotype Analysis

A.I. Poletaev
Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

Mapping and Sequencing with BACs and Fosmids

Melvin I. Simon
California Institute of Technology, Pasadena, California

Towards a Globally Integrated, Sequence-Ready BAC Map of the Human Genome

Melvin I. Simon
California Institute of Technology, Pasadena, California

Generation of Normalized and Subtracted cDNA Libraries to Facilitate Gene Discovery

Marcelo Bento Soares
Columbia University, New York, New York

Mapping in Man-Mouse Homology Regions

Lisa Stubbs
Oak Ridge National Laboratory, Oak Ridge, Tennessee

Positional Cloning of Murine Genes

Lisa Stubbs
Oak Ridge National Laboratory, Oak Ridge, Tennessee

Human Artificial Episomal Chromosomes (HAECS) for Building Large Genomic Libraries

Jean-Michel H. Vos
University of North Carolina, Chapel Hill

*Cosmid and cDNA Map of a Human Chromosome 13q14 Region Frequently Lost at B Cell Chronic Lymphocytic Leukemia

N.K. Yankovsky
N.I. Vavilov Institute of General Genetics, Moscow, Russia

Informatics

BCM Server Core

Daniel Davison
Baylor College of Medicine, Houston, Texas

A Freely Sharable Database-Management System Designed for Use in Component-Based, Modular Genome Informatics Systems

Nathan Goodman
The Jackson Laboratory, Bar Harbor, Maine

A Software Environment for Large-Scale Sequencing

Mark Graves
Baylor College of Medicine, Houston, Texas

Generalized Hidden Markov Models for Genomic Sequence Analysis

David Haussler
University of California, Santa Cruz

Identification, Organization, and Analysis of Mammalian Repetitive DNA Information

Jerzy Jurka
Genetic Information Research Institute, Palo Alto, California

*TRRD, GERD and COMPEL: Databases on Gene-Expression Regulation as a Tool for Analysis of Functional Genomic Sequences

N.A. Kolchanov
Institute of Cytology and Genetics, Novosibirsk, Russia

Data-Management Tools for Genomic Databases

Victor M. Markowitz and **I-Min A. Chen**
Lawrence Berkeley National Laboratory, Berkeley, California

The Genome Topographer: System Design

T. Marr
Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

A Flexible Sequence Reconstructor for Large-Scale DNA Sequencing: A Customizable Software System for Fragment Assembly

Gene Myers
University of Arizona, Tucson

The Role of Integrated Software and Databases in Genome Sequence Interpretation and Metabolic Reconstruction

Ross Overbeek
Argonne National Laboratory, Argonne, Illinois

Database Transformations for Biological Applications

G. Christian Overton, Susan B. Davidson, and **Peter Buneman**
University of Pennsylvania, Philadelphia

Las Vegas Algorithm for Gene Recognition: Suboptimal and Error-Tolerant Spliced Alignment

Pavel A. Pevzner
University of Southern California, Los Angeles, California

Foundations for a Syntactic Pattern-Recognition System for Genomic DNA Sequences: Languages, Automata, Interfaces, and Macromolecules

David B. Searls
SmithKline Beecham Pharmaceuticals, King of Prussia, Pennsylvania

Analysis and Annotation of Nucleic Acid Sequence

David J. States
Washington University, St. Louis, Missouri

Gene Recognition, Modeling, and Homology Search in GRAIL and genQuest

Edward C. Uberbacher
Oak Ridge National Laboratory, Oak Ridge, Tennessee

Informatics Support for Mapping in Mouse-Human Homology Regions

Edward Uberbacher
Oak Ridge National Laboratory, Oak Ridge, Tennessee

SubmitData: Data Submission to Public Genomic Databases

Manfred D. Zorn
Lawrence Berkeley National Laboratory, University of California, Berkeley

ELSI

The Human Genome: Science and the Social Consequences; Interactive Exhibits and Programs on Genetics and the Human Genome

Charles C. Carlson
The Exploratorium, San Francisco, California

Documentary Series for Public Broadcasting

Graham Chedd and **Noel Schwerin**
Chedd-Angier Production Company, Watertown, Massachusetts

Human Genome Teacher Networking Project

Debra L. Collins and **R. Neil Schimke**
University of Kansas Medical Center, Kansas City, Kansas

Human Genome Education Program

Lane Conn
Stanford Human Genome Center, Palo Alto, California

Your World/Our World–Biotechnology & You: Special Issue on the Human Genome Project

Jeff Davidson and **Laurence Weinberger**
Pennsylvania Biotechnology Association, State College, Pennsylvania

The Human Genome Project and Mental Retardation: An Educational Program

Sharon Davis
The Arc of the United States, Arlington, Texas

Pathways to Genetic Screening: Molecular Genetics Meets the High-Risk Family

Troy Duster
University of California, Berkeley

Intellectual Property Issues in Genomics

Rebecca S. Eisenberg
University of Michigan Law School, Ann Arbor, Michigan

AAAS Congressional Fellowship Program

Stephen Goodman
The American Society of Human Genetics, Bethesda, Maryland

A Hispanic Educational Program for Scientific, Ethical, Legal, and Social Aspects of the Human Genome Project

Margaret C. Jefferson and **Mary Ann Sesma**[1]
California State University and [1]Los Angeles Unified School District, Los Angeles, California

Implications of the Geneticization of Health Care for Primary Care Practitioners

Mary B. Mahowald
University of Chicago, Chicago, Illinois

Nontraditional Inheritance: Genetics and the Nature of Science; Instructional Materials for High School Biology

Joseph D. McInerney and **B. Ellen Friedman**
Biological Sciences Curriculum Study, Colorado Springs, Colorado

The Human Genome Project: Biology, Computers, and Privacy: Development of Educational Materials for High School Biology

Joseph D. McInerney and **Lynda B. Micikas**
Biological Sciences Curriculum Study, Colorado Springs, Colorado

Involvement of High School Students in Sequencing the Human Genome

Maureen M. Munn, Maynard V. Olson, and **Leroy Hood**
University of Washington, Seattle

The Gene Letter: A Newsletter on Ethical, Legal, and Social Issues in Genetics for Interested Professionals and Consumers

Philip J. Reilly, Dorothy C. Wertz, and **Robin J.R. Blatt**
The Shriver Center for Mental Retardation, Waltham, Massachusetts

The DNA Files: A Nationally Syndicated Series of Radio Programs on the Social Implications of Human Genome Research and Its Applications

Bari Scott
Genome Radio Project, KPFA-FM, Berkeley, California

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

Maria Sosa, Judy Kass, and Tracy Gath
American Association for the Advancement of Science, Washington, D.C.

The Community College Initiative

Sylvia J. Spengler and Laurel Egenberger
Lawrence Berkeley National Laboratory, Berkeley, California

Genome Educators

Sylvia Spengler and **Janice Mann**
Lawrence Berkeley National Laboratory, Berkeley, California

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

Sara L. Tobin and **Ann Boughton**[1]
Stanford University, Palo Alto, California
[1]Thumbnail Graphics, Oklahoma City, Oklahoma

The Genetics Adjudication Resource Project

Franklin M. Zweig
Einstein Institute for Science, Health, and the Courts, Bethesda, Maryland

Infrastructure

Alexander Hollaender Distinguished Postdoctoral Fellowships

Linda Holmes and **Eugene Spejewski**
Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

Human Genome Management Information System

Betty K. Mansfield and **John S. Wassom**
Oak Ridge National Laboratory, Oak Ridge, Tennessee

Human Genome Program Coordination

Sylvia J. Spengler
Lawrence Berkeley National Laboratory, Berkeley, California

Support of Human Genome Program Proposal Reviews

Walter Williams
Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

Former Soviet Union Office of Health and Environmental Research Program

James Wright
Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

SBIR

1996 Phase I

An Engineered RNA/DNA Polymerase to Increase Speed and Economy of DNA Sequencing

Mark W. Knuth
Promega Corporation, Madison, Wisconsin

Directed Multiple DNA Sequencing and Expression Analysis by Hybridization

Gualberto Ruano
BIOS Laboratories, Inc., New Haven, Connecticut

1996 Phase II

A Graphical Ad Hoc Query Interface Capable of Accessing Heterogeneous Public Genome Databases

Joseph Leone
CyberConnect Corporation, Storrs, Connecticut

Low-Cost Automated Preparation of Plasmid, Cosmid, and Yeast DNA

William P. MacConnell
MacConnell Research Corporation, San Diego, California

GRAIL-GenQuest: A Comprehensive Computational Framework for DNA Sequence Analysis

Ruth Ann Manning
ApoCom, Inc., Oak Ridge, Tennessee

Appendix F: DOE BER Program

Text and photos in this appendix first appeared in a brochure prepared by the Human Genome Management Information System for the DOE Office of Biological and Environmental Research to announce a symposium celebrating 50 years of achievements in the Biological and Environmental Research Program. "Serving Science and Society into the New Millennium" was held on May 21–22, 1997, at the National Academy of Sciences in Washington, D.C. The color brochure and other recent publications related to BER research, including the historically comprehensive A Vital Legacy, *may be obtained from HGMIS at the address on the inside front cover.*

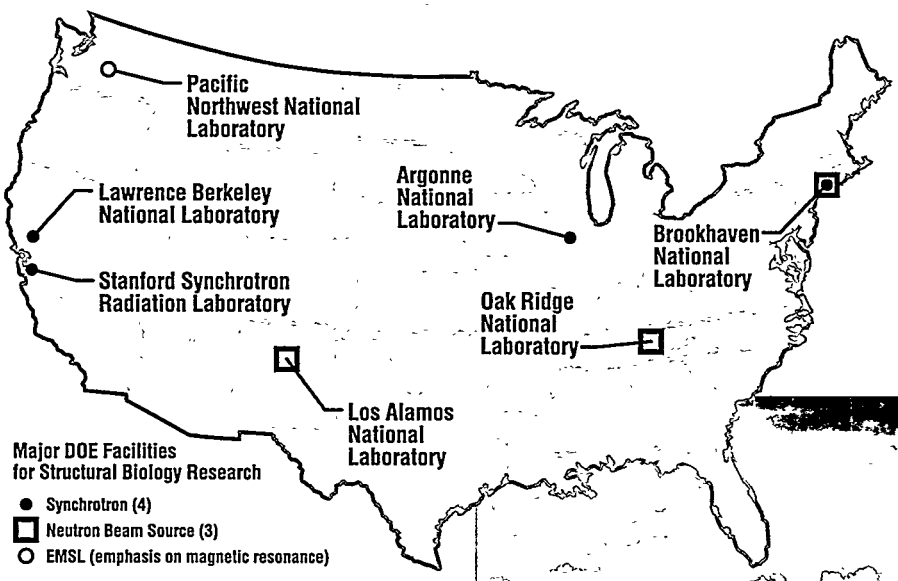
DOE Biological and Environmental Research Program

An Extraordinary Legacy

To exploit the boundless promise of energy technologies and shed light on their consequences to public health and the environment, the Biological and Environmental Research program of the U.S. Department of Energy's (DOE) Office of Health and Environmental Research (OHER) has engaged in a variety of multidisciplinary research activities:

- Establishing the world's first Human Genome Program.
- Developing advanced medical diagnostic tools and treatments for human disease.
- Assessing the health effects of radiation.

Biological and Environmental Research Program
Aristides Patrinos, Ph.D.
Associate Director for Energy Research
for the
Office of Biological and Environmental Research
U.S. Department of Energy
301/903-3251, Fax: 301/903-5051
http://www.er.doe.gov/production/ober/ober_top.html



National User Facilities

Dedicated biomedical resources, such as those maintained by BER at several DOE laboratories, are available at little or no charge. These resources enable scientists to gain an understanding of relationships between biological structures and their functions, study disease processes, develop new pharmaceuticals, and conduct basic research in molecular biology and environmental processes.

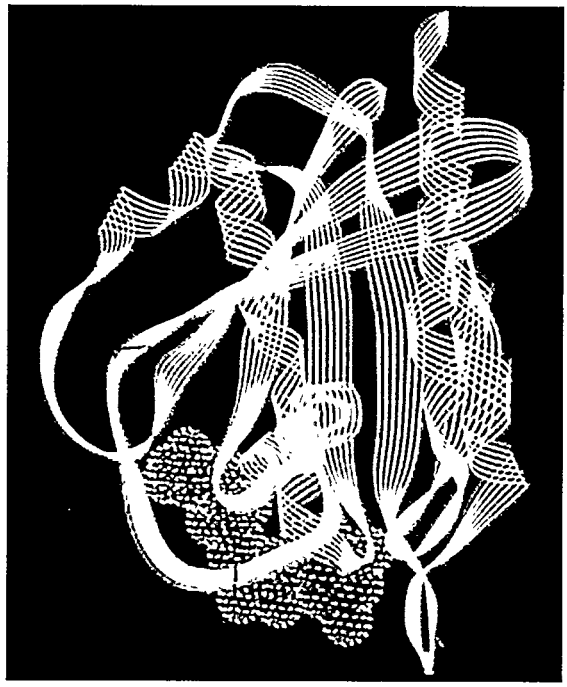
William R. Wiley Environmental Molecular Sciences Laboratory (EMSL) is a national collaborative user facility for providing innovative approaches to meet the needs of DOE's environmental missions.

An Enduring Mandate

DOE is carrying forward Congressional mandates that began with its predecessors, the Atomic Energy Commission and the Energy Research and Development Agency:

Contribute to a Healthy Citizenry

- Develop innovative technologies for tomorrow's biomedical sciences.
- Provide the basis for individual risk assessments by determining the human genome's fine structure by the year 2005.
- Conduct research into advanced medical technologies and radiopharmaceuticals.
- Build and support national user facilities for determining biological structure, and ultimately function, at the molecular and cellular level.



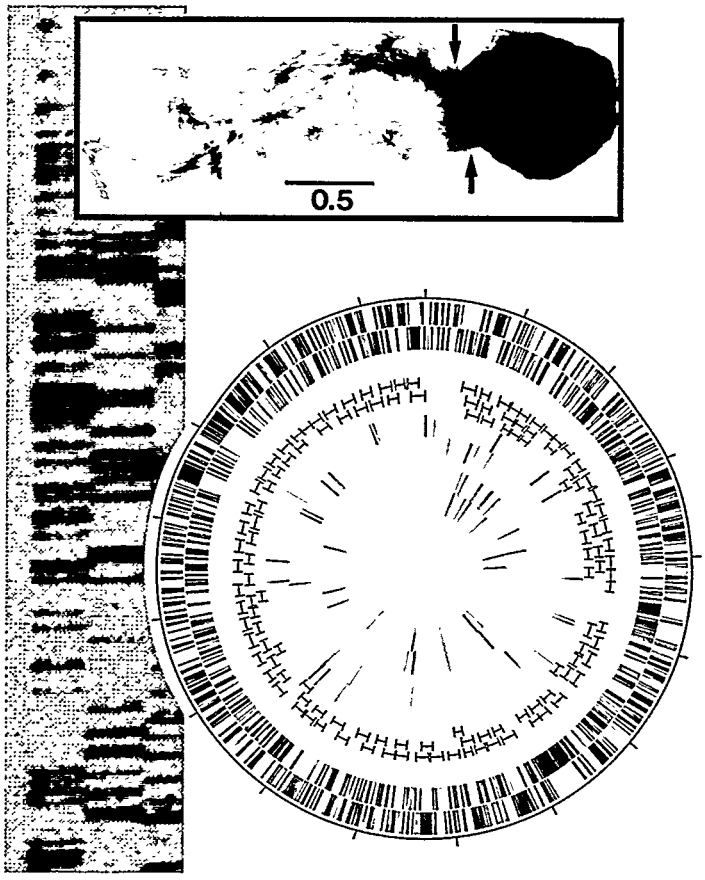
DOE user facilities are revealing the molecular details of life. Knowing the 3-D structure of the ras protein (above), an important molecular switch governing human cell growth, will enable interventions to shut off this switch in cancer cells.

Understand Global Climate Change

Predict the effects of energy production and its use on the regional and global environment by acquiring data and developing the necessary understanding of environmental processes.

Contribute to Environmental Cleanup

Conduct fundamental research to establish a better scientific basis for remediating contaminated sites.



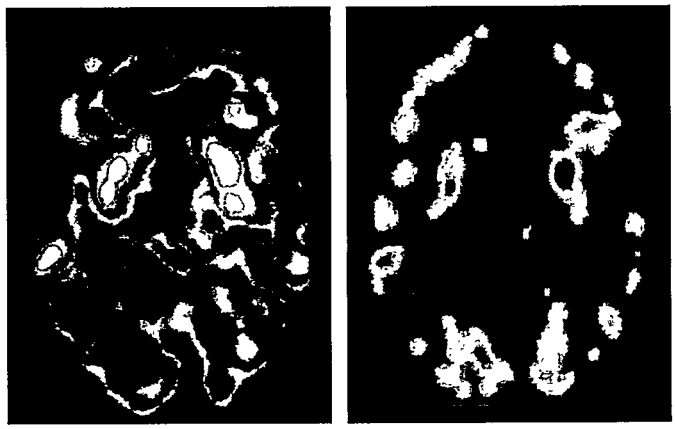
Determining the fine structure—DNA sequence—of the microorganism *Methanococcus jannaschii* (pictured at right, top) and other minimal life forms in DOE's Microbial Genome Program will benefit medicine, agriculture, industrial and energy production, and environmental bioremediation. The circular representation of the single *M. jannaschii* chromosome, which was fully sequenced in 1996, illustrates the location of genes and other important features. (Vertical bar represents a portion of a sequencing experiment.)

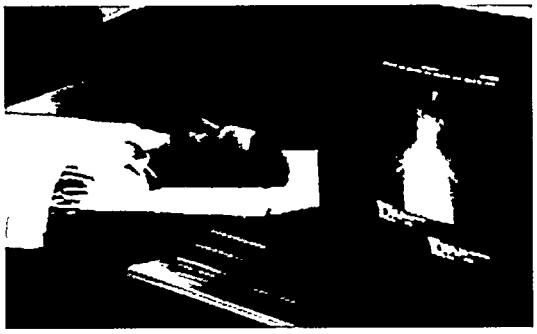
Fifty Years of Achievements. . .
Leading to Innovative Solutions

Tools for Medicine and Research

Radioisotopes developed for medicine and medical imaging are being merged with current knowledge in biology and genetics to discover new ways of diagnosing and treating cancer and other disorders, detecting genes in action, and understanding normal development and function of human organ systems.

- Radioactive molecules used in medical imaging for positron emission tomography (PET) and magnetic resonance imaging (MRI) allow noninvasive diagnosis, monitoring, and exploration of human disorders and their treatments.
- Isotopes and other tracers of brain activity are being used to explore drug addiction, the effects of smoking, Alzheimer's disease, Parkinson's disease, and schizophrenia.
- Technetium-99m is used to diagnose diseases of the kidney, liver, heart, brain, and other organs in about 13 million patients per year.
- Striking successes have been achieved using charged atomic particles to treat thyroid diseases, pituitary tumors, and eye cancer, among other disorders.





One-quarter of all patients in U.S. hospitals undergo tests using descendants of cameras developed by BER to follow radioactive tracers in the body. PET scanning has been key to a generation of brain metabolism studies as well as diagnostic tests for heart disease and cancer. PET studies above reveal brain metabolism differences in recovering alcoholics (left, 10 days, and right, 30 days, after withdrawal from alcohol).

The laser-based flow cytometer developed at DOE national laboratories enables researchers to separate human chromosomes for analysis.

Genome Projects

A legacy of DOE research on genetic effects paved the way for the world's first Human Genome Program. Now new genomic technologies are being applied to environmental cleanup through the DOE Natural and Accelerated Bioremediation Research and Microbial Genome programs, healthcare and risk assessment, and such other national priorities as industrial processes and agriculture.

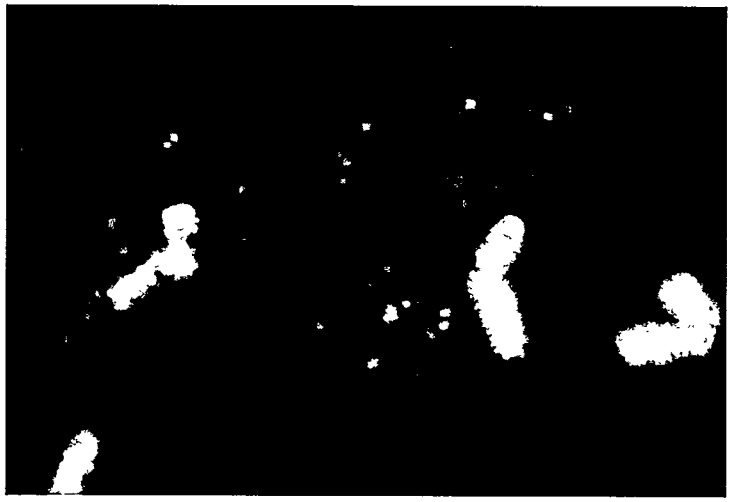
Discover the breadth of current activities and recent accomplishments via the BER Web Site:

http://www.er.doe.gov/production/ober/ober_top.html

Radiation Risks and Protection Guidelines

BER studies have become the foundation for laws and standards that protect the population, including workers exposed to radiological sources:

- Guidelines for the safe use of diagnostic X rays and radiopharmaceuticals.
- Safety standards for the presence of radionuclides in food and drinking water.
- Radiation-detection systems and dosimetry techniques.



Human chromosomes "painted" by fluorescent dyes to detect abnormal exchange of genetic material frequently present in cancer. Chromosome paints also serve as valuable resources for other clinical and research applications.

Finding a Link Between DNA Damage and Cancers

Studies of DNA damage have uncovered similar mechanisms at work in damage caused by radiation exposure, X rays, ultraviolet light, and cancer-causing chemicals. A screening test for such chemicals is now one of the first hurdles a new compound must clear on its way to regulatory and public acceptance.

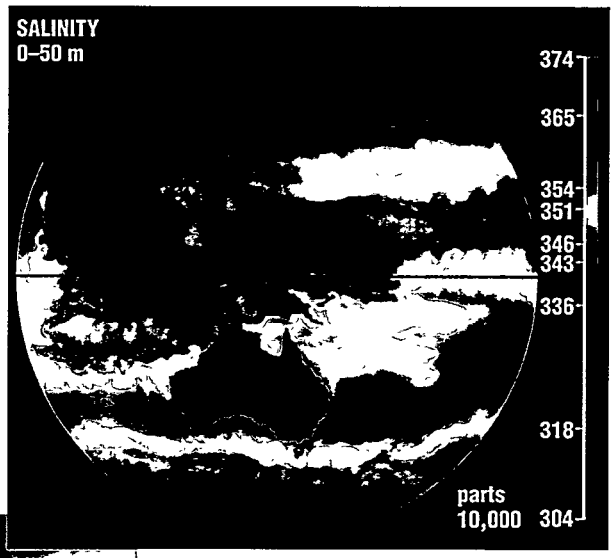
"... (it's) not so much where we stand as in what direction we are moving. [Oliver Wendell Homes, Sr.]"

Tracking the Regional and Global Movement of Pollutants

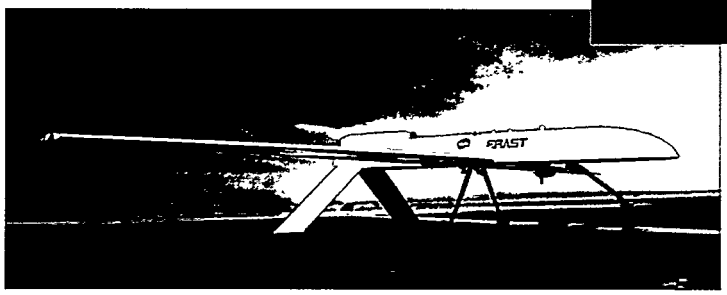
BER research helped to establish the earliest and most authoritative monitoring network in the world to detect airborne radioisotopes. The use of atmospheric tracers has led to the improved ability to predict the dispersion of pollutants.

Understanding Global Change

Important achievements in environmental research have led to enhanced capabilities in studying global change, including more accurate predictions of global and regional climate changes induced by increasing atmospheric concentrations of greenhouse gases.



High-performance computing is promoting faster and more realistic solutions to long-term climate change.



The Unmanned Aerospace Vehicle (above) conducts measurements to quantify the fate of solar radiation falling on the earth.

Creating a New Science of Ecology

BER achievements in using radioactive tracers to follow the movements of animals, routes of chemicals through food chains, decomposition of forest detritus, together with the program's introduction of computer simulations, created the new field of radioecology.

This glossary was adapted from definitions in the DOE *Primer on Molecular Genetics* (1992).

http://www.ornl.gov/hgmis/publicat/primer/intro.html

A

Adenine (A): A nitrogenous base, one member of the base pair A-T (adenine-thymine).

Allele: Alternative form 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).

B

BAC: See bacterial artificial chromosome.

Bacterial artificial chromosome (BAC): A vector used to clone DNA fragments (100- to 300-kb insert size; average, 150 kb) in *Escherichia coli* cells. Based on naturally occurring F-factor plasmid found in the bacterium *E. coli*. Compare cloning vector.

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.

C

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.

Chromosome: The self-replicating genetic structure 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.

Clone: 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 vectors 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 sequence: Nucleic acid base sequence that can form a double-stranded structure by matching base pairs with another sequence; 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: Group of clones representing overlapping regions of a genome.

Contig map: A map depicting the relative order of a linked library of small overlapping clones representing a complete chromosomal segment.

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).

D

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 probe: 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

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 chromosome.

Evolutionarily conserved: See conserved sequence.

Exogenous DNA: DNA originating outside an organism.

Exon: The protein-coding DNA sequence of a gene. Compare intron.

Exonuclease: An enzyme that cleaves nucleotides sequentially from free ends of a linear nucleic acid substrate.

Expressed gene: See gene expression.

F

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 separate chromosomes on the basis of their DNA content.

G

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).

Gene family: Group 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 technology: See recombinant DNA technology.

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 project: Research and technology development effort 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).

H

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.

Homology: Similarity in DNA or protein sequences between individuals of the same species or among different species.

Homologous chromosome: Chromosome containing the same linear gene sequences as another, 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.

I

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.

Intron: The DNA base sequence interrupting the protein-coding sequence of a gene; this sequence is transcribed into RNA but is cut out of the message before it is translated into protein. Compare exon.

In vitro: Outside a living organism.

K

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.

L

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.

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.

M

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 disorder: See polygenic disorder.

Multiplexing: A sequencing approach that uses several pooled samples simultaneously, greatly increasing sequencing speed.

Mutation: Any heritable change in DNA sequence. Compare polymorphism.

N

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.

O

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.

P

P1-derived artificial chromosome (PAC): A vector used to clone DNA fragments (100- to 300-kb insert size; average, 150 kb) in *Escherichia coli* cells. Based on bacteriophage (a virus) P1 genome. Compare cloning vector.

PAC: See P1-derived artificial chromosome.

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 disorder: Genetic disorder 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 chromosome.

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 bodys 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.

R

Rare-cutter enzyme: See restriction enzyme cutting site.

Recombinant clone: Clone containing recombinant DNA molecules. See recombinant DNA technology.

Recombinant DNA molecules: A combination of DNA molecules of different origin that are joined using recombinant DNA technologies.

Recombinant DNA technology: Procedure 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 region or sequence: 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.

Ribonucleotide: 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.

S

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 chromosome: The X or Y chromosome in human beings that determines 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 cell: 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.

T

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 end of a chromosome. This specialized structure is 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.

U

Uracil: A nitrogenous base normally found in RNA but not DNA; uracil is capable of forming a base pair with adenine.

V

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 more than 100,000 transistors on a chip.

Y

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.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1998 433-460/60046

Acronyms and Initialisms

AAAS	American Association for the Advancement of Science
ABI	Applied Biosystems, Inc.
AEC	Atomic Energy Commission
AIDS	acquired immune deficiency syndrome
ANL	Argonne National Laboratory
ATP	Advanced Technology Program
BAC	bacterial artificial chromosome
BCM	Baylor College of Medicine
BDG	Batten's disease gene
BER	Biological and Environmental Research Program
bp	base pair
BSCS	Biological Sciences Curriculum Study
CAE	capillary array electrophoresis
cDNA	complementary DNA
CE	capillary electrophoresis
cM	centimorgan
CRADA	Cooperative Research and Development Agreement
DNA	deoxyribonucleic acid
DOE	Department of Energy
EDS	electronic data submission
ELSI	ethical, legal, and social issues
ERDA	Energy Research and Development Administration
EST	expressed sequence tag
FISH	fluorescence in situ hybridization
FMF	familial Mediterranean fever
GDB	Genome Database
GPI	Genetics and Public Issues Program
GRAIL	Gene Recognition and Analysis Internet Link
GSDB	Genome Sequence DataBase
HAEC	human artificial episomal chromosome
HELSRD	Health Effects and Life Sciences Research Division
HGMIS	Human Genome Management Information System
HGN	*Human Genome News*
HGP	Human Genome Project, Human Genome Program
HLA	human leukocyte antigen
HUGO	Human Genome Organisation
IMAGE	Integrated Molecular Analysis of Gene Expression
JGI	Joint Genome Institute
kb	kilobase
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LLNL	Lawrence Livermore National Laboratory
LTI	Life Technologies, Inc.
MALDI	matrix-assisted laser desorption ionization
Mb	megabase
MS	mass spectrometry
NCGR	National Center for Genome Resources
NIH	National Institutes of Health
NLGLP	National Laboratory Gene Library Project
OBER	Office of Biological and Environmental Research
OHER	Office of Health and Environmental Research
ORNL	Oak Ridge National Laboratory
PAC	P1 artificial chromosome
PATCO	Premier American Technologies Corp.
PLC	programmable logic controller
QDFM	Quantitative DNA Fiber Mapping
R&D	research and development
SASE	sample sequencing
SBH	sequencing by hybridization
SBIR	Small Business Innovation Research
SCAN	Sequence Comparison ANalysis program
SCW	single-chromosome workshop
SHOM	sequencing by hybridization on matrices
STS	sequence tagged site
STTR	Small Business Technology Transfer
TCR	T-cell receptor
UC	University of California
UCB	University of California, Berkeley
USDA	United States Department of Agriculture
UW	University of Washington
WHS	Wolf-Hirschhorn syndrome
YAC	yeast artificial chromosome

HGMIS
Oak Ridge National Laboratory
1060 Commerce Park
Oak Ridge, TN 37830

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

ER-72