

Human Genome news

A Forum for
Human Genome
Information
Exchange

Sponsored by the U.S. Department of Energy

ISSN: 1050-6101

Vol. 7, No. 6, April–June 1996

Research Pace Applauded at Santa Fe '96 Workshop

Genome Remains Priority for DOE Energy Research, says Krebs

Martha Krebs, Director of the DOE Office of Energy Research (ER), welcomed participants to the fifth DOE Human Genome Program Contractor-Grantee Workshop on January 28–February 1, 1996.



Martha Krebs
Director
DOE Office of
Energy Research

Krebs' assessment of the strengths and future directions of the program was well received by almost 400 researchers, program managers, and invited guests who met in Santa Fe, New Mexico.

More than 50 speakers (and posters representing about 200 projects) reported on progress in mapping; resource development; sequencing; informatics; and ethical, legal, and social issues.

The research was carried out at DOE human genome centers at Lawrence Berkeley National Laboratory (LBNL),

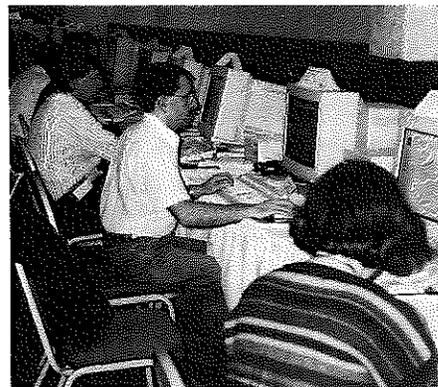
Lawrence Livermore National Laboratory (LLNL), and Los Alamos National Laboratory (LANL) and at about 90 U.S. universities and research organizations.

Krebs applauded the remarkable pace of the program and praised retiring Program Director Dave Smith for his guiding philosophy of "supporting the best science to make the fastest progress."

Commenting on the future of the program, Krebs said "it is an increasing priority for ER and will remain so in the coming months and years because it addresses DOE missions. The program now needs to evolve rapidly to capitalize on opportunities that technological advances are creating, especially with high-throughput sequencing." Krebs stated that for the remainder of the genome project, sequencing efforts will be of primary importance. "We must begin to direct a significant fraction of the project budget into the focused effort necessary for production of large amounts of sequence data," she said.

Krebs conveyed her concern about the long-term funding outlook for science in the perilous budgetary climate now pervading Washington. Her closing message was an urgent one: All researchers must speak out on behalf of science

Santa Fe '96 Highlights Featured on pp. 1–17



investment by discussing its present and future impact on the American people and on U.S. leadership in all fields of science and technology in the 21st century. "We need to make sure that people understand what they will lose if they don't support this kind of effort," she said.

Krebs suggested contacting representatives from the local to national levels, writing letters and pieces for newspaper editorial pages, enlisting the aid of university presidents and deans, and involving professional societies in this crucial task.

by Denise Casey, HG MIS

Meeting highlights begin on next page.

Research Abstracts

Online: http://www.ornl.gov/TechResources/Human_Genome/publicat/96santa/intro.html

Print requests: bkq@ornl.gov (see also p. 20)

Next Contractor-Grantee Workshop

November 1997. Contact: Sylvia Spengler (LBNL, 510/486-4879, sylviaj@violet.berkeley.edu)

In This Issue

1 DOE Santa Fe '96 Workshop 18 Yeast Genome Sequenced 18 ELSI Committee Statement 19 Sequencing Meeting 20 Pilot Sequencing Projects
20 Mapping Milestones 20 In Memorium 21 GDB Forum 21 Electronic Data Access 22 Calendar of Events, Training Calendar
23 ELSI Task Force 23 Prostate Cancer Consortium 23 Funding Information 24 Subscription/Orders, Acronym List

Santa Fe '96

Spreading the Word

ELSI Projects Target Diverse Audience

Worldwide progress toward obtaining a human DNA reference sequence has heightened the urgency of dealing with a host of challenging ethical, legal, and social issues (ELSI) surrounding the data. From the start, the DOE Human Genome Program has devoted up to 3% of its annual budget to addressing these topics, focusing particularly on the privacy and confidentiality of genetic information (including issues of ownership and commercialization) and on the development of educational materials geared toward a diverse public.

Highlights of five engaging ELSI presentations follow.

Crime and Punishment Meet Genomics: Responsibility Reconsidered

In a compelling dramatization of genetics in the courtroom, Franklin Zweig (Einstein Institute), speaking before Judge Rosalyn Bell (Maryland Court of Special Appeals), rendered closing arguments for and against the death penalty in a case involving a convicted murderer who may have had a genetic predisposition toward violence. Zweig then charged the audience to act as a jury and cast their votes for the death penalty or life in prison. Ballots were collected, tallied,

and reported the next day, with a majority voting against the death penalty. [This exercise was based on an actual case. The convicted murderer was executed in April.]

The demonstration effectively drove home Zweig's point: Molecular biology makes trouble for the courts. About 30,000 federal and state judges, representing the pinnacle of the government's authority, want to dispense justice through fair trials. The definitive power is the ability of the state to take the life of a person convicted of a capital crime. Recent years have seen that power rise, fueled in part by a citizenry at wit's end about violence and the victims it creates. Interpretation of genomic research, Zweig asserted, can contribute to the decision about taking a defendant's life.

The question asked of the scientific community by the justice system, he continued, is clear but impossible to answer: Did the criminal act lie in the defendant's genes? If so, should people be penalized for the genes they carry? Scientists' answers could effect a sea change in jurisprudence and redefine justice by contributing to a shift away from historical precedents traditionally used by the courts.

In his Genetics Adjudication Resource Project, Zweig's objectives are to help

judges understand the scientific validity of the genetic-based claims that are rushing into the nation's courtrooms and, to the extent possible, avoid the past decade's confusion in adjudicating forensic DNA technologies. The Hon. Pauline Newman is chair of the Advisory and Review Committee for this adjudication project. Judge Newman serves on the U.S. Court of Appeals for the Federal Circuit, Washington, the highest court under the U.S. Supreme Court.

Genome Radio Project: "The DNA Files"

Bari Scott, Matt Binder, and Jude Thilman of the Genome Radio Project discussed the development of a series of hour-long programs exploring genetic issues, to be aired on public radio next year. Plans are to market these audio tapes with supplementary materials, including a WWW site for educators.

The group played lively excerpts from the pilot program, "DNA and Behavior: Is Our Fate Written in Our Genes?" Included were visits to laboratories studying identical twins separated at birth and obesity in mice. The main theme of the pilot, explained Binder, the senior producer, is that a complex interaction of both nature and nurture determines many human traits. Important subthemes include the media's role in creating perceptions about these issues. Behavioral genetics was chosen as the pilot topic to attract a large audience quickly.

The plan is to tie these programs to call-in, question-and-answer radio sessions with scientific experts from local areas. Some stations are interested in sponsoring talks by scientists in the community. Scott, who is developing a list of individuals, institutions, and laboratories willing to help out at their local public radio station, welcomes further input from scientists (contact: 510/848-6767, ext. 264; Fax: /883-0311; strp@aol.com).

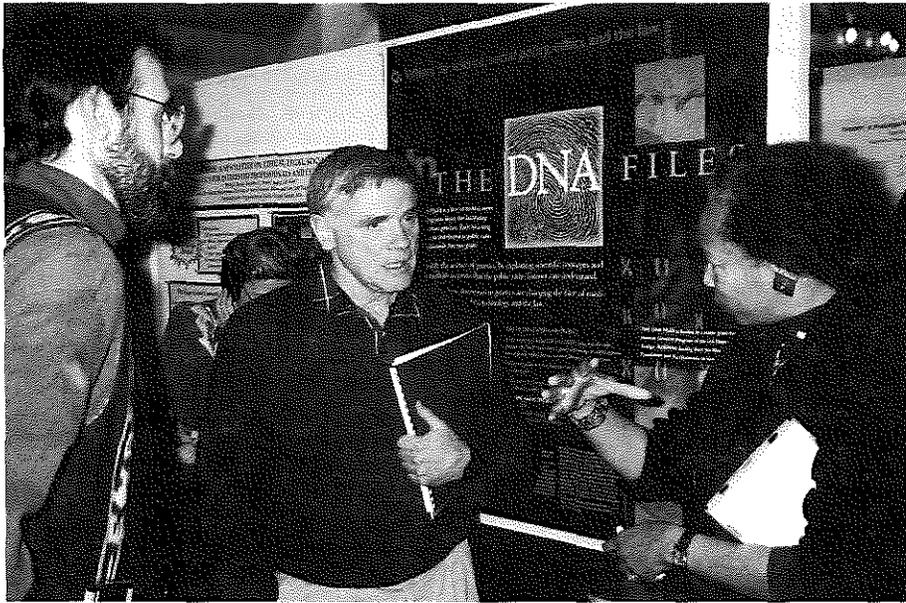
Plain-Language Genomics for Adult Science Literacy

Learning about genomics, asserted Maria Sosa [American Association for the Advancement of Science (AAAS)], makes sense for all Americans. The knowledge could affect their health, help them get jobs in a growing field,

Patenting and Database Controversies—Does the One with the Most Sequence Really Win?

Rebecca Eisenberg (University of Michigan Law School) raised intriguing questions about the proper roles of government and industry in genomic research. One consideration is who stands to benefit (and to lose) from the private appropriation of genomic information. Controversy, she noted, is particularly acute over the intellectual property rights in large-scale cDNA sequencing data.

Genomics research involves a major commitment of both federal and private funds, and both groups have an interest in making use of the wealth of new discoveries, she noted. In addition, the implications for human health make public access crucial. Access also affects the interests of the young biotechnology industry, as well as the more established pharmaceutical companies. Eisenberg analyzed possible motivations behind recent steps taken by Merck & Co., Human Genome Sciences, and Incyte to develop databases of human gene sequences. The take-home lesson, she concluded, may be that the value of a rich public domain to both the public and private sectors has been underestimated. An article on Eisenberg's project will appear in a future issue of *HGN*.⁰



Leroy Hood (center) of the University of Washington, Seattle, and Matt Binder and Bari Scott (right) of the Genome Radio Project.

and even help them with their children's homework. But adults lacking literacy skills are denied access to this information.

Sosa is developing a module, targeted at or above the 6th- to 8th-grade reading level, to provide these underserved people with the background knowledge necessary to understand ELSI issues that may impact their lives. The 2-year project, part of Science + Literacy for Health sponsored by AAAS, will deliver materials to literacy classes, community groups providing health services, and public libraries.

What Should We Teach the Kids?

While "correct" answers to some ELSI questions may not exist, well-reasoned ones begin with a solid understanding of science. Joe McInerney [Biological Sciences Curriculum Study (BSCS)] discussed major challenges to high school biology education that have been crystallized by the Human Genome Project.

These challenges include teaching about the nature and methods of science and what distinguishes scientific explanations for natural phenomena from other views of the world; ELSI issues related to science and technology; and the central role of technology in society. Technology can be fallible, have some associated risks, and sometimes serve the interests of particular individuals or groups.

BSCS is addressing these educational challenges with three modules funded by the DOE ELSI program. The modules deal with the science and ELSI of the Human Genome Project, the importance of informatics in the project and some related ELSI issues involving genomics databases, and nontraditional mechanisms of inheritance. In the last module, BSCS moves away from the usual genetics curriculum (with examples based on single-gene disorders) to present notions of

Special Genome Issue of *Your World/Our World*

"Exploring the Human Genome" is an expanded 24-page issue of *Your World/Our World: Biotechnology and You*. The colorful magazine, developed by the Pennsylvania Biotechnology Association (PBA) for teaching biotechnology to students in grades 7 through 10, is also appropriate for lay audiences. Articles and learning activities explore the Human Genome Project; DNA structure and function; genes, proteins, and genetic disease; mapping; invention of PCR; informatics; Huntington's Disease; and ELSI implications of genome research. This special genome issue is made available in part by DOE, which provided a complimentary copy to every 7th- to 10th-grade U.S. science teacher (about 40,000). The issue is endorsed by James Watson, former director of the NIH National Center for Human Genome Research. Teacher's guide included; subscription packages and back issues available. (PBA, 800/796-5806 or 814/238-4080, Fax: -4081, 73150.1623@compuserve.com) ♦

relative risk, susceptibility, and predisposition. This print module, titled *Changing Concepts of Inheritance: Genetics and the Methods of Science*, should be available in early 1997. For a free copy of this or the second module (print or computer disk), contact Dee Miller, BSCS; 5415 Mark Dabbling Blvd.; Colorado Springs, CO 80918 (Fax: 719/531-9104). ♦

Santa Fe '96

Progress at the DOE Labs

LANL: Scanning the Genome with SASE Sequencing

Researchers at the LANL human genome center are testing a large-scale sequencing approach designed to identify genes quickly while capitalizing on LANL's high-resolution maps of chromosome 16. The dense clone coverage of the chromosome—now at about 98% and mostly in cosmids—provides an ideal framework for sequencing [Doggett et al., *Nature* **377**, 335-65 (1995)].

LANL's strategy, outlined by Darrell Ricke, is to skim through chromosome 16 using a random (or "shotgun")

method, in which they break 40-kb cosmid clones into 3-kb pieces, subclone them, and sequence both ends of each subclone. The sequences are analyzed using a new sequence-comparison system, and interesting regions (such as coding areas) are identified for more detailed, "finished" sequencing.

The focus on regions of immediate interest makes this a low-investment (one-tenth the price of finished genomic sequencing), potentially high payoff strategy. Another advantage of the approach is that identifying the genes and exons via sequence analysis provides more information than simply mapping ESTs. ▶

The front end of this approach, which LANL calls SASE (for sampled sequencing), will allow LANL to rapidly generate aligned sequences along the chromosome 16 map. Sequencing both ends of a 1x sampling of subcloned cosmid fragments, along with cosmid end sequences, yields 70% sequence coverage with 98% clone coverage. The majority of this clone coverage is ordered by the relationships among the subclone end sequences, which are ideal substrates for directed sequencing strategies. At LANL, finished sequencing is done rapidly by parallel primer walking along the original cosmid DNA.

LANL Web Site

<http://www-ls.lanl.gov/DBqueries/QueryPage.html>

All SASE data are deposited in Genome Sequence Data Base and remain readily available for analysis. A notation is placed on particular sequences already targeted for finishing.

The data are sufficient to allow PCR amplification of the sequenced region, eliminating the need for extensive clone archiving and distribution and enabling many laboratories to participate in completing the sequencing of chromosome 16. LANL will SASE sequence cosmids to determine a

minimum tiling set of 3-kb subclones, and LANL will then finish sequencing the targeted LANL cosmids. This collaboration leverages the strengths of both laboratories and increases productivity. LANL is also collaborating with The Institute for Genomic Research in supplying map information for sequencing a portion of 16p.

Ricke acknowledged that some regions of the chromosome would be missed, but an early concentration on potentially interesting areas makes this strategy attractive (see "Telomeres" box below). Genomic areas with lower information content could be finished later, he noted, when sequencing technologies are more cost-effective.

Targets for Finished Sequencing

To identify the genes in SASE data, LANL developed the SCAN (Sequence Comparison ANalysis) program. SCAN integrates the results from BLAST and FASTA searches and will soon add GenQuest and GRAIL servers and Smith-Waterman. When SCAN finds a repeat, vector, *Escherichia coli* homology, or rodent homology, it gives a 1- to 2-line summary. All results are integrated and a summary report generated, with HTML pages that are hotlinked to other databases.

The sequence-analysis and automatic annotation features make SCAN an

efficient program for large-scale projects. Results are displayed in a multiple-sequence alignment that gives a quick overview. Cosmids can be screened within minutes to reveal homologies.

The initial 1-Mb region of 16p completed by SASE has proven to be very gene rich, and investigators plan to finish sequencing the region and continue to adjacent regions of the 4-Mb high-resolution physical map. Projected sequencing throughput would allow complete SASE analysis of the 90-Mb euchromatic arms of this chromosome in just a few years.

SASE Sequence Analysis

David Torney explained a new technique for identifying coding sequences. This technique involves converting DNA sequences into binary sequences of 0s and 1s and then determining the "parities" of subsequences. The parity takes one of two values, depending on whether the number of 1s in the subsequences is even or odd.

"This technique comprehensively captures the features of coding sequences," Torney said. For example, a group of sequences of length n can be characterized completely by the average parities for all 2^n subsequences.

Thus, investigators can make accurate subsequence classifications based on differences between coding and non-coding sequences. Subsequences with the fewest letters were found to be the most discriminating. Using this approach, LANL scientists correctly classified both coding and noncoding 54-base sequences 72.5% of the time. If, in addition, the sequence frame and strand were known, the correct classification rate was 82%.

LLNL: Linking Production Sequencing to the Underlying Biology

At the LLNL human genome center, large-scale chromosome 19 sequencing is coupled with understanding the human genes involved in DNA repair (see "Spell Checking" box, p. 5).

These interests are rooted in DOE's mission to develop better technologies for measuring health effects, particularly mutations. Alterations in DNA

Looking for Genes Near Human Telomeres

Most human genes are hundreds of thousands of bases away from the telomeres (chromosome ends). First identified by the LANL group in 1988, telomeres consist of a series of tandemly repeating DNA sequences such as (TTAGGG) n . The sequences in the subtelomeric regions consist of tandem arrays of more complex repeated sequences, which may act as a buffer to guard against the possible deleterious effects of telomere shortening. Recent studies have suggested that telomere size may be related to aging or the growth of cancer cells. The 7q telomere, however, lacks large blocks of subtelomeric repetitive DNA. This region was chosen as potentially interesting for analysis by SASE and complete sequencing because any genes or exons found here could be potential targets for alteration if telomere shortening or instability should occur.

Han-Chang Chi (LANL) reported on a successful application of SASE combined with parallel primer walking to determine the entire terminal 230-kb region of human chromosome 7q. Cosmid contigs were constructed from a human telomeric YAC clone; 9 overlapping cosmids were sequenced, with a gap resolved by long PCR. All sequences were assembled by either DNA STAR or AUTOASSEMBLER and analyzed by SCAN. SCAN uncovered numerous ESTs localized to this region, as well as 2 exons with 99% homology to the cDNA of a known human gene, vasoactive intestinal polypeptide receptor 2A.0

repair genes can predispose individuals to cancer. LLNL researchers have cloned six different genes involved in repair processes. The focus has been on three of the genes that feed directly into genomic sequencing and the downstream biology aimed at elucidating repair processes. Low-pass sequencing approaches are being developed to minimize redundancy, increase throughput, find genes, and identify candidate regions for higher-redundancy sequencing.

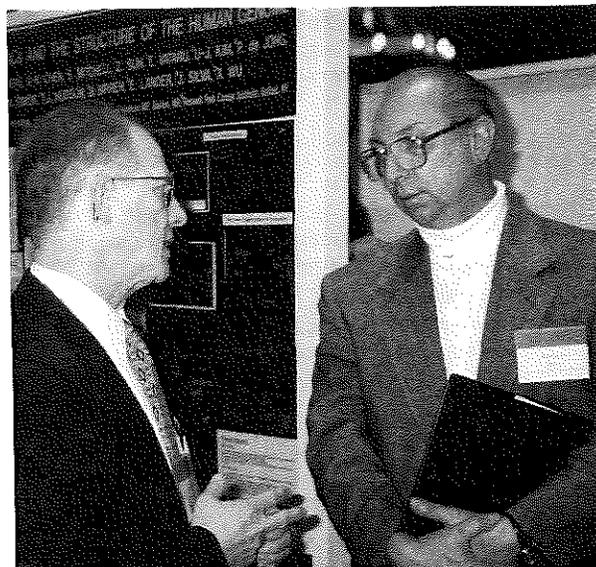
Chromosome 19 sequence data are analyzed and used to generate targeting constructs in making transgenic mouse models and for characterizing the structure and function of these repair genes. LLNL is complementing genomic sequencing with sequencing of full-length cDNAs mapped to chromosome 19 cosmids. The genome center is also performing comparative sequence analyses of the mouse and human genomes, especially in DNA repair gene regions, for elucidating coding structure and identifying putative regulatory regions. These latter projects are being done in collaboration with researchers at Oak Ridge National Laboratory (see Mouse and Human section, p. 8).

With the chromosome 19 map completed, the LLNL genome center was reorganized recently to scale up the sequencing facility and provide high-throughput, high-accuracy sequence for the entire chromosome. Jane Lamerdin discussed some components of the random shotgun strategy, which includes use of a modified LBNL colony picker, 3 Beckman Biomek 1000s, and production of 600 templates a day. Center scientists are now running 10 to 15 gels a day using an ABI autoassembly package and Phred and Phrap software. Lamerdin estimates that, with an 8-fold redundancy and 70% success rate, current capabilities are about 4 Mb of finished sequence a year. Data analysis has been a bottleneck.

Joe Balch is leading an LLNL team to develop a next-generation DNA sequencer based on arrays of micro-channels etched and sealed in a glass substrate as an alternative to arrays of discrete glass capillaries. The plan is to develop a 96-channel array system first and a 384-channel array system later to sequence DNA samples in less than 2 hours.

DNA Repair Gene Analysis

The LLNL genome center has generated some 1.2 Mb of genomic sequence, using a random shotgun strategy with an 8-fold average redundancy and getting complete double-stranded coverage where possible. The primary effort has been targeted to cosmids containing the human DNA repair genes *HHR23A*, *XRCC1*, and *ERCC2* on chromosome 19, *ERCC4* on chromosome 16, *XRCC3* on chromosome 14, and *XRCC2* on chromosome 7, as well as selected rodent homologs. Genomic



Michael Riches (left) of the DOE Office of Health and Environmental Research and Anthony Carrano, Director of the Lawrence Livermore National Laboratory Human Genome Center.

LLNL Web Site

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

Spell Checking the Genome

Left unchecked, damage to human DNA can wreak havoc with biological systems, resulting in diseases such as cancer. DNA molecules are especially vulnerable when copying themselves or when exposed to such environmental insults as ionizing radiation and chemicals.

Fortunately, a system of sensitive DNA repair enzymes monitors the genome for damage and repairs most errors. In fact, human DNA repair systems allow only about 3 mistakes to escape their notice during replication of the 3 billion base pairs in the human genome, a feature that is important in enabling evolutionary changes.

Designating DNA repair enzymes "Molecule of the Year" in 1994, *Science* acknowledged these amazing molecules that "preserve our health, maintain our species, make evolution possible, and contribute to a sound scientific policy on environmental hazards" [266, 1925 (December 23, 1994)].

sequencing is also being used as a gene-discovery method in a chromosome 19 targeted region (19p13.1) associated with olfactory receptors and a congenital kidney disease.

The LLNL group sequenced a total of 76 kb containing human and mouse *XRCC1* genes, identifying coding regions and nine conserved elements. They also completed 54 kb of human sequence encompassing the *ERCC2* gene and 54 kb spanning the syntenic regions in mouse and hamster. A defect in this gene leads to the disorder xeroderma pigmentosum, in which some people have extreme uv sensitivity and are very prone to cancer. Other phenotypic effects include neurological defects and a defect in sulfur metabolism characterized by brittle hair. Sequence analysis of the *ERCC2* gene by Christine Weber detected no single location for mutations leading to a particular defect; all seem to map to the last third of the gene. Structural analysis of the protein sequence may provide interesting clues, including more precise association of mutations with phenotypes.

Researchers found that the human *ERCC2* gene, comprising 23 exons (coding areas), is 98% identical to the rodent homolog at the protein level. They identified two genes flanking *ERCC2*; all three genes and their

orientation are conserved in humans, mice, and hamsters. Gene products of *ERCC2* and *ERCC4* are involved in the nucleotide excision repair pathway that recognizes and removes DNA damage.

Lamerdin noted that these results underscore the power of using a comparative approach to finding genes because all the coding areas in the *ERCC2* target region were not identified by gene-finding software.

Sequencing has been completed on a cosmid and its associated cDNA for the recently cloned human *XRCC3* gene, which appears to play a crucial role in chromosomal stability. The predicted protein shares residue identity with the guanosine 5'-triphosphate binding domain of the *Saccharomyces cerevisiae* rad51 and rad57 proteins involved in recombinational repair. Sequence analyses of several candidate cDNAs for the *XRCC2* gene also show similarity to the same domain in these proteins. Sequence analysis of the *XRCC3*-containing cosmid identified a kinesin light chain gene physically linked to a DNA repair gene.

LBL: Scaling Up a Directed Approach

For the past year, the human genome center at Berkeley Lab has focused on building new mechanisms and technologies for large-scale human genome sequencing. The center consists of four groups: genome sequencing and technology, human physical mapping and biology, automation, and informatics.

The initial sequence target in the human genome consists of a 10-Mb growth-factor-rich region located at 5q31-q35. A collaborative project with the Berkeley-based *Drosophila* Genome Center uses technologies developed at LBNL to physically map and sequence the organism's genome. A new project with the Resource for

Molecular Cytogenetics at LBNL and University of California, San Francisco (UCSF), will sequence a 600-kb region on human chromosome 20.

In 1995 Berkeley Lab completed almost 1.6 Mb of sequence, with over

700 kb primarily from human chromosome 5. Total sequence generated during the last 4 years is 3.7 Mb, with over 1 Mb of human sequence.

All sequence is double stranded, and the error rate is less than 1 in 2500 bases. Researchers found the physi-

cal map based on P1 clones (average insert size, 80 kb) to be an excellent substrate for genomic sequencing. All human sequence was obtained by sequencing 3-kb subclones derived from the P1 physical-mapping clone set that spans the target region on chromosome 5. Sources of P1 include a chromosome 5 map generated by Eddie Rubin and Jan-Feng Cheng at LBNL and a chromosome 20 map by Joe Gray's group at LBNL and UCSF.

Chris Martin, head of the production sequencing group, described scale-up strategies for the directed sequencing approach, in which every sequencing template is first mapped to a resolution of 30 bp. The advantages of this approach include a large reduction in the number of sequencing reactions needed and in the sequence-assembly steps that follow. A key challenge, Martin noted, was the development of management and training structures that could scale up to the level needed.

Directed Sequencing Strategy

Four modular, highly adaptable components have been developed for the directed approach, with data quality monitored at each step. Most decision making has been automated. The approach involves shearing and subcloning the physical-mapping clone, end sequencing 192 of these subclones, and generating a minimal tiling path of subclones using custom software. Subsequent steps consist of generating and mapping transposon inserts in each subclone in the minimum tiling path and sequencing using commercial primer-binding sites engineered into the transposon. The sequence is then assembled using the high-resolution physical-mapping

Chromosome 19 Map: Status and New Applications

Emilio Garcia summarized the state of the LLNL high-resolution chromosome 19 physical map and discussed its value for gene hunting and detailed analyses of genome organization.

Physical Map

The cosmid-based physical map now consists of 32 islands, 3 in the q arm and 29 in the p arm, with average gap sizes of 90 and 70 kb, respectively. A unique map feature is its metric backbone, in which distance between ordered contigs was estimated using FISH; gap distances are known. The contigs are being converted to an *EcoR* I map (now covering 42 Mb, or 83% of the chromosome) that provides the minimum tiling path for reducing redundant sequencing.

Now spanning 90% of chromosome 19 euchromatin, the comprehensive map has been enriched with the addition of over 170 genes, 176 cDNAs, 315 STSs, 135 polymorphic markers, over 400 YACs, and 45 Mb of *EcoR* I restriction mapping.

Applications

In a collaborative project with researchers in France and South Africa, LLNL supplied cosmid clones every 200 kb across a chromosomal region identified in population studies as associated with a cardiac conduction disease (progressive familial heart block). These cosmids will enable researchers to search for polymorphic markers across the region. The LLNL map also provides some potential candidate genes to be tested.

Garcia pointed out that the map can serve as a bridge from cytogenetic knowledge to molecular analysis. For example, researchers have been analyzing a translocation (chromosomal exchange) between chromosomes 6 and 19 that is associated with a hereditary renal dysplasia. Searching large numbers of cosmid clones across a chromosomal region led to a particular translocation being localized fairly quickly to a specific cosmid. The translocation region in the cosmid was sequenced and found to have high homology to the *USF2* (upstream stimulatory factor 2) gene. *USF2* is a ubiquitously expressed factor implicated in the expression of several tissue-specific or developmentally regulated genes. ◊

information produced by the preceding steps.

A faster variation of the process was proposed in which a mostly single-stranded, or scaffold, sequence based on more widely spaced (600- to 1000-bp) transposons will be constructed. The data from the end-sequencing step will then be layered into this mapped, verified, and transposon-based scaffold sequence and used to develop completed sequence at even lower cost.

Sequence Analysis

Sam Pitluck of the informatics group at Berkeley Laboratory described software tools developed in collaboration with Gene Meyers and Susan Larson (University of Arizona). The Fragment Assembly Kernal (FAK, written by Meyers) was chosen for its ability to handle up-front mapping information as constraints. The group built an interface to FAK using SPACE (Sequencing Platform using ACE), a variant of the ACeDB suite of database, analysis, and display software (Durbin and Mieg, 1991). ACeDB, developed originally for the *C. elegans* genome research community, has been used mostly as a database program. In SPACE the capabilities of trace editing, assembly, and fragment and contig display were added. SPACE is now being used in the production environment (<http://www-hgc.lbl.gov/homes/pitluck/spitluck.html>).

Kelly Frazer, a recipient of a DOE Human Genome Distinguished Postdoctoral fellowship, described

LBL Web Site

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

efforts to probe the primary sequence data from the Berkeley production group to discover genes and catalog expression patterns. The

team is now analyzing 1.2 Mb from 5q31 that contain the interleukin cluster growth-factor genes.

Cutting Costs

An ongoing task at the center is to identify and reduce expensive commercial costs and the steps requiring human intervention. Toward these goals, changes include using the ABI catalyst, with its low-volume pipetting abilities, combined with the ABI 377, which can detect low sample amounts. These changes, along with discounted bulk purchases of disposables over the past 2 years, have reduced supply costs by half. Another important approach is the development of custom automated devices that reduce sequencing labor costs. The group also plans to change from commercial to custom size standards that are cheaper and work better with the imaging station. Operating costs are expected to fall to less than \$0.25 per base fairly soon.

Joint Projects with Other DOE Centers

Collaborations between the Berkeley Lab and LANL are expanding. LANL's SASE end-sequencing data will be used to feed Berkeley's path-generation and subclone-sequencing components. Berkeley expects to develop similar collaborations with LLNL for sequencing chromosome 19.◊

Stretched to the Max: FISH Mapping on DNA Fibers

A new technique based on applying FISH techniques to linear stretched DNA molecules may help researchers resolve issues critical to large-scale DNA sequencing efforts.

Heinz-Ulrich Weier (Resource for Molecular Cytogenetics) says the recently developed technique, called quantitative DNA fiber mapping (QDFM), can help researchers construct high-resolution physical maps and minimal tiling paths, assess gap sizes and devise closure strategies, and provide quality-control checks during map- and sequence-assembly steps.

QDFM combines molecular combing techniques to attach and stretch DNA molecules across a glass microscope slide. FISH is used to hybridize fluorescently tagged probes to the straightened DNA fibers. Digital image technology records and analyzes images from the fluorescence microscope and measures the position of the DNA sequence or probe along the DNA fiber.

QDFM is rapid and provides a high spatial resolution of 1 to 2 kb, up to 1 Mb. Throughput of QDFM could be increased dramatically with automated image analysis that includes algorithms for finding the fibers, autofocus-ing, and handling multiple slides.

With as many as 20 clones combed on a single microscope slide, early results look promising, but QDFM's impact on genome research will depend on how well it scales up. The technique is amenable to automation, notes Weier, which could increase its throughput manyfold. Berkeley Lab plans to integrate QDFM into the large-scale sequencing process (<http://rmc-www.lbl.gov/> and <http://www.lbl.gov/~weier/>).◊

DOE Santa Fe '96 Abstracts

Available in Print, Searchable, or Downloadable Formats

The text of some 200 abstracts presented at the 1996 DOE Human Genome Program Contractor-Grantee Workshop (January 28-February 1, 1996) is available in 4 formats:

- hard copy (from HGMIS, see p. 20 for address)
- searchable WWW version with links to specific Web research sites (http://www.ornl.gov/TechResources/Human_Genome/publicat/96santa/intro.html)
- downloadable self-extracting zip-condensed file for PCs (<ftp://infosrv1.ctd.ornl.gov/pub/hgmis/96workshop.exe>; type the command `96workshop.exe -d` to extract the files into their correct directories)
- downloadable nonself-extracting zipped file for platforms other than PC (<ftp://infosrv1.ctd.ornl.gov/pub/hgmis/96workshop.zip>); user must have a utility program to unzip compressed files.◊

Santa Fe '96

Leaping Across Genomes

Comparing Mouse and Human DNA

Luckily for researchers trying to find and understand human genes, nature does not reinvent the wheel. Many genes that are important for basic life functions are spared major evolutionary changes, with similarities retained across species, often even the order of the genes along the chromosomes.

The mouse, having long been used as a model for genetic studies, offers a highly characterized genetic system with many established inbred strains available for study. Within the best-mapped homologous mouse and human regions, the presence and location of specific genes and gene families can be predicted in one species based on mapping results obtained in the other.

Information on gene function derived from analyzing human hereditary traits or mapped murine mutations can be applied from one species to another. Side-by-side genome sequencing enables close comparisons that provide insights into the evolutionary mechanisms underlying overall gene organization.

Some highlights follow of workshop presentations focusing on different aspects of mouse-human comparison studies.

Homologous Regions: Mouse Chromosome 7 and Human Chromosome 19

Lisa Stubbs [Oak Ridge National Laboratory (ORNL)] described collaborations with Lawrence Livermore National Laboratory. New methods are being explored to exploit mouse-human genomic relationships, using LLNL's collection of contiguous cosmid and YAC clones spanning human chromosome 19.

Work has focused initially on one of the largest regions of homology found for the two species: the proximal portion of mouse chromosome 7 and the entire long arm of human chromosome 19. Stubbs described the results of these comparative analyses.

Gene content, order, and spacing are remarkably well conserved throughout the length of this 23-cM to 29-Mb region of mouse-human homology, except for five major rearrangements clustered in two sites. Because of an almost perfect megabase-to-centimorgan relationship, mapping information can be extrapolated between maps of the two species. Recent mapping studies have been extended to include other regions, and work is under way to define borders of mouse-human syntenic segments on a broader, genome-wide scale.

ORNL investigators have also developed a highly efficient method of isolating exons and conserved regulatory sequences, using overlapping human cosmids and parallel sets of mouse P1 or BAC clones. (For reports of collaborative studies on human and mouse DNA repair genes, see the LLNL article, p. 4.)

Human Genome Postdoctoral fellows Evan Eichler (LLNL) and Mark Shannon (ORNL) discussed, respectively, the identification and characterization of three additional zinc finger genes in a 2-Mb cluster on 19p12 and structural and functional analysis of a conserved zinc-finger gene cluster. The cluster is located distal to *XRCC1* in human chromosome 19q13.2 and in the related interval in proximal mouse chromosome 7. Zinc fingers are protein regions that fold around a zinc atom and may be involved with their binding to nucleic acids.

Immune System Genes

Lee Rowen (University of Washington, Seattle) discussed the analysis of over 1 Mb of sequence from T-cell receptor (TCR) beta loci of both human and mouse [see *Science* 272, 1755-62 (June 21, 1996)]. TCRs play a major role in immunity and autoimmune disease. About half the human TCR beta locus is composed of long homologous repeats in which members of multigene subfamilies are embedded; a portion has even been translocated to another chromosome. These repeats suggest a mechanism for divergence of gene function. By contrast, the mouse locus contains far less repeated DNA. TCR beta variable gene segments in human are twice as numerous as in mouse, even though both species have about the same number of subfamilies.

The most surprising result, Rowen said, was finding two gene families inhabiting the same genomic address: the genes for human and mouse TCRs and for pancreatic trypsinogen. She discussed the evolutionary changes and possible origins of the gene families. In contrast with the situation in the variable beta gene segments described above, the mouse locus has undergone a greater expansion in the number and variety of trypsinogen genes than its human counterpart. ◊

Greg Lennon (left) of Lawrence Livermore National Laboratory and Bento Soares of Columbia University.



Santa Fe '96

Homing in on a DNA Sequence

Detection Strategies and Technologies

Several speakers reported on the rapidly expanding work to develop new technologies for detecting specific DNA sequences. These technologies will be useful in identifying disease mutations and genetic typing of genes as well as for resequencing specific genomic regions. Many investigators are currently working on scaleup for diagnostic uses.

Repeat Sequences

Charles Cantor [Boston University (BU)] described hybridization-based methods for isolating and profiling triplet-repeat DNA sequences and for general mutation detection. In one approach developed by Cassandra Smith (BU), target DNA is hybridized to an array of simple repeating sequences immobilized on magnetic microbeads and analyzed for mismatches (bulges) in the target or array. Detection of repeat length is by electrophoresis on an automated fluorescent gel reader. BU scientists are also developing chip-based methods with a detection system based on mass spectrometry, which may prove to be a practical method for comparing short sequences.

Single-Nucleotide Variations

Deborah Nickerson (University of Washington, Seattle) discussed a method for scanning STSs from the physical map to develop polymorphic markers based on single-base changes. These markers will be useful for the high-throughput genotyping of human populations needed for studying complex traits, including diseases. An estimated 5 million single-base changes (1 in 600 bp) in the human genome lead to population diversity.

Nickerson's group screened 154 STSs from the Whitehead-MIT physical map and found that 1 in 4 was polymorphic (single-base variations), with variation frequency (heterozygosity) greater than 30%. To detect sequence variations in STSs, they used Polyphred software, which detects heterozygous positions in automated sequence traces and interfaces with Phil Green's programs Phred, Phrap, and Consed

(stay@u.washington.edu). Nickerson estimates that minimal scanning of long (>250 bp) STSs would yield a new genetic marker every 2 Mb and produce more than 1500 useful diallelic markers.

Resequencing Genomes

Commenting on the 10^7 bp of sequence entered into the public sequence databases for the relatively tiny 10-kb HIV genome, George Church (Harvard Medical School) pointed out that sequencing will not end once the first genome is finished. His laboratory is developing scalable, sensitive, cost-efficient technologies for diagnostic resequencing of genomes, as well as for new sequencing efforts. Using available components, Church's group developed a system that eliminates probing and separates detection from the electrophoresis step. In collaborative efforts, the researchers have developed a rapid method to detect mass tags (using 400 different electrophoreses) on primers and clones or genomic target DNA. This method enables targeting of 100 spots per second.

Genetic Bar Codes

Mary Ann Brow (Third Wave Technologies) described a new thermostable, structure-specific enzyme that can detect mutations in clinically significant genes, including β -globin, p53, and the genes coding for drug-resistance in *Mycobacterium tuberculosis*. The enzyme Cleavase recognizes and cuts secondary structures formed in the single strands of a DNA fragment following denaturation. Presence of a mutation is indicated by a change in fragment pattern near the mutation region. The Cleavase reaction (cleavase fragment length polymorphisms, called CFLPs) can thus detect and localize mutations. Like bar codes, cleavage patterns are reproducible and can be archived electronically.

DNA Chips

Michael Pirrung (Duke University) discussed the preparation of high-

density arrays of short DNA sequences (oligonucleotides) for arrayed primer extension (APEX) in mutation detection and sequencing. Pirrung's group has developed a superior new photoremovable group for light-directed DNA array synthesis. APEX uses analyte DNA as a template and synthesized DNA in an array as a primer. These arrays can be used effectively for comparison sequencing with APEX and for analysis of gene expression with mRNA templates and reverse transcriptase. Substitution, deletion, and insertion mutations have been detected. ◊



Robert Sutherland (left) of Los Alamos National Laboratory and Carol Harger of National Center for Genome Resources.

Library Screening Services

- A copy of the 3 \times coverage RPCI-1 human genomic PAC library, developed by Pieter de Jong (Roswell Park Cancer Institute) under DOE sponsorship, has been deposited at BIOS Laboratories. The library is available as a screening service individually or in combination with FISH mapping and chromosome localization with somatic cell hybrid panels. The dual mapping results confirm gene localizations to specific human cytogenetic bands. (BIOS Laboratories: 800/678-9487 or 203/773-1450, Fax: 800/315-7435, <http://www.bioslabs.com/>)
- Genome Systems offers custom PCR or hybridization screening of three human libraries. The PAC library is from de Jong's laboratory; P1 and BAC libraries were developed using DNA from an anonymous source. Genome Systems also offers preparation and sequencing of genomic clones and chromosomal localization via FISH. (Genome Systems: 800/430-0030, Fax: 314/692-0044, genome@mo.net, <http://www.genomesystems.com/>) ◊

Santa Fe '96

Packaging the Genome

Community Resources for Mapping, Sequencing, Finding Genes

Collections of cloned DNA pieces (libraries) provide the essential starting material for genome researchers. Workshop speakers discussed improvements to the depth and quality of a virtual alphabet of clone libraries—BACs, PACs, cDNAs, HAECs, and TAR-YACs—and on their usefulness for mapping, sequencing, and functional analysis. The newer stable, large-insert vectors reduce chimerism and allow relatively easy DNA isolation and manipulation. Many groups are beginning to use these clones as sequencing substrates as well.

Progress toward creating a single, widely applicable, mapped resource was also reported (see "BAC-PAC" box, p. 11); early applications attest to its usefulness for studying the human genome from sequence to global levels.

Highlights of these presentations follow.

BACs

Mel Simon (California Institute of Technology) suggested thinking of BACs as large cosmids, or small YACs without problems. Simon spoke about the advantages of this large-insert cloning system, the current human and mouse BAC libraries, and their usefulness as mapping and sequencing reagents. BACs are versatile, have a low incidence of chimerism, and are easy to manipulate, he said. Isolation of the circular DNA and removal of host DNA is simple.

The human-insert BAC library now stands at about 280,000 clones (average insert size, 140 kb), representing 10× coverage; 170,000 currently are available from Research Genetics (800/533-4363; <http://www.resgen.com/>). This summer, 15× coverage of the human genome is expected, with 20 to 25× coverage by the end of the year. Plans are to increase the depth of the library to 30× to enable construction of optimal contig maps that can be used to select minimally overlapping BAC sets for genomic sequencing.

A mouse BAC library, made from the 129 mouse-strain embryonic stem cells used for producing transgenic mice, contains 230,000 clones (average insert size, 140 kb) and can be obtained from Research Genetics.

Mapping and Sequencing Applications. BAC libraries can be screened to obtain a reliable set of clones corresponding to a specific marker and can be used for walking along a chromosome. On a larger scale, the libraries can be pooled on plates and probed to obtain a series of clones to generate physical maps. In collaboration with the Sanger Centre (U.K.), Simon's team has been using BACs to translate chromosome 22 YAC maps into BACs and

then using ESTs to assemble the BAC clones into contigs. More than 600 BAC clones (average insert size, 145 kb) have been selected and mapped using a variety of markers (e.g., cDNAs, ESTs, STSs, cosmids); 90% of the markers gave hits in the BAC library. The clones were assembled into 120 contigs that are being verified by fingerprinting; gaps are closed by screening deeper into the BAC library with markers and BAC end probes.

ESTs from the radiation hybrid–YAC framework maps can be used as landmarks for rapidly assembling BACs to generate genome-wide BAC contig maps. Simon's group is planning to construct such BAC-EST maps, initially using 30,000 mapped ESTs or cDNAs. The resulting maps will provide high-resolution gene maps and, more important, entry points for gene finding and large-scale genomic sequencing.

BACs can be used in a variety of ways for sequencing. Some groups have been successful in sequencing BAC ends (see "Early Successes" box, p. 13). These ends are used for making STSs or primers to screen the library further to extend or verify contigs or obtain the next BAC to sequence in a walk along a chromosome. An 8 to 10× BAC library could be used to develop a library array for sequencing the human genome distributively, keeping track of where in the array a BAC fits, and choosing a minimally overlapping group of BACs. Alternatively, BAC ends can be used to walk, and from these arrays a minimally overlapping set can be selected, verified, and sequenced (<http://www.tree.caltech.edu/>).

PACs

Pieter de Jong (Roswell Park Cancer Institute) described his newer human PAC libraries based on the slightly modified bacteriophage P1 vector (pCYPAC2) he has been using for the last 2 years. The library consists of more than 440,000 individual clones arrayed in over 1200 384-well plates that have been prepared in 4 sections designated RPCI-1, 3, 4, and 5; total coverage is 16×. Stability of clones is high, and chimerism is low to nonexistent.

The RPCI-1 segment (3× coverage, 120,000 clones) has been distributed

Merck Gene Index Update

Keith Elliston (Merck & Co.) reported on the analysis of sequence data produced in the Washington University (WU)–Merck EST sequencing project. Sequencing was done on cDNA clones from Bento Soares' high-quality normalized libraries that were arrayed and distributed by the IMAGE consortium [*HGN* 6(6), 3 and 7(5), 1–2, 7]. The project's goal is to develop a nonredundant resource, the Merck Gene Index, composed of one cDNA clone per expressed human gene. More than 245,000 EST sequences from this project have been submitted to public databases by WU, representing over 91 Mb of DNA from about 150,000 clones.

Elliston described in some detail the work begun at Merck to develop the set of nonredundant sequences. The latest results on creating the preliminary index are available via WWW [IMAGE (<http://www-bio.llnl.gov/bbrpl/image/image.html>), WU (<http://genome.wustl.edu/esthmpg.html>)]. EST coverage of known transcripts is high (65% of known genes from GenBank), 5' end coverage of genes hit is relatively high (about 30%), and 3' ESTs extend known transcripts (27%). The rate of chimerism is extremely low. Commenting on the initial data-analysis results, Elliston characterized the project as informative, high quality, and worthwhile.◊

to more than 40 genome centers worldwide; library screening results in an average of 3 positive PACs per probe or marker. In situ hybridization of 250 PAC clones demonstrates little or no chimerism or instability.

RPCI-1 is available from several screening companies and noncommercial resource centers [BIOS Laboratories, Inc. (800/678-9487, Fax: /315-7435, <http://www.bioslabs.com/>); Genome Systems (314/692-0033, Fax: -0044, <http://www.genomesystems.com/>)]. In Toronto, Lap Chee Tsui's laboratory is providing free screening services for Canadians (cfdata@sickkids.on.ca); in Europe, filters can be obtained from Hans Lehrach (lehrach@mping-dahlen.mpg.de) and pools and filters from the U.K. Resource Centre (<http://www.hgmp.mrc.ac.uk/>).

Distribution of RPCI-3 (3x, 78,000 clones) is under way. RPCI-4 and 5 will be available upon request. High-density colony membranes are being distributed at cost, mainly to groups having a copy of the PAC library. Washington University and the Sanger Centre have the complete collection (16x redundancy). De Jong's group is now generating a similar PAC library from the 129 mouse strain.

Human cDNA Libraries

Bento Soares (Columbia University) described his group's latest approach to normalizing cDNA libraries, which has enabled them to

- lower the frequency of prevalent mRNAs while increasing the representation of rarer transcripts,
- preserve the representation of full-length clones, and

- minimize the representation of internal priming events within mRNAs during synthesis of first-strand cDNA.

Soares also discussed using subtractive hybridization methods to remove all sequenced clones from the collections of clones to be sequenced.

Soares shared the podium with Keith Elliston (Merck & Co.), who discussed results of sequence data analysis derived from Soares' normalized libraries (see box on Merck Gene Index).

HAECs and TAR-YACs

Human cells may prove to be more practical hosts for cloning stretches of the human genome that are unclonable or unstable in the bacterial cells used for the libraries described above.

(see HAECs, p. 12)

A Versatile Mapped BAC-PAC Resource

Julie Korenberg (Cedars-Sinai Research Institute) described the practical value of having a single resource of mapped clones for studying human genetics at the sequence, gene, chromosome, and whole-genome levels. She reported on progress toward creating such a resource and its early applications for mapping, sequencing, gene isolation, molecular cytogenetics, and genome-organization analysis.

Ultimately, the Mapped BAC-PAC Resource will represent 0.8 to 1.2x of the human genome (about 18,000 to 20,000 clones, covering over 70%) in a stable framework resource. The Resource will be based on BACs and PACs from Melvin Simon and Pieter de Jong and integrated at 1000 to 5000 loci with the radiation hybrid, genetic, and STS maps. Korenberg chose BACs and PACs because they are relatively stable, easy to use, and useful for sequencing and probes.

By using FISH, almost 4000 BACs and PACs (3638 and 242, respectively) have been assigned to regions of 2 to 6 Mb (average insert sizes, 125 to 150 kb). These BACs and PACs represent 17 to 20% of the human genome. The rate of chimerism is

very low (8% for BACs and 2.5% for PACs).

BACs have been obtained for all 24 chromosomes and 17 of the centromeres. Specific BACs for 11 telomeres and nonspecific BACs for 19 telomeres are available, and 100 BACs have been mapped to single bands in the mouse genome.

Tom Hudson (Whitehead-MIT) is screening 5000 markers against the subset of nearly 18,000 BACs; so far, agreement with MIT chromosome and region assignment has been 80% for 216 markers. All common markers are also mapped on STS and radiation hybrid maps.

Applications

- The Cedars-Sinai group's analysis of BACs from a chromosome 7 region usually deleted in Williams syndrome suggests a novel genomic structure involving clustered low-copy repetitive sequences whose arrangement may predispose to deletions. This may be a model for other conditions, such as cancers, characterized by deleted or rearranged regions. People with Williams syndrome have heart disease (due to a lack of elastin) and mental retardation.

Resource Information

<http://www.csmc.edu/genetics/korenberg/korenberg.html> and jkorenberg@mailgate.csmc.edu

- The Resource has been useful for finding areas not covered by STSs, as illustrated by megabase-size regions in a 3- to 4-Mb contig constructed by Korenberg's group for the chromosome 21 Down syndrome region.
- Several collaborative projects are supplying BACs to disease-gene hunters who use them to fill gaps in YAC maps for target regions.
- LANL researchers localized 13% of their unlinked contigs by hybridizing to half the 90 BACs supplied by Korenberg's group. Her team also sent BACs to scientists studying chromosomes 12, 19, and 22.
- Ultimately, the Resource can be used as a framework for genome sequencing or for anchoring end-sequence projects.
- For molecular cytogenetic studies, the Resource can replace chromosomal banding patterns, with a very high resolution of 1 BAC every 700 kb. These can be made into regional paints and used interchangeably to study contigs. ◊

Completely Sequenced Genomes*

- *Haemophilus influenzae* (bacterium), 1.9 Mb, completed 5/95 by TIGR
- *Mycoplasma genitalium* (bacterium), 0.580 Mb, completed 5/95 by TIGR
- *Methanococcus jannaschii* (Archaea), 2 Mb, completed 12/95 by TIGR
- *Saccharomyces cerevisiae* (eukaryote), 12 Mb, completed 3/96 by international team

*Does not include viruses.

HAECs (from p. 11)

Jean-Michel Vos (University of North Carolina) spoke about his group's second-generation vector for cloning DNA in human cells as human artificial episomal chromosomes (HAECs) with an insert range of 80 to 350 kb. This system may also be advantageous for maintaining large DNA regions as single fragments for studying human gene function and regulation of other critical genomic regions that may span large areas.

Vos' team is testing their resource for isolating and sequencing refractory contig gaps. (To obtain the resource: angemari@gibbs.oit.unc.edu; Vos' WWW site: <http://www.med.unc.edu/lccclvoslab/>).

Vladimir Larionov and Natalya Kouprina (visiting scientists from the Institute of Cytology, St. Petersburg, Russia) and Michael Resnick (National Institute of Environmental Health Sciences) reported a new approach to YAC cloning that exploits the high recombination tendency of DNA molecules during transformation into yeast cells. Transformation-associated recombination (TAR) events are used to clone out specific large regions of DNA, such as repeat sequences. With no chimerism and an average-size insert of 250 kb, TAR-YACs could prove useful for cloning gene families and specific genes from total-genome DNA as well as for filling gaps in chromosome maps. The group demonstrated simple purification of circular molecules and selective isolation of chromosomal, subchromosomal, and gene family DNA. ◊

Santa Fe '96

Sequencing: Moving Toward Production

Last year's triumphs in sequencing entire microbial genomes [*HGN* 7(1), 5 (May–June 1995)] left little doubt that the era of large-scale sequencing had begun. The latest whole-genome sequencing feat was presented at the workshop, this time for the heat-loving, methane-producing *M. jannaschii*.

These remarkable accomplishments represent an important step toward developing and optimizing the technologies and strategies needed to fully sequence the 3 billion bases of human DNA. Optimism runs high that the first human genome reference sequence can be obtained on time, even without revolutionary technical advances. However, significant improvements still are needed for increasing the accuracy and efficiency and reducing the cost of conventional gel-electrophoretic methods.

Current large-scale genome sequencing focuses on using gel-based instruments with random (shotgun) or directed strategies, often combining elements of both approaches. Pilot projects for production sequencing of human DNA concentrate on identifying and analyzing areas of known biological importance. (See articles on LANL, LLNL, and LBNL, pp. 3, 4, and 6, respectively; on the comparative analysis of human and mouse, p. 8; and on the NIH program to establish six large-scale pilot human genome sequencing centers, p. 20.)

Support for the whole-genome and other microbial sequencing efforts described below comes from DOE's new Microbial Genome Program, which aims to sequence the genomes of microbes with potential industrial, environmental, and economic importance.

In addition to microbial production-sequencing reports, speakers presented progress on technology improvements with current gel-based systems; improved gel technologies, especially capillary electrophoresis (CE); and alternative, potentially high payoff technologies such as mass spectrometry (MS). Many of these DNA-analysis technologies will feed into the rapidly expanding biotechnology industry,

where they will have broader applications in clinical diagnostics, environmental testing, industrial process monitoring, forensics, and agriculture.

Highlights of presentations follow.

Microbial Genome Sequencing

M. jannaschii

Carol Bult [The Institute for Genomic Research (TIGR)] described the whole-genome shotgun-sequencing strategy used to obtain the complete genomic sequence of *M. jannaschii*, an organism first isolated from a deep-sea hydrothermal vent in 1982. This genome is the first to be completed from the Archaea domain of life, a group of unique microbes that are genetically distinct from both bacteria and eukaryotes. The Archaea, which include methanogens, thermoacidophiles, and extreme halophiles, may represent some of the earliest forms of living cells.

For more information on TIGR projects, organisms, or software, contact Owen White (annotation coordinator, owhite@tigr.org), Granger Sutton (TIGR Assembler, grange@tigr.org), or Tony Kerlavage (bioinformatics director, arkerlav@tigr.org).

Bult identified several aspects critical to the group's success. These factors include the availability of a random genomic 2.5-kb-insert plasmid library from Gary Olsen (University of Illinois) and a representative 20-kb-insert lambda library for building a genome scaffold; high-quality sequence data from both ends of the plasmid and lambda clones (using ABI 373 and 377 automated sequencers with fluorescent technologies); and a robust sequence fragment-assembly engine (TIGR Assembler, discussed below by G. Sutton). Sequence coverage was obtained for the entire genome. Bult emphasized the importance of tightly integrating data production with tools for managing and analyzing data. Sequence annotation is now complete, and all data and clones will be available by early summer (see TIGR WWW page, <http://www.tigr.org/>).

Pyrococcus furiosus

Robert Weiss (University of Utah) described a project to sequence the 2-Mb genome of *P. furiosus*, another of the hyperthermophilic Archaea. Investigators are using a multiplexed, transposon-based directed approach with an end-sequencing strategy. In both the mapping and sequencing phases, automated devices detect enzyme-linked fluorescence from DNA hybrids on nylon membranes.

Early Successes with Sequencing BACs

BAC clones are showing early promise as sequencing substrates, and some teams are using BAC end-sequence data to build the physical maps—called scaffold sequence maps—for further sequencing.

Cecilie Boysen (California Institute of Technology) discussed her early successes in using BAC clones to obtain the complete, contiguous, 1.1-Mb sequence of the human T-cell receptor alpha/delta locus.

BACs perform with high fidelity; the entire sequence is easily obtained using the shotgun method; contamination with *Escherichia coli* sequence can be kept low; and repeat areas, as well as assembly and editing steps, are handled well by Phil Green's Phred and Phrap programs (see "Sequence Finishing," p. 15).

Boysen also discussed a method to obtain insert and sequence information directly from BAC DNA. This method resulted in an almost 100% success rate, averaging 495 bp with few errors.

She briefly outlined a strategy developed by Leroy Hood (University of Washington, Seattle), Craig Venter (TIGR), and Hamilton Smith (Johns Hopkins University). In this strategy, scientists sequence the ends of BACs from a 15- to 20-fold library to generate a sequence scaffold, pick seed clones every 10 Mb or so, and sequence them completely. The sequenced clones are then compared to the database containing all the end-sequence information, and clones that overlap the least are chosen for sequencing. ♦

Summarizing mapping progress, Weiss reported that 1.4 Mb (two-thirds of the genome) had been processed; 2500 mapped clones, representing minimal-set coverage of 0.76 Mb, have been sequenced. Mapping transposons, he noted, is much simpler than performing the base-calling, assembly, and editing steps required by other sequencing approaches, and the integrity of the transposon map encourages scaleup. The group is now in the end-sequencing phase and will piece together the genome by aligning end sequence with nucleated transposon maps and walking their way to continuity. They have about 500 kb of consensus sequence.

Borrelia burgdorferi

John Dunn [Brookhaven National Laboratory (BNL)] and colleagues are sequencing the 935-kb genome of *B. burgdorferi*, the spirochete that causes Lyme disease, to develop methods for DNA sequencing by primer walking with a hexamer library. The goal is to eliminate up-front mapping and enable a tenfold reduction in the template preparations required for a typical shotgun project. The approach involves using end sequencing and either hexamer strings or ligated hexamers (18-mers) to walk down the molecule. As a test of the method, a 35-kb *Borrelia* fasmid clone (based on the Fos vector) was sequenced by this approach. Dunn believes the group can achieve fourfold redundancy, with both strands completely sequenced. The next stage in the project, he said, is to bring online a high-volume CE system to allow faster sequencing (<http://www.bio.bnl.gov/>).

Sequencing Technologies

Researchers at the workshop reported on efforts to optimize current approaches in scaling up for multi-megabase sequencing. Goals are to minimize or eliminate some of the bottlenecks at each step in the sequencing process: template isolation, sequencing reactions, fragment separation and detection, and data collection and analysis.

Emphasis is on developing fully automated, integrated, modular systems that can process a sample from template isolation to data analysis with little or no human intervention.

Front-End Automation: The "Sequatron"

The "Sequatron" described by Trevor Hawkins (Whitehead-MIT; <http://www-genome.wi.mit.edu/~tlh/>) uses available components to automate and integrate the tasks of DNA isolation, setup of sequencing reactions, thermal cycling, and sample purification and concentration for separation on gels. The major component is an articulated CRS 255A robotic arm. Solid-phase reversible immobilization is used to isolate and manipulate the DNA on magnetic particles throughout the process.

Current throughput is 80 microtiter plates of samples (about 8000) from M13 phage supernatants or crude PCR products to sequence-ready samples every 24 hours. New enzymes, energy-transfer primers, and higher-density microtiter plates may increase throughput to 25,000 samples.

Speeding Up DNA Separation and Detection: Capillary Gel Systems

Researchers reported surprising and exciting results over the past year in developing newer CE methods for dramatically faster and higher-resolution DNA fragment separation. Advantages of CE include longer read lengths; improved heat transfer in the long, thin, polymer-filled capillaries vs that in standard slab gels; automatability; and online fragment detection.

Edward Yeung (Iowa State University) described a CE system based on novel separation, detection, and imaging techniques for real-time monitoring that he believes will soon enable sequencing of 40 Mb of DNA in a single day. A first-generation system has been constructed that uses new fluid-gel matrices in 100 capillaries that are read simultaneously in real time, compared with the much slower consecutive-reading technology available with current instruments. Most processes are automated. The team is scaling up the technology to allow parallel sequencing in up to 1000 capillaries.

The new system, the ESY9600 Multiplexed Capillary Electrophoresis DNA Sequencer, is scheduled for release this year by Premier American Technologies Corporation, which licensed the technology from DOE's Ames Laboratory at Iowa State. The system

Santa Fe '96

Capturing the Data and Making It Useful

Redesigning GDB and GSDB

The explosive growth of information and the challenges of acquiring, representing, and providing access to data pose new and monumental tasks for the large public databases. Ken Fasman [Genome Database (GDB)] and Gifford Keen [Genome Sequence Data Base (GSDB)] discussed the restructuring of GDB and GSDB to handle the flood of data and make it useful for downstream biology.

GDB

Observing that one can't scroll or BLAST through 3 billion base pairs in a meaningful way, Fasman defined GDB's future role as the coordination site for the complete electronic description of the human genome. The map, he asserted, provides an ideal framework for jumping into the sequence (<http://gdbwww.gdb.org/>).

Sequencing (from p. 15)

Future plans include developing a post-processor version of BEAUTY and providing access to information from and direct links to other databases, including organism-specific databases. The BCM group is also furnishing external analysis services to the Genome Sequence Data Base sequence annotator. Human Genome Postdoctoral fellow Mark Graves (BCM) reported on a simple database-management system for biologists to use in designing their own laboratory databases (mgraves@bcm.tmc.edu).

Gary Stormo (University of Colorado) described an approach for predicting coding regions in genomic DNA; it uses multiple types of evidence, combines them into a single scoring function, and returns both optimal and ranked suboptimal solutions. The approach is robust to substitution errors but sensitive to frameshift errors (software and information: <http://beagle.colorado.edu/~eesnyder/GeneParser.html>). Stormo's group is now exploring methods for predicting other classes of sequence regions, especially promoters. ◊

Fasman described the extensive changes made to GDB over the last 2 years that have culminated in the enhanced representation of genomic maps and gene information in GDB V6.0, which was released early this year [*HGN* 7(3-4), 13-14 and 7(5), 15].

Redesign of the database schema and front-end interfaces now provide true graphical genetic and physical map representation; direct community editing and curation, including third-party annotation; and an improved model for gene information that includes links to databases describing function, structure, products, expression, and associated phenotypes. A user can create a link from any GDB object to any other entity on the Internet. GDB plans to become the focal point for accessing information about the human genome.

Under the Hood

New technologies used in developing V6.0 include an object-oriented data model, object broker, data-driven WWW interface, and graphical interfaces for the most popular computer platforms. The new GDB architecture depends heavily on OPM developed by Victor Markowitz and colleagues at LBNL (see "GDB-LBNL" box, p. 17). GDB 6.0 data representation is captured in a schema file that drives all other pieces of software. This new architecture will enable GDB to adapt more quickly to changes in biological knowledge and representation of maps, genes, and other structures.

At the heart of the system is a Sybase database server that communicates in SQL, the relational query language. Everything from that point forward deals in complex objects, rather than in the rows and tables of a relational database.

Goals

Future enhancements will include improved map editing, an integrated editing environment, improved polymorphism and mutation representation, and integration with the specialized GSDB Sequence Annota-



Richard Mural of Oak Ridge National Laboratory and Michèle Durand of the French Consulate in San Francisco.

tor and Mouse Genome Database interfaces. To tie GDB to the evolving sequence databases, an interface is being developed to represent gene structure maps (maps of introns, exons, and regulatory regions associated with genes).

GSDB

Keen identified data acquisition, representation, and access as major issues for sequence databases.

Capturing and Annotating Data

Data acquisition is a two-part challenge, he said. Vast quantities of sequence data will be captured with custom software for bulk-submission processes; future plans include direct database-to-database communication for direct downloading of data from laboratories into GSDB. The more difficult task in data acquisition, he noted, is capturing the follow-on sequence annotation, which is usually published in print journals and subsequently "lost." This data will be crucial for studying gene expression, variation, and function. GSDB Annotator, a graphical browser and editor, is being developed to facilitate community annotation of the database. Researchers are also working to provide access to such common analysis algorithms as BLAST and GRAIL.

Data Representation: Building Whole Chromosomes

In addition to captured sequences and annotations, information needs to be generated about relationships between sequences. The data must be

maintained in a form capable of supporting complex, ad hoc queries. GSDB is working toward a model within the near future of 24 sequences for humans, one for each chromosome. As data comes in, it will be aligned to the representative sequence, which initially will have many gaps. Keen drew an analogy of GSDB as a community laboratory information-management system supporting what is essentially a multiyear,

multilaboratory, multiorganism shotgun-assembly process. Feature accession numbers will enable separation of annotation from sequences.

Data Access

Although GSDB has the tools and the structure (normalized and atomized data) to answer such robust queries as annotation relationships, problems with data quality and consistency do not allow this to be done well. GSDB

is now mounting a major effort to develop software for rationalizing the data stream as it enters the database.

GSDB has also developed an object-oriented access library that sits on top of the database. Almost all GSDB applications and the software that imports data from other databases work through this object layer. GSDB will make the object libraries and an application programming interface available to the public. Programmatic access will be through assigned accounts, and the database can be accessed either through the object libraries or directly on the table, row, and column level.

Availability

The new GSDB schema is complete and should be operational later this year. After fairly extensive alpha and beta testing, GSDB Annotator should be released at the same time on Mac and Sun, with Windows to follow. Software will be available via ftp from NCGR's Web site (<http://www.ncgr.org/>). ♦



Chris Fields of Genome Sequence Data Base and Arian Smith of the University of Washington, Seattle.

GDB-LBNL Collaboration Adds Up to 6.0

The recent overhaul of the database architecture is not only a major achievement in making GDB a richer, more complete representation of the current scientific understanding of genomic information; it also represents the culmination of a successful scientific collaboration. In an intense 2-year effort, teams at Lawrence Berkeley National Laboratory (LBNL) and GDB used LBNL's Object Protocol Model (OPM) software to develop the GDB 6.0 schema and underlying Sybase databases.

OPM provides a single high-level interface to GDB's underlying genomic databases. "OPM makes the underlying database transparent, so you don't have to worry about the database management system used for implementing the database," says Victor Markowitz, leader of the LBNL group.

Markowitz and GDB's Director of Informatics Ken Fasman began discussing a collaboration in February 1994. Even though GDB staff members were impressed with OPM's potential, they were concerned that it had been used only on smaller database projects, not something as complex as GDB with its many different

types of object classes. However, they had a strong incentive to reorganize the database to meet future needs.

"Initially, the research community using GDB was small, directly involved, and familiar with the early genomics research and its representation within GDB," says GDB's Bob Cottingham. "As the genome project expanded, the community of researchers interested in GDB grew rapidly and the need arose for a high-level data model and interface expressed in terms that were clearer to biological researchers."

Adding to the problems expected with implementing the OPM software were the challenges of coordinating a project across a continent. Electronic communication sped up the collaboration but stymied teamwork, making personal interactions important.

"We had regular meetings, a video conference, and many phone calls. When requirements changed, sometimes communications got scrambled. However, the commitment on both sides, as well as patience and trust, kept us going," says Markowitz.

Cooperation between the two organizations extended past the Fasman-Markowitz connection.

"Although Victor and I both worked very hard to make this collaboration a reality," says Fasman, "most important were the dedication and mutual respect in the working relationships among GDB staff members Stan Letovsky, Peter Li, and Krishna Palaniappan and OPM project members Amy Chen and Ernie Szeto."

Earlier this year at the DOE Santa Fe contractor-grantee workshop, GDB staff demonstrated the newly released Version 6.0. One of the primary advantages is its flexibility in allowing minor software changes to be made continuously and transparently. Previously, most changes had to be held for major new releases.

As expected, GDB 6.0 is requiring some user adjustments. "For GDB users, switching to the new version is like going from driving a car to piloting an F-16," Cottingham said. "It will take awhile for people to adapt to the new paradigms. Previously, our database modeling was all done directly at the relational level, which made it difficult to understand. OPM is a major improvement." [Murray Browne, HGMIS] ♦

International News

Yeast Genome Sequenced

Human Genome Project Reaches Major Goal

An international consortium of scientists announced at the end of April that they had achieved a major goal of the Human Genome Project—the complete sequence of a eukaryote, the single-celled *Saccharomyces cerevisiae* strain S288C. The 16 yeast chromosomes were sequenced from tip to tip with no gaps, and both strands of the DNA double helix were analyzed, resulting in an accuracy rate higher than 99.99%. The biggest surprise of the project was that more than half the genes uncovered during sequencing were previously unknown, despite decades of intense scrutiny by yeast geneticists. Another unexpected discovery was the degree of redundancy in the genome, with several genes often appearing to have homologous sequences and functions.

The full yeast sequence has been publicly available since the end of April, but access formerly was limited to the laboratories involved in the sequencing project, those in the follow-up functional analysis program (Eurofan), and companies in the Yeast Industry Platform, Brussels.

With some 12 million base pairs and 6000 genes, the eukaryotic (nucleus-containing) yeast has already provided biologists with a valuable resource for determining the function of individual human genes involved in such medical problems as cancer, neurological disorders, and skeletal disorders. "Now we know for the first time all the genes it takes to make a simple eukaryotic cell," said Mark Johnston [Washington University, St. Louis (WUSL)]. "As the human genome is sequenced, we will be able to compare human genes with those of yeast. When a similar gene is located, its function in humans [a more complex eukaryotic organism] can be deduced through experiments with yeast, which is much more amenable to genetic manipulation."

Over the next few years, scientists in the United States and Europe will piece together for the first time a comprehensive look at how all the genes function as an integrated cell system.

The quest to interpret the yeast genome began in the 1950s when Robert Mortimer (University of California, Berkeley) began genetic mapping of all the organism's genes. In the early 1980s, Maynard Olson (then at WUSL) created a physical map of the yeast genome by cloning overlapping DNA fragments. This map provided the starting point for the sequencing phase, which was initiated in 1989 when Andre Goffeau (Catholic University of Louvain-La-Neuve, Belgium) organized a group of European laboratories to take on the task.

As part of the Human Genome Project, laboratories applied large-scale automation, helping to finish the work some 2 years sooner than the scientists themselves had predicted.

Yeast Data Web Sites

- <ftp:mips.embnet.org> (directory /yeast)
- <ftp:ebi.ac.uk> (directory /pub/databases/yeast)
- <genome-ftp.stanford.edu> (directory /yeast/genome seq)
- BLAST and FASTA searches: <http://genome-www.stanford.edu/Saccharomyces/>

The project, which cost about \$30 million, involved more than 100 laboratories in the European Union (EU), the United States, Canada, the United Kingdom, and Japan. "In 1993, we made a gentleman's agreement not to compete but to divide the work among us in order to complete the sequence rapidly with as little duplication as possible," said Goffeau, who coordinated the 70-laboratory EU initiative. "We agreed not to stake out any territory and, on several occasions, DNA fragments to be sequenced were redistributed according to the respective capabilities of the sequencing teams."

Europeans sequenced 55% of the genome; the Sanger Centre, 17%; WUSL, 15%; Stanford University, 7%; McGill University in Canada, 4%; and the Institute of Physical and Chemical Research (RIKEN) in Japan, 2%.

Sanger Centre and Institut Pasteur are now sequencing *Mycobacterium tuberculosis* with support from The Wellcome Trust. ◊

Large-Scale Functional Analysis of Yeast

RFA HG-96-001

The NIH National Center for Human Genome Research (NCHGR) and the National Cancer Institute (NCI) have issued an RFA for research projects that will rapidly, comprehensively, and efficiently enrich the yeast sequence with biological information. Proposed projects should take advantage of the complete *Saccharomyces cerevisiae* sequence in new global approaches to studying biological phenomena important for human health and disease. These studies should be based on technologies that use and add value to the complete DNA sequence and are efficient, cost-effective, and scalable to the entire yeast genome.

Contacts

General Genomic Research: Elise Feingold, NCHGR (301/496-7531, Fax: /480-2770, elise_feingold@nih.gov)

Cancer-Related Research: Cheryl Marks, NCI (301/496-7028, Fax: /402-1037, cheryl_marks@nih.gov)

Due Dates

- Letter of Intent: August 9, 1996
- Application: September 6, 1996 ◊

ELSI
Committee
Releases
Statement

The *Statement on the Principled Conduct of Genetic Research*, developed by the Ethical, Legal, and Social Issues Committee of the Human Genome Organisation (HUGO), was approved by the HUGO Council on March 21. The statement is based on four tenets: recognition that the human genome is part of humanity's common heritage; adherence to international norms of human rights; respect for the values, traditions, culture, and integrity of participants; and the acceptance and preservation of human dignity and freedom. The full statement is available via WWW (<http://hugo.gdb.org/conduct.htm>) and from the HUGO Americas office (301/654-1477, Fax: /652-3368, hugo@gdb.org). ◊

International Large-Scale Sequencing Meeting

About 50 scientists from countries publicly supporting large-scale human genome sequencing attended an international meeting in Bermuda on February 25-28, 1996. The meeting was designed to coordinate, compare, and evaluate human genome mapping and sequencing strategies; consider the potential role of new technologies in sequencing and informatics; and discuss scenarios for data release. Attendees included representatives from and scientists funded by The Wellcome Trust, U.K. Medical Research Council, Généthon, NIH National Center for Human Genome Research, DOE, German Human Genome Programme, European Commission, Human Genome Organisation (HUGO), and the Human Genome Projects of France and Japan.

Sessions were held on sequencing strategies and resources; issues related to large-scale, high-throughput sequencing; informatics; and international data dissemination. Almost every major sequencing center in the world was represented, and initial discussions focused on sequencing strategies used by each laboratory. Most groups reported pursuing highly redundant shotgun strategies followed by directed closure, although some discussed the merits of an approach in which sequencing of any region

would be done initially using one-pass or low-pass sequencing. Considerable discussion led to the consensus that the final product representing the first human genome sequence should be done at high accuracy.

Sequencing

Only a few human chromosomes (16, 19, 21, 22, and X) clearly have enough sequence-ready clones to initiate immediate sequencing, and only partial maps are available for some of these. Thus many major groups felt that building sequence-ready maps and clones must be integrated with plans for large-scale sequencing. In many cases this would involve converting low-resolution YAC maps to sequenceable clones such as BACs, PACs, or cosmids. For chromosomes without clonal coverage, BAC maps could be generated. Both the mapped BAC and cosmid substrates were considered suitable starting points for large-scale sequencing. Attendees developed a folder of each laboratory's resources available for contribution to the public domain.

Informatics

The informatics session revealed that a variety of software tools provide alternatives to commercially available packages for sequence-data assembly and analysis. These programs, which offer substantial advantages over previous packages, generally are available free to universities and nonprofit institutes via ftp servers. Many investigators are using the new programs Phred and Phrap by Phil Green (University of Washington, Seattle), which have greatly improved accuracy and editing capabilities.

Dissemination Guidelines

The following principles were endorsed unanimously by the attendees.

- All human genomic sequence data generated by centers funded for large-scale human sequencing should be freely available and in the public domain to encourage research and development and to maximize the benefit to society.
- Sequence assemblies should be released as soon as possible; in some centers, assemblies larger

Summary of Principles

<http://hugo.gdb.org/bermuda.htm>

than 1 kb would be released automatically on a daily basis.

- Finished annotated sequence should be submitted immediately to public databases.
- These principles should apply to all human genomic sequences generated by public large-scale sequencing centers to avoid having such centers establish a privileged position in exploitation and control of human sequence information.
- To promote coordination, large-scale sequencing centers should inform HUGO of their intention to sequence particular regions of the human genome. HUGO would present this information on its WWW page and link to individual centers for more detailed information regarding specific regions. Centers could thus declare their interactions in a general framework while allowing detailed interrogation at the local level.

Attendees agreed that meetings of this type should be held annually to discuss new technologies and data dissemination in coordinating human genome sequencing worldwide.

A final poll was conducted at the end of the meeting to determine the sequencing targets and goals of each laboratory. If funding requested from granting agencies is forthcoming, the projected total of human sequence over the next 3 to 5 years would be about 1 billion bases or one-third the human genome. This goal depends on substantial technology improvements and cost reductions. Some laboratory representatives also indicated that they would sequence targeted regions simultaneously in human and mouse genomes. [David Smith (former director, DOE Health Effects and Life Sciences Research Division, dsmith9336@aol.com), and Anthony Carrano (Lawrence Livermore National Laboratory, carrano1@llnl.gov)] ◇

Automated DNA Sequencing Newsgroup

AUTOMATED-SEQUENCING/bionet.genome.autosequencing is a moderated newsgroup for discussion and assistance on issues related to automated DNA sequencing and analysis of such fragments as microsatellites and single-strand conformational polymorphisms. To reach any BIOSCI/bionet newsgroup via WWW (<http://www.bio.net/>), click on "Access the BIOSCI/bionet Newsgroups." then on the desired newsgroup. Users can read, reply to, or post new messages through the Web site if their browsers are configured properly to send e-mail. To subscribe by e-mail from the Americas or Pacific rim, send a message to biosci-server@net.bio.net; leave the subject line blank and enter *subscribe autoseq* in the message body. To subscribe from Europe, Africa, or central Asia, send *help* in the message body to mxt@dl.ac.uk to retrieve general server-usage instructions. ◇

NCHGR Initiates Sequencing Pilot Projects

The NIH National Center for Human Genome Research (NCHGR) recently announced a pilot study to explore the feasibility of large-scale sequencing of human DNA. This initiative, which is budgeted at \$60 million over 3 years, involves six U.S. research centers and is projected to produce the sequence of about 3% of human DNA in the first 2 years.

The pilot study is designed to show whether large-scale sequencing can be done rapidly, accurately, and cost-effectively using current strategies and variations. Groups participating in the pilot project will strive for an error rate of no more than 1 per 10,000 bases, or 99.99% accuracy, in all regions of the genome. The immediate challenge is to refine strategies needed to determine the order of the 3 billion bases in the genome, analyze the information, and present it to the rest of the biomedical research community.

Principal investigators, first-year grants, and individual project goals follow.

Mark Adams (The Institute for Genomic Research): Sequence human DNA on the chromosome 16 short arm; create DNA libraries and develop software for sample tracking, data management, and automation of sequence data assembly.

Richard A. Gibbs (Baylor College of Medicine): Explore the structure of chromosome X regions of high and low gene density and test a novel strategy to reduce the number of sequencing reactions needed to complete a region of DNA with high accuracy.

Eric Lander (Whitehead Institute for Biomedical Research): Develop an exportable robotic system, operated by a relatively small team, with the capacity to sequence human DNA rapidly, accurately, and cost-effectively; focusing first on chromosomes 9 and 17, develop automation to convert the physical map to the map required for sequencing.

Richard Myers (Stanford University): Test a directed strategy that

requires more up-front mapping but less complex computation to sequence regions of chromosomes 4 and 21; in collaboration with industrial partners, develop enzymes to improve up-front mapping and DNA chips to verify the sequence.

Maynard Olson (University of Washington, Seattle): Apply critical technologies identified by the group to sequencing regions of chromosome 7. In a complementary parallel project sponsored by DOE, Olson's group will develop additional technology and will sequence regions of other chromosomes.

Robert Waterston (Washington University, St. Louis): Sequence 60 to 100 Mb of human chromosomes 7, 22, and X; test modular management structure for large-scale genome sequencing. Collaborate with the Sanger Centre to increase productivity through improved mechanization, increased automation, and the creation of new software; and decrease costs while maintaining high accuracy.

NCHGR will require grantees to release sequencing data quickly via the WWW. Since the end of 1992, all U.S. Human Genome Project investigators have been held to a rapid data-release standard based on a set of sharing guidelines developed by DOE and NIH. This philosophy was also much in evidence at the recent sequencing meeting sponsored by HUGO in Bermuda, where researchers from major sequencing centers around the world agreed on the importance of continuing the rapid-release policy (see p. 19).◊

Researchers Reach Two Mapping Milestones

Human Genome Project researchers reached the project's large-scale genetic-mapping goals with the publication of two maps in the March 14, 1996, issue of *Nature*: a dense genetic map of the laboratory mouse constructed by Eric Lander and colleagues at Whitehead Institute-MIT and a separate linkage map of the human genome by Jean Weissenbach's group at Généthron.

The mouse map (<http://www-genome.wi.mit.edu/>) contains 7377 markers scattered along 20 pairs of chromosomes, or about 1 every 400,000 nucleotide bases. The human genetic map is composed entirely of 5264 microsatellite markers (1 every 0.7 cM) that can be assayed rapidly in large numbers by PCR.

With genetic maps for both human and mouse now essentially complete, the second phase of the Human Genome Project is concentrating on sequence-ready map production and large-scale sequencing projects.◊

Human Genome news

This newsletter is intended to facilitate communication, help prevent duplication of research effort, and inform persons interested in genome research. Suggestions are invited.

Human Genome Management Information System
Oak Ridge National Laboratory
1060 Commerce Park, MS 6480
Oak Ridge, TN 37830
423/576-6669, Fax: /574-9888,
<http://www.ornl.gov/hgmis/>

Managing Editor
Betty K. Mansfield bkq@ornl.gov

Editors/Writers/Designers
Anne E. Adamson
Denise K. Casey
Sheryl A. Martin
Judy M. Wyrick

Production Assistants
Murray Browne
Marissa D. Mills
Laura N. Yust

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

Health Effects and Life Sciences Research Division, OHER
Marvin E. Frazier, Acting Director

http://www.er.doe.gov/production/oher/hug_top.html
Contact: Daniel W. Drell
301/903-6488, Fax: -8521
Daniel.Drell@er.doe.gov or genome@er.doe.gov



In Memorium

Verne M. Chapman
Roswell Park Memorial Institute
Eugene L. Lawler
University of California, Berkeley
Nat L. Sternberg
E.I. du Pont de Nemours & Co.
John J. Wasmuth
University of California, Irvine



Genome Meetings Web Site Initiated

The DOE Human Genome Program recently initiated a WWW home page (<http://www.ornl.gov/meetings/>) to provide information on future and past conference proceedings of interest to all genome researchers. The complete evolution of some meetings could include electronic registration and abstract submission, agenda and abstract postings, selected figures, and full reports.

HGMIS will link to or incorporate text for genome-related meetings. DOE grantees who are organizing genome meetings are encouraged to contact HGMIS to add their links or text to the meetings site (see p. 20 for HGMIS addresses). ◊

Other URLs

Chromosome Abnormality Database

http://www.hgmp.mrc.ac.uk/local-data/Cad_Start.html

Ninth International Mouse Genome Conference (November 1995)

<http://www.ornl.gov/Tech.Resources/meetings/reports/meisler.html>

DOE Workshop on BACs (December 1995), Abstracts

<http://www.ornl.gov/meetings/bacpac/95bac.html>

HUGO-GDB Chromosome Editors

<http://hugo.gdb.org/editors.htm>

National Center for Genome Resources

<http://www.ncgr.org/>

Mouse Chromosome Reports

The 1995 Mouse Chromosome Committee reports are accessible via ftp as self-extracting Macintosh files (<ftp://bioweb.princeton.edu> and <ftp://informatics.jax.org>). To ftp from the MGD page (<http://www.informatics.jax.org/mgd.html/>), select "Informatics ftp server" under Other Resources and point to the "chr committees" listing (direct URL, <http://www.informatics.jax.org/doc/cc.html>). ◊

New Encyclopedia Release

Encyclopedia of the Mouse Genome version 1.0a17 for the Macintosh is now available from the Mouse Genome Informatics (MGI) Web site (<http://www.informatics.jax.org/>). New features and enhancements include support for MGD-Encyclopedia interaction via Netscape, a "zoom box" for zooming in on a mouse map, and "Preferences" dialog changes that enable users to specify an MGD host or select a browser for interaction. In addition, several problems with previous Macintosh versions have been corrected. [MGI User Support: 207/288-6445, mgi-help@informatics.jax.org] ◊

Genetic Voluntary Groups Directory

The *Directory of National Genetic Voluntary Organizations and Related Resources*, produced by the Alliance of Genetic Support Groups (AGSG), lists support groups for specific genetic conditions as well as organizations that are more broadly based. The second-edition, 178-page directory is available on the Web (<http://medhlp.netusa.net/agsg/agsgup.htm>) and in hard copy from AGSG; 35 Wisconsin Circle, Ste. 440; Chevy Chase, MD 20815 (800/336-GENE or 301/652-5553, Fax: /654-0171, alliance@capaccess.org). Cost, \$20 per copy plus \$2 postage prepaid; discounts. ◊

GDB Access Via WWW

The GDB Web server is available directly at the following URLs:

- **United States** <http://gdbwww.gdb.org/>
- **Australia** <http://morgan.angis.su.oz.au/gdb/gdb/top.html>
- **France** <http://gdb.infobiogen.fr/>
- **Germany** <http://gdbwww.dkfz-heidelberg.de/>
- **Israel** <http://inherit1.weizmann.ac.il/gdb/docs/gdbhome.html>
- **Japan** <http://gdb.gdbnet.ad.jp/gdb/docs/gdbhome.html>
- **Netherlands** <http://www-gdb.caos.kun.nl/gdb/docs/gdbhome.html>
- **Sweden** <http://gdb.embnet.se:443/gdb/docs/gdbhome.html>
- **United Kingdom** <http://www.hgmp.mrc.ac.uk/gdb/docs/gdbhome.html>

GDB User Support Offices

UNITED STATES

Baltimore, Maryland
help@gdb.org

AUSTRALIA

Sydney
bucholtz@angis.su.oz.au

FRANCE

Villejuif
gdb@infobiogen.fr

GERMANY

Heidelberg
gdb@dkfz-heidelberg.de

ISRAEL

Rehovot
lsprilus@weizmann.weizmann.ac.il

JAPAN

Tokyo
mika@gdb.gdbnet.ad.jp

NETHERLANDS

Nijmegen
post@caos.caos.kun.nl

SWEDEN

Uppsala
help@gdb.embnet.se

UNITED KINGDOM

Cambridge
admin@hgmp.mrc.ac.uk

GDB Links to Mammalian Homology, Enzyme Function Data

Genome Database (GDB) has reestablished links from over 1500 GDB human gene entries to mammalian homology data within the Mouse Genome Database (MGD) and to over 400 enzyme entries within the ENZYME database. Both types of links are accessible from gene entries in GDB.

Although its primary focus is the mouse, MGD's homology data also includes gene symbols, chromosomal locations, and citations regarding numerous mammalian species. Mammalian homology data can be accessed from GDB by querying for a specific human gene (e.g., *SOD1*) and following the homology link to the relevant MGD entry.

ENZYME database entries include the reaction catalyzed,

cofactors, links to the PROSITE database (protein sites and patterns), and detailed SWISS-PROT entries. ENZYME database links are from protein products rather than from gene entries themselves. For example, the enzyme link for the human *SOD1* gene can be found by querying for this gene, choosing "Protein SUPEROXIDE DISMUTASE" from the Gene *SOD1* entry, and then selecting the "EC: 1.15.1.1" link. ◊

Relevant URLs

- MGD, <http://www.informatics.jax.org/mgd.html>
- ENZYME, <http://expasy.hcuge.ch/sprot/enzyme.html>
- PROSITE, <http://expasy.hcuge.ch/sprot/prosite.html>
- SWISS-PROT, <http://expasy.hcuge.ch/sprot/sprot-top.html>

This newsletter is prepared at the request of the DOE Office of Health and Environmental Research by the Biomedical and Environmental Information Analysis Section of the Health Sciences Research Division at Oak Ridge National Laboratory, which is managed by Lockheed Martin Energy Research Corp. for the U.S. Department of Energy, under Contract DE-AC05-96OR22464. ◊

Calendar of Genome and Biotechnology Meetings*

August 1996

19-20. First Fungal Genome Workshop; Stillwater, OK [D. Brooks, 800/452-2787, Fax: 405/744-6992, dbrooks@okway.okstate.edu, <http://www.okstate.edu/artsci/microffgw/ffgw.htm>]

September 1996

6-8. Chromosome 9 Workshop; Oxford, U.K. [M. Povey, +44-171/387-7050, ext. 5043, Fax: -3496, sue@galton.ucl.ac.uk]

8-12. 4th *E. coli* and Small Genomes Meeting; Lake Arrowhead, CA [J.H. Miller, 310/825-8460, Fax: /206-3088, jhmiller@ewald.mbi.ucla.edu]

16-17. NIH Natl. Advisory Council for Human Genome Res.; Washington, DC [J. Ades, 301/402-2205, Fax: -2218, ja51b@nih.gov]

18-22. 2nd Biennial Workshop in Mouse Molecular Neurogenetics; Bar Harbor, ME [Jackson Lab., 207/288-6260, Fax: -8254, training@jax.org, <http://www.jax.org/>]

19-20. Bioinformatics and the Effective Use of the Internet for Rapid Drug Discovery; Seattle [IBC, 508/481-6400, Fax: -7911, inq@ibcusa.com, <http://www.io.org/~ibc/>]

25-29. Gene Therapy; Cold Spring Harbor, NY (abs. deadline: July 10) [CSHL, 516/367-8346, Fax: -8845, meetings@cshl.org, <http://www.cshl.org/>]

26-28. Tuskegee Univ. Conf. on the Human Genome Project; Tuskegee, AL [W. Sapp, 334/727-8961, Fax: -2069, wsapp@acd.tusk.edu, <http://www.ornl.gov/meetings/>]

30-Oct. 1. BioWest'96; San Francisco [Bio-Conf Intl., 800/524-6266, Fax: 914/245-1211]

30-Oct. 1. Gene Mutations: Detection and Analysis; Baltimore [CHI, 617/630-1300, Fax: -1325, chi@healthtech.com, <http://www.healthtech.com/conferences/>]

30-Oct. 2. German Conference on Bioinformatics; Leipzig, Germany [GCB'96, +49-341/971-6100, Fax: -6109, GCB96@imise.uni-leipzig.de]

October 1996

1-4. Chromosome X Workshop; Cambridge, U.K. [D. Bentley, +44-1223/494-891, Fax: -919, drb@sanger.ac.uk]

2-5. Intl. Forum for Genetic Engineering; Dornach, Switzerland [Conf. office, +41-61/706-4444, Fax: -4446, 100716.1756@compuserve.com, <http://www.peak.org/~armstroi/>]

3-5. 6th Intl. Workshop on Identification of Transcribed Sequences; Edinburgh [R. Mural, muralrj@ornl.gov, A. Brooks, tony@alpha.medgen.uu.se or K. Gardiner, gardiner@eri.uchsc.edu; <http://www.ornl.gov/meetings/>]

5-8. 8th Intl. Genome Sequencing and Analysis Conf.; Hilton Head, SC [TIGR, C. Sadler, 301/838-3509, Fax: -0229, seqconf@tigr.org]

6-8. Chromosome 18 Workshop; Boston [G. Silverman, 617/355-6416, Fax: -7677, silverman_g@a1.tch.harvard.edu]

8-11. 10th Intl. Mouse Genome Conf.; Paris [D. Miller, 716/845-4390, Fax: -8169, dmiller@mcbio.med.buffalo.edu]

18-20. Mapping Structure and Function of Telomeres and Centromeres; Oxford, U.K. [N. Spurr, +44-171/269-3846, Fax: -3802, spurr@icrf.icnet.uk]

20-24. Genetics and Molecular Biology of Industrial Microorganisms; Bloomington, IN [S.I.B., 703/691-3357, Fax: -7991, simhq@aol.com]

21-24. 35th Annu. Hanford Symp. on Health and Environment: Microbial Genome Research and its Applications; Richland, WA [F.E. Metting, 509/372-0317, Fax: /376-9650, fb_metting@pnl.gov]

23-24. 2nd Annual Genetic Vaccines; IBC, Washington, DC [see contact: Sept. 19-20]

25-31. Molecular Genetic Approaches to the Treatment of Genetic Diseases; Lake Tahoe, NV [Cambridge Symposia, 617/630-1399, Fax: -1395, symposia@cambridge.org, <http://www.cambridge.org/symposia/>]

26-28. Chromosome 8 Workshop; San Antonio, TX [R. Leach, 210/567-6947, Fax: -6781, leach@uthscsa.edu]

26-29. NSGC 15th Annu. Educ. Conf.; San Francisco [B. Leopold, 610/872-7608, Fax: -1192, beansgc@aol.com]

26-30. 1996 AMIA Annual Fall Symposium (formerly SCAMC); Washington, DC [AMIA, 301/657-1291, Fax: -1296, denise@amia2.amia.org, <http://lamia2.amia.org/96ctoc.html>]

28-29. Gene Functional Analysis; CHI, San Diego [see contact: Sept. 30-Oct. 1]

29-Nov. 2. ASHG; San Francisco [M. Ryan, 301/571-1825, Fax: /530-7079]

November 1996

4-5. Gene Localization; CHI, San Diego [see contact: Sept. 30-Oct. 1]

4-6. 2nd IEEE Symposia on Intelligence in Neural and Biological Systems; Washington, DC [N.G. Bourbakis, 607/777-2165, Fax: -4464, bourbaki@binguns.cc.binghamton.edu]

7-8. Natl. Conf. on Preparing Schools for the Genetic Revolution; Lincoln, NE (abs. deadline: April 1) [G. Wright, 402/472-8881, Fax: -8412, gwright@unl.edu, <http://nncf.unl.edu/confcall.html>]

14-17. 4th Meeting of European Working Group on Human Gene Transfer and Therapy; Leiden, Netherlands (abs. deadline: Sept. 15) [H. van Gennep, +31-71/514-8203, Fax: /512-8095, r.c.hoeben@biochemistry.medfac.leidenuniv.nl]

21-22. European Biotechnology Symposium '96; Cologne, Germany [BioConferences Intl., 301/652-3072, Fax: -4951]

December 1996

7-11. 6th Intl. Congress on Cell Biol. and 36th American Society for Cell Biology Annu. Meeting; San Francisco (abs. deadline: June 14) [ASCB 301/530-7153, Fax: -7139, congress@ascb.faseb.org]

January 1997

6-9. Pacific Symposium on Biocomputing; Kapalua, Maui, Hawaii (abs. deadline: July 1) [F. De La Vega, 525/747-7000, ext. 5355, Fax: -7100, fvega@gene.cinvestav.mx, <http://www.cgl.ucsf.edu/psb/>]

12-16. Plant and Animal Genome V Conf.; San Diego, CA (abs. deadline: Nov. 4) [D. Scherago, 212/643-1750, Fax: -1758, pag5@schera.com]

20-22. 1st Annual Intl. Conf. On Computational Molecular Biology; Santa Fe, NM (abs. deadline: July 1) [S. Istrail, 505/845-7612, Fax: -7442, scistra@cs.sandia.gov, <http://www.cs.sandia.gov/recomb97>]

March 1997

6-8. HGM '97; Toronto [HUGO, 301/654-1477, Fax: /652-3368]

23-27. Electrophoresis '97; Seattle [D. Wiley, 913/843-1221, Fax: -1274, dwiley@allenpress.com]

31-April 3. 11th Intl. Conf. on Math. and Computer Modeling & Scientific Computing; Washington, DC (abs. deadline: Oct. 31) [X.J. Avula, 573/341-4585, Fax: /364-3351, avula@umr.edu]

April 1997

4-5. 15th Annual SERGG Meeting; Atlanta [M. Lane, 404/727-5844, Fax: -5783, mrl@rw.ped.emory.edu, <http://www.cc.emory.edu/PEDIATRICS/sergg/meeting/meeting.htm>]

14-19. Genes and Gene Families in Medical, Agricultural, & Biological Research: 9th Intl. Congress on Isozymes; San Antonio, TX [Barr Enterprises, J. Cunningham, 301/898-3772, Fax: -5596] ◊

Training Calendar*

September 1996

15-27. Methods in Molecular Biology; Corvallis, OR [K. Field, 541/737-1837, Fax: -0496, fieldk@bcc.orst.edu]

October 1996

10-23. YACs in Structural and Biological Genome Analysis; Cold Spring Harbor, NY [CSHL, 516/367-8346, Fax: -8845, meetings@cshl.org, <http://www.cshl.org/>]

15-29. Computational Molecular Biology; Chapel Hill, NC [W. Litaker, 919/966-1730, Fax: -6821, litaker@med.unc.edu]

November 1996

6-19. Molecular and Cell Biology of *S. Pombe* and Other Yeasts; CSHL, Cold Spring Harbor, NY [see contact above: Oct. 10-23]

15-16. Practical Biotechnology for the Teaching Laboratory; Norwalk, CA [J. Boyle, 310/860-2451, ext. 2682, Fax: /467-5005, jsbHITEK@aol.com]

January 1997

24-25. Applications of Biotechnology for Society; Norwalk, CA [see contact: Nov. 15-16] ◊

Extended calendars and a list of organizations offering training are available at <http://www.ornl.gov/hgmis> or from HGMIS (see p. 20 for contact information).

*Dates and meeting status may change; courses may also be offered at other times and places; check with contact person.

**Attendance is either limited or restricted.

ELSI Task Force Seeks Genetic-Testing Experiences

The Task Force on Genetic Testing, part of the DOE-NIH Working Group on Ethical, Legal, and Social Implications of Human Genome Research, is examining the development and provision of genetic tests in the United States. The task force seeks to learn about specific experiences, both good and bad, of laboratories providing tests and of physicians, genetic counselors, nurses, consumers, and others who order or receive test results.

Topics include but are not limited to informed consent, laboratory quality, use of genetic tests by nongeneticists such as primary care physicians, communication and counseling about test results, conflicts of interest in ordering tests, and the role of institutional review boards in using tests. Specific names of laboratories, organizations, and individuals should be omitted. [Contact: Neil Holtzman; 550 N. Broadway, Ste. 511; Baltimore, MD 21205 (410/955-7894, Fax: -0241, holtzman@welchlink.welch.jhu.edu)]

Interim Principles on Web

The task force has developed a set of interim principles in three areas: scientific validation of new tests, laboratory quality, and education and counseling related to delivery of test results. These interim principles, which are being made public to give interested parties the opportunity to comment, are on the Web (<http://linfonet.welch.jhu.edu/policy/genetics/>). Further information may be obtained from staff attorney Joshua Brown (address above, jbrown@welchlink.welch.jhu.edu).

Guidelines on Ethical Issues

The correct contact for obtaining the *Guidelines on Ethical Issues in Medical Genetics and the Provision of Genetics Services* [HGN 7(5), 14 (January-March 1996)] is Victor Boulyjenkov (Fax: +41-22/791-0746, boulyjenkov@who.ch). The 117-page document was published in 1995 by the Hereditary Diseases Programme of the World Health Organization (WHO) but is not a formal WHO publication. ◊

Genome Researchers Seek Cancer Gene

In a national effort, researchers in the Prostate Cancer Consortium, sponsored by the CaP CURE foundation, are building on the progress, accomplishments, and infrastructure of the Human Genome Project to find the gene for the most common cancer of American men. Prostate cancer caused an estimated 40,400 deaths in 1995, and 244,000 new cases were diagnosed.

Headed by Leroy Hood [University of Washington, Seattle (UWS)], an investigator in the Human Genome Project, the consortium's goals are to differentiate various forms of prostate cancer, determine the most effective methods of treatment for each, and eventually find a cure.

CaP CURE (Cure of Cancer of the Prostate) was founded in 1992 by Michael Milken after the former financier was diagnosed with the disease.

The consortium will identify and study families with an abnormally high incidence of prostate cancer, acquire and store high-quality prostate tissues for research, and support genetic and linkage studies leading to identification of the prostate cancer genes. Xenograft models will be used to develop extensive cDNA libraries of potential diagnostic and therapeutic compounds.

Participating consortium institutions and their research functions are listed below.

Tissue Banks: Dana Farber Cancer Institute, Boston; University of Texas M.D. Anderson Cancer Center, Houston; Washington University School of Medicine, St. Louis; and UWS

Epidemiology: Fred Hutchinson Cancer Center, Seattle

Genetic Mapping: Fred Hutchinson Cancer Center; Whitehead Institute-MIT Center for Genome Research

Expression Mapping: UWS; University of California, Los Angeles

For more information on the consortium, contact Leroy Hood (206/616-5014, Fax: /685-7301, tawny@washingt.edu) [Anne Adamson, HGMIS] ◊

U.S. Genome Research Funding

Investigators wishing to apply for funding are urged to discuss projects with agency staff before submitting proposals.

DOE Office of Health and Environmental Research (OHER) Human Genome Program

- Contact for funding information or general inquiries: genome@er.doe.gov or 301/903-6488
- Relevant documents: http://www.er.doe.gov/production/ohher/hug_top.html

Alexander Hollaender Distinguished Postdoctoral Fellowships (DOE)

Research opportunities are available in energy-related life, biomedical, and environmental sciences, including human genome, global change, and supporting disciplines.

- Next deadline: January 1997
- Contact: Barbara Dorsey, Oak Ridge Institute for Science and Education (423/576-9975, Fax: /241-5219)

NIH National Center for Human Genome Research (NCHGR)

Program announcements are listed in *NIH Guide for Grants and Contracts* (<http://www.nih.gov/>).

- NCHGR Program Contact: 301/496-7531, Fax: /480-2770, <http://www.nchgr.nih.gov/> (see p. 20)
- ELSI: 301/402-4997

Small Business Innovation Research (SBIR) Grants

DOE and NIH invite small business firms (less than 500 employees) to submit grant applications addressing the human genome topic of SBIR programs. The two agencies also support the Small Business Technology Transfer (STTR) program to foster transfers between research institutions and small businesses. Contacts:

- Kay Etzler; c/o SBIR Program Manager, ER-16; DOE; Washington, DC 20585 (301/903-5867, Fax: -5488)
 - Bettie Graham (see contact, NCHGR). NIH SBIR due April 15, August 15, and December 15. STTR, December 1
- National SBIR/STTR conferences: Washington, DC (Oct. 28-30, 1996); Anaheim, CA (Nov. 13-15, 1996). Conference hotline: 407/791-0720; electronic registration: 203/379-9427. ◊

NIST ATP Call for Proposals

The National Institute of Standards and Technology has issued a call for proposals in the Advanced Technology Program.

Contacts: 1/800-ATP-FUND, Fax: 301/926-9524 or /590-3053, arp@mtcf.nist.gov, <http://www.atp.nist.gov/>

Due Date: September 18, 1996. ◊

