Lawrence Berkeley National Laboratory Human Genome Center

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

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

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

Joyce Pfeiffer
Administrative Assistant

Michael Palazzolo*
Director, 1996–97

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

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

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

Sequencing

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

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

Instrumentation and Automation

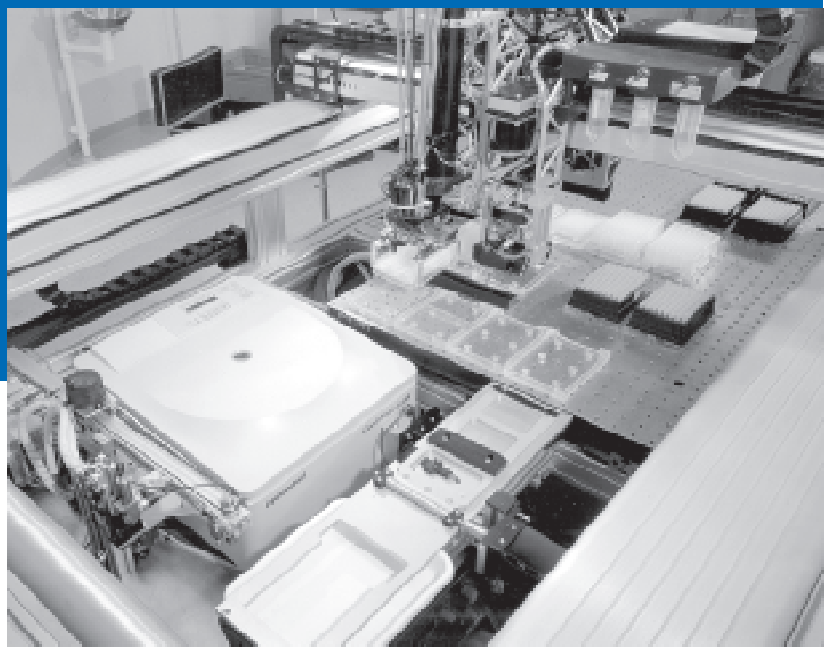
The instrumentation and automation program encompasses the design and fabrication of custom apparatus to facilitate experiments, the programming of laboratory robots to automate repetitive procedures, and the development of (1) improved hardware to extend the applicability range of existing commercial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

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

Update

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

*Now at Amgen, Inc.



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

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

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

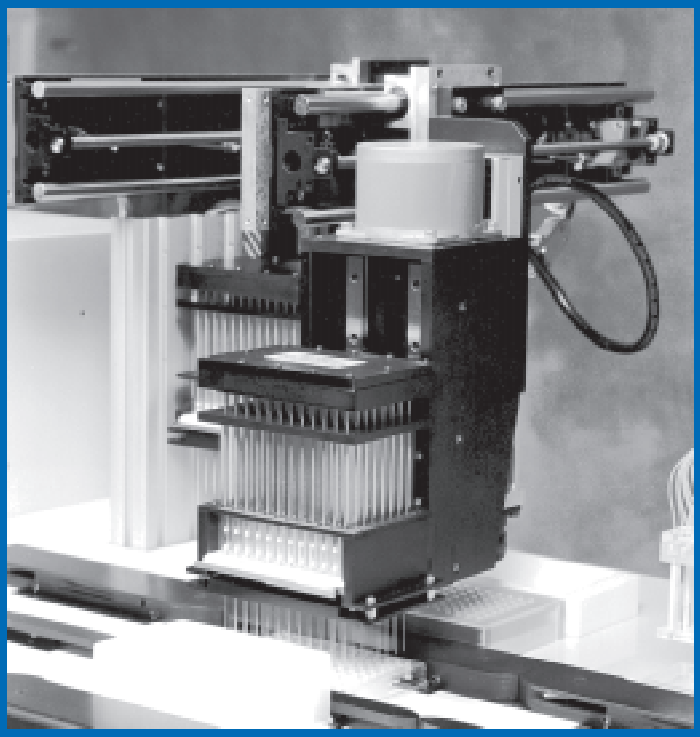
Berkeley Lab's near-term needs are for 960 samples per day of DNA extracted from overnight bacteria growths. The DNA protocol is a modified boil prep prepared in a 96-well format. Overnight bacteria growths are lysed, and samples are separated from cell debris by centrifugation. The DNA is recovered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

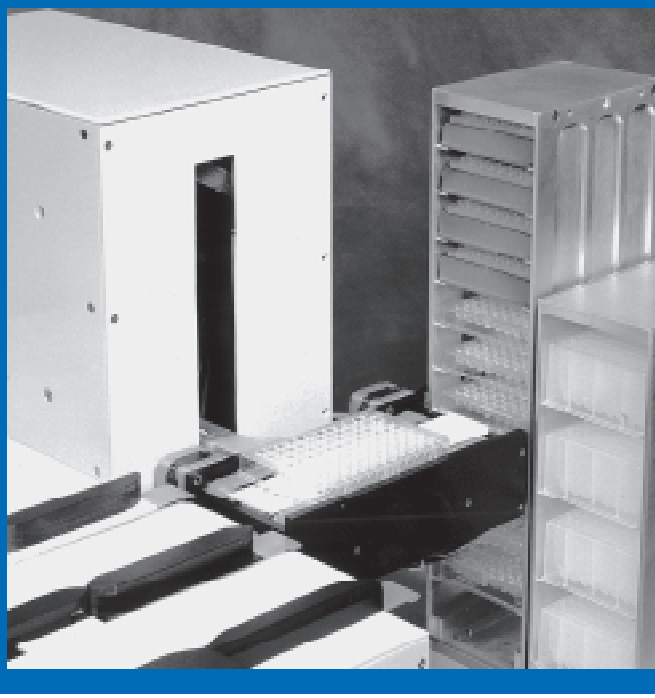
Informatics

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

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

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





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

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

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

Progress to Date

Chromosome 5

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

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

Chromosome 21

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

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

Transgenic Mice

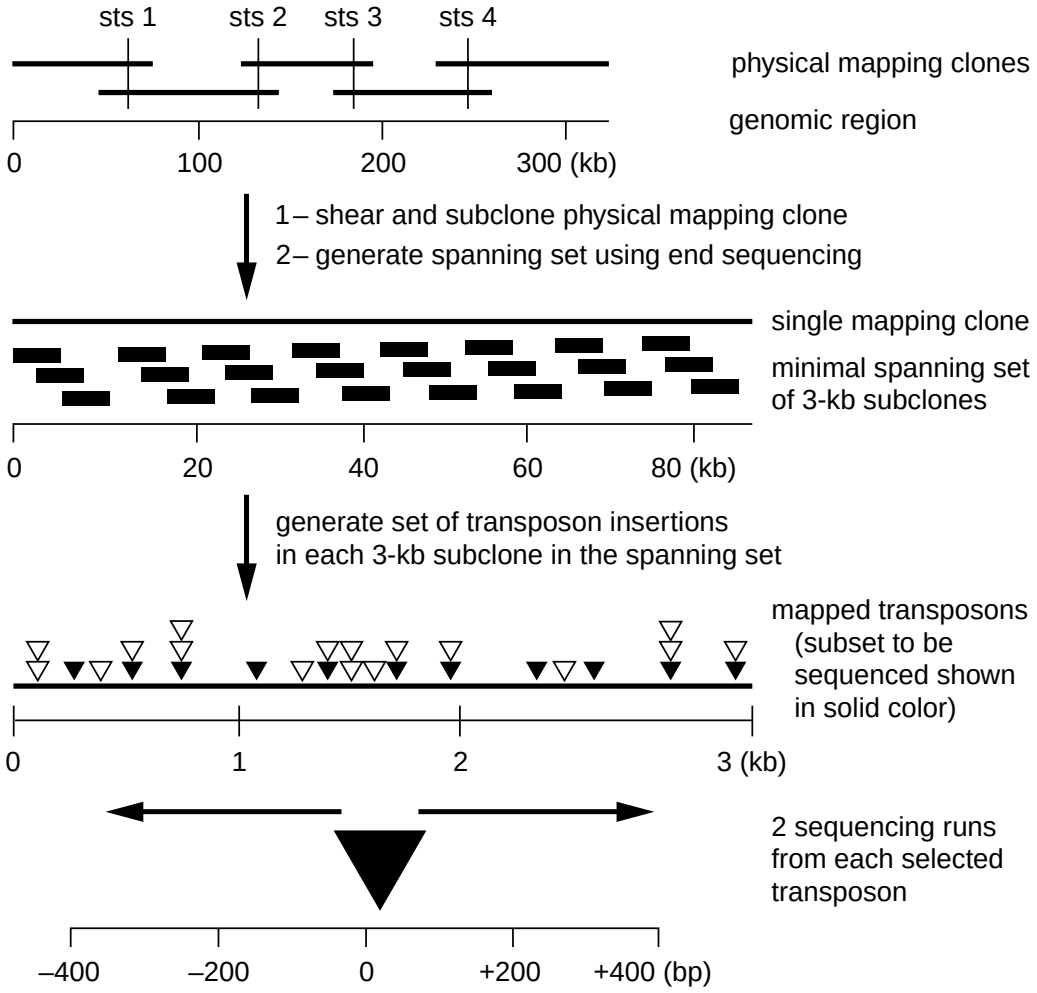
One of the approaches for determining the biological function of newly identified genes uses YAC transgenic mice. Human sequence harbored by YACs in transgenic mice has been shown to be correctly regulated both temporally and spatially. A set of nonchimeric overlapping YACs identified from the 5q31 region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

Resource for Molecular Cytogenetics

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

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

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



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

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

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

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

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

Production Sequencing

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

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

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

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

Technology for Large-Scale Sequencing

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

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

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

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

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

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

High-Resolution Physical Mapping

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

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

Ultimate Goal

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

Genome Database

http://www.gdb.org

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

Stanley Letovsky
Informatics Director

Robert Cottingham
Operations Director

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

David Kingsbury
Director, 1993–97*

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

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

A New Data Model

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

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

Direct Community Data Submission and Curation

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

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

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

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

*Now at Chiron Pharmaceuticals, Emeryville, California

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

Improved Map Representation and Querying

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

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

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

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

A Variety of Searching Strategies

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

Querying by Object Directly from GDB's Home Page

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

Querying by Region of Interest

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

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

Querying by Polymorphism

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

Work in Progress

Mapview 2.3

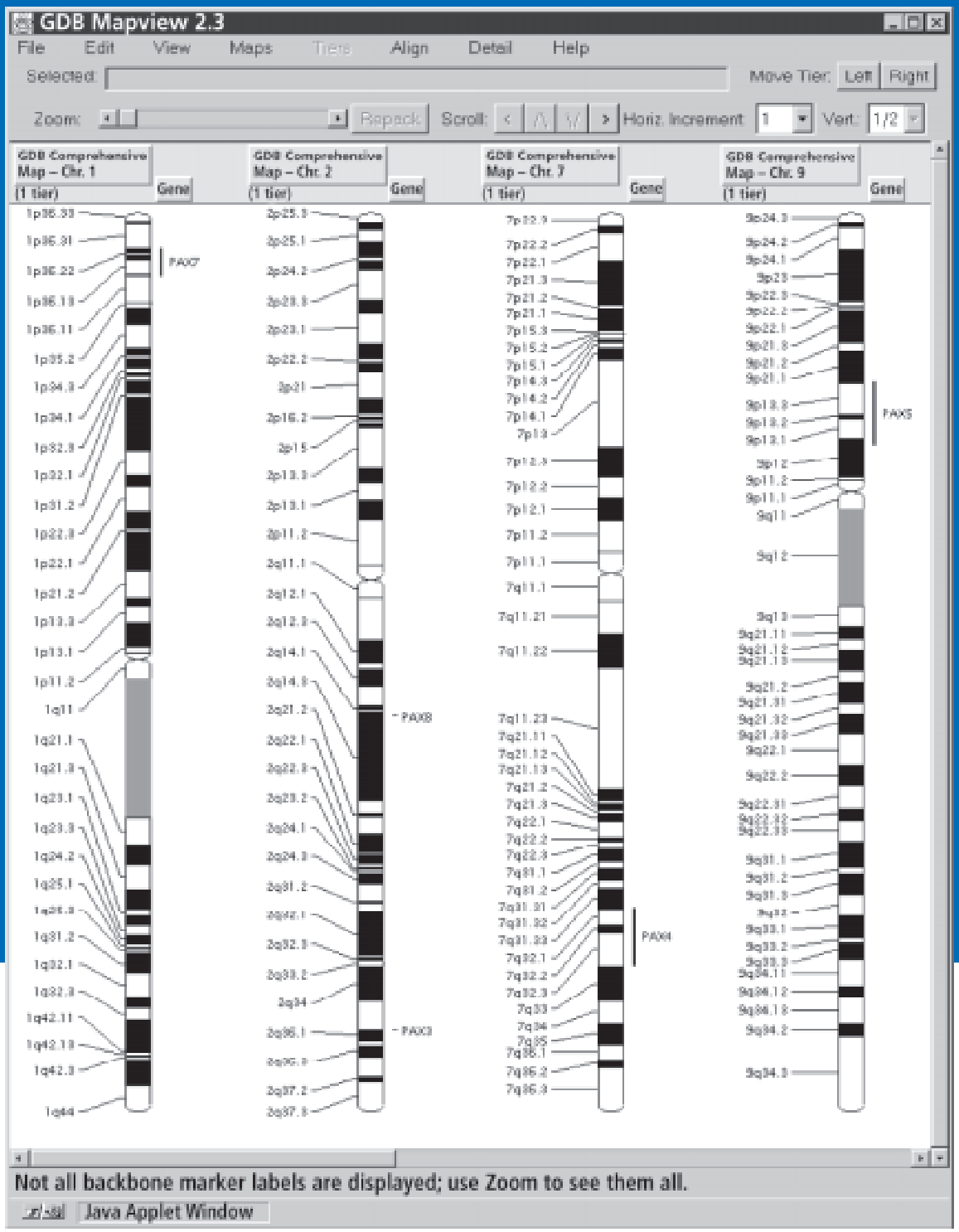
Mapview 2.1, the next generation of the GDB map viewer, was released in March 1997. The latest version, Mapview 2.3, is available in all common computing environments because it is written in the Java programming language. Most important, the new viewer can display multiple aligned maps side by side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

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

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

Variation

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

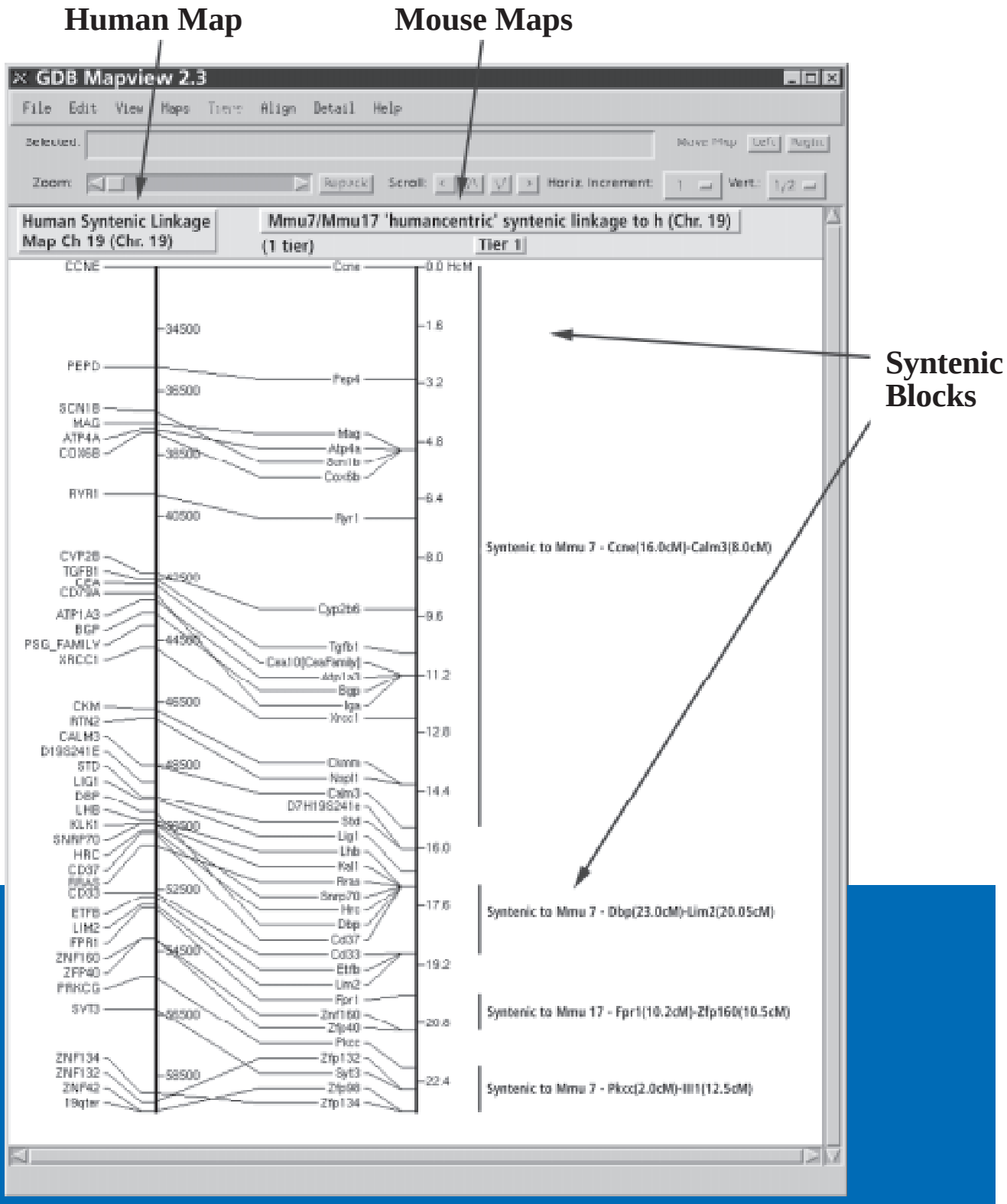


Graphical Display of Results of Query for Genes with Names matching "PAX*." [Source: Robert Cottingham, GDB]

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

Mouse Synteny

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



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

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

Collaborations

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

Sequencing Progress

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

Outreach

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

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

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

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

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

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

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

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

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

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

Genome Sequence DataBase

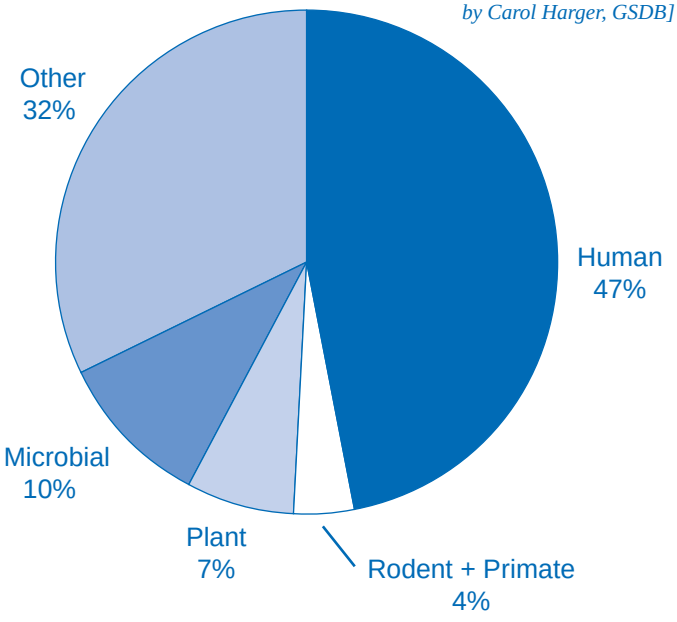
GSDB is a relational database that contains nucleotide sequence data (see pie chart) and its associated annotation from all known organisms (*http://www.ncgr.org/gsdb*). All data are freely available to the public. The major goals of GSDB are to provide the support structure for storing sequence data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

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

GSDB Enhancements

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

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



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

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

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

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

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

New Tools

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

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

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

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

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

Future Plans

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

Microbial Genome Web Pages

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

Genetics and Public Issues Program

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

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

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

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