

GDB: the Human Genome Database

Stanley I. Letovsky*, Robert W. Cottingham, Christopher J. Porter and Peter W. D. Li

Division of Biomedical Information Sciences, Johns Hopkins University School of Medicine, Baltimore, MD 21205-2236, USA

Received October 2, 1997; Accepted October 3, 1997

ABSTRACT

The Genome Database (GDB, <http://www.gdb.org>) is a public repository of data on human genes, clones, STSs, polymorphisms and maps. GDB entries are highly cross-linked to each other, to literature citations and to entries in other databases, including the sequence databases, OMIM, and the Mouse Genome Database. Mapping data from large genome centers and smaller mapping efforts are added to GDB on an ongoing basis. The database can be searched by a variety of methods, ranging from keyword searches to complex queries. Major functionality extensions in the last year include the ongoing computation of integrated human genome maps, called Comprehensive Maps, and the use of those maps to support positional queries and graphic displays. The capabilities of the GDB map viewer (Mapview) have been extended to include map printing and the graphical display of ad hoc query results. The HUGO Nomenclature Committee continues to curate the proposed and official gene symbols and related data in collaboration with GDB. As genome research shifts its emphasis from mapping to sequencing and functional analysis, the scope of the GDB schema is being extended. We are in the process of adding representations of gene function and expression, and improving our representation of human polymorphism and mutation.

CONTENT AND CURATION

Consistent with GDB's historical focus on mapping, the main classes of data in the database are maps, genes, amplimers, clones and polymorphisms, as well as supporting data such as references. The total database size is now upwards of 8 gigabytes. The content of GDB comes primarily from two sources: the scientific literature and submission from data producers. This latter category can be divided into unsolicited submissions and collaborative loads of what are typically large datasets. Additional curation is performed by outside specialists, including such groups as the HUGO (7) Gene Nomenclature Committee (8) and Chromosome Committee (9).

With the assistance of the Nomenclature Committee, GDB continues to curate information about human genes, including official HUGO gene symbols. On average, between 70 and 80

new genes are added to the database each month, with the total number of named genes (i.e., genes with a known function, phenotype or product) standing at almost 7200 on October 1, 1997. The database now also stores information about genes for which the function or phenotype is unknown, representing these as putative genes; this category is being used to store information about genes identified by cDNA sequencing and ESTs, including the 35 000 UniGene (10) clusters which do not contain a known gene. The UniGene clusters have been loaded into GDB, which allows them to be included in the GDB comprehensive maps. Each UniGene cluster is linked back to its parent entry on the NCBI web site.

The number of amplimers (STSs and other PCR markers) in GDB also continues to grow, both through direct submission and through large data loads. In the year ending October 1, 1997, this number grew from 47 759 to 77 450. Many groups use the same amplimers in their maps, but refer to them by different names; this complicates the task of map integration enormously. GDB is attempting to curate a complete list of amplimer names that includes all known synonyms; we hope that this will alleviate the problem. Large amplimer loads have come from the Radiation Hybrid Database (11), the Whitehead Institute (12), the Stanford Human Genome Center (13) and the Sanger Center (14). We periodically compare our amplimer data to dbSTS (15), load any amplimers not in GDB, and update links to dbSTS and the sequence databases.

The growth in the number of clones in the database (from 567 204 to 2 476 893 in the year to October 1, 1997) results largely from the loading of large-insert clone libraries covering the entire genome. These libraries are being used for large-scale sequencing projects at the sequencing centers, and in the BAC-end sequencing projects at TIGR (16) and the University of Washington (17). These clones will be assembled into contigs and linked to their entries in the sequence databases as the data become available. The libraries currently in GDB include the Caltech BAC libraries, the RPCI PAC and BAC libraries, the Genome Systems BAC library and the DuPont/Merck P1 library. We are also loading chromosome-specific genomic libraries, often in cosmid vectors, used contig construction.

The other large source of clones in GDB is the ongoing series of data loads from the I.M.A.G.E. consortium (18). These partial cDNA clones are a major source of EST sequences. The I.M.A.G.E dataset in GDB now also includes clones from the NCI Cancer Gene Anatomy Project (19).

*To whom correspondence should be addressed. Tel: +1 410 614 5636; Fax: +1 410 614 3200; Email: letovsky@gdb.org

ESTs from the dbEST database (20) are now loaded into GDB, and linked to source clones, amplimers, and known genes or transcript clusters. Users can search GDB for ESTs using their name or sequence accession number. Mapping information from the Whitehead RH maps and the RH consortium whole genome transcript maps allows the ESTs to be placed on the GDB comprehensive maps.

GDB continues to load both whole genome maps and maps of smaller regions. Whole genome maps include the Genethon and CHLC linkage maps, radiation hybrid maps from the Whitehead Institute and the Stanford Human Genome Center, the Whitehead Institute Contig maps, and the RH consortium transcript maps, and are updated in GDB as new versions become available. Large regional maps include the X chromosome contigs recently published from Washington University in St Louis, and the chromosome 7 contigs from NHGRI. We collect maps of smaller regions from the literature, and from direct submissions. These smaller maps often give more detailed information about a region of the genome, and add value to the larger maps. The information from all maps in the database is used to compile the GDB comprehensive map.

MAP INTEGRATION

GDB stores many maps of each chromosome, generated by a variety of mapping methods. When one is interested in a region such as the neighborhood of a gene or marker, it is useful to be able to see all maps that have data in that region, regardless of whether they contain the desired marker. To support querying the database by region of interest, we have developed an integrated map of the human genome which combines data from all the maps of each chromosome. These are called Comprehensive Maps. Every locus that appears on any primary map of a chromosome normally appears on the comprehensive map of that chromosome. (Exceptions to this rule arise when a primary map does not have enough markers in common with the other primary maps to allow it to be aligned to the comprehensive map.) Queries for all loci in a region of interest are processed against the comprehensive map, thereby searching all relevant maps. The comprehensive maps are also useful for display purposes since they organize the content of a region by class of locus (e.g., gene, marker, clone) rather than source of the data. This approach yields a much less complex presentation than an alignment of numerous primary maps. We continue to provide access to aligned displays of primary maps since some of the information in these maps is not always captured in the comprehensive map, including detailed orders, order discrepancies between maps, and non-linear metric relations between maps.

A comprehensive map of a chromosome is generated by selecting one map as a standard map. The distances in this map will be the basis for the comprehensive map distances, so for example if a radiation hybrid map is used as the standard the comprehensive map coordinates will approximate to centiRays. Every other map of the chromosome is warped in a non-linear way so as to bring as many of its markers as possible into coincidence with the same markers on the standard map. As each map is incorporated into the comprehensive map its elements then become available for use as common markers with the next map. The warping process involves the construction of a piecewise-linear warping function through a longest monotonic (i.e., order-consistent) chain of common markers; evaluation of the

dispersion of standard coordinates assigned to the same marker from different maps confirms that this process yields good alignments.

After the warping has been carried out, every element of every map is located in the space of the standard map. The same locus may therefore occur several times as a result of having been on several maps. Hopefully these occurrences will be near each other, although chimerism, mapping errors or data entry errors can cause discrepancies between the positions; we have developed scripts to detect these discrepancies and construct reports of mapping problems in each chromosome. The different map elements for a locus are subsequently merged in a conservative manner: coarse localizations that are consistent with finer ones are thrown out, the remaining fine localizations are merged to produce an uncertainty interval. This interval defines the comprehensive map position for the locus.

QUERYING

This section describes some of the more common queries which can be answered using GDB. Detailed instructions for carrying out these and other queries can be found at <http://www.gdb.org/nar98.html>

Querying by name, keyword or accession number

The simplest search methods available are:

- find an object with a known GDB accession number
- find objects associated with a known sequence database accession number
- find objects having a specified name. The name can include wildcard symbols (e.g., AFM147xb*), and will automatically match synonyms as well as primary names.
- find objects that contain one or more keywords anywhere in their 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, or any other mapped object. Given a region of interest, the comprehensive map is searched to find all loci that fall within it. These loci can be displayed in a table, or graphically as a slice through the comprehensive map, or as slices through a chosen set of primary maps. The comprehensive map slice shows all loci in the region, including genes, ESTs, amplimers and clones.

A region can also be specified as a neighborhood around a single marker of interest. Figure 1 shows a display of 3% of the comprehensive map centered on the TSC1 gene. The display shows the mapping of genes and amplimers in this region.

Retrieving a graphical view of locus position

The results of queries for genes, amplimers, ESTs or clones can be displayed on the GDB comprehensive map. If the results are spread across several chromosomes, several chromosomes are displayed. Figure 2 shows how a query for all the PAX genes (specified as symbol=PAX* on the gene query form) retrieves genes on multiple chromosomes (five of nine chromosomes are visible in the figure). Double-clicking on one of these genes brings up the detailed information for the gene in the Web browser.

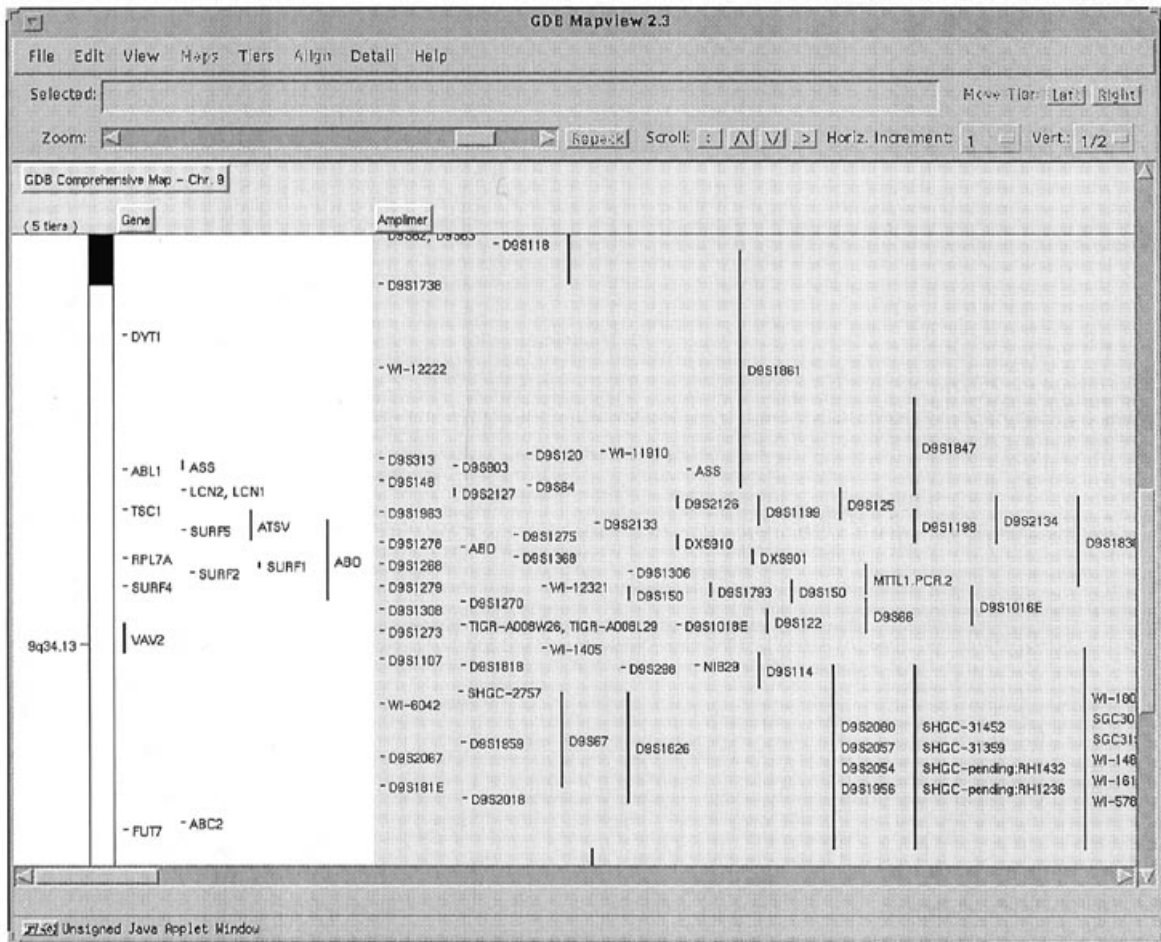


Figure 1. Neighborhood around the TSC1 gene on the Comprehensive Map of Chromosome 9.

Polymorphism queries

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

REPORTS

A number of useful reports are computed regularly from the GDB content and made available from the Web site. This list of reports is likely to grow as users request new reports which might be of use to others. The reports currently available are:

- Genetic Diseases by Chromosome: a list, for each chromosome, of genes with links to disease entries in OMIM (21). The report has active links to OMIM and GDB entries for each gene, and an option to view a 'disease map' of the chromosome.
- Lists of Genes: an alphabetical list of all genes mapped to each chromosome, and alphabetical lists of all genes in GDB.

- GDB statistics: showing the number of clones, amplicons, genes, etc. in the database.
- Human/Mouse synteny plots: comparative mapping plots showing mouse map regions that correspond to each human chromosome.
- Reports for HUGO Committees: these include lists of genes without approved symbols, genes with unreviewed cytogenetic localizations, and order conflicts between maps in the database.

To request other reports send mail to help@gdb.org

FUTURE DIRECTIONS

GDB is extending and improving its representation of the following areas:

Variation

Since its inception GDB has been a repository for polymorphism data. Currently there are >18 000 polymorphisms in GDB. A collaboration with the Human Gene Mutation Database (HGMD) (22), based in Cardiff, UK, and headed by David Cooper and Michael Krawczak, has been initiated. HGMD's extensive collection of human mutation data covers many disease-causing loci and includes sequence level characterization of the mutations.

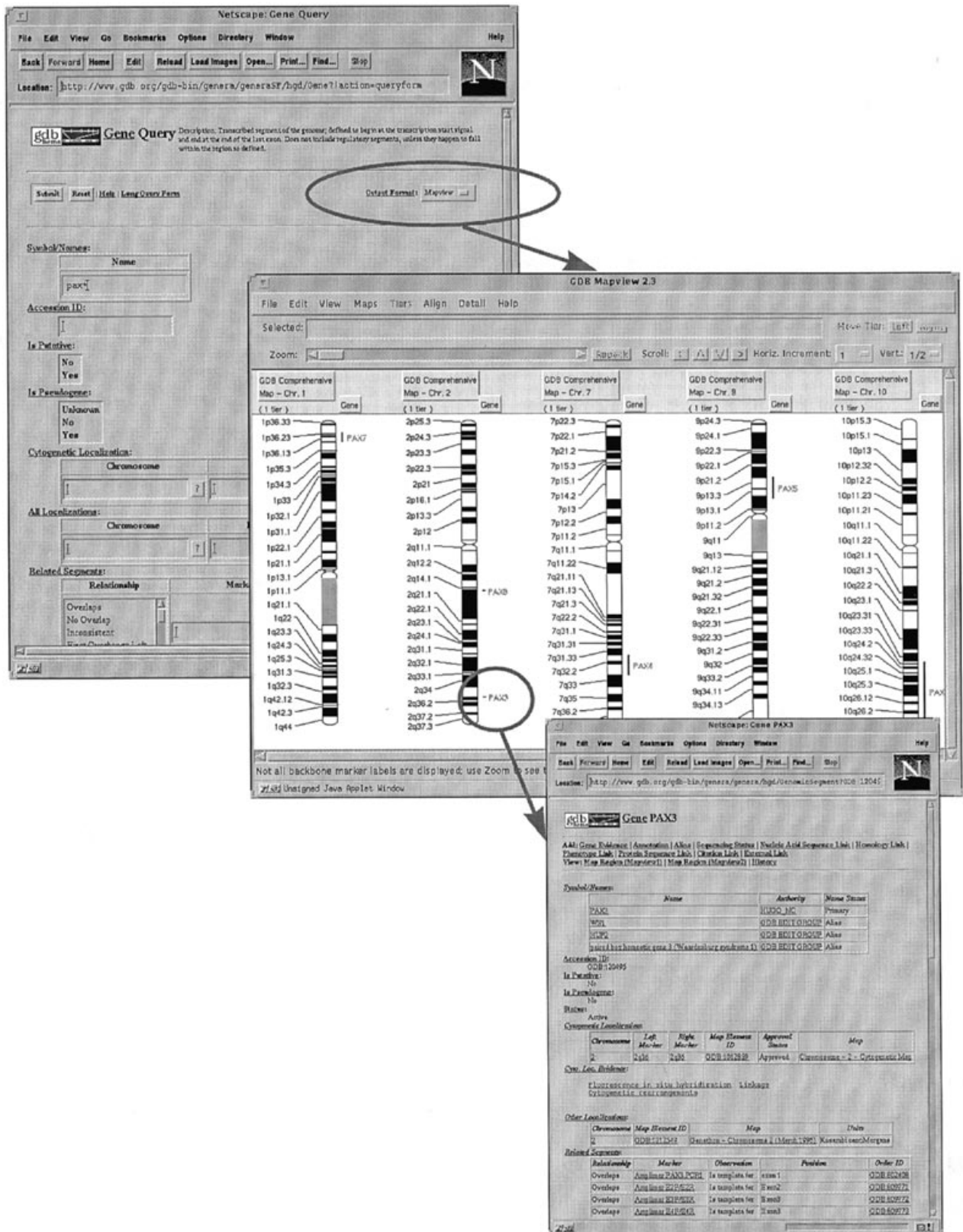


Figure 2. Graphic display of results of query for genes with names matching 'PAX**'.

Human Map Mouse Map(s)

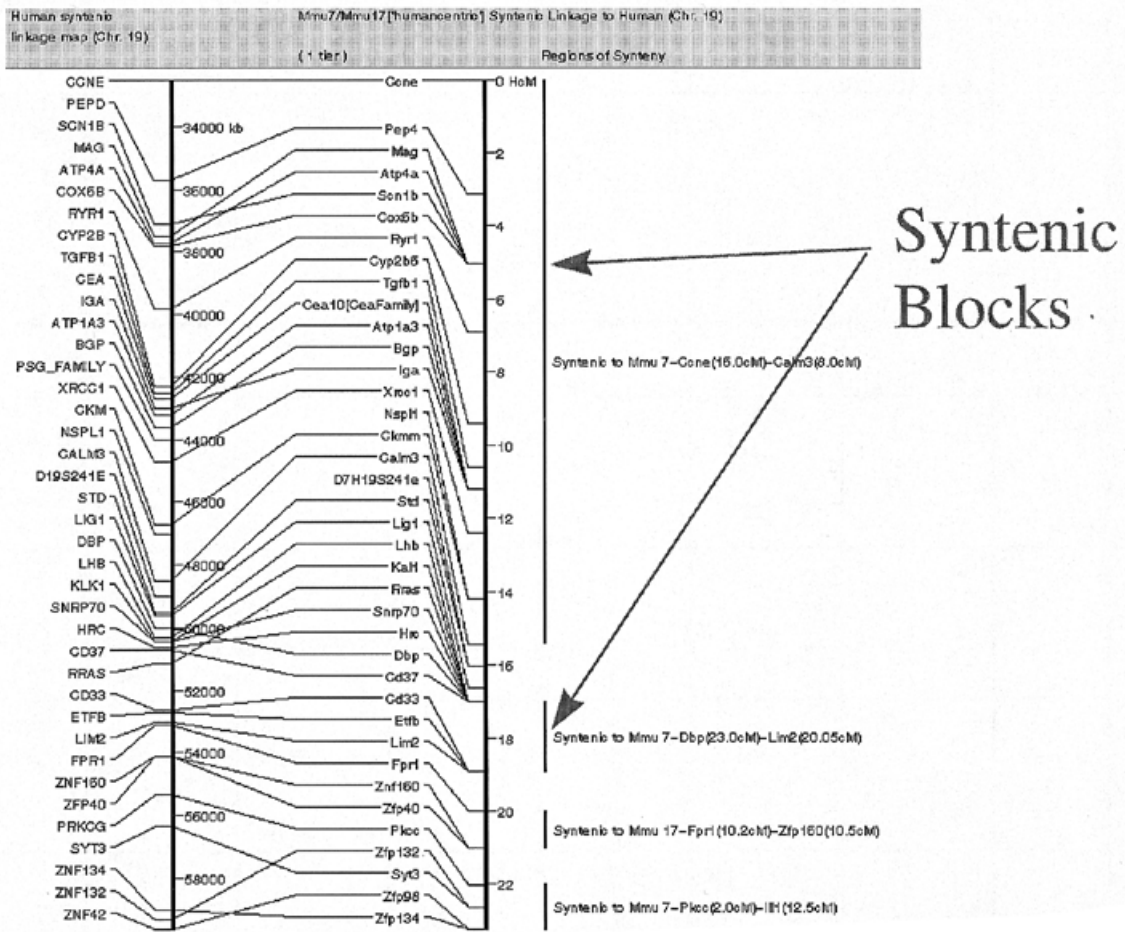


Figure 3. Rearranged mouse map aligned against human chromosome.

This dataset will be included in GDB, and updated from HGMD on an ongoing basis. The HGMD team will also provide advice on the representation of genetic variation in GDB. The representation of variation in GDB is currently 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 (SNPs), which are expected to be a major source of information on human genetic variation in the near future.

Mouse synteny

Since genomic relationships between mouse and man provide important clues regarding gene location, phenotype and function, one of GDB's goals is to enable direct comparisons between these two organisms, in collaboration with the Mouse Genome Database (MGD) at The Jackson Laboratory (23). GDB is making the additions to its schema to represent this information so that it can be displayed graphically with Mapview (see Fig. 3). In addition, algorithmic work is underway to automatically identify regions of conserved synteny between mouse and man from mapping data; these algorithms will allow the synteny maps to be updated on a regular basis. 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 (24), and similar data will be available in the database itself in the spring of 1998.

Genome Annotation Consortium

GDB is a participant in the Genome Annotation Consortium (GAC) project (25), whose goal is to produce high quality, automatic annotation of genomic sequences. Currently GDB is developing a prototype mechanism to transition from GDB's Mapview display to the GAC sequence level browser over common regions of the genome. The GAC will also establish a reference sequence of the human genome which 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 can now be recorded in GDB. The GAC will determine what regions of the genome

have been completed based on submissions to the sequence databases. GDB will also be collaborating with the European Bioinformatics Institute (EBI) and HUGO to maintain a single shared Human Sequence Index which will record commitments and status for sequencing clones or regions. As a result it will be possible to display the sequencing status of any region alongside the other mapping data from GDB.

DATA SUBMISSION

Users can edit the database directly over the World Wide Web by obtaining a user account. To obtain an account, fill out the form at <http://www.gdb.org/gdb-bin/gdb/regmail>, or contact help@gdb.org. Editing help is available online; questions may also be directed to help@gdb.org, or at the address, phone and fax numbers listed below.

For help with large submissions of data, or with producing files which can be loaded directly into GDB, contact data@gdb.org.

CONTACTING GDB

Baltimore

Many questions about GDB can be answered by documents on our Web server (<http://www.gdb.org>), or those of GDB's international sites (see below). Additional inquiries can be directed to: GDB Users Services, Johns Hopkins University School of Medicine, 2024 East Monument Street, Suite 1-200, Baltimore, MD 21205-2236, USA. Tel: +1 410 955 9705; Fax: +1 410 614 0434; Email: help@gdb.org. (Similar services are provided by each of the international sites. Please check their Web servers for local contact information.)

GDB international sites

Access to GDB is also available at the following replication sites. These sites receive daily updates via the internet. All of the data and documentation discussed in this article are available at these URLs as well.

Australia:	ANGIS, University of Sydney http://gdb.angis.su.oz.au WEHI, Melbourne http://wehih.wehi.edu.au/gdb
France:	INFOBIOGEN, Paris http://gdb.infobiogen.fr/
Germany:	DKFZ, Heidelberg http://gdbwww.dkfz-heidelberg.de/
Israel:	Weizmann Institute of Science, Rehovot http://gdb.weizmann.ac.il/
Italy:	TIGEM, Milan http://gdb.tigem.it
Japan:	JST Corp., Tokyo http://www.gdb.gdbnet.ad.jp/

Netherlands:	CAOS/CAMM, University of Nijmegen http://www-gdb.caos.kun.nl/gdb
Sweden:	Uppsala Biomedical Center http://gdb.embnet.se/gdb/
UK:	HGMP Resource Center, Hinxton http://www.hgmp.mrc.ac.uk/gdb

CITING THE GENOME DATABASE

When citing the Genome Database in the literature, please reference this article as:

Letovsky, S.I., Cottingham, R.W., Porter, C.J. and Li, P.W.D. (1998) GDB: The Human Genome Database. *Nucleic Acids Res.*, **26**, 94-99.

ACKNOWLEDGEMENTS

The GDB Human Genome Database is an international project funded by a grant from the US Department of Energy (DE-FC02-91ER61230) with additional support from the US National Institutes of Health and the Japan Science and Technology Agency.

REFERENCES

- Pearson, P.L. (1991) *Nucleic Acids Res.*, **19**, 2237-2239.
- Pearson, P.L., Matheson, N.W., Flescher, D.C. and Robbins, R.J. (1992) *Nucleic Acids Res.*, **20**, 2201-2206.
- Cuticchia, A.J., Fasman, K.H., Kingsbury, D.T., Robbins, R.J. and Pearson, P.L. (1993) *Nucleic Acids Res.*, **21**, 3003-3006.
- Fasman, K.H., Cuticchia, A.J. and Kingsbury, D.T. (1994) *Nucleic Acids Res.*, **22**, 3462-3469.
- Fasman, K.H., Letovsky, S.I., Cottingham, R.W. and Kingsbury, D.T. (1996) *Nucleic Acids Res.*, **24**, 57-63.
- Fasman, K.H., Letovsky, S.I., Li, P., Cottingham, R.W. and Kingsbury, D.T. (1997) *Nucleic Acids Res.*, **25**, 72-80.
- Human Genome Organization (HUGO), <http://hugo.gdb.org/>
- HUGO Nomenclature Committee, <http://www.gene.ucl.ac.uk/nomenclature/>
- HUGO Chromosome Committees, http://www.gdb.org/gdb/hugo_eds.html
- <http://www.ncbi.nlm.nih.gov/UniGene/>
- Radiation Hybrid Database (RHdb), <http://www.ebi.ac.uk/RHdb/>
- Whitehead Institute, <http://www.genome.wi.mit.edu/>
- Stanford Human Genome Center, <http://www.shgc.stanford.edu/>
- Sanger Center, <http://www.sanger.ac.uk/>
- <http://www.ncbi.nlm.nih.gov/dbSTS/>
- TIGR BAC end sequencing project, <http://www.tigr.org/tdb/humgen/humgen.html>
- University of Washington BAC end sequencing project, <http://serac.mbt.washington.edu/>
- I.M.A.G.E. Consortium, <http://www-bio.llnl.gov/bbrp/image/image.html>
- NCI Cancer Gene Anatomy Project, <http://www.ncbi.nlm.nih.gov/ncigap/>
- <http://www.ncbi.nlm.nih.gov/dbEST/>
- On-Line Mendelian Inheritance in Man (OMIM), <http://www3.ncbi.nlm.nih.gov/omim/>
- Human Gene Mutation Database (HGMD), <http://www.cf.ac.uk/uwcm/mg/hgmd0.html>
- Mouse Genome Database (MGD), <http://www.informatics.jax.org/>
- <http://www.gdb.org/~letovsky/maps/reports.html#synteny>
- Genome Annotation Consortium (GAC), <http://compbio.ornl.gov/CoLab/>