*Nucleic Acids Research, 1993, Vol. 21, No. 13 3003–3006*

# The GDB™ human genome data base anno 1993

A.Jamie Cuticchia, Kenneth H.Fasman, David T.Kingsbury, Robert J.Robbins and Peter L.Pearson
Johns Hopkins University School of Medicine, Baltimore, MD 21205-2100, USA

## ABSTRACT

**Version 5.0 of the Genome Data Base (GDB™) was released in March 1993. This document describes some of the significant changes to the types of data which are stored within the GDB. In addition to handling a wider scope of data, the GDB 5.0 application software now supports the X-Windows protocol. Although the GDB software still remains the most widely utilized method for accessing the data, alternate methods of access are now available, including direct SQL (Structured Query Language) queries, FTP (Internet File Transfer Protocol), WAIS (Wide Area Information Server), and other tools produced by third-party developers.**

## INTRODUCTION

The Howard Hughes Medical Institute (HHMI) and the Johns Hopkins University School of Medicine formed the Genome Data Base in June 1989 with the mission to provide informatics support for the mapping data resulting from the Human Genome Initiative. GDB Version 1 went online in September 1990. (1) Since then, the GDB has successfully obtained U.S. federal funding from the Department of Energy and the National Institutes of Health. It has also received some support from abroad.

Since its inception, the Genome Data Base has been implemented as a relational database. The database is produced under the Sybase RDBMS (relational database management system) and presently operates on Sun workstations and servers under the UNIX operating system.

### Database content and organization in GDB 5.0

Since the publication of the Genome Data Base Anno 1992 (2), the GDB software has evolved to track the needs of the genome community. As before, the data is accessed through a set of linked data managers. Figure 1 shows the relationship among the GDB data managers and OMIM (the online version of Victor A.McKusick's catalog of human genes and genetic disorders, *Mendelian Inheritance in Man)* (3). The number of data managers has doubled from five to ten in GDB 5.0. The new data managers are:

Mutation — This manager permits storage of data on individual mutations which are linked to loci and the mendelian phenotypes cataloged in OMIM.

Cell Line — This manager provides information to describe the cell lines available for breakpoints and chromosome fragments useful in mapping.

Library — This manager provides specific information on which libraries were used to generate probes. It is maintained to assist GDB users in obtaining reagents.

Polymorphism — This manager provides polymorphism data previously only available through the locus to which they were associated. Now individual polymorphisms can be retrieved directly according to a definable criteria.

Population — This manager provides information on the populations in which a particular mutation or polymorphism occurs.

Significant enhancements were also introduced for two previously existing managers:

Map — This manager provides distance information (including overlap information) and confidence limits (e.g. lod scores) are provided. There is no longer any practical limit to the number of objects in a map.

Probe — This manager now represents amplification conditions for PCR primer sets. Reagent–reagent interactions are supported in order to provide information such as which STSs are associated with which YACs.

### Data acquisition efforts

The GDB must provide data to the scientific community in a timely manner, and at the same time insure a high level of data integrity. To this end, the data within the GDB come from a variety of sources (i.e., paper and direct electronic submissions, interactive data entry into the database) and are reviewed for scientific merit by the GDB editors. The editors are chosen by the chromosome communities and appointed by the Human Genome Mapping Committee (HGMC) for their scientific knowledge of a particular area of the genome. Each chromosome has its own group of editors, as does mitochondria, nomenclature, DNA, clinical disorders, comparative mapping, and linkage.

Before the implementation of GDB 5.0, data required the approval of the GDB editors before becoming generally available. To balance the need to maintain data quality and the need to display newly submitted data as quickly as possible, the GDB has now adopted the policy of displaying data which has not yet been 'approved' by the editors (while still allowing the users the

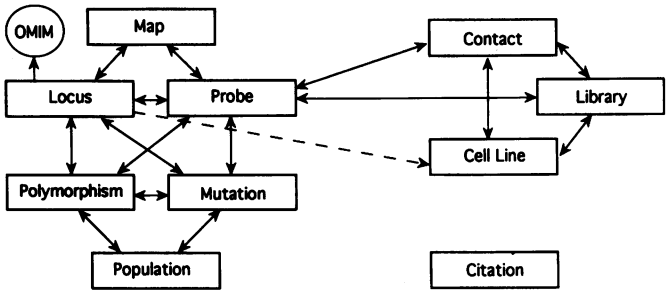


**Figure 1.** Organization of Data in the GDB. This diagram shows the connections among the ten data managers (within the GDB) and OMIM. Note that the Locus and Cell Line managers are connected only by breakpoint set information. Data in all GDB managers with the exception of those data in Contact Manager are linked to the Citation Manager. The OMIM database may be called from the Locus Manager. Once you have entered a data manager and then called another one, it is impossible to return to the initial data manager by selecting it directly.

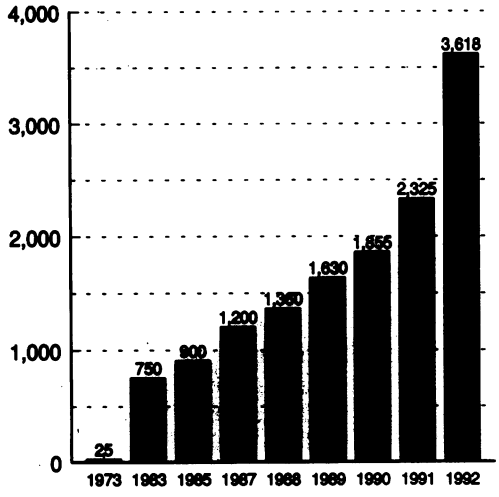


**Figure 2.** Number of known human genes. In 1992 more human genes were added to the GDB than during HGM 10 (Human Gene Mapping Meeting), HGM 10.5, and HGM 11 (1989–1991) combined.

option to examine only approved data). Direct submitters may ask that their data be kept confidential for a period not exceeding six months to allow for conventional publication of their findings.

The increase in known human genes since the first Human Gene Mapping meeting is illustrated in Figure 2. Table 1 shows the increase in data contained in the GDB since March 1991. The accelerated growth in the amount of data places the responsibility upon the GDB of allocating the resources, both personnel and software tools, to insure that the data is processed in a timely manner. The acquisition process involves data from many sources. Data comes to the GDB from the research community through both paper submissions and electronic files and are processed by the central support staff. Additionally, the central support staff are involved in the entry of data resulting from the Single Chromosome Workshops (SCWs) in a coordinated effort with the GDB editors.

Much of the information that appears in the literature without previously having been submitted to the GDB has been entered into the database through the efforts of the GDB editors. This

**Table 1.** GDB Data Statistics

| | Mar 91 | Mar 92 | Mar 93 |
|---|---|---|---|
| **Loci** | | | |
| Total Genes | 2,217 | 3,029 | 3,849 |
| Total D-segments | 5,369 | 9,435 | 14,598 |
| Mapped Genes | 1,883 | 2,332 | 2,735 |
| Mapped D-segments | 5,369 | 7,195 | 12,599 |
| Mapped Fragile Sites | 113 | 113 | 116 |
| Mapped Breakpoints | 0 | 53 | 250 |
| Map Sets | 0 | 254 | 598 |
| Total Mapped Loci: | 7,365 | 9,947 | 16,298 |
| Total Loci: | 7,699 | 12,844 | 19,414 |
| **Probes** | | | |
| PCR | 519 | 1,426 | 5,379 |
| ASO | 432 | 434 | 444 |
| Clones | 14,032 | 20,399 | 27,035 |
| Total Probes: | 14,983 | 22,259 | 32,858 |
| **Polymorphisms** | | | |
| Genes | 521 | 653 | 740 |
| D-segments | 2,145 | 2,929 | 4,116 |
| Total Polymorphisms: | 4,435 | 5,898 | 7,500 |
| **Citations** | | | |
| Journal Articles | 15,467 | 19,831 | 25,974 |
| Personal Communications | 5,508 | 5,910 | 6,267 |
| Books | 12 | 24 | 40 |
| Theses | 1 | 1 | 2 |
| Total Citations: | 20,988 | 25,766 | 32,283 |
| **Contacts** | | | |
| Probe Contacts | 1,922 | 2,562 | 3,273 |
| **Disease Loci, Mendelian Phenotypes, and Cloned Genes (Data taken from OMIM)** | 5,248 | 5,710 | 6,157 |

is due in part to the fact that most journals allow the publication of human mapping and disease data without requiring that the data be submitted to the public databases. The editors in the course of their own scientific investigations come across articles with data relevant to the GDB and place that information into the database.

As the data continues to accumulate with increasing speed, it is unrealistic to ask the GDB editors to be involved with the time-consuming process of data entry. Therefore, starting this year 'curators' will become the data entry contacts for the GDB editors. They will insure the flow of data from the journals to the database, as well as gather those data which are only available to researchers working in a particular area. Curators will be assigned specific areas of the human genome for which they are responsible for collecting data. Curators will also assist the organizers of SCWs by insuring that all reagents and maps discussed in conjunction with the meeting are placed into the database. By introducing curators to the data acquisition process, many of the tasks previously delegated to the GDB editors will now be carried out by the GDB staff and the role of the editors will be one of overseeing that the data accurately reflects advances in their areas of expertise.

## Distribution mechanisms

The data collected by the GDB (as well as the data contained in OMIM) are available through a number of distribution mechanisms. In the past, the data were most readily obtained by users from the main node in Baltimore through a modem connection. However, this year's usage statistics in Baltimore

indicate that most users of the GDB Baltimore node obtain access through the Internet. Currently the mechanisms for access to the GDB are: (1) utilization of the GDB in Baltimore; (2) utilization of the GDB through remote sites (GDB nodes) throughout the world; (3) downloading the data using anonymous FTP; (4) direct SQL access to the data; (5) downloading the data using Internet data gathering tools such as WAIS and Gopher; (6) accessing GDB data through software produced by other parties.

## 1. Utilization of the GDB in Baltimore

Online access to GDB is available through both the Internet and modem connections. Although users are required to obtain an account to access the software, registration is free. For those users within the U.S. and Canada who do not have Internet access, the GDB is accessible through the SprintNet (formerly Telenet) modem communications network. The GDB provides communications software for both the IBM-PC and Macintosh computers for use in accessing the database if the user does not already possess a communications package supporting VT100 terminal emulation. Users of the GDB are encouraged to access the database through the Internet. In addition to gaining increased speed of data transfer, users on the Internet have the ability to define retrievals and have the resulting report sent to them via electronic mail. In GDB 5.0 it is possible to define a set of queries that are performed on a periodic basis and have the reports mailed automatically, thus keeping the investigator informed on advances in an area of interest.

The use of VT100 terminal emulation is still supported in GDB 5.0 as a means to access the software. However the X-Windows protocol is also supported for users accessing GDB via the Internet. The use of X-Windows allows one to use a mouse within the GDB application. Although an X-Windows terminal may connect directly to Baltimore it is **strongly advised** that users wishing to run X-Windows sessions to access the GDB obtain the GDB front-end software for installation on a local workstation. By running the software locally, only the data associated with a retrieve are sent across the Internet between the user and Baltimore. A single local front-end is commonly used to support entire universities, allowing X-windows connections at local area network speeds. GDB User Support is prepared to answer questions concerning the hardware requirements for running the front-end software. There is a small fee associated with the software to defray shipping and support costs and Sybase license requirements (if the site does not already possess such a license).

## 2. Accessing the GDB software through other remote sites

Although worldwide networking continues to expand, in many areas of the world it is still more useful to establish a local copy of the GDB called a GDB node. These nodes provide better network access for their users than that available directly from Baltimore. In addition, they can offer support for their users that may not be available from Baltimore due to time differences and language barriers. It is the goal that all GDB sites offer the same level of commitment to providing adequate network support. All nodes operate under a formal agreement with the GDB which defines mutual responsibilities and obligations. An address list for GDB nodes is provided at the end of this article.

## 3. Downloading the data using anonymous FTP

The GDB provides a separate computer for accessing data and other related information using FTP (Internet File Transfer Protocol). All the data contained within the GDB is available in table dumps of the database on this server. By using the anonymous FTP server listed at the end of this paper, users may obtain diagrams of the GDB relational data structures along with a detailed data dictionary. With the aid of this documentation users can design subset databases for their personal use. All GDB documentation and data submission forms are also available through this server.

## 4. Direct SQL access to the data

Those users who need to construct customized complex queries not supported through the GDB front-end software have the opportunity to access a copy of the database for direct SQL queries. The GDB schema diagrams and data dictionary are provided on the anonymous FTP server (see section 3). GDB User Support can assist interested parties in obtaining access to the interactive SQL server.

## 5. Downloading the data using WAIS and Gopher

Wide Area Information Server (WAIS) software was developed by a joint effort of Thinking Machines, Apple Computer, and Dow Jones. Based on an existing Internet data exchange protocol, it allows a user to perform keyword searches of many databases simultaneously. The databases themselves are usually organized as a series of flat file 'documents.' A new flat file version of the GDB has been created to provide access to the data via WAIS client software. The flat file database is updated weekly from the primary copy of the relational database in Baltimore.

Gopher was originally developed in 1991 by the University of Minnesota Microcomputer Workstation Networks Center to help users of the campus network find answers to commonly asked questions. Gopher has since grown into an information system used by a large number of sites throughout the world. In order to access Gopher a client program must be run on the user's computer. Although some public logins are available for users not wishing to install the software locally, they lack many of the basic features (such as saving a file to disk or printing) available when the software is installed locally. Gopher software is freely available by anonymous FTP at many sites including the GDB FTP server. Both GDB and OMIM are available for searching via Gopher by pointing the client software to the Gopher 'hole' and selecting the 'Search Databases at Hopkins' menu.

## 6. Accessing GDB data through software produced by others

The GDB staff is collaborating with a number of organizations which are developing their own application software for accessing the database. These efforts include, among others:

Cold Spring Harbor—Thomas Marr and Corprew Reed have produced GDB Accessor, Macintosh software for accessing GDB and GenBank.

Los Alamos National Laboratory—Michael Cinkosky and James Fickett have produced SIGMA, the System for Integrated Genome Map Assembly.

Lawrence Livermore National Laboratory—Thomas Slezak and Elbert Branscomb have produced map assembly software.

German Cancer Research Center (DKFZ)—Otto Ritter has produced IGD, the Integrated Genome Database which allows one to use Acedb to view some of the data which is stored in GDB.

The GDB is also working to provide GDB and OMIM information via CD-ROM. OMIM will be available on CD-ROM before the end of this year. Additionally, the files utilized in the

WAIS version of GDB are being examined for the production of a prototype GDB CD-ROM.

## USER INFORMATION

Potential users can obtain information for accessing the data from:
GDB Human Genome Data Base
Johns Hopkins University School of Medicine
2024. East Monument Street
Baltimore, MD 21205-2100, USA
General Information: (410) 955-9705
User Support: (410) 955-7058
Fax: (410) 614-0434
E-mail: help@welch.jhu.edu

FTP server address: mendel.welch.jhu.edu (128.220.59.42)
FTP login: anonymous
FTP password: your e-mail address

Gopher Hole: merlot.welch.jhu.edu (128.220.59.18)

For the United Kingdom:
Ms. Christine Bates
Human Gene Mapping Program Resource Centre
CRC
Watford Road
Harrow Middlesex HA1 3UJ
UK
Phone: 44-81-869-3446
Fax: 44-81-869-3807
E-mail: cbates@mrc-crc.ac.uk

For continental Europe:
Dr. Otto Ritter
German Cancer Research Centre (DKFZ)
Department of Molecular Biophysics
Im Neuenheimer Feld 280
D-6900 Heidelberg
Germany
Phone: 49-6221-42-2372
Fax: 49-6221-42-2333
E-mail: dok261@cvx12.dkfz-heidelberg.de

For Australia:
Dr. Alex Reisner
Australian National Genomic Information Service (ANGIS)
Department of Electrical Engineering
Building J03
University of Sydney
Sydney, NSW 2006
Australia
Phone: 61-2-692-2948
Fax: 61-2-692-3847
E-mail: reisner@angis.su.oz.au

For Sweden:
GDB User Support
Biomedical Centre
Box 570
S-751 23 Uppsala
Sweden
Phone: 46-18-17-40-57
Fax: 46-18-52-48-69
E-mail: help@gdb.embnet.se

For the Netherlands:
CAOS/CAMM Center
Faculty of Science
University of Nijmegen
PO Box 9010
6500 GL Nijmegen
The Netherlands
Phone: 31-80-653-391
Fax: 31-80-652-977
E-mail: chaft@caos.caos.kun.nl

## ACKNOWLEDGEMENTS

The authors would like to thank Dan Jacobson and Meg Wright for their contributions to this article.

## REFERENCES

1. Pearson, P.L. (1991) *Nucleic Acids Res.*, 19-Supplement, 2237–9.
2. Pearson, P.L., Matheson N.W., Flescher D.C., and Robbins R.J. (1992) *Nucleic Acids Res.*, 20-Supplement, 2201–2206.
3. McKusick, V.A. (1992) *Mendelian Inheritance in Man*. Baltimore: Johns Hopkins Press.