NCBI News

National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

Winter 2000

New Entrez Genomes Views: A Fresh Look at Human Chromosomes

A set of new Entrez Genomes display formats provides an improved vantage point on the human chromosomal sequences in GenBank, facilitating convenient access to finished contig data and more supplementary information than ever before. Figure 1 shows a partial Entrez Genomes view of human chromosome 22, illustrating many of the features highlighted here.

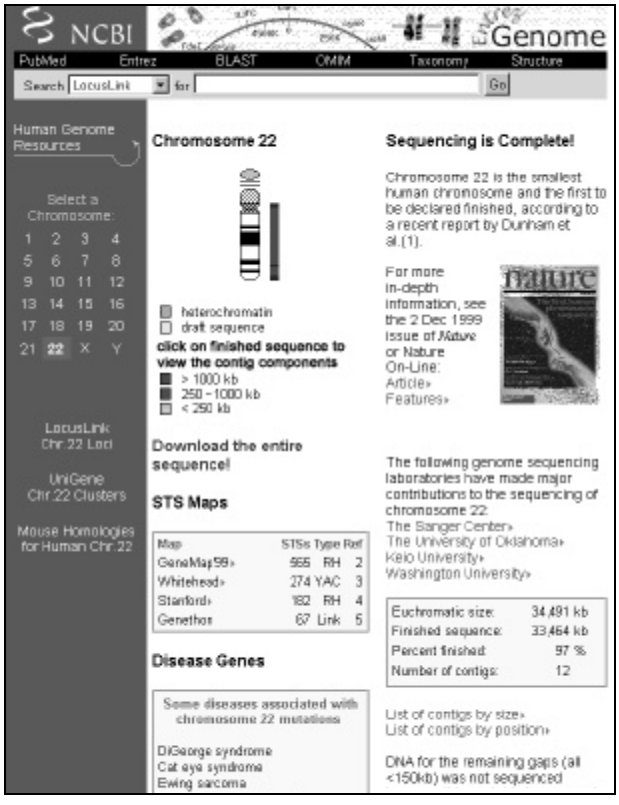

Figure 1: *Entrez Genomes View of Human Chromosome 22.*
[Reprinted by permission from Nature *402(6761):489-95, copyright 1999 Macmillan Magazines Ltd.]*

For each chromosome, the new Entrez Genomes views now provide:

A graphic depicting sequencing progress with finished regions shown in red: Clicking on a finished region leads to the corresponding contig data.

Sequencing progress statistics: These include euchromatic size, amount of finished sequence given in kilobases or as a percentage of the chromosome, and the number of contigs available for the chromosome.

Links to contig lists sorted by size and position.

Links to various STS maps.

Links to the coordinating chromosome sequencing center: Other sequencing centers that are working on the chromosome are listed as well.

Links to the Mitelman summary of chromosome aberrations associated with cancers.

A selection of disease genes that map to the chromosome: Links are also given to the Genes and Disease Web site, OMIM, and OMIM Morbid Map.

Links to references giving information on the chromosomal sequence: For example, the Entrez Genomes view for chromosome 22 gives a link to the full text of the paper by Dunham et al. (*Nature* (1999) 402, 489-95) that reports the completion of chromosome 22 sequencing.

To see the human chromosome views, go to www.ncbi.nlm.nih.gov/genome/guide.—*RM, DW*

In this issue

1 Entrez Genomes
2 IgBLAST
2 BLAST 1.4
3 PubMed Central
4 News Briefs
4 Mitochondria Energize RefSeq
4 PSI-BLAST Profiles
5 Frequently Asked Questions
6 Textbooks Linked to PubMed
6 BankIt 3.0
6 Mouse and Rat in LocusLink
7 BLASTLab
8 Malaria Menace Mapped

NCBI News

NCBI News is distributed four times a year. We welcome communication from users of NCBI databases and software and invite suggestions for articles in future issues. Send correspondence to *NCBI News* at the address below.

NCBI News
National Library of Medicine
Bldg. 38A, Room 8N-803
8600 Rockville Pike
Bethesda, MD 20894
Phone: (301) 496-2475
Fax: (301) 480-9241
E-mail: info@ncbi.nlm.nih.gov

Editors
Dennis Benson
Barbara Rapp

Contributors
Renata McCarthy
Jo McEntyre
Margaret McGhee
Liz Pope
Jian Ye

Writer
David Wheeler

Editing, Graphics, and Production
Marla Fogelman
Veronica Johnson
Jennifer Vyskocil

Design Consultant
Gary Mosteller

In 1988, Congress established the National Center for Biotechnology Information as part of the National Library of Medicine; its charge is to create information systems for molecular biology and genetics data and perform research in computational molecular biology.

NIH Publication No. 00-3272

ISSN 1060-8788
ISSN 1098-8408 (Online Version)

IgBLAST: An Immunoglobulin-Specific Search Tool

Immunoglobulin BLAST (IgBLAST), a specialized variant of BLAST that is designed for the analysis of human or mouse immunoglobulin sequence data, is now available online from NCBI. IgBLAST performs blastn or blastp searches of a non-redundant (nr) database of germline V genes and returns the best three V-gene matches, and the best two D-gene and J-gene matches. IgBLAST also annotates the query sequence, based on the domain information drawn from an alignment between the top-matched germline V gene and the query sequence, using the nomenclature of the Kabat Database of Sequences of Proteins of Immunological Interest. Such features as framework regions (FWR1, for instance) or complementarity-determining regions (CDR1, for instance) are delineated.

IgBLAST may also be used to search the standard BLAST nr database. In this case, the best matches to the germline gene database are aligned with the query as before, followed by the best hits to the nr database. One additional feature of IgBLAST is its ability to flag the returned matches from the nr database according to their germline V-gene origins. This allows one to distinguish easily between the results that use the same germline V-gene as the query and those that do not.

Although IgBLAST also supports protein searches using blastp, only V-gene matches will be reported. D-gene and J-gene matches are reported only for DNA searches with blastn.

The germline V-gene database currently contains Igh, Ig kappa, Ig lambda, and D and J genes from both human and mouse. The database can be viewed in the form of an annotated multiple sequence alignment by clicking on the appropriate database link in the sidebar of the IgBLAST page.

Try IgBLAST at www.ncbi.nlm.nih.gov/igblast.—*DW, JY*

Old BLAST 1.4 Network Server Retired

The old BLAST 1.4 network server, which accounts for less than 3% of all BLAST jobs performed at NCBI, was retired on March 1, 2000. It was replaced by BLAST 2.0, which was designed to handle the increasing size and complexity of the sequence databases.

The following NCBI programs used BLAST 1.4 and will no longer function: blastcl2, blastcli, and PowerBlast outside of Sequin. Blastcl3, available at ftp://ncbi.nlm.nih.gov/blast/network/netblast, should be used instead. MacVector versions prior to 6.5.3 are also affected by this change.

This change will *not* affect the NCBI BLAST Web pages, e-mail server, GCG connection, blastcl3 client, PowerBlast within Sequin, or MacVector at release 6.5.3 and higher.

PubMed Central Archive Developed at NCBI

PubMed Central is a Web-based repository established at the National Institutes of Health (NIH) to provide barrier-free access to primary research reports in the life sciences. So named because of its natural integration with the existing PubMed retrieval system, PubMed Central will serve as a host for scientific publishers, professional societies, and other groups to archive, organize, and distribute their research articles at no cost to the user.

Content

The scope of PubMed Central includes the broad life sciences, encompassing plant and agricultural research as well as biology and medicine. Participating journals and editorial groups may submit peer-reviewed reports from journals, as well as screened, although not formally peer-reviewed, reports from recognized editorial boards.

Participants' Roles

Contributing publishers, societies, and other editorial groups independent of the NIH have complete responsibility for their input. An international PubMed Central advisory committee establishes criteria for certifying groups that may submit material.

NIH's responsibilities, to be carried out by NCBI, pertain to developing, maintaining, and providing access to the repository. This includes facilitating the submission of SGML-tagged content; developing technology for enhanced retrieval, presentation, and navigation; and archiving the content to guarantee accessibility in the future.

How to Access

PubMed Central began accepting research reports this year and is now accessible on the Web at www.pubmedcentral.nih.gov.

From the home page, select a title to see the table of contents of the most recent issue available. From there, select specific articles or click on the **Archive** box for back issues. Enhanced search capabilities are currently under development. Links to publishers' sites are also included.

Two journals, *Molecular Biology of the Cell* and *PNAS: The Proceedings of the National Academy of Sciences* are currently available. Journals currently in process include *Biochemical Journal*, *Canadian Medical Association Journal*, *Frontiers in Bioscience*, and five journals from BioMed Central. Many additional journals have expressed interest and are preparing content for submission.

How to Participate

Guidelines covering acceptance criteria, media formats, copyright, and data formats are available from the PubMed Central home page under the **About PubMed Central** link.

Organizations interested in depositing content are urged to contact us at pubmedcentral@nih.gov.— *MM*

Selected Recent Publications by NCBI Staff

Boguski, MS. Biosequence exegesis. *Science* 286(5439):453–5, 1999.

Gerlach, VI, **L Aravind,** G Gotway, RA Schultz, **EV Koonin,** and EC Friedberg. Human and mouse homologs of *Escherichia coli* DinB (DNA polymerase IV), members of the UmuC/ DinB superfamily. *Proc Natl Acad Sci USA* 96(21):11922–7, 1999.

Grishin, NV. Phosphatidylinositol phosphate kinase: a link between protein kinase and glutathione synthase folds. *J Mol Biol* 291(2):239–47, 1999.

Koonin, EV, L Aravind, K Hofmann, J Tschopp, and VM Dixit. Apoptosis. Searching for FLASH domains. *Nature* 401(6754):662–3, 1999.

Panchenko, A, A Marchler-Bauer, and **SH Bryant.** Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins Suppl* 3:133–40, 1999.

Spouge, JL, A Marchler-Bauer, and **S Bryant.** The combinatorics and extreme value statistics of protein threading. *Ann Combinatorics* 3:81–93, 1999.

Su, X, MT Ferdig, Y Huang, **CQ Huynh,** A Liu, J You, **JC Wootton,** and TE Wellems. A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* 286(5443):1351–3, 1999.

Tatusova, TA, I Karsch-Mizrachi, and **JA Ostell.** Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics* 15(7): 536–43, 1999.

Wheelan, SJ, MS Boguski, L Duret, and **W Makalowski.** Human and nematode orthologs—lessons from the analysis of 1,800 human genes and the proteome of the *Caenorhabditis elegans*. *Gene* 238(1):163–70, 1999.

Wolfsberg, T, and I Makolowska. Web alert. Pattern formation and developmental mechanisms. *Curr Opin Genet Dev* 9(4):385–6, 1999.

News Briefs

SNP Consortium Makes First dbSNP Contribution

The SNP Consortium, including the Wellcome Trust and 11 other pharmaceutical companies, has recently made its first contributions totaling 7,396 SNPs to NCBI's database of single nucleotide polymorphisms (dbSNP). The SNP consortium was formed to create a public repository of SNP information useful in the study of disease. For more information or to search dbSNP, see www.ncbi.nlm.nih.gov/SNP/.

GenBank Release 116 Posts Largest Increase

With 5.8 billion base pairs from more than 5.6 million sequences, the recent GenBank Release 116.0 outweighs version 115.0 by 1,151 megabase pairs (Mbp), posting the largest single-release increase ever by GenBank and retiring the previous record of 812 Mbp. Uncompressed, the Release 116.0 flatfiles require roughly 21,350 MB (sequence files only) or 23,300 MB (including the "index" files). Release 116 is now available for downloading via ftp at ftp://ncbi.nlm.nih.gov/genbank.

UniVec Vector Screening Database Available by FTP

The NCBI UniVec database, used by the VecScreen Web service for identifying DNA sequence segments that may be of vector origin, is now available for downloading at ftp://ncbi.nlm.nih.gov/pub/UniVec.

A second database, UniVec_Core, is also available. UniVec_Core is a subset of sequences from the full UniVec database, chosen to minimize the number of false positive hits.

Both UniVec and UniVec_Core sequences are in the FASTA format. They are suitable for processing by the formatdb program of the Standalone BLAST package (available at ftp://ncbi.nlm.nih.gov/blast/executables/), so that each database can then be searched locally by the blastall program.

VecScreen and further information regarding UniVec can be found at www.ncbi.nlm.nih.gov/VecScreen/.

NCBI on Exhibit

NCBI will be exhibiting at the meetings listed below. The conference list is also available from NCBI's home page under **About NCBI.** For further information, contact NCBI at info@ncbi.nlm.nih.gov.

American Association for Cancer Research (AACR)
San Francisco, California
April 1-5

Human Genome Meeting—HUGO
Vancouver, British Columbia
April 9-12

Immunology 2000
Seattle, Washington
May 12-16

American Society for Microbiology (ASM)
Los Angeles, California
May 22-24

American Society for Biochemistry and Molecular Biology (ASBMB)
Boston, Massachusetts
June 4-9

Special Libraries Association (SLA)
Philadelphia, Pennsylvania
June 10-15

Endocrine Society (ENDO 2000)
Toronto, Canada
June 21-24

Mitochondrial Genomes Energize RefSeq

NCBI now offers a collection of 145 eukaryotic organelle genome sequences as part of RefSeq. These include 123 complete mitochondrial sequences as well as 16 complete plastid sequences. The animal (metazoan) mitochondrial records are considered Reviewed; that is, they have been manually curated by the NCBI staff and include standardized gene, protein, and RNA names. Other mitochondrial and chloroplast genome records are still Provisional RefSeq entries and are therefore presented as found in the source GenBank records used to create them. Visit www.ncbi.nlm.nih.gov/PMGifs/Genomes/organelles.html for more information.

Search Database of PSI-BLAST Profiles with IMPALA

The NCBI BLAST group has developed IMPALA—software to match a protein sequence against a library of score matrices stored from PSI-BLAST. A standalone version of the IMPALA suite of programs is included within the standalone BLAST distributions found at ftp://ncbi.nlm.nih.gov/blast/executables/.

A Web-based IMPALA search is available on the BLOCKS server at the Fred Hutchinson Cancer Center (blocks.fhcrc.org/blocks/impala.html). Any protein can be searched against a library of score matrices derived from the BLOCKS database.

Q & A Frequently Asked Questions

Q.	A.
How can I create a Definition line for my sequence submission automatically from within Sequin?	An easy way is to use the **Annotate>Generate Definition Line** menu item. This option creates Definition lines based on the source and feature annotations you have made to the record, and conforms to GenBank style guidelines. An existing sequence title (i.e., DEFINITION line) can be edited by double-clicking on it and making changes in the editing window that pops up.
What is the new Entrez Genomes equivalent of the alphabetical list containing chromosome-specific markers that I used to see in the old Entrez Genomes views of human chromosomes?	Although the old alphabetical marker list is no longer available, any marker that used to be in an Entrez Genomes human chromosome view can now be found using the new STS searching page (www.ncbi.nlm.nih.gov/genome/sts/query.cgi?). The STS search page includes all markers from dbSTS and all the human maps used in Entrez Genomes and GeneMap '99.
There does not seem to be a "print" function in Cn3D. How can I print Cn3D images?	You can print a Cn3D image by first exporting the image as a GIF file using Cn3D's File/Export/GIF function. You can then print this GIF image using most image-viewing programs.
Where can I get a very basic summary of my GenBank data access options?	For a basic summary of data access routes, take a look at www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html.
Is NCBI News *available in PDF format?*	Yes, in addition to HTML access to *NCBI News*, you can print a PDF version of the current and back issues at www.ncbi.nlm.nih.gov/About/newsletter.html.
If I run a BLAST search against only the nr database, am I likely to miss anything important?	Yes. The BLAST htgs (High Throughput Genomic Sequence) database is excluded from nr and must be searched separately; the same is true of the BLAST EST and GSS databases. The **Microbial Genomes: Finished and Unfinished** link on the main BLAST page provides access to data on 68 finished and unfinished microbial genomes, which are also not contained in the nr database. Researchers interested in BLASTing against human contig data should access these data by using the **Human Genome BLAST** link from the main BLAST page, rather than searching nr.

PubMed Abstracts Linked to Textbook Information

In collaboration with book publishers, the National Center for Biotechnology Information (NCBI) is adapting textbooks for the Web and linking them to the PubMed literature database. These textbooks will provide background information to PubMed users for exploring concepts they encounter in PubMed citations.

The first textbook to be included online is *Molecular Biology of the Cell*, 3rd ed., 1994, by Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson, published by Garland Publishing, Inc. *Molecular Biology of the Cell* is one of the most widely used undergraduate textbooks in molecular and cell biology. This textbook provides background information on central topics of molecular and cell biology. Books covering other topics and approaches to biology and medicine will appear in the future.

Linking the Textbooks with PubMed

Textbook information is accessed from the **Books** link shown on the abstract display of most PubMed records. Selecting this link causes the abstract to be redisplayed, with some words and phrases highlighted as links to corresponding textbook sections. Users may navigate within these sections, accessing text, figures, tables, and references. Rather than mirroring the print version of the textbooks, we provide logical units of content based on the book's native organization of chapters, sections, and subsections. The size of a unit and its interconnection with other parts of the book depend on both the native organization of the book and the intention of the publisher.

From the textbook sections, references are also linked back to their PubMed abstracts. This gives the textbook reader a starting point to further explore the literature using PubMed's **Related Articles** function. Because books usually contain established knowledge, their reference citations are often several years old; using the **Related Articles** link, readers can move forward in time to find articles that are similar, but more recent, than those cited in the book.

For More Information

Authors, editors, and publishers interested in linking a book to PubMed should contact books@ncbi.nlm.nih.gov.

For more information on the textbook project and how to access book information, select the **Books** link from the NCBI Literature Databases home page at www.ncbi.nlm.nih.gov/Literature.— *MM, BR, DW*

BankIt 3.0 Offers New Validation Features

A new version of NCBI's popular online sequence submission tool, BankIt, is now available. BankIt 3.0 is designed to allow for the validation and annotation of sequence data before submission to GenBank. BankIt 3.0 automatically checks sequences and their features to confirm that they are complete, biologically correct, and free of contaminating sequence, such as vector or linker sequence. This latter contamination check is made using VecScreen, available at www.ncbi.nlm.nih.gov/VecScreen/.

BankIt 3.0 also provides an expanded source organism modifier list and automatically translates all coding regions (CDS features) when either the CDS nucleotide interval or the resulting protein sequence is provided by the submitter. Try the new BankIt 3.0 at www.ncbi.nlm.nih.gov/BankIt/.

Mouse and Rat Now in LocusLink

LocusLink has now been expanded to include mouse and rat genes, in addition to human, containing records for more than 13,946 human loci, 13,014 mouse loci, and 2,268 rat loci. LocusLink provides a gateway of access points to information on gene maps, phenotypes, nomenclature, reference sequences, and related resources. Genome locus seekers should visit www.ncbi.nlm.nih.gov/LocusLink.

How to Search Huge Local Databases

The amount of public sequence data is growing at an exponential rate and is likely to continue to do so for the foreseeable future. With this data growth comes the problem of transferring, formatting, and searching gigabase-scale databases. Given current data transfer rates and computational resources, including CPU speeds and memory configurations, a "divide and conquer" approach has been implemented by NCBI in the current version of standalone BLAST. Features of standalone BLAST (blastall) and formatdb allow one to create and search arrays of smaller databases rather than having to search a single huge database. This allows efficient searches of databases with effective sizes far in excess of the RAM available on most small computer systems.

Standalone BLAST is able to search several databases sequentially with a single query using a syntax such as

```
blastall -i infile -d "part1 part2 part3" -p blastn -o out
```

In this case, the databases "part1", "part2", and "part3" have been created in the usual manner using formatdb with a syntax such as

```
formatdb -i part1 -o T -p F
```

The ability to name multiple databases in the blastall command line gives the user the flexibility to search an arbitrary group of databases that may be derived either from the division of a single huge source database or from several separate source databases. However, since each database must be formatted in a separate step, this process may become cumbersome if many databases are to be created.

A recent feature of formatdb streamlines the formatting process by creating several smaller database "volumes" automatically from a single huge source file. Furthermore, searches of these volumes are performed without explicitly naming each volume on the blastall command line.

To create a set of database volumes from a single source file, with a filename of "huge", use formatdb with a syntax such as

```
formatdb -i huge -o T -p F -v 1000000000
```

This command line will create a number of database "volumes," each containing one billion base pairs or fewer, as specified by the "-v" option, from the source database file. The volumes will have names consisting of the root database followed by a two-digit volume extension, followed by the usual BLAST database extensions. These smaller databases can be searched as if they were a single entity using

```
blastall -i infile -d huge -p blastn -o out
```

In this case, BLAST recognizes that the database "huge" has been partitioned into several volumes because it detects a file with the name of the root database followed by an extension of "nal" (for protein databases, the extension is "pal"). This file specifies a database list to be searched when the root database name is specified to BLAST. BLAST sequentially searches each database listed in this "nal" file and generates output that is indistinguishable from that of a single database search. A sample "nal" file, resulting from formatting the datafile "huge" into three volumes, is given below. The "DBLIST" line can also be edited to specify additional databases to be searched.

```
#
# Alias file created Tue Jan 18 13:12:24 2000
#
#
TITLE huge
#
DBLIST huge.00 huge.01 huge.02
#
#GILIST
#
#OIDLIST
#
```

The "nal" and "pal" files can also be used to simplify searches of multiple databases created separately as in the first example. For instance, a file called "multi.nal" containing the following lines could be created from scratch using a text editor.

```
#
# Alias file created Tue Jan 18 13:12:24 2000
#
#
TITLE multi
#
DBLIST part1 part2 part3
#
#GILIST
#
#OIDLIST
#
```

The "multi.nal" file would allow the three databases, "part1", "part2", and "part3", to be searched by specifying a single database name, "multi", on the blastall command line as follows:

```
blastall -i infile -d multi -p blastn -o out
```

The BLAST Lab feature is intended to provide detailed technical information on some of the more specialized uses of the BLAST family of programs. Topics are selected from the range of questions received by the BLAST Help Group.

Malaria Menace Mapped

Biologists at the National Institute of Allergy and Infectious Diseases, in collaboration with researchers at NCBI, have produced a genetic map of the human malaria parasite, *Plasmodium falciparum*. Maps, markers, and recombination data of linkage groups corresponding to the 14 *P. falciparum* nuclear chromosomes are available at NCBI's Malaria Genetics & Genomics page. Figure 1 illustrates the map for *P. falciparum* linkage group 1, which shows the relative position of several STS markers. Clicking on one of the STS markers leads to the appropriate GenBank sequence record. Crossover counts, crossover locations, and genotype segregation proportions for each of the linkage groups are also available.

To access these mapping data or download sequence data for each of *P. falciparum*'s 14 chromosomes, visit www.ncbi.nlm.nih.gov/Malaria/.—*DW*

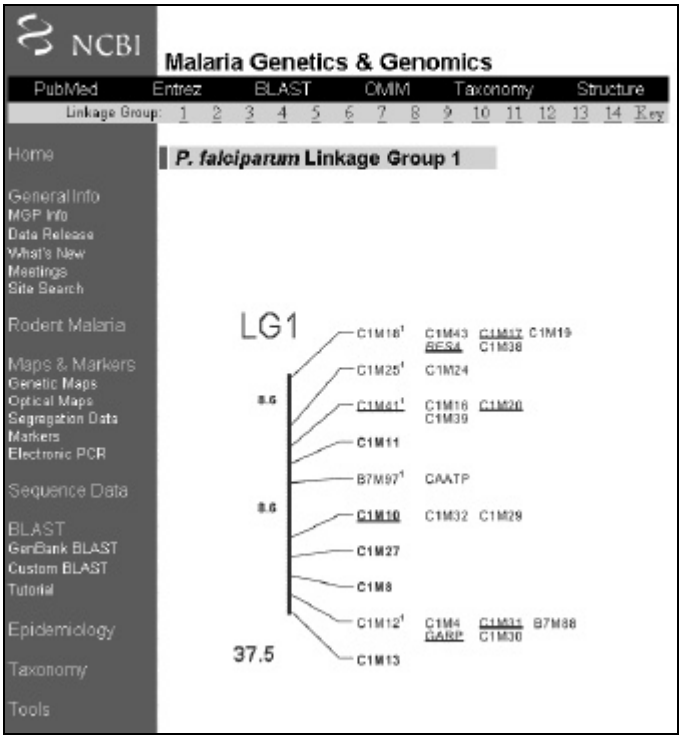

Figure 1: *Map for* P. falciparum *linkage group 1.*

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service, National Institutes of Health
National Library of Medicine
National Center for Biotechnology Information
Bldg. 38A, Room 8N-803
8600 Rockville Pike
Bethesda, Maryland 20894