NCBI News

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Department of Health and Human Services

Fall/Winter 2002

Plasmodium falciparum Arrives in GenBank

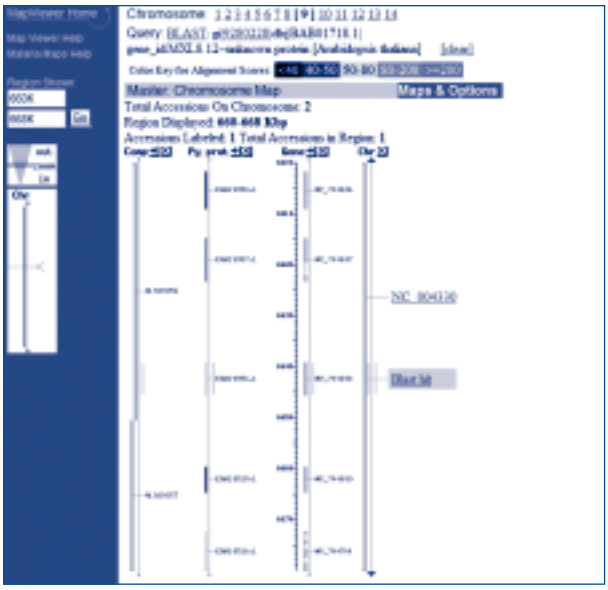

Figure 1: Entrez Map Viewer display of BLAST hit to chromsome 9 of *P. falciparum* for unknown *Arabidopsis* protein BAB01718. The BLAST hit coincides with *P. falciparum* protein NP_704698, as seen on the Gene map, and predicted protein EAA19956 from *P. yoelii*, as seen on the "Py prot" map.

The recently completed genome of the malaria parasite *Plasmodium falciparum* is available in GenBank and displayed in the Entrez Map Viewer. As reported in the October 3rd, 2002 issue of Nature (search in PubMed for PMID 12368864), *P. falciparum* clone 3D7 consists of a 23-megabase nuclear genome of 14 chromosomes assembled from finished high-throughput genomic sequence (HTGS, Phase 3). The genome was sequenced by members of the Malaria Genome Sequencing Consortium including The Institute for Genomic Research (TIGR), in collaboration with the US Naval Medical Research Center, the Sanger Institute, and the Stanford Genome Technology Center at Stanford University.

The sequence of each *P. falciparum* chromosome is represented in the Entrez Genomes database as a separate RefSeq record (with accession number of the form NC_XXXXXX) and may be found by the query "*Plasmodium falciparum* 3D7[organism]" in the Entrez Genomes.

The Entrez document summary for the sequence of chromosome 9, returned by the above query, appears as follows:

NC_004330
Plasmodium falciparum 3D7
chromosome 9,
complete sequence[99]
gi|23613523|ref|NC_004330.1|

continued on page 3

Third Party Annotation Database Debuts at GenBank

As the amount of publicly available sequence data rapidly increases, third party annotation will become increasingly important. The Third Party Annotation (TPA) database, created by GenBank and its international partners DNA Data Bank of Japan (DDBJ) and European Bioinformatics Institute (EBI), accepts third party annotation of genomic sequence, or computationally derived/predicted sequences. TPA submissions must use sequence data that is already represented in GenBank, and the analysis upon

continued on page 5

In this issue

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
David Wheeler

Writers
Vyvy Pham
David Wheeler

Editing and Production
Robert Yates

Graphic Design
Tim Cripps
Gary Mosteller
Joe Vuthipong

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. 03-3272

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

Entrez Map Viewer Gets a Home Page

The Entrez Map Viewer displays combinations of physical, genetic and sequence-based chromosomal maps for a variety of organisms ranging from yeast and *Plasmodium falciparum* to plants such as *Arabidopsis thaliana*, insects such as *Drosophila melanogaster* to mammals such as *Mus musculus* and *Homo sapiens*. Map Viewer displays for over a dozen organisms can now be accessed from a single Map Viewer home page using either a pull-down menu or a phylogenetic diagram. Links to genomic BLAST searches, in which BLAST results are displayed in their genomic context using the Map Viewer, are also provided. Searches for genetic loci for any of the genomes displayed in the Map Viewer may be launched directly from the home page. The new Map Viewer home page is accessible via the "Map Viewer" link under "Hot Spots" on the NCBI home page.

Drosophila Release 3 in GenBank and on Display in the Map Viewer

Release 3 of the annotated genomic assembly of the euchromatic arms of the six *Drosophila melanogaster* chromosomes is now available in GenBank. Annotated on this assembly are 13,500 genes, encoding at least 18,000 proteins.

The genome may be viewed in the Map Viewer using an array of 5 maps including a "Band" map showing detailed chromosomal banding patterns; a "contig" map, showing the position of NCBI NT contigs on the chromosome; a "component" map, showing the disposition of the components that were assembled to construct the contigs; a "transcript" map showing the alignment of transcripts to the genomic sequence; and a "genes_seq" map, showing the positions of annotated genes. Links to the NCBI Sequence Viewer (sv) and Sequence Downloader (seq) are available when the "genes _seq" map is the master. When the "transcript" map is made the master, links to best hit *Anopheles gambiae* proteins are shown along with links to BLAST2 Sequences alignment displays.

The *Drosophila* Map Viewer can be accessed from the Map Viewer home page. The version 3 *Drosophila* GenBank chromsomal records with annotations are available at:

ftp.ncbi.nih.gov/genomes/Drosophila_melanogaster

In addition, a new Whole Genome Shotgun (WGS) submission of *Drosophila* contigs sequenced at Celera is available under WGS project accession AABU00000000.

New Map in Mouse Map Viewer and Updated Genomic Assembly

The mouse Map Viewer has been updated with NCBI's new annotation of the Mouse Genome Sequencing Consortium's Whole Genome Shotgun assembly (MGSCv3).

This update includes many more contigs assembled from finished BAC sequences, which are displayed in a new "Strain" map that shows those portions of the genome that are covered by sequence arising from various mouse strains. The strain whose sequence is currently on display is indicated by a blue line on the strain map; alternative strain sequences are indicated by orange lines. The current strain is set by clicking on the strain name on the strain map after which the Map Viewer display will be refreshed with the display of the selected strain sequence.

Plasmodium falciparum

continued from page 1

The accession number in the document summary is a link to a Map Viewer display of the assembled sequence data, gene annotations, and genetic linkages.

Sequence maps available in the Map Viewer for *P. falciparum* include Chromosome, Component, and Gene maps. The chromosome sequence map shows the genomic sequence assembled from smaller components, visible on the components map. The genes sequence map displays genes that have been annotated on the chromosome including known and putative genes placed using blastx alignments of reference sequence proteins to the chromosome sequence. Another informative map is the "Py Protein" map, which presents reciprocal best hits between predicted proteins encoded in the *P. falciparum* genome and those predicted to be encoded in the whole genome shotgun (WGS) sequences of *P. yoelii*, a rodent malaria parasite. The order of the hits shown on the "Py protein" map reflects the order of the corresponding genes on the *P. falciparum* "gene_seq" map.

A Genome BLAST against the *P. falciparum* genome using unidentified *A. thaliana* protein BAB01718 leads to the genome view shown in the figure on page 1. Note the hit to the *P. falciparum* genome in an area encoding a protein that is a reciprocal best hit to a *P. yoelii* protein.

A variety of *Plasmodium* data has been incorporated into a malaria-related resource called Malaria Triad: Genetics & Genomics that is available from NCBI at:

www.ncbi.nlm.nih.gov/projects/Malaria/

The site combines information on the organisms involved in malaria research and provides access to services for data retrieval, genetic and genomic analysis of the sequence data, and malaria-related news. The *Plasmodium* genomes data (*P. falciparum*, *P. yoelii* and *P. vivax*), as well the genome of *Anopheles gambiae*, can be accessed from the malaria page along with the results of pre-computed sequence and literature searches. BLAST services have been tailored to search against malaria-related organisms and the *Plasmodium falciparum* and *Plasmodium yoelii* genomes have been incorporated into Genome BLAST (see figure 1), in which BLAST results are displayed in their genomic context using the Entrez Map Viewer.

The *P. falciparum* genome can be downloaded from the NCBI FTP site at:

ftp.ncbi.nih.gov/genbank/genomes/P_falciparum/

What's the Longest Sequence in GenBank? How About the Largest Protein?

The Entrez search system makes it relatively easy to determine the answers to both of these questions. A bit of trial and error yields:

```
25000000:50000000[Sequence Length]
NOT "srcdb refseq"[Properties]
```

This query, which ensures that the sequence we find is in the primary database, GenBank, and is not a derivative record from the NCBI RefSeq database, picks up a single record:

```
Locus       AE014297   27890790 bp
            DNA  linear  CON 18-SEP-2002
Definition  Drosophila melanogaster
            chromosome 3R, complete
            sequence.
```

This sequence is part of the recently deposited build 3 of the *Drosophila melanogaster* genome visible in the Map Viewer.

The longest protein, found using...

```
30000:50000[Sequence Length]
```

...turns out to be human Titin, NP_596869, which is an astounding 34,350 amino acids in length. Titin is a muscle protein that binds to the Z-disc region and the Z-line and M-line of the sarcomere, respectively, so that a single titin molecule spans half the length of a sarcomere. As part of its processing of this RefSeq, NCBI has identified 274 "Immunoglobulin" and 264 "Fibronectin" domains within this isoform of titin.

Graphical Structure Summaries

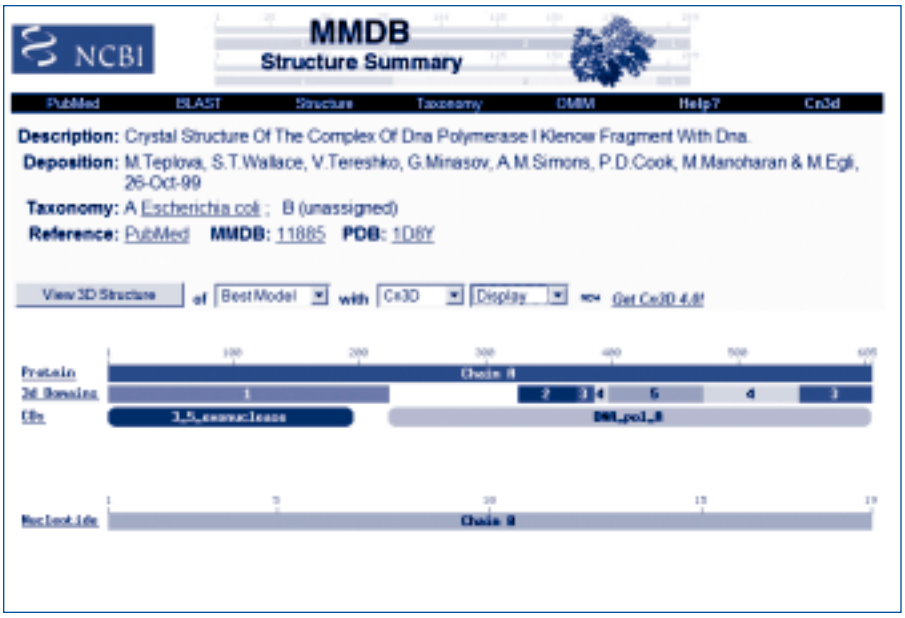

Figure 1: Structure summary for MMDB structure 11885, derived from PDB record 1D8Y: DNA polymerase I complexed with DNA. Protein and nucleotide chains as well as 3D domain and Conserved Domain assignments are represented by clickable colored bars.

Molecular Modeling Database (MMDB) Structure Summaries and Vector Alignment Search Tool (VAST) reports now use graphical overviews to depict macromolecular chains, domains, and regions of structural alignment to other proteins. An MMDB Structure Summary for the Klenow fragment of DNA Polymerase I in complex with a short piece of single-stranded DNA is given in Figure 1.

Graphical Overviews

The graphical overview of Figure 1 consists of two blocks; one for the protein chains represented in the structure file, and another for the nucleotide chain. The first block is comprised of three tiers of colored bars, the first of which, labeled "Protein", indicates that the structure contains a single protein chain of 605 amino acids. The next tier, labeled "3d-Domains", includes seven colored bars showing the disposition of the 3D domains assigned by NCBI on the basis of structural compactness. Five 3D domains have been detected with the first beginning at the amino terminal end of the protein chain, labeled "1", and the last ending at the carboxy-terminal end, labeled "3". Note that 3D domains "3" and "4" are comprised of two discontiguous segments of sequence within the amino-acid chain and are therefore shown using two separate bars each. The domain color scheme used in the Structure Summary follows that used by Cn3D, NCBI's macromolecular structure viewer, when structures are viewed and colored by "Domain" so that the two-dimensional depiction of the structure in the Structure Summary is easily correlated with its 3D representation.

The "CDs" tier shows the mapping of Conserved Domains detected within the protein via a Reverse-PSI BLAST comparison to NCBI's Conserved Domain Database (CDD). Although Conserved Domain definitions are based on sequence alignments and 3D domain assignments are based on structural compactness, there is often a good correspondence between the two, as can be seen in Figure 1 by the superposition of 3D domain "1" and the 3'→5'exonuclease Conserved Domain.

Links from the Structure Summary

Each colored bar in a tier is a link to further structural information. Clicking on the bar labeled "Chain A" leads to the VAST 3D neighbor report from which structural alignments can be downloaded and viewed using Cn3D. The 3D domain bar labeled "1" is a link to the VAST 3D neighbors report for the first 3D domain, the 3'→5' exonuclease domain, shown in Figure 3. The CDD bar labeled "3_5_exonuclease" leads to a report showing CDD sequence alignments which also include the "query" sequence; in this case, the sequence of 1D8Y, chain A. "Protein" and "Nucleotide" links to the left of the bars lead to Entrez nucleotide and protein sequence records, respectively.

Clicking on the "View 3D Structure" button in the Structure Summary shown in Figure 1 invokes Cn3D to display the structure shown in

continued on page 5

Graphical Structure

continued from page 4

Figure 2 in which the 3'→5' exonuclease domain is shown with cylinders representing alpha helices. The structure is colored by domain so that the color-scheme matches that used in the graphical domain bars shown in the Structure Summary report of Figure 1.

Explore the NCBI structure database at:

www.ncbi.nlm.nih.gov/Structure/

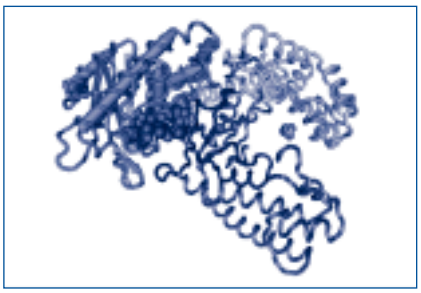

Figure 2: Cn3D rendering of DNA polymerase I. The 3'→5' exonuclease domain, to the left, is shown using "3D objects" to highlight helices and strands. The other four domains are shown in a "tube-worm" representation. The single-stranded DNA is shown in a space-filled representation, as is the phosphate group appearing in the cleft between the "fingers" and "thumb" domains of the enzyme.

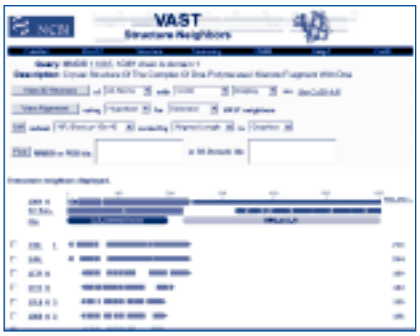

Figure 3: VAST 3D neighbor report for 3D domain "1" corresponding to the 3'→5' exonuclease Conserved Domain of DNA polymerase I. The portions of each 3D neighbor aligning with 3D domain "1" are indicated graphically. Links from the aligned-region bars display the detailed structural alignment and 3D superposition.

Third Party Annotation

continued from page 1

which the annotations are based must appear in a peer-reviewed scientific journal. Those wishing to add a feature annotation, such as a gene, to an unannotated genomic sequence or, wanting to combine two or more records, such as a set of ESTs, to create a longer transcript sequence, can submit their analysis or assembly to the TPA database. Trace data sequences or Whole Genome Shotgun (WGS) may be used as the basis of a TPA submission, but data from secondary sources such as NCBI Reference sequences or primary data from proprietary databases may not be used.

Third parties can submit annotations using either Sequin or BankIt. If using BankIt, choose "NO" when asked whether the submission is primary data in order to initiate the TPA option. Those making TPA submissions via Sequin should indicate this in their email message to NCBI and provide accession numbers for the primary sequence (s) used in their analysis. Instructions for making TPA submissions are found at:

www.ncbi.nlm.nih.gov/Genbank/index.html

TPA records can be located with Entrez using the TPA term within the Properties field; for example:

TPA [prop]

As of November 2002, there were 104 TPA records in the Entrez database. An example of a TPA record is shown below. The "Primary" field shows how the sequence in the TPA record was constructed from existing database sequences. In the case below, four GenBank database sequences were combined to produce the sequence upon which the submission is based. For instance, bases 1 through 503 in the TPA sequence were derived from bases 3 through 505 in GenBank sequence AQ655575.1.

Locus BK000167 561 bp DNA linear INV 19-OCT-2002
Definition TPA: Trypanosoma brucei GRIP domain containing protein gene, partial cds.
Accession BK000167
Version BK000167.1 GI:24137384
Keywords Third Party Annotation; TPA.
.....

PRIMARY	TPA_SPAN	PRIMARY_IDENTIFIER	PRIMARY_SPAN	COMP
	1-503	AQ655575.1	3-505	
	323-561	AL465640.1	1-239	
	335-561	AQ638516.1	1-227	
	404-561	AZ213060.1	435-592	c

PubMed Central Tops 100 Journal Mark

PubMed Central (PMC) is an online archive of full-text peer-reviewed research articles, editorials and essays encompassing topics in the life sciences. The PMC service debuted in February 2000 with content from the Proceedings of the *National Academy of Sciences* (PNAS) and *Molecular Biology of the Cell*. Presently over 100 journals deposit the full text of their articles in PMC. A list of currently available and forthcoming journals is accessible from the PMC homepage. PMC is integrated with PubMed, such that articles retrieved in PubMed will contain links to the PMC versions, when available.

The online format of articles in PMC is consistent for all journals. The journal's logo serves as a hyperlink to the publisher's site, providing journal-specific content, and some journals may direct PMC to display the full-text articles from the journal site. A search box on the PMC home page offers free text searches of the journals in PMC.

PMC is also searchable as an Entrez database. In addition to full text searching, organism names in the full text are recognized and indexed, to provide links to the taxonomy database. PMC can be accessed from the PubMed homepage, by choosing PMC from the Entrez drop-down menu on any page, or directly at:

www.pubmedcentral.gov/

The NCBI Handbook is on the Bookshelf

NCBI now offers a handbook available in electronic form in the NCBI Books database. The handbook is geared towards people who want to get the most from the NCBI webpages or who want a detailed introduction to NCBI resources with extensive links to help documentation. The 23 chapters of the NCBI Handbook, organized into four sections, cover topics ranging from the processing of sequence submissions to the assembly of the human genome.

Each chapter in the Handbook is available for online searching and browsing and also in PDF format for printing.

The Table of Contents for the Handbook is given below:

Part 1. The Databases

1. GenBank: The Nucleotide Sequence Database
2. PubMed: The Bibliographic Database
3. Macromolecular Structure Databases
4. The Taxonomy Project
5. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation
6. The Gene Expression Omnibus (GEO): A Gene Expression and Hybridization Repository
7. Online Mendelian Inheritance in Man (OMIM): A Directory of Human Genes and Genetic Disorders
8. The NCBI BookShelf: Searchable Biomedical Books
9. PubMed Central (PMC): An Archive for Literature from Life Sciences Journals
10. The SKY/CGH Database for Spectral Karyotyping and Comparative Genomic Hybridization Data

Part 2. Data Flow and Processing

11. Sequin: A Sequence Submission and Editing Tool
12. The Processing of Biological Sequence Data at NCBI
13. Genome Assembly and Annotation Process

Part 3. Querying and Linking the Data

14. The Entrez Search and Retrieval System
15. The BLAST Sequence Analysis Tool
16. LinkOut: Linking to External Resources from Entrez Databases
17. The Reference Sequence (RefSeq) Project
18. LocusLink: A Directory of Genes
19. Using the Map Viewer to Explore Genomes
20. UniGene: A Unified View of the Transcriptome
21. The Clusters of Orthologous Groups (COGs) Database: Phylogenetic Classification of Proteins from Complete Genomes

Part 4. User Support

22. User Services: Helping You Find Your Way
23. Exercises: Using Map Viewer

Other new books on the bookshelf are:

The Human ATP-Binding Cassette (ABC) Transporter Superfamily.
Dean, Michael, Bethesda (MD): NCBI, National Library of Medicine (US).

Basic Neurochemistry, Molecular, Cellular, and Medical Aspects. 6th ed.
Siegel, George J.; Agranoff, Bernard W.; Albers, R. Wayne; Fisher, Stephen K.; Uhler, Michael D., editors, Philadelphia, Pennsylvania: Lippincott,Williams & Wilkins.

caaatccgttcttgatcgtacatagcgcatgtcagncaaatc
caaatccattcttgatcgtacatggcacatgtcagtcaaatc

Searching the Trace Archive with Discontiguous MegaBlast

Modern sequencing technology has facilitated the production of a huge and growing volume of raw, unannotated nucleotide sequence for a variety of organisms. The rapidly expanding NCBI Trace Archive contains over 100 billion base pairs of such sequence from dozens of organisms. Making use of this unannotated sequence data requires the ability to compare these sequences to others, in particular, to the annotated sequences in the GenBank database. The sheer volume of data, however, makes it a challenge to perform sensitive comparisons quickly.

To maximize sensitivity when comparing coding sequences between organisms, translated searches are the best choice since they convert nucleotide sequences to their more tightly conserved protein translations before the comparisons are made. Because of the need to perform translations and comparisons in 6 reading frames, however, translated searches are very time-consuming.

Untranslated searches are more rapid but much less sensitive because codon usage differences between organisms allow similar proteins to be encoded by dissimilar nucleotide sequences. To facilitate sensitive untranslated searches, NCBI has developed a program called Cross Species MegaBLAST.

MegaBLAST uses an exact contiguous nucleotide or "word" match of length 28 as the starting point for constructing alignments. However, as the identity between the sequences to be compared dips below 80%, the requirement for a contiguous word hit leads to the omission of many statistically significant alignments with the concomitant generation of many short random alignments. Cross-Species MegaBLAST works on the principle that if alignments are initiated not with an exact contiguous word match, but with the match of an equivalent number of noncontiguous positions within longer segments of the sequence, fewer words are found, but a greater fraction of those found produce statistically significant alignments.[1]

An example of a discontiguous word is the "coding" template given below:

```
110110110110110110
```

This template gives optimal results when comparing coding sequences across species. The positions occupied by the twelve "1"s must match between the two sequences to be compared in order for MegaBLAST to begin an alignment. The template allows every third position to vary in accordance with the third "wobble base" of the genetic code. This discontiguous word is more sensitive than the corresponding contiguous word of length 12 for comparisons of coding regions across species where sequence divergence is high.

Figures 1 and 2 below show MegaBLAST graphical overviews for two searches of the mouse trace sequences using the transcript sequence of the human HEXA gene as the query. Both searches return results in similar lengths of time. However, the Cross-Species search is far more sensitive and is able to detect matches over about 75% of the human HEXA sequence.

Cross-Species MegaBLAST is found as a search option on the Trace Archive page under the "Cross Species Comparison" link. Both the "coding" template and an "optimal" discontiguous template for searches using non-coding sequences are supported.

[1] Ma, B., Tromp, J., Li, M., "PatternHunter: faster and more sensitive homology search", *Bioinformatics* 2002 Mar;18(3):440-5

Figure 1: Graphical overview of the results of a MegaBLAST search using the transcript sequence of the human HEXA gene as a query against the mouse data in the Trace Archive.

Figure 2: Graphical overview of the results of a Cross-Species MegaBLAST search using the transcript sequence of the human HEXA gene as a query against the mouse data in the Trace Archive.

New Microbial Genomes in GenBank

Shigella flexneri 2a str. 301	4,607,203 bp	AE005674	NC_004337	Nucleic Acids Res. 2002. Oct 15; 30(20):4432-41
Leptospira interrogans serovar lai str. 56601	4,691,184 bp	AE010300, AE010301	NC_004342, NC_004343	Chinese National Human Genome Center
Streptococcus agalactiae 2603 V/R	2,160,267 bp	AE009948	NC_004116	Proc. Natl. Acad. Sci. USA 2002. 99 (19): 12391-6
Oceanobacillus iheyensis	3,630,528 bp	BA000028	NC_004193	Nucleic Acids Res. 2002. Sep 15; 30(18):3927-35
Escherichia coli CFT073	5,231,428 bp	AE014075	NC_004431	Proc. Natl. Acad. Sci. USA 2002. 99(26):17043-8

GenBank Release 133

GenBank Release 133, including 28.5 billion bases from 22.3 million sequences, is now available at:

www.ncbi.nih.gov/Genbank

Uncompressed, the Release 133 flatfiles comprise 107 gigabytes. The ASN.1 version comprises roughly 84 gigabytes. As of GenBank Release 134 in February of 2003, the cumulative GenBank Update (GBCU) products will be discontinued. For further details, see the GenBank release notes at:

ftp.ncbi.nih.gov/genbank/gbrel.txt

GenBank mirrors are available at:

genbank.sdsc.edu/pub

bio-mirror.net/biomirror/genbank

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

Official Business
Penalty for Private Use $300

FIRST CLASS MAIL
POSTAGE & FEES PAID
PHS/NIH/NLM
BETHESDA, MD
PERMIT NO. G-816