NCBI News

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

Spring 2001

Tour the Human Genome with Map Viewer

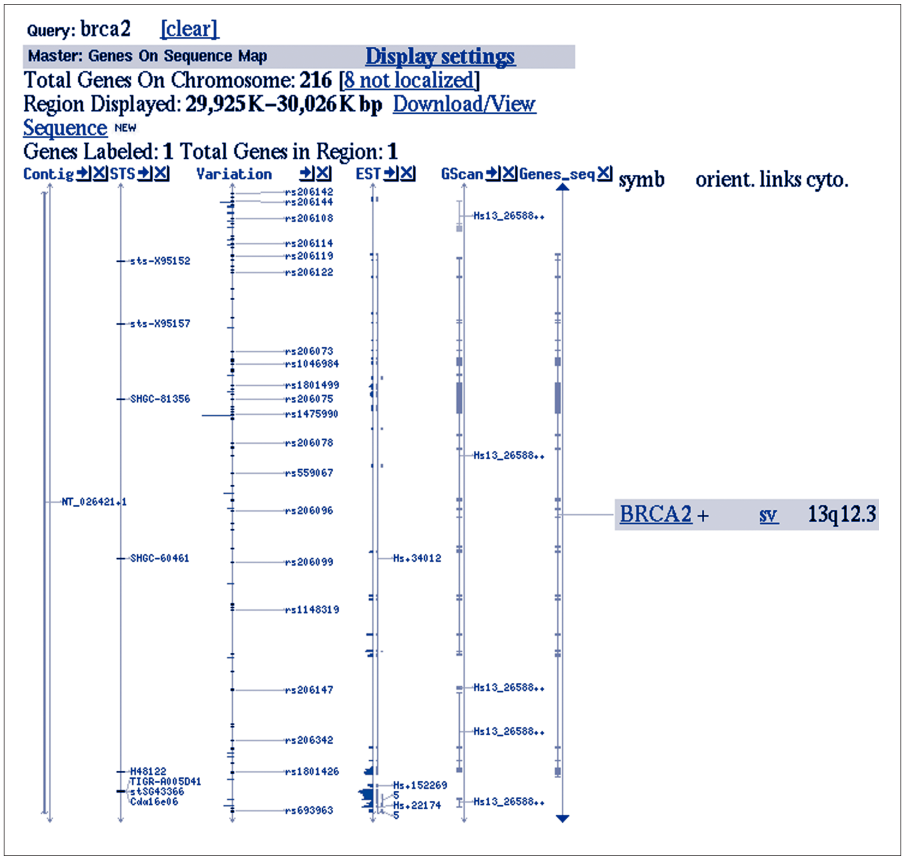
NCBI is committed to an ongoing program of incorporating new data and annotation into its human genome resources and producing updated assemblies on a regular basis. The data used to generate NCBI's assembly of the human genome include draft and finished sequence deposited in GenBank by the Human Genome Project sequencing centers, as well as by individual contributors. New and updated sequence data continues to be submitted from both sources.

The Human Genome Map Viewer provides integrated access to the genome data through a collection of genetic, physical, and sequence maps. Views of the data range from specific genes to whole genomic regions of interest.

A total of 7 maps may be chosen for simultaneous display from a set of 21, including clone and FISH-mapped clone maps, several radiation hybrid maps, an EST map, and the Genethon and Marshfield genetic maps. The Map Viewer can display a particular region of the genome centered on a gene or marker of interest, or a region defined by an arbitrary base range.

Figure 1 shows six selected maps, with the display centered on the 100Kb area surrounding the BRCA2 gene. The rightmost is called the

continued on page 2



Query: brca2 [clear]

Master: Genes On Sequence Map Display settings

Total Genes On Chromosome: 216 [8 not localized]

Region Displayed: 29,925 K–30,026 K bp Download/View

Sequence NEW

Genes Labeled: 1 Total Genes in Region: 1

Contig STS Variation EST GScan Genes_seq symb orient. links cyto.

BRCA2 + sv 13q12.3

Figure 1: *Map Viewer display of six parallel maps for human BRCA1*

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
Barbara Rapp

Contributors
Colleen Broder
Vyvy Pham

Writer
David Wheeler

Editing, Graphics, and Production
Jennifer Carson Vyskocil
Cheryl Richardson
Adrienne Kitts

Design Consultants
Gary Mosteller
Tim Cripps

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.

The contents of this newsletter may be reprinted without permission. The mention of trade names, commercial products, or organizations does not imply endorsement by NCBI, NIH, or the U.S. Government.

NIH Publication No. 01-3272

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

Map Viewer

continued from page 1

Master Map, which shows the greatest detail and provides additional links. The other maps provide summary information only.

In this example, the initial search was for the term “brca2”. Two BRCA2 records were found on chromosome 13; one for the disease associated with the gene and one for the gene itself. Following the gene link, the default display showed two maps — a sequence map labeled “Genes_seq” and a cytogenetic map labeled “Genes_cyto”. To view the alternative set of maps shown in Figure 1, the Display Options and Zoom features were used. Each of the maps is discussed below.

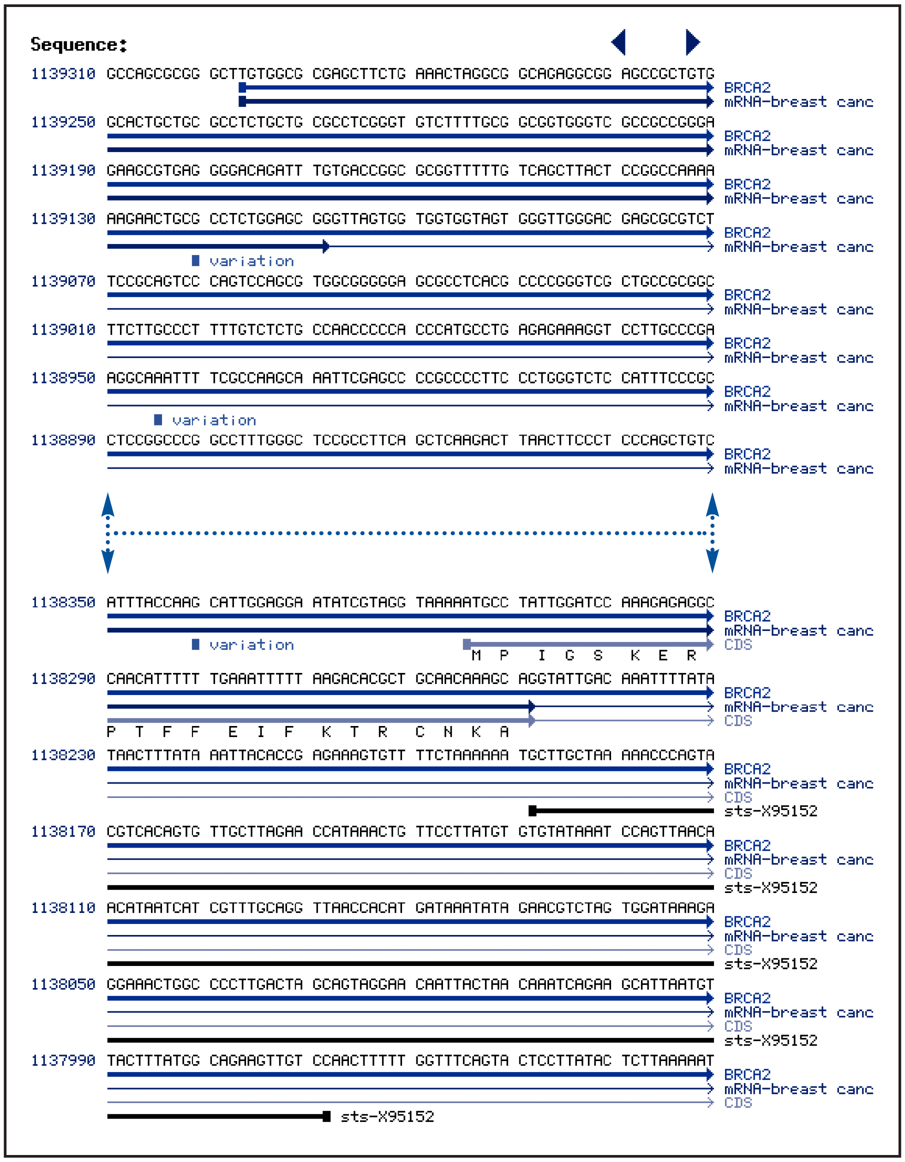
The Master Map is designated by the user through the Display Options function and, in this example, it is the sequence map labelled “Genes_ seq”. The exon structure of the BRCA2 gene is shown as a set of thick lines superimposed upon the thin line that represents the entire gene.

First on the left, the Contig map shows the location of BRCA2 on a particular contig created by the NCBI genome assembly process. The entire contig may be downloaded from the map page, or viewed directly as a GenBank contig record. This contig record contains links to all features found on the contig, such as STSs, SNPs, FISH-mapped clones, annotated genes and CDSs.

Second from the left in Figure 1 is the STS map, which indicates the positions of sequence-tagged sites in the region of BRCA2. Four STS markers are shown within the range of the BRCA2 gene given on the Genes_seq map. Next, the Variation map shows many SNPs within the BRCA2 gene, with a particularly dense clustering in the region of the largest exon of BRCA2.

Figure 1 also includes two recent additions to the palette of maps available in the human Map Viewer. The EST map shows where ESTs align well to the genomic sequence, with histograms indicating the mapping density. Including links to UniGene, the EST map can be used to identify undocumented exons or to identify the prominent splice variants of genes. The GScan map shows *ab initio* gene models derived from GenomeScan, a program related to the popular GenScan gene prediction tool. This map includes links to the protein similarities, discovered via a blastx search of the protein databases, that were used to support the gene predictions.

Returning to the Genes_seq map, the Master Map in this example, two links to additional information are provided. The BRCA2 link leads to the corresponding Locus Link record, which provides a complete summary of information relating to the gene plus links to related resources. The “sv” link leads to the NCBI Sequence Viewer, which provides the most detailed sequence-level view of the human genome.



```
Sequence:
1139310 GCCAGCGCGG GCTTGTGGCG CGAGCTTCTG AAACTAGGCG GCAGAGGCGG AGCCGCTGTG
                                                                BRCA2
                                                                mRNA-breast canc
1139250 GCACTGCTGC GCCTCTGCTG CGCCTCGGGT GTCTTTTGCG GCGGTGGGTC GCCGCCGGGA
                                                                BRCA2
                                                                mRNA-breast canc
1139190 GAAGCGTGAG GGGACAGATT TGTGACCGGC GCGGTTTTTG TCAGCTTACT CCGGCCAAAA
                                                                BRCA2
                                                                mRNA-breast canc
1139130 AAGAACTGCG CCTCTGGAGC GGGTTAGTGG TGGTGGTAGT GGGTTGGGAC GAGCGCGTCT
                                                                BRCA2
                                                                mRNA-breast canc
           variation
1139070 TCCGCAGTCC CAGTCCAGCG TGGCGGGGGA GCGCCTCACG CCCCGGGTCG CTGCCGCGGC
                                                                BRCA2
                                                                mRNA-breast canc
1139010 TTCTTGCCCT TTTGTCTCTG CCAACCCCCA CCCATGCCTG AGAGAAAGGT CCTTGCCCGA
                                                                BRCA2
                                                                mRNA-breast canc
1138950 AGGCAAATTT TCGCCAAGCA AATTCGAGCC CCGCCCCTTC CCTGGGTCTC CATTTCCCGC
                                                                BRCA2
                                                                mRNA-breast canc
       variation
1138890 CTCCGGCCCG GCCTTTGGGC TCCGCCTTCA GCTCAAGACT TAACTTCCCT CCCAGCTGTC
                                                                BRCA2
                                                                mRNA-breast canc

1138350 ATTTACCAAG CATTGGAGGA ATATCGTAGG TAAAAATGCC TATTGGATCC AAAGAGAGGC
                                                                BRCA2
                                                                mRNA-breast canc
           variation                                            CDS
                                         M  P   I  G  S   K  E  R
1138290 CAACATTTTT TGAAATTTTT AAGACACGCT GCAACAAAGC AGGTATTGAC AAATTTTATA
                                                                BRCA2
                                                                mRNA-breast canc
                                                                CDS
        P  T  F  F   E  I  F   K  T  R   C  N  K  A
1138230 TAACTTTATA AATTACACCG AGAAAGTGTT TTCTAAAAAA TGCTTGCTAA AAACCCAGTA
                                                                BRCA2
                                                                mRNA-breast canc
                                                                CDS
                                                                sts-X95152
1138170 CGTCACAGTG TTGCTTAGAA CCATAAACTG TTCCTTATGT GTGTATAAAT CCAGTTAACA
                                                                BRCA2
                                                                mRNA-breast canc
                                                                CDS
                                                                sts-X95152
1138110 ACATAATCAT CGTTTGCAGG TTAACCACAT GATAAATATA GAACGTCTAG TGGATAAAGA
                                                                BRCA2
                                                                mRNA-breast canc
                                                                CDS
                                                                sts-X95152
1138050 GGAAACTGGC CCCTTGACTA GCAGTAGGAA CAATTACTAA CAAATCAGAA GCATTAATGT
                                                                BRCA2
                                                                mRNA-breast canc
                                                                CDS
                                                                sts-X95152
1137990 TACTTTATGG CAGAAGTTGT CCAACTTTTT GGTTTCAGTA CTCCTTATAC TCTTAAAAAT
                                                                BRCA2
                                                                mRNA-breast canc
                                                                CDS
                          sts-X95152
```

Figure 2: *Sequence-level view of the human BRCA2 gene.*

The Sequence Viewer display, a small portion of which is shown in Figure 2, shows the location of exons, CDSs, and STS markers along the BRCA2 gene. This example includes the DNA sequence for the initial region of the gene, showing the first two exons and intervening intron. One can see that the start codon is located within the second exon and that translation of this CDS indicates an initial amino acid sequence of "MPIGS..." for the BRCA2 gene product. Directly following is the STS marker sts-X95152, which is also visible on the STS map shown in Figure 1. The ends of the BRCA2 gene are marked by sts-X95152 and H48122. Note that the BRCA2 gene is encoded on the reverse complement of contig NT_009984, such that sts-X95152, which is near the beginning of the gene, is found at the bottom of the Map Viewer display. The Sequence Viewer also pinpoints the location of three SNPs, two in the first intron and one in the second exon. All three SNPs are found in noncoding regions.

Any portion of the sequence shown in a Map Viewer view may be downloaded for further analysis using the Download/View feature. A tabular report for the displayed region can also be generated, including the coordinates of all Map Viewer features shown on the displayed maps, with links to each feature.

An advanced search feature is also offered. Search terms can be restricted to particular maps; specific features such STSs, clones, or contigs; certain subsets of SNPs (such as SNPs in coding vs. noncoding regions, or SNPs with various degrees of heterozygosity); or particular chromosomes. It is also possible to search for several items together so as to generate a Map View containing several features, or to search a particular region defined by a set of markers. A syntax such as "H48122 OR sts-X95157 OR brca2" will generate three hits on chromosome 13, which can be viewed together.

The latest NCBI human genome assembly can be downloaded from the Genome side of the FTP site at ftp.ncbi.nih.gov/genomes/H_sapiens.

The FTP site includes sequence data for each chromosome, as well as the mRNA and protein sequences generated by the NCBI annotation project. The data used by the Map Viewer to display the various integrated maps is contained within the "maps" subdirectory. A README file contains more details. —*VP, DW*

Investigator Profile Eugene V. Koonin

Comparative Genomics: *From Sequence to Evolution to Function*

The recent explosion in genome sequencing has led to a rapid enrichment of the protein databases, both in terms of number and variety of protein sequences. The function(s) of the majority of these proteins remains unknown. By classifying proteins according to their degree of sequence similarity, which generally reflects evolutionary (homologous) relationships, computational biologists are able to predict the three-dimensional structure and a likely function for many proteins, and determine their evolutionary origin.

COGs: A Tool for Whole Genome Comparative Analyses

The database of Clusters of Orthologous Groups of proteins (COGs), developed by NCBI investigators **Tatusov, Koonin,** and **Lipman,** is designed to classify proteins from completely sequenced genomes on the basis of orthologous relationships. The first version of this database was released in 1997, and contained proteins from seven genomes and consisted of 720 **COGs**. Today, the database includes proteins from 34 complete genomes and consists of 2,885 COGs. A companion program, **COGNITOR**, was developed to fit new proteins into COGs. COGNITOR may also be used to annotate newly sequenced genomes and allows researchers to predict the function(s) of individual proteins or protein sets.

Over a period of years, NCBI investigators developed and refined computational approaches that allowed them to detect previously unnoticed but potentially important protein sequence similarities. Using this strategy to search various protein databases, researchers are able to compare the genomes of different organisms and identify conserved protein families and key protein pathways that are modified, or absent, in an organism. Comparative genome analyses also provide fundamental insights into the organization and evolution of highly diverged species and are instrumental in identifying other biological features that may confer a distinct evolutionary advantage to an organism.

Galperin and **Koonin** addressed the issue of detecting targets for anti-bacterial drugs using a comparative-genomic approach. For this purpose, one needs to identify genes that are likely to be essential for the survival of bacterial pathogens, but are absent in the host. One method for predicting essential genes, sometimes called genome subtraction, capitalizes on the COGs database, which includes conserved protein families represented in at least three phylogenetically distant organisms. The ability to query the database and retrieve a list of all COGs with a particular phylogenetic pattern allows researchers to identify genes that are present in the genomes of all or most pathogenic bacteria, but absent in its eukaryotic host, delineating a potential drug target.

Experimental approaches may then be used to validate the essentiality of the selected genes for bacterial survival and characterize their cellular functions. Using the COGs database as a tool to search for drug targets in microbial genomes demonstrates the potential of comparative genomics in accelerating the drug discovery process.

Predicting New Components of Known Molecular Complexes and Pathways

In another study, **Aravind, Koonin** and colleagues compared 4,344 protein sequences from fission yeast with all available eukaryotic protein sequences. They identified protein sequences that were common to both fission yeast and non-fungal eukaryotes, but that were missing or significantly different in baker's yeast. These two species of yeast are evolutionarily close enough such that direct counterparts among their genes are readily detectable, but distant enough to support substantial gene differences. Analysis of the combined data showed that since its radiation from a common ancestor with fission yeast, baker's yeast had lost about 300 genes and approximately 300 additional genes had diverged significantly. The most notable feature of the set of genes lost in baker's yeast was the co-elimination of functionally connected groups of proteins, such as, for example, proteins involved in post-transcriptional gene silencing. By examining patterns of coordinated gene loss, in combination with a careful analysis of conserved domains, researchers can reconstruct functional interactions between and among proteins and predict previously unknown pathways.

Selecting Target Proteins for Structure Determination

The determination of a protein's three-dimensional structure is key to unlocking biologic function. At this time, it still not feasible technically to determine the structures of all of the proteins encoded in the human genome, or even in a typical smaller prokaryotic genome. However, considerable information relating to a protein's structure may be gleaned from studying its sequence. This is because there exists a limited number of distinct protein building blocks, or folds. Proteins with similar sequences tend to have similar folds and hence, similar structures. This suggests that for each sequence, researchers should be able to identify a homologous protein with a known structure that may serve as a model for the structural characterization of other proteins.

To explore this concept, **Wolf** and **Koonin** constructed a protein-fold recognition procedure based on a method for iterative searching of sequence databases. Using this approach, they determined that the distribution of the most common protein folds is similar in bacteria and archaea, but distinct in eukaryotes, demonstrating the ability of this method to detect subtle relationships between proteins from various phylogenetic lineages that were previously only detectable by structure-structure comparisons. Based on these results, investigators felt this method was both a sensitive and reliable procedure for determining potential targets, or a representative set of protein folds that would allow researchers to predict the structures for the rest of the proteins encoded in an organism's genome with confidence and in reasonable detail.

The next step was to determine the number of structures needed to obtain characterized representatives for nearly all folds. **Wolf, Grishin,** and **Koonin** devised a mathematical model that described the distributions generated by randomly sampling from the universal population of protein folds and families. They used this equation to estimate the number of folds and families in the protein universe and in complete genomes. The total number of folds in globular, water-soluble proteins was estimated at approximately 1,000, with structural information available for about one-third of these proteins. The number of protein families that show significant sequence conservation was estimated to be between 4,000 and 7,000, with structures available for about 20 percent of these. To cover all folds, one needs to structurally characterize approximately 85 percent of the protein families, as many folds contain only one or two families. Yet, the current number of structurally characterized protein families is only between 15 and 25 percent of the required number. These data emphasize the need to carefully select targets for protein structure determination so as to maximize the chance of obtaining structures from new folds.

Eukaryotic Genomes

Wolf, Kondrashov, and **Koonin** used comparative genomics to further our understanding of the origins of introns, the sequences that interrupt eukaryotic genes and comprise the most important feature that distinguishes eukaryotic genes from prokaryotic ones. They compared the protein-coding sequences of the roundworm, a multicellular eukaryote, against a complete, non-redundant protein database. Results demonstrated that a large number of the eukaryotic proteins showed significantly greater similarity to bacterial homologs than to archaeal ones and that some proteins even had a greater resemblance to their bacterial counterparts than to those from other eukaryotes. In addition, approximately 1,300 "ancient" genes were identified—genes that were more or less conserved in both archaea and bacteria. Next, they estimated and compared the average intron density in roundworm "ancient" and "bacterial" genes as it has been hypothesized that the protein-coding genes of the last universal common ancestor contained introns. If this were true, then the genes of ancient and bacterial origin should differ in their intron densities because genes acquired from bacteria had only a limited time to accrue introns. Yet data did not show a statistically significant difference in intron density between these two gene categories, lending credence to a second theory that postulates that introns invaded genes after the emergence of eukaryotes.

These brief research highlights demonstrate the impact that molecular analysis of genomic data, combined with modern computational and theoretical approaches, can have on furthering our understanding of the evolutionary, fundamental and practical problems facing biomedical researchers today.These studies also show the present utility and future potential of complete genome comparisons in identifying gene products produced by a particular organism and in predicting their structure and function.Using this approach, one can also identify a gene that is common to all organisms within the three domains of life, as well as a gene that is unique to a particular domain, thereby gaining meaningful insights into the organization and evolution of biological systems. —*CB*

Mouse Genome Resources at NCBI

The Mouse Sequencing Consortium (MSC), an international public-private effort to accelerate the sequencing of the mouse genome, recently announced that it has achieved its goal to generate three-fold coverage of the mouse DNA sequence.

Rapid Access to Trace Data

The MSC used a whole genome shotgun sequencing approach, generating 95% of the sequence of the mouse genome, albeit in small, unordered fragments. The shotgun reads are available from NCBI's Trace Archive, a novel type of database established to make the individual raw sequence reads publicly available. To date, the MSC has deposited more than 15 million individual unique mouse sequence traces, searchable by BLAST.

Providing rapid access to raw data, the Trace Archive serves as a repository for all trace data generated at the major centers involved in sequencing efforts of various organisms. Ancillary information further describing each of the traces is also available. In order to ensure that the public databases remain current and comprehensive, NCBI exchanges data regularly with the Ensembl Trace Server located at the Sanger Center. The Trace Archive can be searched at www.ncbi.nlm.nih.gov/Traces/.

Mouse Genome Map Viewer

The mouse genome sequencing project will next utilize larger stretches of DNA of known map position, and assemble the fragmentary pieces of sequence into the finished, highly accurate sequence of the mouse genome. To accommodate this new sequence, NCBI has created a Mouse Genome Map Viewer, similar to that used for viewing the human genome.

The map viewer currently displays a genetic linkage map, generated from data available from the Mouse Genome Database, and a radiation hybrid map from the Whitehead Institute and MRC-Harwell that includes genetic loci, gene-bases STSs, and simple sequence length polymorphisms. The data are searchable by map position, gene symbol, gene name, or marker name. The Mouse Map Viewer can be accessed from the Genomic Biology page at www.ncbi.nlm.nih.gov/Genomes/.

Other resources include a special BLAST form that facilitates BLAST searches of finished mouse genome sequence (available from the Mouse Genome Sequencing page) and a Human-Mouse Homology Map at www.ncbi.nlm.nih.gov/Homology/.
—CB, BR

UniSTS Integrates Markers From Multiple Sources

The UniSTS database provides more detailed information about Sequence Tagged Sites (STSs) shown on the STS map of the Human Genome Map Viewer. For each STS marker, UniSTS displays primer sequences, product size, mapping information, and cross references to LocusLink, dbSNP, RHdb, GDB, MGD, and the Map Viewer. The report also lists any GenBank or RefSeq records containing the primer sequences, as determined by NCBI's electronic-PCR service.

UniSTS integrates marker and mapping data from several public resources including eight human maps, three mouse maps, and GenBank. It can be searched using a marker name, accession number, gene symbol or text terms from gene descriptions.

GenBank Mirror Sites

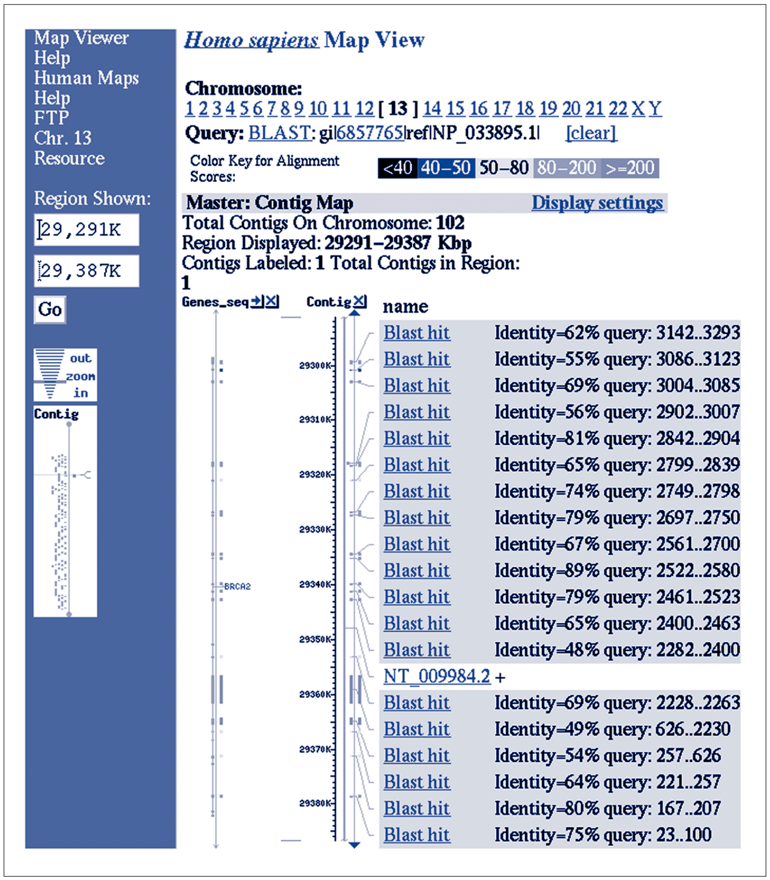
Alternate distribution sites are available for downloading full releases and daily updates of GenBank. Complete mirrors of ftp://ncbi.nlm.nih.gov/genbank/ are offered by the San Diego Supercomputer Center (SDSC) and the Bio-Mirror project.

The SDSC site is located at: ftp://genbank.sdsc.edu/pub/.

The Bio-Mirror site is located at: ftp://bio-mirror.net/biomirror/genbank/.

BLAST Your Way Into the Human Genome

BLAST searches can be run against the latest NCBI Human Genome assembly using Human Genome BLAST available from www.ncbi.nlm.nih.gov/BLAST/. In this BLAST example, we will try to locate the human homologue of the mouse Brca2 gene. Because amino acid sequence is more strongly conserved across species than nucleotide sequence, we will perform a tblastn search using the protein sequence for mouse Brca2 (NP_033895) as our query. Next, we will look for a region of the human genome which, when translated, matches the protein sequence of mouse Brca2. Here will use the default database "genome" in order to search against the human genomic sequence. Other database options include "mrna" and "protein", which allow searches of NCBI-predicted mRNA and protein sequences respectively. Searches of the human genome are filtered for low-complexity and repetitive elements by default, and there is no need to change this setting for this search.



Region Shown: 29,291K 29,387K

Homo sapiens Map View

Chromosome:
1 2 3 4 5 6 7 8 9 10 11 12 **[13]** 14 15 16 17 18 19 20 21 22 X Y

Query: BLAST: gi|6857765|ref|NP_033895.1| [clear]

Color Key for Alignment Scores: <40 40–50 50–80 80–200 >=200

Master: Contig Map Display settings

Total Contigs On Chromosome: **102**
Region Displayed: **29291–29387 Kbp**
Contigs Labeled: **1** Total Contigs in Region: **1**

name	
Blast hit	Identity=62% query: 3142..3293
Blast hit	Identity=55% query: 3086..3123
Blast hit	Identity=69% query: 3004..3085
Blast hit	Identity=56% query: 2902..3007
Blast hit	Identity=81% query: 2842..2904
Blast hit	Identity=65% query: 2799..2839
Blast hit	Identity=74% query: 2749..2798
Blast hit	Identity=79% query: 2697..2750
Blast hit	Identity=67% query: 2561..2700
Blast hit	Identity=89% query: 2522..2580
Blast hit	Identity=79% query: 2461..2523
Blast hit	Identity=65% query: 2400..2463
Blast hit	Identity=48% query: 2282..2400
NT_009984.2 +	
Blast hit	Identity=69% query: 2228..2263
Blast hit	Identity=49% query: 626..2230
Blast hit	Identity=54% query: 257..626
Blast hit	Identity=64% query: 221..257
Blast hit	Identity=80% query: 167..207
Blast hit	Identity=75% query: 23..100

Figure 1: *Human Genome BLAST Hits: Genome View*

The human genome BLAST search results returned are much like those returned from a conventional BLAST search, with the exception that the hits are only to NCBI-constructed contig sequences rather than to user-submitted GenBank sequences. A button is provided that allows a BLAST user to view hits on the human genome. When this button is pressed, a Human Genome Map View is generated, showing the positions of the BLAST hits on the Contig and Genes_ seq maps, as shown in Figure 1. On the Genes_ seq map, one sees a thin line representing the BRCA2 gene, interspersed with thick segments, representing exons. To the right of this line is an array of line segments representing BLAST hits, color coded by quality as indicated on the scale at the top of the display. Note that these BLAST hits appear to track closely the exons of BRCA2. Links to "BLAST hit" lead to a conventional pairwise alignment display.

The two most extensive matches, receiving the largest BLAST scores, are to two large exons found towards the bottom of the display. However, the percent identities for these two hits are 49% and 54% respectively, two of the lowest percentages shown. These two regions are also well-populated with SNPs, as seen in Figure 1 of "Tour the Human Genome" in this issue. Although SNP sampling bias in these regions cannot be ruled out, the variation within the human gene in this region is consonant with the greater degree of variation between the mouse and human in these regions.

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.

Recent BLAST Enhancements, New BLAST Features

The BLAST home page has been restructured so that searches using the principal varieties of BLAST—blastn for nucleotides and blastp for proteins—are separated from those involving the translation of a DNA sequence—blastx, tblastn, and tblastx. Special BLAST pages that use pre-set parameters optimized for finding short nucleotide or short peptide matches have also been created. MegaBLAST offers an alternative to blastn large nucleotide queries or batches of multiple query sequences.

Another important enhancement to the BLAST service is the ability to limit searches to a database subset defined by an Entrez query. For example, to limit a blastp search to viral capsid proteins, enter the following query into the new "Limit by Entrez query" box: Viruses [Organism] AND capsid [Protein Name]

To facilitate searches using unique parameter sets, custom parameters for BLAST searches may be saved within a BLAST URL, then bookmarked. Choose the desired parameters on the BLAST page, press the Get URL button, and a link to a new page with your parameters set in the URL will be generated.

As a new output option, XML has now been added to HTML, Plain-Text, and ASN.1.

Web PSI-BLAST has been enhanced to accept a Position Specific Score Matrix (PSSM) that can be pasted into the upload box. A reciprocal output option returns PSI-BLAST results as a PSSM, rather than a sequence alignment. The "bioseq" format returns results in the ASN.1 format.

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