# NCBI News

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

Volume 14, Issue 2


## Open Mass Spectrometry Search Algorithm (OMSSA)

Many years ago proteins were the biological molecules most often identified, sequenced and quantified by biologists. However, the advent of rapid methods of nucleic acid sequencing and measurement made DNA and RNA the center of attention—until recently. But proteins are back, now that newer mass spectrometer technologies allow the identification and analysis of proteins from complex biological samples. This current proteomics boom requires efficient computational methods of protein identification since analysis can involve thousands of peptide mass spectra derived from the coupled liquid chromatography–mass spectrometry of biological samples. NCBI's Open Mass Spectrometry Search Algorithm (OMSSA)[1] is a free search engine for analyzing and identifying peptides from tandem mass spectrometry (ms/ms) peptide spectra. The OMSSA algorithm scores peptide hits using a probability-based method that compares experimental fragments with those calculated from libraries of known protein sequences. The statistical model used by OMSSA is similar to the one used in the BLAST algorithm. OMSSA uses an expected value significance threshold, familiar to users of BLAST, to discriminate true matches from those that may be due to

*continued on page 4*

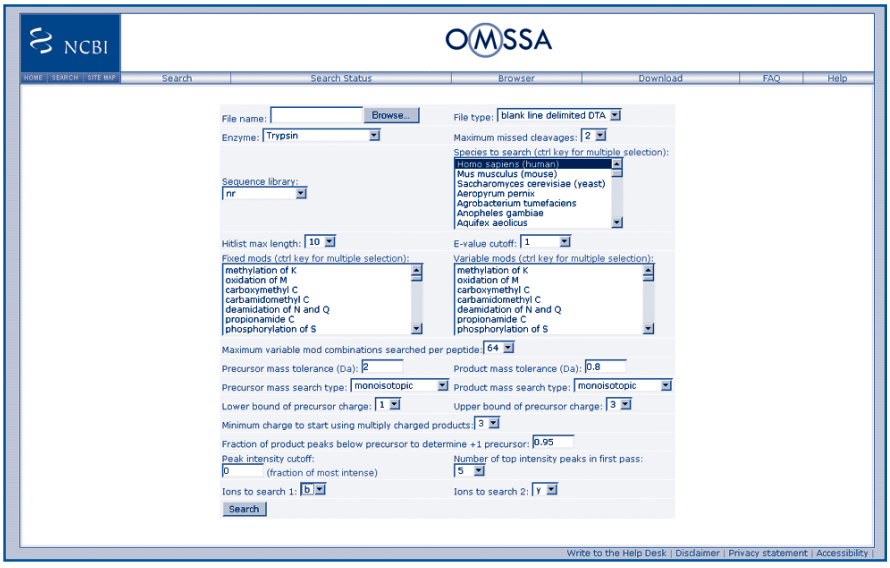


**Figure 1.** The Web interface to the OMSSA search tool. Cleavage and modification conditions, database and species selection as well as various settings, including filtering options and stringency, can be specified.

## The Probe Database Makes its Debut in Entrez

Nucleic acid probes are molecules that complement or hybridize to a specific mRNA or DNA sequence and are designed to target a gene so that information about its location, structure and function can be obtained. Probes are important reagents used in a wide range of biomolecular studies. In many cases, however, the molecular probe sequences have not been deposited in primary nucleotide sequence databases such as GenBank because these sequences are derived from longer sequences already present in the database. Moreover, many impor-

*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. To subscribe to NCBI News, send your name and address to either the street or E-mail address below.

NCBI News
National Library of Medicine
Bldg. 38A, Room 3S-308
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

*Contributors*
Susan Dombrowski
Emir Khatipov
Monica Romiti
Eric Sayers
Tao Tao

*Writers*
Peter Cooper

*Editing and Production*
Robert Yates

*Print & Web Design*
Robert Yates

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. 05-3272

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


## New Related Structures Link from Proteins in Entrez

One of the first steps in modeling the 3D structure of a protein is to find solved structures of proteins that have a high degree of sequence similarity to a target sequence. Because only about 1% of the protein sequences in the Entrez protein database are derived from protein structures, it will usually be necessary to find sequence-related structures in order to link a protein to a 3D structure. Entrez now displays new Related Structures links that perform this function with a single click of the mouse. Each protein sequence in Entrez Protein that has a BLAST hit to a structure-derived sequence now has a Related Structure link in the Links menu to the right of the record in the Entrez display; almost 36% of the 7.3 million protein sequences in Entrez have such links. Following this link displays the BLAST alignments of the related structures to the query either graphically or as a table, and these results can be sorted by BLAST score, E-value, aligned length, or sequence identity. Clicking on any of the alignment bars displays a detailed pairwise alignment and allows the alignment to be loaded into Cn3D for viewing. Figure 1 displays the non-identical related structures to NP_690059, the NCBI RefSeq for coagulation factor VII in rat. Sequence-similar structures align to both the catalytic heavy chain and the calcium-binding light chain of the protein. Figure 2 shows alignment of the query to 1KLI_H, the heavy chain of the human homolog displayed in Cn3D.

*—ES*

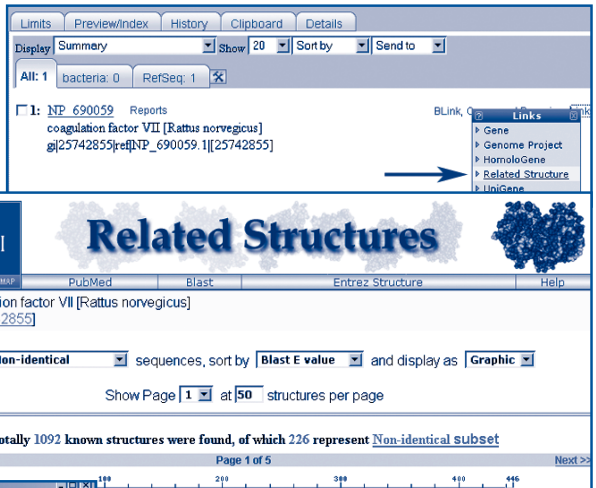


**Figure 1.** The Entrez links menu for the NCBI RefSeq for rat coagulation factor VII showing the "Related Structure link. This link leads to the graphical display of alignments of sequence-similar proteins from the structure database. Clicking on the graphic alignment with 1KLH_H displays the sequence alignment that can then be mapped onto the structure using Cn3D.

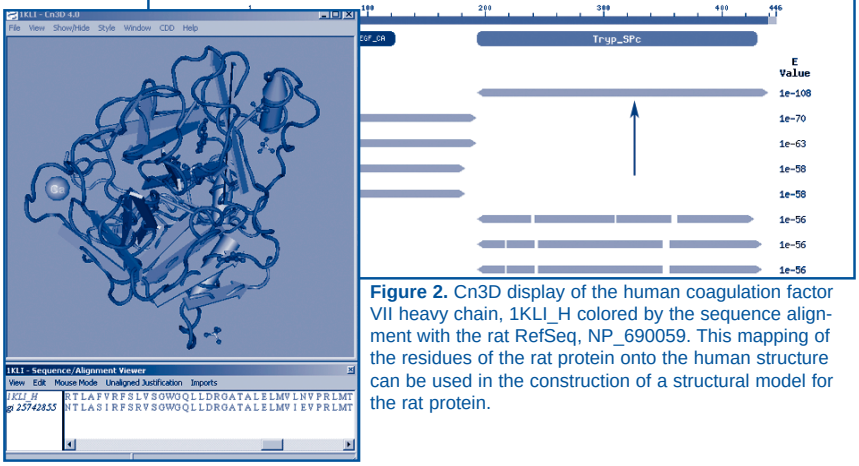


**Figure 2.** Cn3D display of the human coagulation factor VII heavy chain, 1KLI_H colored by the sequence alignment with the rat RefSeq, NP_690059. This mapping of the residues of the rat protein onto the human structure can be used in the construction of a structural model for the rat protein.

## Update: Downloading Preformatted BLAST databases

Databases for use with BLAST have traditionally been generated from text files of sequences in FASTA format using the BLAST utility program 'formatdb'. As the sequence databases have grown, standard BLAST databases in FASTA format have become increasingly unwieldy to download and store locally. To help alleviate the burden of maintaining these large FASTA files locally, NCBI provides preformatted BLAST databases at:

ftp.ncbi.nih.gov/blast/db

Although these preformatted databases have been available for a few years, many people continue to download the FASTA formatted databases. This is quickly becoming a Herculean feat as the largest of these FASTA files now exceeds 40 Gigabytes. Using preformatted files has several advantages — shorter download times, no need for running formatdb, and inclusion of taxonomy information for database entries. If needed, the entire FASTA formatted version of the database or any subset of sequences can quickly be generated from the formatted database using the BLAST utility fastacmd (See BLAST Lab, NCBI News, Summer 2003). For instance,

fastacmd nr -D1

extracts the entire preformatted nr database in FASTA format.

Larger formatted databases such as nt, est and wgs are broken into smaller one-Gigabyte volumes. A PERL script is available to simplify this task of keeping preformatted BLAST database current.

www.ncbi.nlm.nih.gov/blast/docs/update_blastdb.pl

The script can be run regularly to help maintain local copies of the BLAST databases.

More detailed information about using the preformatted databases is available from the BLAST ftp site.

ftp.ncbi.nih.gov/blast/db/blastdb.txt

For questions regarding the BLAST or services, please contact the BLAST help desk.

blast-help@ncbi.nlm.nih.gov

## Selected New Microbial Genomes in GenBank®

| Organism | GenBank \| RefSeq Accession Numbers |
|---|---|
| *Erlichia ruminantium str. Gardel* | CR925677 \| NC_006831 |
| *Wolbachia endosymbiont strain TRS of Brugia malayi* | AE017321 \| NC_006833 |
| *Sulfolobus acidocaldarius DSM 639* | CP000077 \| NC_007181 |
| *Rickettsia felis URRWXCal2* | CP000053—55 \| NC_007109—11 |
| *Corynebacterium jeikeium K411* | CR931997 \| NC_007164 |
| *Psychrobacter arcticum 273-4* | CP000082 \| NC_007204 |
| *Candidatus Pelagibacter ubique HTCC1062* | CP000084 \| NC_007205 |
| *Xanthomonas oryzae pv. oryzae KACC10331* | CP013598 \| NC_006834 |
| *Vibrio fischeri ES114* | chromosome: CP000020—21 \| NC_006840—41<br>plasmid: CP000022 \| NC_006842 |
| *Neisseria gonorrhoeae FA 1090* | AE004969 \| NC_002946 |
| *Bacteroides fagilis NCTC 9343* | chromosome: CR626927 \| NC_003228<br>plasmid: CR626928 \| NC_006837 |
| *Chlamydophila abortus S26/3* | CR848038 \| NC_004552 |
| *Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67* | chromosome: AE017220 \| NC_006905<br>plasmid: |
| *Brucella abortus biovar 1 str. 9-941* | chromosome: AE017223—4 \| NC_006932—3 |
| *Pesudomonas syringae pv. syringae B728a* | CP000075 \| NC_007005 |
| *Xanthomonas campestris pv. campestris str. 8004* | CP000050 \| NC_007086 |
| *Pseudomonas fluorescens Pf-5* | CP000076 \| NC_004129 |
| *Staphylococcus haemolyticus JCSC1435* | AP006716 \| NC_007168 |
| *Pseudomonas_syringae_phaseolicola_1448A* | chromosome: CP000058 \| NC_005773<br>plasmid: CP000059-60 \| NC_007274—5 |
| *Mycoplasma hyopneumoniae J* | AE017243 \| NC_007295 |
| *Mycoplasma synoviae 53* | AE017245 \| NC_007294 |
| *Dechloromonas aromatica RCB* | CP000089 \| NC_007298 |
| *Streptococcus pyogenes MGAS5005* | CP000017 \| NC_007297 |
| *Streptococcus pyogenes MGA6180* | CP000056 \| NC_007296 |
| *Mycoplasma hyopneumoniae 7448* | AE017244 \| NC_007332 |
| *Thermobifida fusca YX* | CP000088 \| NC_007333 |

For more detailed information, see the online version of the August 2005 NCBI News, or use the GenBank or RefSeq Accession Number to search the Entrez "Genome" database using the query box on the NCBI Home Page.

## *OMSSA*
*continued from page 1*

chance. OMSSA is very effective at identifying spectra from standard protein cocktails at high speed. The tool works with data from ion traps employing traditional charge associated dissociation, and electron transfer dissociation technologies.

OMSSA uses pre-formatted BLAST protein libraries from NCBI as the source of the calculated spectra, including the popular NCBI non-redundant protein (nr) and the NCBI reference protein (RefSeq) sets. Custom libraries of FASTA formatted sequences can also be used in the standalone version after processing with the NCBI BLAST utility, "formatdb". OMSSA is compatible with common data formats for experimental mass spectra including the "dta", "pkl", and "mgf" formats.

### The Web Interface

The Web interface to OMSSA, shown in Figure 1, is linked to the OMSSA homepage at:

pubchem.ncbi.nlm.nih.gov/omssa

Experimental spectra can be uploaded and compared to calculated spectra from the nr and RefSeq protein libraries limited to selectable organisms, making identification easier. Common digestion methods and post-translational modifications can be selected as well as stringency thresholds. Currently, the Web version can search up to 2000 mass spectra at a time.

Output can be viewed in the Web browser or can be saved in the "csv" format suitable for import into spreadsheets or in the OMSSA "omx" xml or "oms" ASN.1 formats that can easily be parsed by computer programs. The latter two formats can be viewed locally in the OMSSA browser as described below. A sample set of results is available for spectra from a standard mixture of four proteins. Figure 2 shows the results for the analysis of one spectrum from this sample set identifying chicken lysozyme C.

### The Standalone Interface

The standalone version of both OMSSA and the OMSSA browser can be downloaded from the NCBI ftp site at:

ftp.ncbi.nlm.nih.gov/pub/lewisg/omssa

Archives containing binaries for Linux, Mac OS X, or Windows are available. The standalone version offers options not available with the Web interface, such as the use of custom sequence databases. Output of the Web and standalone versions can be displayed with the OMSSA browser program that provides a convenient browser and graphical display for OMSSA output that is saved locally.

Additional information about OMSSA including help documentation and an extensive FAQ can be found on the OMSSA Homepage at:

pubchem.ncbi.nlm.nih.gov/omssa

Users can also subscribe to the OMSSA email list or browse through archived discussion threads from the list.

[1]Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. 2004. Open mass spectrometry search algorithm. *J. Proteome Res.* (5):958-64. PMID: 15473683

*—EK*

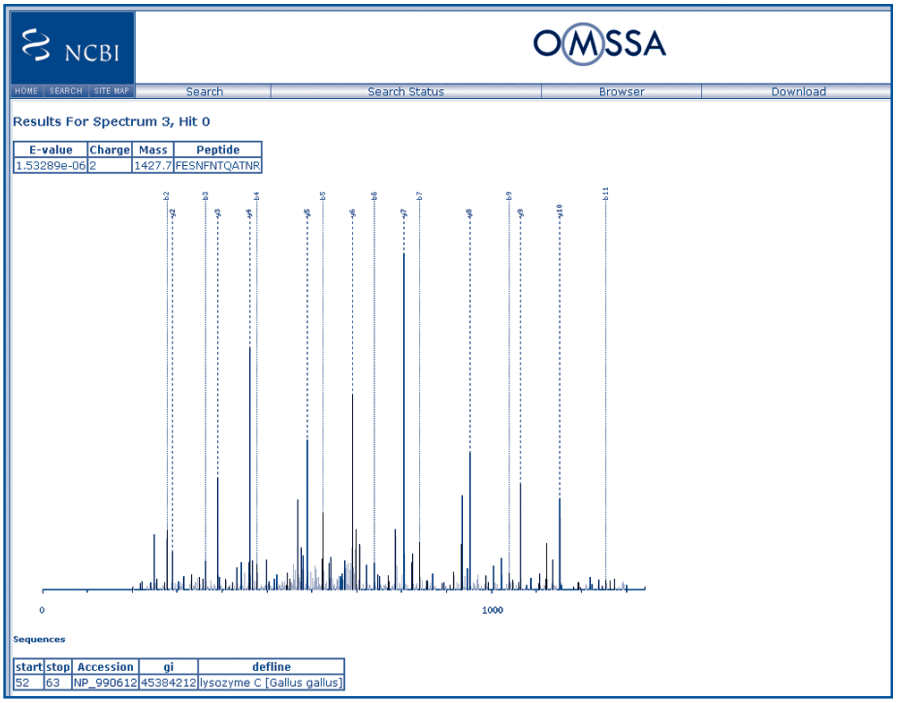


**Figure 2.** OMSSA identifies chicken lysozyme C from one of the spectra derived from a standard mixture of four proteins (chicken lysozyme, bovine serum albumin, bovine carbonic anhydrase II and horse heart myoglobin). The experimental spectrum is shown in black. The RefSeq protein library is the source of the calculated ions (numbered lines).The results also give peptide sequence and mass information, E value, m/z values of the ions, and provides the link to the source NCBI record.

## *Probe db*
*continued from page 1*

tant details about the probes, such as experimental application, the efficacy of the probe in that application, and information on obtaining these reagents are not normally available in traditional sequence records. A number of high-throughput genome analysis projects, among them ENCODE, HapMap, and GENSAT, are producing large numbers of molecular probes that have been used in a variety of experimental applications. The information obtained from these and other molecular studies will have a profound impact in the areas of functional genomics, diagnostic medicine, and agriculture. In order for researchers to have access to these and other molecular probes, NCBI has established the Probe database as a public repository for probe reagents.

Probe contains sequences of oligonucleotide probes and results from a wide variety of applications and technologies, including gene silencing experiments (RNAi), gene expression assays, variation analysis, genome mapping, nucleic acid detection, and genotyping studies. Currently, more than one million records are in the Probe database and include reagents and results for experiments in human, mouse, rat, fruit fly, and *C. elegans*.

### Querying the Database

The Probe database is now searchable as a part of the Entrez database system at NCBI

www.ncbi.nlm.nih.gov/entrez query.fcgi?db=probe

As with all other Entrez databases, search terms can be restricted to a specific field and combined using Boolean logic. A list of terms indexed under each field can be found by going to the Preview/Index tab on the Entrez Probe home page, selecting the appropriate search field and clicking "Index". For example, the following Entrez Probe query find probes that have been used to study the gene mutated in ataxia telangiectasia, ATM.

ATM [gene name]

This search returns a list of probes shown in Figure 1. The returned results contain probes designed for expression, gene silencing and variation studies. Results are easily filtered by application using the "Expression", "Silencing", or "Variation" tabs at the top of the display. Clicking on the probe ID displays an Entrez Probe summary report. Depending on the application or technology, the report includes the sense and/or antisense probe sequence, a link to order the probe from the manufacturer, supporting literature references with links to PubMed and a list of other potentially cross-reacting probes in the database, their location on the target mRNA and the experimental results.

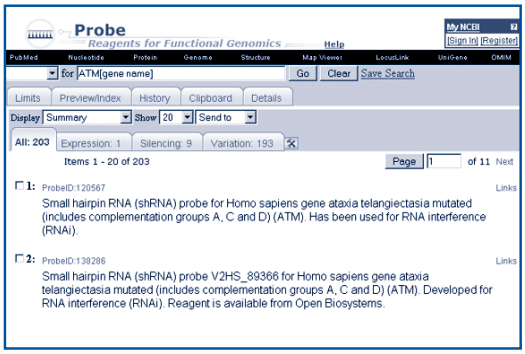


**Figure 1.** Summary results for a search for probes specific to the ATM gene.

*continued on page 6*


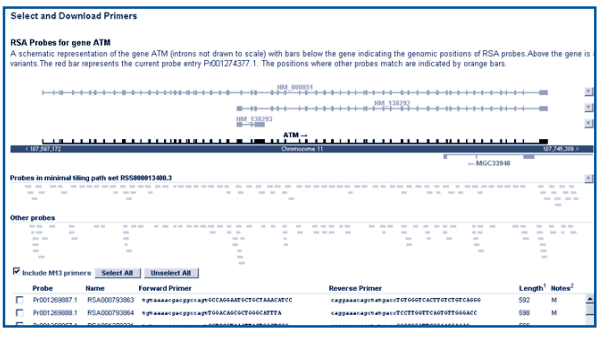


**Figure 2.** Samples of probe reports. A) a portion of the report for an expression probe for the ATM gene used in the GenSat project. The summary shows an image of an in situ hybridization experiment using this probe to label a whole mouse embryo section. B) A summary of the RSA probes that cover the exons of the human ATM gene.

An expression probe for ATM, #196090, was used in an in situ hybridization study of gene expression in the developing mouse brain These data are part of the Gene Expression Nervous System SAtlas (GENSAT) database reported in the May 2005 NCBI News. The summary for probes such as #196090 from GENSAT contains previews of corresponding *in situ* hybridization images in Entrez GENSAT (Figure 2 A). The gene silencing probes for ATM are RNA interference (RNAi) reagents used in sequence specific gene silencing experiments. These include small-hairpin (shRNA) and small interfering (siRNA) probes which target the ATM gene in

## *Probe db*
*continued from page 5*

human. An RNAi Homepage is also available for researchers interested in gene silencing.

www.ncbi.nlm.nih.gov/genome/RNAi

Queries on this page return only gene silencing reagents. The RNAi home page also contains LinkOuts to supporting references and reviews on RNAi technology, RNAi projects, probe design tools, companies specializing in RNAi technology, and recent primary literature and reviews from PubMed.

The remaining probes retrieved for ATM are used to study sequence variation. Probe #1274377 is a resequencing amplicon (RSA) - a set of primers designed for PCR amplification and resequencing of the exons of human genes. The RSA probes are intended for use in SNP discovery and are available for a large fraction of human genes. The summary for #1274377 shows the alignment of all RSA probes for the ATM gene and its transcript variants and includes a link to select and download primers (Figure 2 B). Another class of variation probes are those associated with the human haplotype mapping projects (HapMap).These are not directly associated with specific genes and are not found by searching for ATM in the probe database. HapMap probes that map to ATM region are easily found by searching for the ATM gene in the Entrez SNP database and following the links to Entrez Probe.

Researchers who wish to submit their unpublished or published data to the probe database should send an email request to

probe-admin@ncbi.nlm.nih.gov

Specific guidelines for submitting the results from RNAi studies are available online:

www.ncbi.nlm.nih.gov/genome/probe/StaticHTML/submission.html

Questions regarding Entrez Probe should be sent to

info@ncbi.nlm.nih.gov

*—SD*

## Entrez Nucleotide Database Splits

NCBI's traditional Nucleotide database has been divided into three component databases: "EST" containing Expressed Sequence Tags (ESTs); "GSS" containing Genome Survey Sequences (GSS); and "CoreNucleotide" containing the remaining nucleotide sequences. This division will facilitate more tightly focused searches.

A statistics line shows the results for the three component nucleotide databases and is shown in Figure 1. The new component databases are included within the Entrez linking scheme and links within and between databases can be selected as usual.

Entrez features such as the Limits, Preview/ Index, History, and MyNCBI tabs are maintained separately. Because unique search fields are available for each new Nucleotide component database more precise searches are possible. See the Entrez help document for more information and for searching details:

www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html

A more detailed article on the split of the Entrez Nucleotide database will appear in the next issue of the NCBI news.

*—MR*

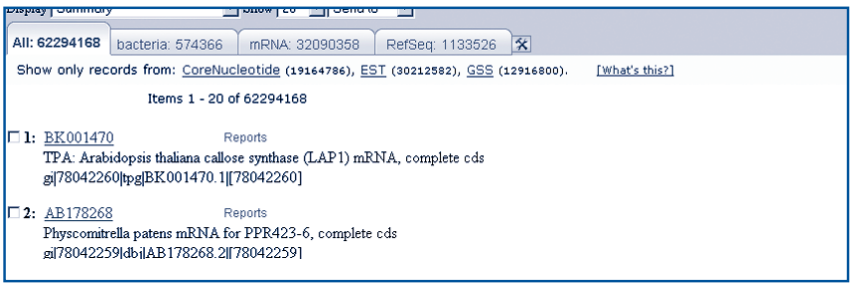


**Figure 1.** Results of the search "All[filter]" is shown in the statistics line below the filter tabs, indicating a total of 62.2 million records in the Entire Nucleotide division of the database info.

## NCBI Courses

NCBI 4-Pack: Practical Web-based Analysis. March 15-16, 2006 at the NCBI, Bethesda, MD

The National Center for Biotechnology Information (NCBI) presents NCBI 4-Pack: Practical Web-based Analysis, a 3-day course including both lectures/demonstrations and computer hands-on sessions on practical applications of bioinformatics resources.

The course is intended for both researchers and educators. Researchers will learn how to apply NCBI resources to their research.

To register, and for more information:

www.ncbi.nlm.nih.gov/Class/PowerTools/pack/application.html

## RefSeq Release 14

RefSeq Release 14 is now available by anonymous FTP at:

ftp.ncbi.nih.gov/refseq/release

Release 14 includes genomic, transcript, and protein sequences available as of November 23, 2005, from 3,198 organisms. The number of RefSeq accessions in Release 114 and their combined lengths is given in the shaded box.

RefSeq releases are posted every two months, and the next release is scheduled for January, 2006. Release notes documenting the scope and content of the database are provided at:

ftp.ncbi.nih.gov/refseq/release/release-notes

For more information, visit the NCBI RefSeq Web Site at:

www.ncbi.nih.gov/RefSeq

| | # of Accessions | # of Basepairs/Residues |
|---|---|---|
| Genomic | 547,997 | 46,370,766,763 |
| RNA | 589,641 | 994,188,604 |
| Protein | 2,135,138 | 763,761,075 |

## New Organisms in UniGene

UniGene now covers 55 animals and plants and can be searched using the Entrez search system where it is linked to nucleotide records. Recent additions to UniGene include *Schistosoma japonicum* (*bilharzia* or blood fluke) with 57,175 transcript sequences in 4,799 clusters, *Phytophthora infestans* (potato late blight pathogen) with 40,312 transcript sequences in 3,330 clusters, and *Macaca mulatta* (rhesus macaque) with 27,540 transcript sequences in 4,516 clusters.

## GenBank® Passes the 100 Gigabase Mark

With the August 2005 release of GenBank, the combined primary nucleotide database produced by GenBank and the collaborating European Molecular Biology Database (EMBL) and DNA Database of Japan (DDBJ) now exceeds 100 billion base pairs. The primary nucleotide data continues to grow at an exponential rate. During the period between August 1997 and August 2005 the database has grown 100 fold with an average doubling time of around 14 months. Improvements in sequencing technology and throughput indicate that the explosive growth of the primary data is likely to continue. In fact, another milestone was reached with release 149: the number of bases derived from whole genome shotgun (WGS) sequencing projects now exceeds the number of bases in the traditional divisions of GenBank (Figure 1). The WGS portion of the primary data is undergoing extremely rapid growth with the number of bases increasing more than ten fold in the past three years. There are 261 WGS projects in release 149 of GenBank including projects for human, mouse, rat, dog, numerous bacteria, and assemblies from environmental samples. With the sequencing of complete genomes becoming routine, genome sequence data will increasingly dominate the primary sequence data. The task of maintaining this data as a comprehensive and accurate resource is a primary goal of the NCBI.

The National Library of Medicine's press release provides additional information and commentary on the 100 Gigabase milestone:

www.nlm.nih.gov/news/press_releases/dna_rna_100_gig.html

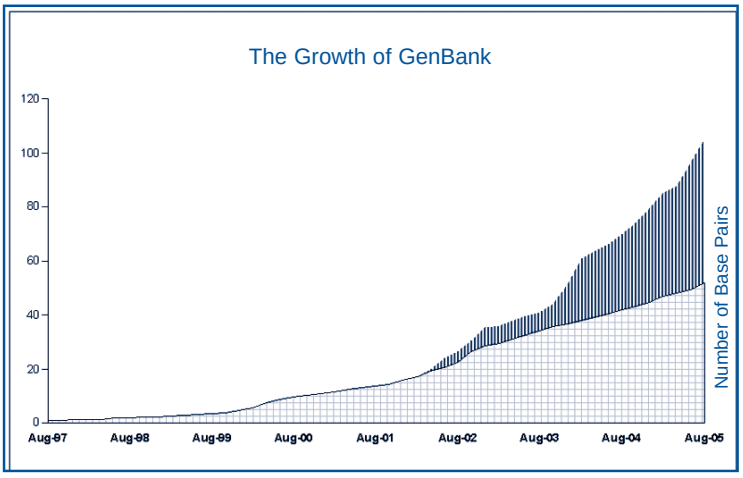


**Figure 1.** The growth of GenBank. The dark area shows the total number of bases including those from whole genome shotgun sequencing projects (WGS). The checkered area shows only the non-WGS portion. With release 149, the number of WGS bases exceeded the number of bases in the traditional GenBank divisions.

## New BLAST Formatter for BLAST Web Services—Old Formatter Retired

The enhanced BLAST formatter first announced as an option in the Summer/Fall 2004 NCBI News has now replaced the old formatter for the BLAST Web service. The new formatter has several features that make BLAST output more informative and easier to interpret. These include improvements to the graphical overview, reporting of features in database sequences, and clearer displays of low-complexity and other filtered regions in the query sequence.

The redesigned graphical overview is assembled within cells of an HTML table, allowing most browsers to easily print it. The new overview connects hits from the same database sequence with a crisp, thin line, making features like exon-intron structure or repeated domains more apparent. The overview of Figure 1 shows two regions of alignment, corresponding to exon sequences, between chimpanzee beta-2-microglobulin mRNA and the human genomic sequence query.

The new formatter also simplifies the interpretation of hits to large database sequences bearing many annotated features by giving links to the features that lie within or close to the match. These links, given for database sequences in excess of 200Kb in length, highlight associations between regions of alignment and biological features such as genes and repeat regions. The BLAST output of Figure 2 shows the match to the albumin gene in a whole genome shotgun supercontig from the dog genome. Use the link to serum albumin to generate a display of the relevant portion of NW_876257, the two megabase supercontig.

Perhaps the biggest improvement provided by the new formatter lies in its handling of masked, low-complexity regions within the query sequence. Low-complexity regions are those with biased amino acid or nucleotide compositions and are usually masked prior to a search in order to provide more meaningful alignments. In traditional BLAST output, one-letter codes for masked amino acids and nucleotides are replaced with X's and n's, respectively. The new formatter allows masked residues, to be displayed in lower case, in order to preserve the identities of the masked residues and in color for better highlighting. Matches within filtered regions are now taken into account when computing the percent identity for an alignment. The lower case option and the mask color selection are available in the 'Format' section of the BLAST submission or formatting pages. Figure 3 shows the new masking displays for the default replacement masking and the lower case masking that retains the original query sequence residues.

For questions regarding the BLAST services, please contact the BLAST help desk.

blast-help@ncbi.nlm.nih.gov

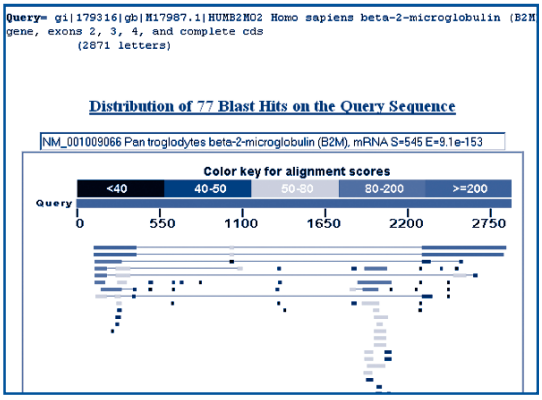


**Figure 1.** The new BLAST graphical overview. Thinner lines connect the two matches to the human beta-2-microglobulin of (query) exons from the chimpanzeee mRNA sequence.

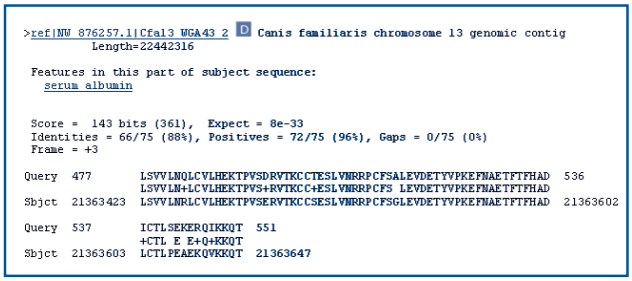


**Figure 2.** Alignment from a translating (tblastn) search of the human albumin protein against the dog genome. The new formatter display indicates that this hit lies within the annotation for the albumin gene on the supercontig, NW_87627.


**Figure 3.** BLAST protein alignments containing low complexity sequence. The upper alignment shows the default replacement masking. The lower alignments shows the lower-case masking option that preserves the query sequence in the output.

## Splign Transcript to Genomic Alignment Tool on the Web

One of the most reliable ways to identify genes is to align transcript sequences to a genomic sequence. Local alignment tools such as BLAST can quickly identify exons but do not include the nonaligning intronic segments in the alignment and lack precision at splice junctions. To produce accurate eukaryotic gene models from transcript alignments, a tool is needed that combines local and global alignment algorithms and accurately tracks splice junctions.

The new NCBI spliced-alignment tool, Splign, includes these design features and is used to help annotate higher eukaryotic genomes at NCBI. Now, in addition to being made available as a standalone tool, Splign is available through a Web interface. The Web interface to Splign as well as a link to download the standalone application and help documentation are available from the Splign homepage.

www.ncbi.nlm.nih.gov/sutils/splign

Splign generates transcript (cDNA) to genomic alignments that include detailed information about exon-intron boundaries, splice-junctions, potential frameshifts and other sequence discrepancies. Splign can also produce alternative models when there is more than one possibility. The Web version of Splign provides an interactive graphical view of the alignment or complete table of results. Figure 1 shows the results of a comparison between a transcript for the fruit fly "pxt" gene, given in GenBank record AF238306, and the sequence of the right arm of fruit fly chromosome 3, given in NCBI RefSeq NT_033777. In the figure, the fifth exon has been selected and the alignment for this segment is displayed. The intron-exon borders for the gene are easily identified both graphically and within the text sequence alignment. Statistics such as the length of each sequence in the alignment, the nucleotide positions of the exons, and the alignment coverage are provided. Sequence mismatches and insertion or deletions are color-coded for easy identification. In this segment, three mismatches and one small deletion in the transcript have been identified.

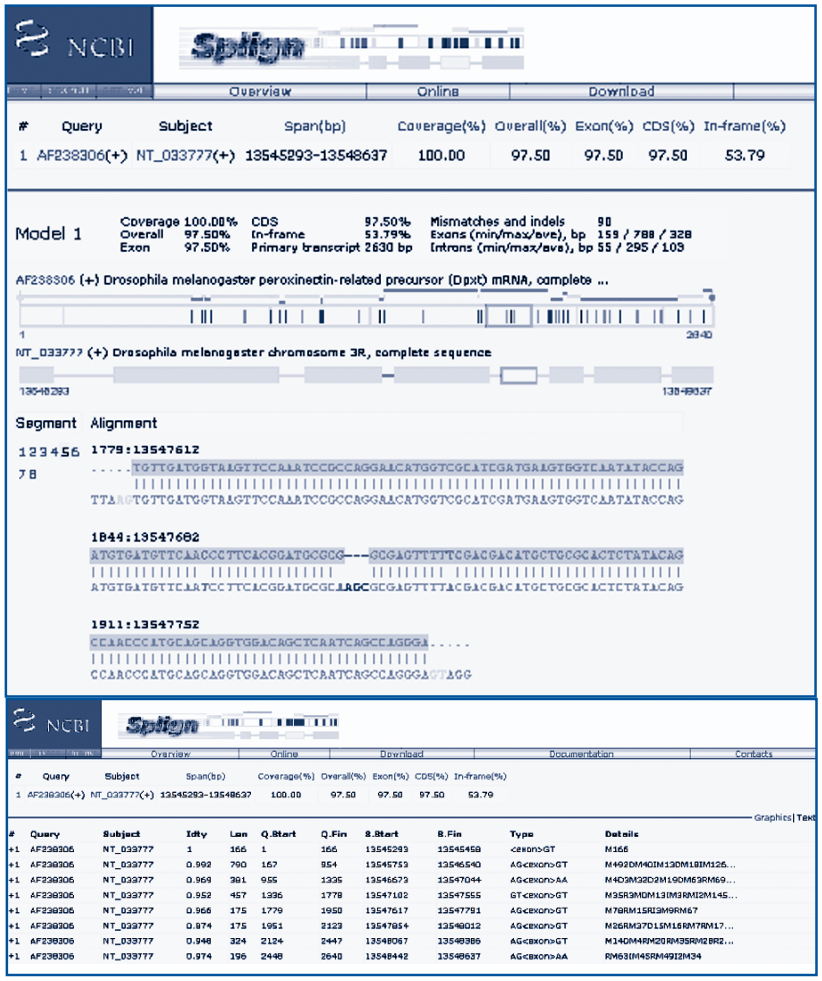


**Figure 1.** Results of mRNA (cDNA) to genomic alignment created by Splign. A. Graphical view the alignment between Drosophila melanogaster sequences AF238306 and NT_033777. B. Tabular format for the same alignment. Boundaries of the aligned regions of the query and subject sequences are shown along with the identified base pairs associated with the intron-exon splice junctions.

## GenBank® Release 150

GenBank Release 150 (October 2005) contains over 46 million sequence entries totaling more than 51 billion base pairs. Release 151 is expected in December. GenBank is accessible via the Entrez search and retrieval system. The flatfile and ASN.1 versions of the Release are found in the "genbank" and "ncbi-asn1" directories respectively at:

ftp.ncbi.nih.gov

Uncompressed, the Release 150 flatfiles consume about 205 Gigabytes while the ASN.1 version consumes about 164 Gigabytes. The data can also be downloaded at a mirror site:

bio-mirror.net/biomirror/genbank

## Genome Builds and Map Viewer Displays

### Map Viewer Highlights

Four important eukaryotic genome sequence assemblies and annotations are available in the NCBI Map Viewer. Classical model organisms, the zebrafish, the purple sea urchin *Strongylocentrodus pupuratus*, bread mold *Neurospora crassa* are displayed along with fungal pathogen *Cryptococcus neoformans* (also called *Filobasidiella neoformans*). This broadens the taxonomic diversity of genomes in the Map Viewer; the zebrafish is the first fish genome, and the urchin and the bread mold add new animal and fungal phyla (echinoderms and basidiomycetes).

The zebrafish genome build 1.1 is NCBI's assembly and annotation of the version 4 (Zv4) sequence produced by the Zebrafish Genome Project. This is 5.7X coverage sequence generated by whole genome shotgun and fingerprinted BAC clone sequencing. Genome features, markers, assembly contigs and components are anchored to the 25 zebrafish chromosomes. The zebrafish Map Viewer graphically displays these on the zebrafish genome sequence assembly. Available sequence maps include NCBI contigs (the "Contig" map), the WGS sequences (the "Component" map), and the location of genes, STSs, ESTs, UniGene clusters (the "ugDr" map), and Gnomon predicted gene models. The current annotation places 26,533 genes and their transcripts onto the zebrafish sequence. The zebrafish Map Viewer continues to display the genetic maps (GAT, HS, MGH, MOP, and ZMAP) and radiation hybrid maps (LN54 and T51) that are maintained in collaboration with ZFIN and members of the zebrafish research community.

Purple sea urchin (*Strongylocentrodus pupuratus*) build 1.1 is NCBI's annotation of the 6X WGS assembly produced by the Human Genome Sequencing Center at the Baylor College of Medicine. Unlike the zebrafish genome, none of the sea urchin assembled contigs are placed on chromosomes. The same sequence features and components that are available for the zebrafish genome can be displayed on the unplaced contigs in the map viewer by searching for markers on the sea urchin Map Viewer page or through the linked sea urchin genome BLAST page. The current build places 20,544 genes and their transcripts on the assembled sequence.

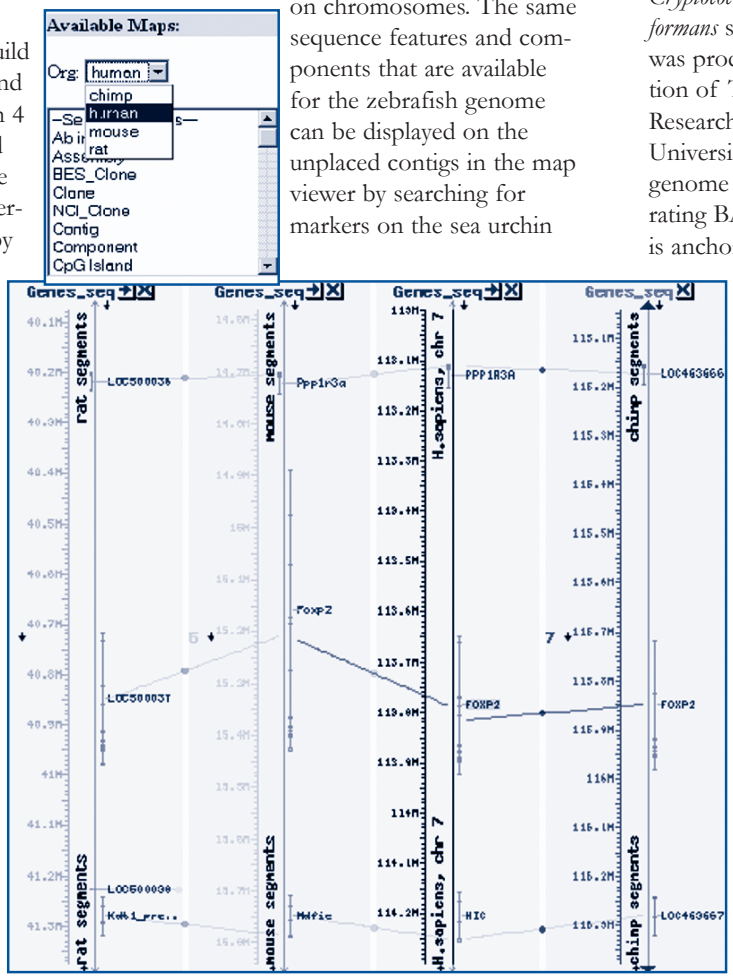


**Figure 1.** Comparative mammalian gene maps. Top: A portion of the "Maps and Options" dialog box showing the organism selection list. Bottom: Map View display of the region surrounding the speech and language gene, FOXP2, in rat, mouse, human and chimpanzee. Lines between maps connect homologous genes as identified in NCBI's HomoloGene resource.

The two new fungal genomes in the Map Viewer are assemblies and annotations provided to the NCBI by the respective sequencing centers. The *Neurospora crassa* genome (strain OR74A) is a 10X whole genome shotgun sequence produced by the Broad institute. The sequence is anchored to the seven *Neurospora* chromosomes. The current annotation contains 10,082 genes and their predicted transcripts. The *Cryptococcus* genome (*Cryptococcus neoformans* strain JEC21—serotype D) was produced through the collaboration of The Insititute for Genomic Research (TIGR) and Stanford University. This is a 10.5X whole genome shotgun assembly incorporating BAC clone end sequence and is anchored to the 14 *Cryptococcus* chromosomes. The *Cryptococcus* build maps 6,617 genes and their predicted transcripts onto the sequence. Available sequence maps in the Map Viewer for both of these fungal genomes include the assembled WGS supercontigs (the "Contig" map), the WGS sequences (the "Component" map), and the location of genes, STSs, ESTs, and UniGene clusters.

### Updated Genomes

The dog genome reported in the previous NCBI news (Volume 14, Issue 1) has been updated to build 2.1 which corresponds to the second version of the boxer genome assembly (CanFam2.0) by the Broad Institute and Agencourt

*continued on page 12*

# Submission Corner

## Submitting Sequence Polymorphisms to NCBI's dbSNP

Small genetic variations at specific positions in the genome, called single nucleotide polymorphisms or SNPs, are often responsible for phenotypic differences. The identification and analysis of SNPs in human and other complex genomes has become one of the major themes of biomedical research since the completion of the human genome sequence. The NCBI database of Single Nucleotide Polymorphisms (dbSNP) provides a public repository for this rapidly growing set of primary data and now contains over 40 million submitted SNPs from 33 different species. SNPs are submitted using a specialized protocol which involves the generation and transmission of a set of files to the NCBI SNP submission group. A brief outline of the SNP submission protocol and file types needed is presented below.

Submitters should begin by navigating to the detailed 'Quick Start' section on the SNP submissions page:

www.ncbi.nlm.nih.gov/SNP/get_html.cgi?whichHtml=how_to_submit

The SNP submission process is modeled on that of the GenBank bulk divisions - sequence tagged site (STS), genome survey sequence (GSS) and expressed sequence tag (EST). In fact, it is possible to simultaneously submit polymorphism data as a STS and a SNP. In all submission scenarios, the submitter creates a text file made up of a combination of required and optional sections — Contact, Publications, Method, Population Description and Assay, among others — for different types of information. Each section of the file is broken up into a set of fields identified by colon-delineated capitalized tags for the various types of data. The SNP submissions page mentioned above provides more information, including examples of the submission file format, and shows the possible sections and fields.

The submission file can be created using any standard text editor. However, electronic spreadsheet software can also be used to prepare the submission and can make the process easier. Figure 1 shows a portion of a spreadsheet used to generate a SNP submission. Data for each field are placed on the same row as the Field tag, and can also be entered on subsequent lines. Before submission the file must be saved in plain text format from the spreadsheet software. The completed submission file should be emailed to:

snp-sub@ncbi.nlm.nih.gov

SNP submissions can be made for either published or unpublished data.

Each submitted SNP is assigned an identifier of the form ss#, where "#" represents an integer identifier. The ss identifier serves the same purpose as an accession number for a GenBank sequence. NCBI also builds a non-redundant Reference SNP (RefSNP) database. Each RefSNP cluster, which is given an identifier of the form rs#, contains polymorphisms that map to the same position in the genome. RefSNPs are available as part of the Entrez database system and are linked to the primary SNP records as well as sequence, gene, genome, structure and functional information.

An example of a submitted SNP record that includes population and other detailed information for the human gene alcohol dehydregnase 2 can be seen on the following Web page:

www.ncbi.nlm.nih.gov/SNP/snp_retrieve.cgi?subsnp_id=3177110

Questions concerning snp submissions should be directed to:

snp-admin@ncbi.nlm.nih.gov

*—MR*

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | TYPE: | CONT | | | | | |
| 2 | HANDLE: | MYHANDLE | | | | | |
| 3 | NAME: | Lon Phan | | | | | |
| 4 | FAX: | 301 480 9241 | | | | | |
| 5 | TEL: | 301 594 8087 | | | | | |
| 6 | EMAIL: | lonphan@ncbi.nlm.nih.gov | | | | | |
| 7 | LAB: | Center for Medical Genetics | | | | | |
| 8 | INST: | National Institutes of Health | | | | | |
| 9 | ADDR: | 9600 Rockledge Dr., Bethesda, MD 20892, USA | | | | | |
| 10 | \|\| | | | | | | |
| 11 | TYPE: | PUB | | | | | |
| 12 | HANDLE: | MYHANDLE | | | | | |
| 13 | PMID: | 96172835 | | | | | |
| 14 | TITLE: | CpG islands of chicken are concentrated on microchromosomes | | | | | |
| 15 | AUTHORS: | McQueen,H.A.; Fantes,J.; Cross,S.H.; Clark,V.H.; Archibald,A.L.; Bird,A.P. | | | | | |
| 16 | JOURNAL: | Nat. Genet. | | | | | |
| 17 | VOLUME: | 12 | | | | | |
| 18 | PAGES: | 321-4 | | | | | |
| 19 | YEAR: | 1996 | | | | | |
| 20 | STATUS: | 4 | | | | | |
| 21 | \|\| | | | | | | |
| 22 | TYPE: | METHOD | | | | | |
| 23 | HANDLE: | MYHANDLE | | | | | |
| 24 | ID: | NCBI-Sequencing | | | | | |
| 25 | METHOD_CLASS: | Sequence | | | | | |
| 26 | SEQ_BOTH_STRANDS: | YES | | | | | |
| 27 | TEMPLATE_TYPE: | DIPLOID | | | | | |
| 28 | MULT_PCR_AMPLIFICATION | NO | | | | | |
| 29 | MULT_CLONES_TESTED: | NO | | | | | |
| 30 | METHOD: | The consensus DNA sequence for this polymorphism was obtained from separate, | | | | | |
| 31 | | overlapping sequences defined by the accompanying GenBank Accession Numbers. | | | | | |
| 32 | | The consensus sequence was confirmed by DNA sequencing. | | | | | |

**Figure 1.** A section of a spreadsheet for creating a SNP submission showing the Contact, Publications, Method and Population sections.

**Map Viewer**
*continued from page 10*

Bioscience. The current NCBI build shows 19,907 genes placed on the whole genome shotgun assembly.

### Other access to data

In addition to access through the Map Viewer, sequences of the genome assemblies, transcripts, protiens and gene models for zebrafish, sea urchin, *Cryptococcus*, Neurospora and the dog are available through the NCBI RefSeq database. The WGS assemblies are also available in GenBank under the following accessions: sea urchin, AAGJ00000000; zebrafish, CAAK00000000; *Neurospora*, AABX00000000; *Cryptococcus*, AE017341-AE017356; dog, AAEX02000000. GenBank and RefSeq records are available for searching in the Entrez text search system and NCBI's Web BLAST services where they are extensively integrated with other resources and databases.

### New chimpanzee comparative maps available

The human, mouse, rat and chimpanzee genomes can now be displayed side-by-side in the Map Viewer. This feature is available in the "Maps and Options" dialog box from the Map Viewer display of any of the four mammalian genomes as shown in the figure. The available maps can be changed to another mammal by choosing from the "Org" pull-down list. These tracks can be added to the current display. The corresponding or syntenic regions in other mammalian genomes are determined from gene homology relationships provided by NCBI's HomoloGene. Figure 1 shows the regions surrounding the transcription factor "foxp2", popularly known as the "speech and language gene", for all four mammalian genomes displayed in the human map viewer.

**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 3S308
8600 Rockville Pike
Bethesda, Maryland 20894

*Official Business*
*Penalty for Private Use $300*

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