NCBI News

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

Volume 15, Issue 2
Fall /Winter 2006/07

New PubMed AbstractPlus Display: Instant Access to Related Links

PubMed's new AbstractPlus display shows the titles of the top five related Articles and is now the default display for single records. The new format provides seamless access to the powerful pre-computed similarities available as Related Articles in the PubMed database.

These Pre-computed related items, sometimes called Neighbors, are records in the same Entrez database that are similar based on automated computational comparisons. Traditionally the Neighbors have been available through an item on the "Links" menu on displayed records. The display in Entrez ranks the Neighbors in descending order by the similarity metric as produced by the relevant comparison algorithm. In the molecular sequence and structure databases the comparison or Neighboring algorithms are familiar to most biologists: NCBI BLAST for sequences and VAST for Structure records. In PubMed, an algorithm compares meaningful words and phrases from the title, abstract, and Medical Subject Headings (MeSH) to calculate a set of PubMed citations that share these characteristics. In PubMed, this provides a rapid and powerful means of expanding a search and can find relevant articles that may be missed by the initial search. The utility of the new AbstractPlus display is easy to see when viewing the abstract of a 2006 article on improvements to the NCBI BLAST services by Jian Ye, Scott McGinnis, and Tom Madden (PMID: 16845079). This article along with another closely related article is

continued on page 6

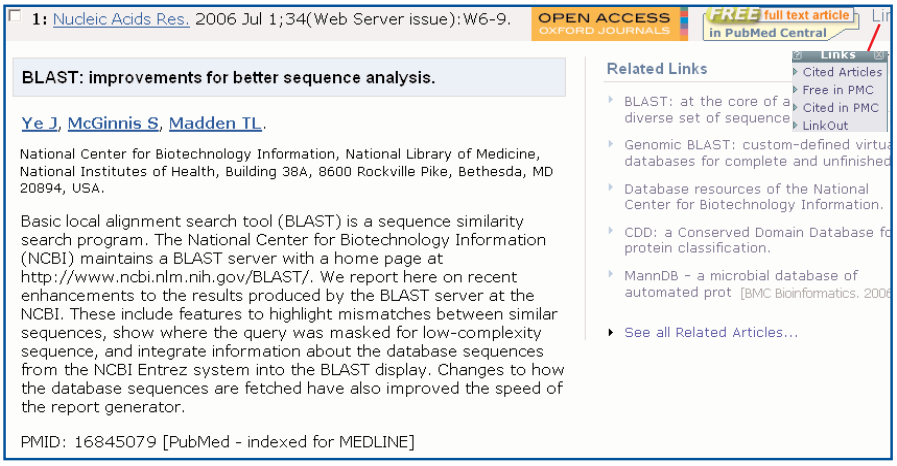

Figure 1. AbstractPlus display of a recent article on improvements to the NCBI BLAST services found using the query of: mcginnis[auth] AND BLAST. The top five articles are highly relevant to the topic but only one is found by the original query.

CDTree and Cn3D 4.2: New Tools for Analyzing Protein Domains

NCBI has released two software packages, CDTree 3.0, a completely new program for manipulating and analyzing Conserved Domain (CD) records, and Cn3D 4.2, an updated version of the NCBI 3D structure viewer that is now closely integrated with CDTree. CDTree can be used either as a helper application for a web browser or as a stand-alone application for creating and modifying CDs and CD hierarchies. Together these two programs form a powerful suite for defining and investigating protein domains based on both sequence and structural alignments. Instructions for downloading and installing the software package containing CDTree and

continued on page 4

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
Monica Romiti

Writers
Peter Cooper
Rana Morris
Eric Sayers
Robert Yates

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. 07-3272

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

The Growth in Number and Diversity of Whole Genome Shotgun Sequencing Projects at the NCBI

The number of Whole Genome Shotgun (WGS) sequencing projects with data in GenBank continues to grow at a rapid pace. Projects include not only single genomes from individual organisms but also metagenomes-- whole genome shotgun sequences from biological communities. There are now more than 400 projects listed in the WGS directory on the GenBank ftp site as of January 24, 2007.

ftp.ncbi.nlm.nih.gov/genbank/wgs

This article highlights some of the recent genome sequences and metagenomes and shows how to access these important data using the Entrez system and the NCBI BLAST services.

WGS sequence in both GenBank and the Entrez system are organized by project, and each project is assigned a master accession that begins with a unique four letter prefix. All sequences belonging to the same project have accessions that share the same four letter prefix. Examples of WGS master accessions can be seen in Table 1 below and Table 2 on page 8. As described in the Summer/Fall 2004 issue of the NCBI news, the Whole Genome Shotgun Projects page that provides a list of WGS projects and accessions is available on the NCBI Website:

www.ncbi.nlm.nih.gov/projects/WGS/WGSprojectlist.cgi

The Entrez Genome Projects Database

The Entrez Genomes Projects Database provides convenient access to all WGS genomes.

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

The following query in the Genome Projects page retrieves more than 400 summaries of projects with WGS data.

has wgs[Properties]

Each of these Project Summaries links to a Project Overview page that provides links to the sequencing center involved and explains the motivation for the project. Figure 1 shows the Project Overview page for the European rabbit. The data are linked through the "Project data" menu.

continued on page 6

Taxonomic Group	Number	Examples	Master Accession
Bacteria	269	*Desulfovibrio vulgaris* subsp. vulgaris DP4	AATW01000000
Archaea	4	*Thermofilum pendens* Hrk 5	AASJ00000000
Fungi	49	Amphibian chytrid	AATT01000000
Green Plants	4	Black cottonwood	AARH01000000
Animals	65		
Nematodes	2	*C. remanei* strain PB4641	AAGD00000000
Mollusks	1	California sea hare	AASC00000000
Insects	25	Yellow fever mosquito	AAGE00000000
Echinoderms	1	Purple sea urchin	AAGJ02000000
Sea Squirts	2	*Ciona savignyi*	AACT01000000
Fishes	5	Three-spined stickleback	AANH01000000
Birds	1	Chicken	AADN02000000
Mammals	30	African elephant	AAGU01000000

Table 1. Whole genome shotgun assemblies of single organism genomes with selected recent examples.

New Database and View Options for Nucleotide BLAST Services

The nucleotide-nucleotide BLAST pages linked to the BLAST homepage now offer the human and mouse transcript and genomic reference sequences as database options (Figure 1). A new output display for all searches provides a more organized presentation of the search results and has powerful new sorting options (Figure 2). In addition to the traditional sorting by Expect Value, results can now be sorted by Maximum Score, Total Score, Percent Query Coverage, and Maximum Percent Identity. The Box below provides definitions for these metrics. Sorting options are also available for multiple hits within the alignments for each subject (database) sequence with the additional option of sorting by query or subject position-an option especially useful for ordering potential exons on genomic matches (Figure 2 B). Matches to genomic sequences in the new genome and transcript databases can now be displayed in the mouse and human Map Viewers as with the specialized genomic BLAST services (Figure 2 C).

These new database and display options provide rapid access to the most popular annotated genomes at the NCBI and expand the power of BLAST as a genome annotation tool.

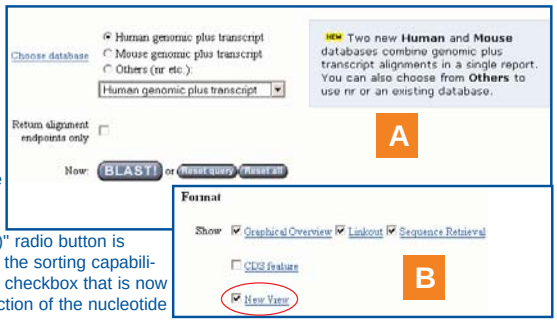

Figure 1. New database and display options for NCBI Web BLAST services. A. The human and mouse genomic and transcript databases can now be selected using the radio buttons on the BLAST form. The traditional BLAST databases are available through the pull-down list once the "Others (nr etc.)" radio button is selected. B. The improved display with the sorting capabilities is invoked through the "New View" checkbox that is now selected by default on the "Format" section of the nucleotide BLAST forms.

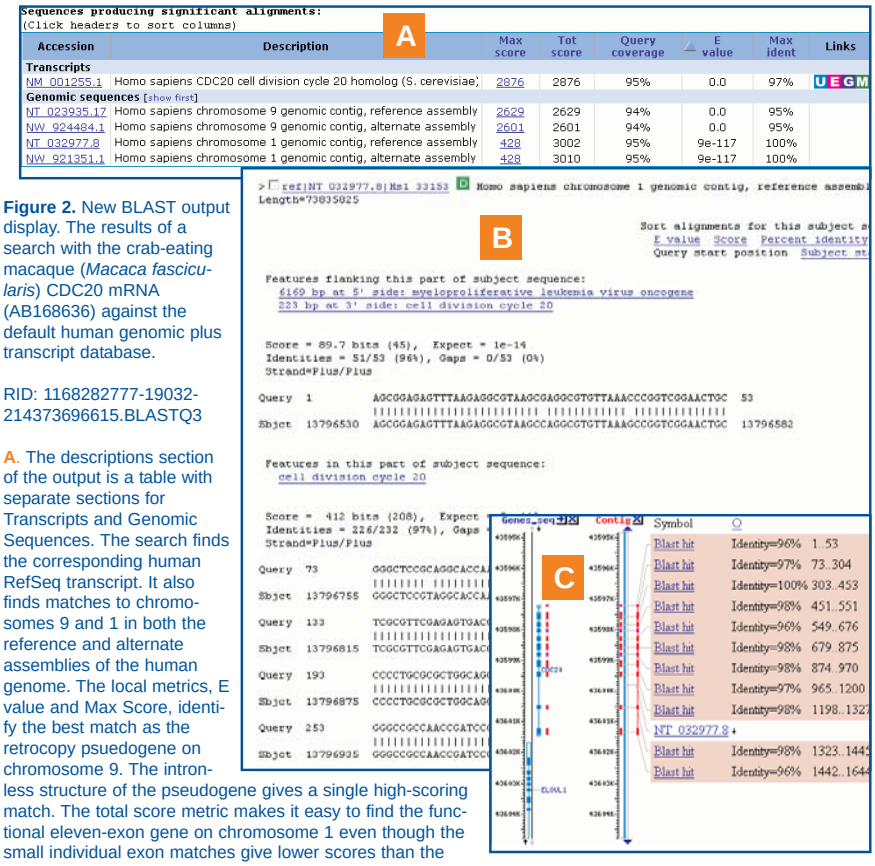

Figure 2. New BLAST output display. The results of a search with the crab-eating macaque (*Macaca fascicularis*) CDC20 mRNA (AB168636) against the default human genomic plus transcript database.

RID: 1168282777-19032-214373696615.BLASTQ3

A. The descriptions section of the output is a table with separate sections for Transcripts and Genomic Sequences. The search finds the corresponding human RefSeq transcript. It also finds matches to chromosomes 9 and 1 in both the reference and alternate assemblies of the human genome. The local metrics, E value and Max Score, identify the best match as the retrocopy psuedogene on chromosome 9. The intronless structure of the pseudogene gives a single high-scoring match. The total score metric makes it easy to find the functional eleven-exon gene on chromosome 1 even though the small individual exon matches give lower scores than the pseudogene. B. The alignment of the macaque mRNA to a contig on chromosome 1 from the reference human genome. Segments are sorted by "Query start position" to show the matches to the exons in order of position rather than E-value. C. Alignments displayed in the human map viewer are available by following the linked genomic sequence identifiers.

Local Extreme Metrics

These measures treat each aligned segment independently. Where there are multiple matches to the same subject (database) sequence, only the metric for the best match is considered. The ***E(xpect) Value*** is the traditional BLAST statistic used to sort output by significance.

E(xpect) Value — the number of alignments expected by chance with a particular score or better. The expect value is the default sorting metric and normally gives the same sorting order as ***Max Score***.

Max(imum) Score — the highest alignment score of a set of aligned segments from the same subject (database) sequence. The score is calculated from the sum of the match rewards and the mismatch, gap open and extend penalties independently for each segment. This normally gives the same sorting order as the ***E Value***.

Max(imum) Identity — the highest percent identity for a set of aligned segments to the same subject sequence.

Total Metrics

These metrics are summed over or include all aligned segments for the same subject sequence. These are most useful for analyzing BLASTmatches to genomic sequences.

Tot(al) Score — the sum of alignment scores of all segments from the same subject sequence. This sorting order may help promote the position of mRNA matches to genomic sequences where there are multiple exons. The Total Score is useful for distinguishing hits to functional multi-exon genes from those to the corresponding intronless retrotransposed psuedogenes (Figure 2).

Query Coverage — the percent of the query length that is included in the aligned segments. This is calculated over all segments as with the ***Tot Score***.

CD Tree
continued from page 1

Cn3D 4.2 are available on the CDTree homepage.

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

CDTree Displays CDs

The most basic use of CDTree is to view one of the pre-defined CD hierarchies accessible on the summary page for a curated CD record that, for example, can be found at

www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=cd02156

Above the Sub-family Hierarchy display is now a button labeled "Interactive Display with CDTree" that can launch CDTree either with either the current CD only or the entire hierarchy (Figure 1). When launched from this page, CDTree by default opens four windows: the Indent Tree, the main window showing the CD hierarchy; the Sequence Tree, showing a phylogenetic tree of the proteins within the CD hierarchy; and the Taxonomy Tree, showing the taxonomy nodes of the same proteins.

Figure 1 shows several CDTree and Cn3D 4.2 displays for the CD hierarchy (Figure 1 a) of the nucleotidyl transferase superfamily (cd02156). Sequence and Taxonomy Tree views of one child CD of this hierarchy, the catalytic core of the arginyl-tRNA synthetase (ArgRS_core, cd00671), are shown in Figure 1 b & c.

CDTree analyzes CD records

When launched from a CD web page, the initial CDTree views provide straightforward taxonomic and phylogenetic analysis through a variety of selection commands that highlight subsets of sequences. Highlighted sequences become red, and this highlight automatically propagates to the same sequences in all open viewers. In Figure 1, one branch of the phylogenetic tree was selected by clicking on the appropriate tree node, highlighting the corresponding sequences in the Indent Tree and Taxonomy Trees. The transferred highlights show that all of these sequences are from eubacteria and represent 17 of the 43 total sequences in the alignment.

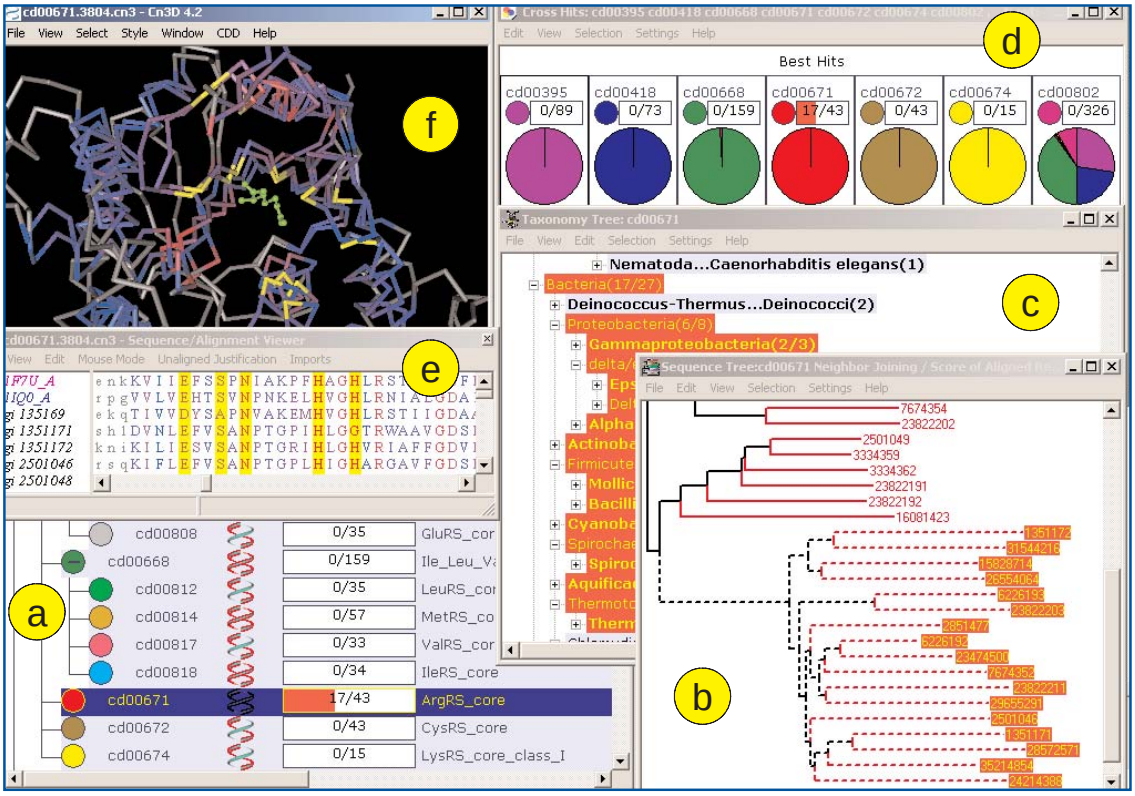

Figure 1. Hierarchy display for the CD family containing cd00671, the catalytic core domain of the arginyl-tRNA synthetase. Clicking the "Interactive Display with CDTree" button provides the option of loading either the single domain (cd00671) or the entire hierarchy into CDTree. Selected displays in CDTree and Cn3D 4.2 for the CD family containing cd00671. a) the Indent Tree display, showing the CD hierarchy with cd00671 selected; b) the Sequence Tree display, showing a neighbor joining tree for cd00671 with one branch highlighted; c) the Taxonomy Tree display, showing the taxonomic nodes represented by the sequences in cd00671; d) the Cross Hits viewer, showing the distribution of best BLAST hits for the six immediate child domains of the overall parent, cd00802; e) Alignment Viewer in Cn3D 4.2, showing the cd00671 alignment with the active site residues highlighted in yellow; f) Structure Viewer in Cn3D 4.2 showing the two aligned structures in cd00671 with the active site residues shown in yellow and the bound arginine in green. The red highlighting of the 17 sequences selected in panel b is automatically propagated to panels a, c and d.

Other analysis tools include the Cross Hits viewer, the CDART viewer, and the CDTree Validator. The Cross Hits viewer shows at a glance the quality of the hierarchy itself. Each CD is represented by a pie chart where the colored wedges indicate the proportion of sequences in that CD that has the best BLAST match to the CD that corresponds to that color. Thus, a well defined hierarchy will look much like the one in Figure 1 D, where each pie for a child CD contains only its own color, while the parent contains representatives for each of its children. The CDART viewer functions much like the existing CDART web service, showing the domain architectures of sequences in selected CDs. It allows, for example, the selection of sequences based on a shared architecture. The CDTree Validator performs several consistency checks on the alignments in a CD hierarchy, requiring that child CDs maintain the alignment of the parent and that no sequence violates the block alignment model of the CD.

CDTree updates CD records

Once a CD has been loaded, CDTree can automatically update the alignment with new sequences using Position Specific Iterative BLAST (PSI-BLAST) or standard protein-protein BLAST. After retrieving the BLAST results from NCBI, CDTree distributes the sequences to their best matching CDs, and these new sequences become "pending" rows, in contrast to the "aligned" rows already part of the block model.

CDTree uses Cn3D 4.2 to View and Edit CD Alignments

A mouse click on the double-helix icon next to any CD in the Indent Tree window launches Cn3D 4.2, where the alignment can be viewed and edited (Figure 1E). This function is available even for alignments that have no 3D structures. If the alignment does contain structural data, then these will be shown as a template to help guide the alignment process (Figure 1 f). Functional features annotated by CDD curators can also be highlighted on both the alignment and structures, as is the active site in Figure 1 E & F. If the CD contains pending sequences, these will be placed in the Imports window in Cn3D, where they can be aligned to the CD block model using any of the alignment algorithms in Cn3D. After this is complete, the Cn3D Save function exports this new alignment back into CDTree. By continuing this cycle, an entire hierarchy can be updated, one CD at a time. The colored box to the right provides a list and description of several new features of Cn3D 4.2 in addition to those related to the CDTree functions mentioned above. More information on Cn3D is available on the Homepage for the program.

www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml

CDTree creates CD records and CD hierarchies

One of the most powerful features of CDTree is that it can create *de novo* CD records, either as the beginning of a new hierarchy or a new child in an existing hierarchy. Quite simply, any group of sequences can be selected and assigned to a new CD, which then can be assigned a parent if appropriate. Alternatively, CDTree can start with a single protein sequence, run a BLAST search, and use the resulting alignment as the nucleus of a new CD record. A utility named "fa2cd" that comes bundled with CDTree can create CDs from legacy alignment data in FASTA format, which can then be imported into CDTree as well.

The combination of CDTree and Cn3D 4.2 provides a powerful software platform that integrates literature, sequence, taxonomic, phylogenetic, functional, and structural data for a protein domain in an interlinked set of views, allowing researchers both to visualize these data more fully and to identify new domains and domain features in poorly annotated protein families.

Selected New Features in Cn3D 4.2

Stereoview - Cn3D 4.2 can display customizable stereo images

PSSM Export - Cn3D 4.2 can export the Position Specific Score Matrix (PSSM) calculated for the current alignment in NCBI scoremat format, which can be added to an RPS-BLAST database or used to begin a PSI-BLAST search.

New Alignment Algorithms - Cn3D 4.2 now can apply the Block Aligner to the entire set of pending sequences at once, and can also align a pending sequence using BLAST to the most similar sequence in the alignment, rather than the master sequence. Cn3D 4.2 also contains a new alignment refiner that sequentially realigns each row in the existing alignment.

New Sequence Conservation Coloring Options - Cn3D 4.2 offers three new coloring options to assist in analyzing the quality of a block alignment model:

—Block Fit - colors all blocks across the entire alignment by how well they fit to the PSSM

—Normalized Block Fit - like Block Fit except that the colors are computed separately for each block, thereby showing which sequences have the best/worst scores for each block

—Block Row Fit - like Block Fit except that colors are computed separately for each row, thereby showing which block has the best/worst score for each sequence

Improved Highlighting - Cn3D 4.2 can highlight by blocks, can extend existing highlights to aligned columns, and can restrict existing highlights to a single row. Highlights can also be cached, or saved, and then can be recalled later. Users may appreciate that clicking in the white space of the sequence window no longer removes the highlighting, thereby saving some frustration! Finally, when launched from within CDTree, Cn3D 4.2 can exchange highlight messages with CDTree. For example, clicking on the double-helix icon next to a CD in the Indent Tree window will send highlights from CDTree to Cn3D, if that CD has been opened with the Cn3D viewer already.

PubMed AbstractPlus
continued from page 1

easily found by the following search in PubMed.

McGinnis S [Author] AND BLAST

Figure 1 shows the AbstractPlus display of this 2006 Nucleic Acids Research article. The first four of the top five "Related Links" are recent articles on NCBI databases and services, including BLAST services. The first of these is also found directly in this sample search and is cited in the current article. But the other four highly relevant articles would be missed unless other searches are performed. Moreover, the links to the AbstractPlus displays of interesting related articles quickly expands the search into other areas. For example, the top link quickly leads to articles about software packages for managing BLAST output or software interfaces for managing local installations of BLAST and other bioinformatics software.

Having the top five related articles on the same page in the AbstractPlus display saves valuable time and enhances the power of PubMed as a discovery system.

Whole Genome Shotgun
continued from page 2

Genome specific resources including, in some cases, a genome-specific BLAST service are linked under "Resource Links".

Individual Genome Sequences

WGS projects include over 250 bacteria and more than 120 eukaryotes from a wide range of taxonomic groups as shown in Table 1. In addition to the Rhesus macaque, reported in the last issue of the NCBI news, notable additions in the animals include the African elephant, the rabbit, the guinea pig, the shrew, the hedgehog, and the domestic cat. There are also a number of first genomes from an interesting array of taxonomic groups: a beetle, *Tribolium castaneaum*; a marsupial, the short-tailed opossum; and a monotreme, the duck-billed platypus, and the first tree species, the black cottonwood. Low coverage sequences of large genomes like that of the elephant may contain nearly a million individual sequences in GenBank. Master GenBank records that collect all of the WGS sequences for a particular project are useful for retrieving these projects using the Entrez system. The genome project overview pages, described above, provide access to the WGS data through these Master records. Master records can be used in the Entrez nucleotide database to retrieve the sequences for WGS projects. For example, the following query retrieves the master records for all mammalian WGS projects from the nucleotide database.

wgs_master[Properties] AND mammals[Organism]

Searching biological features in WGS genomes

The prokaryotic genomes and several WGS projects for higher eukaryotes are available as annotated genomes with reference sequences, gene records and, for the eukaryotes, sequence maps in the Map Viewer. The Rhesus macaque the *Tribolium castaneum*, and the *Populus trichocarpa* are three recent examples of organisms with fully annotated WGS-based genomes. Features of these annotated genomes may be searched effectively using the Entrez Gene, Protein and Nucleotide databases. The assemblies, Reference Sequence mRNAs and proteins, and other sequence collections for annotated genomes are available for BLAST searches through the genomic BLAST pages linked to the Map Viewer homepage.

www.ncbi.nlm.nih.gov/mapview

Many of the other WGS projects for eukaryotes are not currently available with annotated genes or proteins. These can be searched for these fea-

continued on page 8

Figure 1. The Genome Project summary page for the European rabbit. The page provides access to the WGS data through the "Project Data" menu and links to genome BLAST pages under "Resource Links".

New Genome Builds and Map Viewer Displays

New to Map Viewer

Genetic maps for twenty four new plants, the first assembled tree genome for black cottonwood (*Populus trichocarpa*), and two assembled protozoan parasite genomes are now available in the Map Viewer.

Plant Genetic Maps

The new plant genetic maps include important food and economic crops such as the common bean, *Brassica oleracea*—the source of cabbage, broccoli, cauliflower and Brussels sprouts—rye, rapeseed, cocoa, almond, and alfalfa. Also added are maps of close relatives and probable wild ancestors of some of these and other important crops. A complete list of the new plants with genetic maps can be found in Table 1. These genetic maps are important platforms for the future development of physical maps and complete genome sequences for several of these plants.

Genome Sequences

The Map Viewer now hosts the black cottonwood (*Populus trichocarpa*) genome assembly. This is the first woody plant genome assembly. The black cottonwood genome build 2.1 is the Joint Genome Institute's assembly of the 7.5 X whole genome shotgun sequence. Available sequence maps include NCBI contigs (the "Contig" map), the WGS sequences (the "Component" map), and the alignment of black cottonwood RNA sequences (the "PthRNA" map), UniGene clusters (the "PthUniGene" map), and Gnomon predicted gene models (the "ab-intio" map). The black cottonwood is not only an important experimental model species but is an economically important forestry resource. The poplar genome is well suited to serve as basis for comparative genomic studies of other broad leaf forest trees. Analysis of this genome has already provided important insights into the evolution of this group of plants into trees.

The genome assemblies and annotations for *Trypanosoma brucei*, the causative agent of African sleeping sickness, and *Theileria parva*, the causative agent of east coast fever--an African cattle disease, can now be viewed and searched in the map viewer.

The *Theileria* genome is based on a chromosome-by-chromosome whole genome assembly produced by The Institute for Genomic Research (TIGR) and the International Livestock Research Institute. The sequence is anchored to the four *Theileria* chromosomes. The current annotation contains 4,089 genes and their predicted transcripts. *Theileria* belongs to the Apicomplexa an important group of human and animal pathogens that includes the malaria parasites (*Plasmodium*). The *Theileria* genome now joins two other Apicomplexan genomes, *Plasmodium falciparum* and *Cryptosporidium parvum*, in the Map Viewer. Comparative analysis of these genomes will continue to provide insights into the parasite life cycles, host-parasite interactions and potential drug and vaccine targets.

The *Trypanosoma* genome is the assembly and annotation produced by an international consortium that includes the Sanger Center and TIGR. The sequence is of the eleven large chromosomes and does not include smaller mini and intermediate chromosomes. The mapped sequence and annotation represent 10,252 genes and their transcripts. The *T. brucei* genome is the first trypanosome genome in the Map Viewer. The sequencing and analysis of the genomes of *T. cruzi*, the causative agent of chagas disease and *Leishmania major*, responsible for leishmaniasis, together with the *T. brucei* genome should provide insights into the biology of these parasites that will help combat these diseases.

Other access to data

In addition to access through the Map Viewer, access to sequences of the genome assemblies, transcripts, proteins and gene models for black cottonwood, *Theleiria parva* and *Trypanosoma brucei* is provided through the RefSeq and Gene databases. GenBank and RefSeq records are available for searching in Entrez or via NCBI's Web BLAST services where they are extensively integrated with other resources and databases.

New Plant Species in Map Viewer

Aegilops tauschii
Aegilops umbellulata
Allium cepa (onion)
Brassica juncea (Indian mustard)
Brassica napus (oilseed rape)
Brassica Nigra (black mustard)
Brassica oleracea (cabbage and others)
Brassica rapa (field mustard)
Capsicum annuum (cayenne pepper)
Eragrostis tef (tef)
Medicago sativa (alfalfa)
Phaseolus vulgaris (French bean)
Populus trichocarpa (black cottonwood)
Prunus dulcis (almond)
Secale cereale (rye)
Setaria italica (foxtail millet)
Solanum lycopersicoides
Solanum melongena (eggplant)
Solanum peruvianum (Peruvian tomato)
Sorghum bicolor (broomcorn)
Theobroma cacao (cacao)
Triticum turgidum (English wheat)
Vigna radiata (mung bean)

Table 1. New Plants with genetic maps in Map Viewer.

Whole Genome Shotgun
continued from page 6

tures through sequence similarity using the NCBI BLAST services. In some cases genome-specific BLAST pages are linked to WGS projects (Figure 1). In all cases, WGS data are available as the "wgs" nucleotide database in the pull-down list on nucleotide-nucleotide BLAST forms linked to the BLAST homepage.

www.ncbi.nlm.nih.gov/BLAST

Figure 2 shows the result of a BLAST search with the platypus beta-2-microglobulin mRNA (AY125948) against the platypus wgs sequence. The search identifies two WGS records containing the four exons of the platypus microglobulin gene. This genomic sequence is not available in any other form at NCBI.

Environmental Sequences: Metagenomics

An interesting application of WGS techniques involves obtaining sequence information from entire biological communities rather than individual species. Acquiring and analyzing sequences obtained from biological communities without isolation of individual clones has been termed 'Metagenomics'. Whole Genome Shotgun metagenomic studies or metagenomes are important for assessing microbial diversity in all ecosystems because the majority of microbial species are unknown and probably unculturable. Communities from unusual or extreme environments seem particularly likely to be rich sources of unknown organisms that may have evolved interesting or useful adaptations. Metagenomes may provide clues to the genetic and biochemical adaptations of these organisms. Perhaps more importantly, in the same way that single organism genomes can provide insights into the specific metabolic pathways in the organism, metagenomes may provide important insights into community metabolism.

Metagenome projects have added sequences to GenBank from unusual environments and communities including acid mine drainage impacted streams, open water and deep sea ocean communities, communities associated with whale falls in the deep ocean and the chemoautotrophic symbiotic community associated with an annelid worm lacking a digestive or excretory system. There are also data from less exotic though still largely unknown communities such as those in the human gut and farm soil. In addition, Metagenomic techniques have been applied to obtain sequences of extinct organisms including Woolly Mammoth and Neanderthal man from well preserved remains.

Source	Master Accessions
Neanderthal Man	CAAN01000000
Wooly Mammoth	CAAM01000000
Gutless Worm Symbionts	AASZ00000000
Bioreactor Sludge	AATO, AATN00000000
Human Gut Biome	AAQL, AAQK00000000
Grey Whale Carcasses	AAGA, AAFZ, AAFY00000000
Farm Soil	AAFX01000000
Acid Mine Drainage Biofilm	AADL00000000, AAWO00000000
Global Ocean Sampling Expedition	AACY00000000
Mouse Gut Metagenomes	AATA, AATB, AATC, AATD, AATE, AATF00000000

Table 2. Environmental sequencing projects (metagenomes) with data in GenBank. Environmental sequencing projects (metagenomes) with data in GenBank.

Environmental Genome Projects: Metagenomes

The simplest access to the metagenomes is through the Environmental Projects link on the right hand side of the Entrez Genome Projects Homepage. More than 30 projects are available at the time of this writing. The Environmental Projects can also be retrieved with the following query from the Genome Projects homepage or from the search box on the NCBI homepage:

type_environmental[Properties]

These metagenome projects have varying amounts and types of data; some projects listed are in progress and have no data yet at NCBI, some have only Trace Archive sequences available and several have WGS data available. Table 2 shows the environmental projects with WGS sequence at NCBI. As described above, all Project Overview pages in the Genome Projects database provide access to the data and other linked resources. In addition, each of the Environmental Projects has links to two specialized BLAST services; one that can search the nucleotide sequences and, in some cases,

continued on next page

New Organisms in UniGene

The Entrez UniGene database now offers over 1,844,162 transcript clusters, linked to nucleotide records, for over 70 animals and plants. Recent additions to UniGene include: *Aedes aegypti* (yellow fever and dengue virus mosquito) with 241,102 transcript sequences in 15,182 clusters, *Aquilegia formosa x Aquilegia pubescens* (hybrid columbine) with 72,522 transcript sequences in 7,675 clusters, *Gossypium hirsutum* (upland cotton) with 83,321 transcript sequences in 10,845 clusters, *Macaca fascicularis* (crab-eating monkey) with 62,745 transcript sequences in 7,488 clusters, *Oryctolagus cuniculus* (rabbit) with 10,827 transcript sequences in 3,766 clusters, *Pimephales promelas* (fathead minnow) with 237,026 transcript sequences in 18,541 clusters, and *Tribolium castaneum* (red flour beetle) with 27,233 transcript sequences in 6,328 clusters.

Whole Genome Shotgun
continued from page 8

proteins for the specific metagenome, and one that allows searches against all or subsets of the metagenomes at once.

Example: The Gutless Worm Metagenome

The sediment dwelling marine annelid, Olavius, entirely lacks a digestive system and has a highly reduced excretory system. This worm depends on a consortium of at least four bacterial symbionts to provide its nutritional and excretory needs. The metagenome for this consortium has provided important insights into relationships among the organisms in this symbiotic community. The consortium contains two sulfur-oxidizing gamma proteobacteria (δ1 and δ3) and two sulfate- reducing delta proteobacteria (δ2 and δ4). The complementary metabolisms of the two sets of bacteria provide each other with appropriate electron donors and acceptors and the worm with organic carbon and other nutrients while processing the worm's nitrogenous waste.[1] The genome-specific BLAST page allows searches against the metagenome of the chemoautotrophic bacterial symbionts. Translating BLAST searches quickly confirm the presence of the two sets of bacterial symbionts. Figure 3 shows the results of a translating BLAST search with a sulfite reductase (**YP_387022**) from Desulfovibrio, a delta proteobacterium. The sulfite reductase finds matches to all four of the symbionts. The best match, **AASZ01000485**, is a section of the partial assembly of the delta 1 symbiont (**DS021230**).

Continued growth

The rapid influx of WGS data in the form of organism and community genomes will continue for the foreseeable future. This new kind of data provides challenges for analyses and completely new perspectives as the scope increases beyond individual organisms to community genomes, and pathways—even of extinct communities. This growing and vast amount of largely unannotated sequence will continue to be most effectively searched through the NCBI BLAST services. The Entrez Genomes Project database will continue to provide the most convenient mechanism to access the WGS and other genomes at NCBI.

1 Woyke T, et al. 2006 . Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, 443(7114):950-5. PMID: 16980956

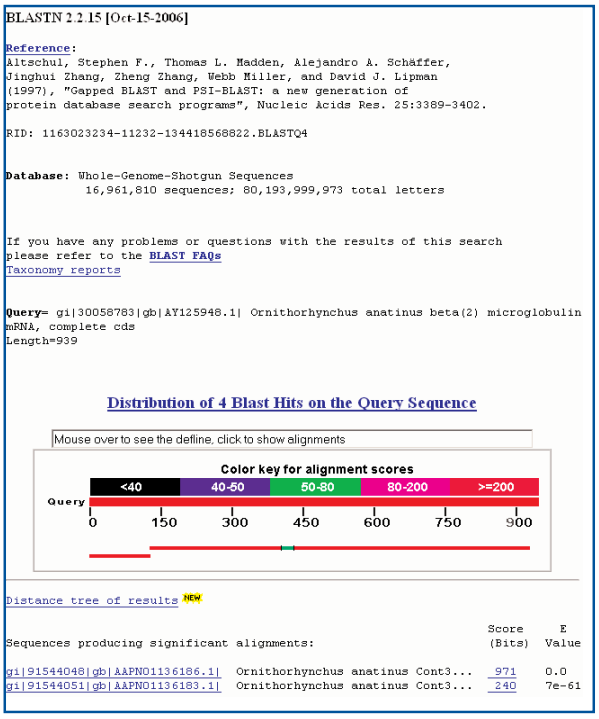

Figure 2. A nucleotide-nucleotide BLAST search against the wgs database using the platypus beta-2-microglobulin mRNA as a query. The database was limited with the Entrez query platypus[Oganism]. The four hits to the two WGS sequences identify the four exons of the platypus microglobulin gene. Use the following RID to retrieve live results: 1163023234-11232-134418568822.BLASTQ4.

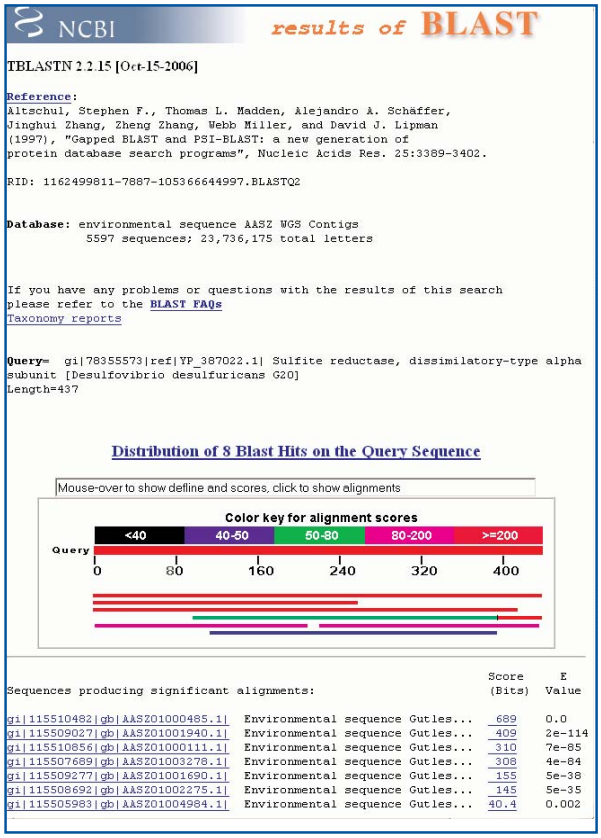

Figure 3. Results of a translating BLAST search (tblastn) against the Olavius metagenome using a Desulfovibrio sulfite reductase protein sequence (YP_387022) as a query. The results contain hits from all four of the symbionts, the two delta and two gamma proteobacteria. The most similar sequence is from the delta 1 symbiont . Use the following RID to retrieve live results: 1162499811-7887-105366644997.BLASTQ2.

RefSeq Release 22

RefSeq Release 22 is now available by anonymous FTP at:

ftp.ncbi.nih.gov/refseq/release

Release 22 includes genomic, transcript, and protein sequences available as of March 5, 2007, from 4,187 organisms. The number of RefSeq accessions in Release 22 and their combined lengths is given in the shaded box.

RefSeq releases are posted every two months, and the next release is scheduled for May, 2007. 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	840,657	80,808,752,887
RNA	929,109	1,632,375,659
Protein	3,438,099	1,215,085,694

GenBank Release 158

GenBank Release 158 (February 2007) contains over 67 million sequence entries totaling more than 71 billion base pairs. Release 159 is scheduled for April 2007. 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 158 flatfiles are 252 Gigabytes and the ASN.1 version is about 217 Gigabytes. The data can also be downloaded at a mirror site:

bio-mirror.net/biomirror/genbank

NCBI Courses

NCBI Courses are a great way for researchers, students, librarians, and teachers to keep up with the ongoing enhancements made to NCBI's molecular biology resources.

The courses are offered free of charge at NCBI and at universities and research institutes throughout the United States. For detailed course descriptions and schedules of upcoming courses, see the NCBI Education Homepage at:

www.ncbi.nlm.nih.gov/Education

Enhanced Field Guide Course

March 29—30, 2007 9 AM—5 PM

NCBI, Bethesda, MD

This course is intended for biologists who desire a more in-depth look at the NCBI resources than can be provided in the standard Field Guide.

To register, and for more information:

www.ncbi.nlm.nih.gov/Class/FieldGuide/FGPlus

A Field Guide to GenBank and NCBI Molecular Biology Resources

April 25, 2007—at the Rochester Institute of Technology, Rochester, NY

May 3, 2007—at the University of California at San Francisco, San Francisco, CA

May 10, 2007—9 AM—5PM at the National Library of Medicine, Bethesda, MD

The Field Guide is a lecture and hands-on computer workshop on GenBank and related databases covering effective use of the Entrez databases and search service, the BLAST similarity search engine, genome data and related resources.

To register, and for more information:

www.ncbi.nlm.nih.gov/Class/FieldGuide/

NCBI Mini Course

May 2—3, 2007 at the Wadsworth Center, Albany, NY

May 10—11, 2007 at MIT, Cambridge, MA

May 15—16, 2007 at Virginia Polytechnic Institute, Blacksburg, VA

NCBI bioinformatics mini-courses are either problem based, such as "Identification of Disease Genes" or NCBI resource based such as "BLAST Quick Start". The courses are 2.5 hours in length with the first hour-and-a- half devoted to an overview that is followed by a one hour hands-on session.

To register, and for more information:

www.ncbi.nlm.nih.gov/Class/minicourses/minischedule.html

NCBI Powerscripting Course

April 24—27, 2007 at NCBI, Bethesda, MD

This 4-day course including both lectures and computer workshops on effectively using the NCBI Entrez Programming Utilities (E-utilities) within scripts to automate search and retrieval operations across the entire suite of Entrez databases.

To register and for more information:

www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/course.html

Submission Corner

Submitting Third Party Annotation Records Using Standard Submission Tools

The Third Party Annotation (TPA) database allows researchers to submit records that provide additional biological annotation but are based on existing records in GenBank. As pointed out in the recent NCBI News (Vol. 15, issue 1) these TPA records can now include a wider range of submission types including inferential findings. This article explains how to submit these TPA records to NCBI through the two standard GenBank direct submission tools, BankIt and Sequin. Before beginning either submission route, one or more sequences from one of the collaborating nucleotide sequence databases, DDBJ, EMBL, or GenBank must be identified as the source data for the TPA record. TPA records cannot be based on Reference Sequences or other non-primary sequence data. All TPA annotations must have some direct or indirect experimental support. The acceptable kinds of experimental evidence are described on the TPA submissions page.

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

Submitting TPA Records Using BankIt

The popular Web submission tool BankIt provides a simple interface for direct submission to GenBank and also allows submission to the TPA database.

www.ncbi.nlm.nih.gov/BankIt

As with GenBank submissions, the BankIt TPA submission route is best for simple submissions of small numbers of records. The Sequin submission route, described below, is a better choice for large or complicated TPA submissions. To submit a TPA record using BankIt, begin as usual by entering the length of the sequence and pressing the "New" button under "GenBank Submissions by WWW" section of the BankIt page. This launches the Bankit Web submissions form that is normally used to prepare GenBank submissions. To specify a TPA submission, set the radio button to "No" to answer the questions on the form: "Is This Primary Sequence Data? Have you determined and annotated the sequence you are submitting?". Follow the instructions on the form and enter the DDBJ/EMBL/GenBank accession numbers of source data and a concise description of the supporting experimental evidence for the biological annotation in the specified text areas. The rest of the submission process is the same as that for standard GenBank data. Once the submission process is complete, the preliminary flatfile view reports, in a temporary COMMENT field, that this is a record from the third party annotation database and lists the primary database accession numbers of the source data. The COMMENT will be removed as the record is processed by the submission staff at NCBI. The BankIt acknowledgement message returned by e-mail, contains the interim BankIt number, designates the submission as a TPA, and includes the accessions and concise evidence information.

Submitting TPA Records Using Sequin

The stand-alone submission and annotation tool, Sequin, also allows submission of TPA records. Sequin provides a powerful platform for propagating features across large and complicated sets of data and saves time and effort for batches or large records. The Sequin Homepage provides more information on downloading and using the Sequin program.

www.ncbi.nlm.nih.gov/Sequin

Begin a GenBank or TPA submission with Sequin by clicking the "Start New Submission" button on the main window of the program. Enter contact and author information and navigate through the dialog windows using the "Next" and "Previous" buttons as usual. Select the radio button for "Third Party Annotation" on the "Sequence Format" dialog window. Enter a description of the biological experiments used as evidence for the TPA in the "TPA Evidence" dialog window prompts. After importing the sequence(s) and adding remaining features, enter the DDBJ/EMBL/GenBank accession number(s) used in the as data sources in the columns on the "Assembly Tracking" dialog. Like the BankIt submission, the preliminary flatfile view in Sequin will show the temporary COMMENT field that indicates that the submission is to the TPA database and lists the primary accession number. Send the completed Sequin TPA submission in an e-mail to

gb-sub@ncbi.nlm.nih.gov

to complete the submission process.

TPA records that have been submitted to NCBI using BankIt or Sequin are released into the NCBI public databases once the TPA data is published in a peer-reviewed journal. At NCBI, TPA records are a part of the sequence databases available for searching through the Entrez and the BLAST services.

—MR

PubChem Grows to 15 Million Substances

With recent deposits from Specs and Thomson Pharma, the PubChem Substance database has now passed the 15 million mark, including deposits from 53 organizations, including commercial, government, and academic institutions. PubChem Substance currently contains 15,450,812 records representing 10,138,100 unique chemicals in PubChem Compound. The PubChem BioAssay database has expanded to 362 assays from 21 depositors, with more than 10 of these depositors new in 2006.

Detailed listings of the sources of PubChem data are available at

pubchem.ncbi.nlm.nih.gov/sources

Top Ten Sources for:

PubChem Substance

Source	No. of Records	Source	No. of Records
ZINC	3,813,892	ChemBank	413,586
ChemDB	3,564,938	ChemIDplus	383,789
DiscoveryGate	2,739,765	Asinex	362,469
Thomson Pharma	2,230,474	National Cancer Institute	268,696
ChemBridge	433,971	Specs	200,744

PubChem BioAssay

Source	No. of BioAssays
National Cancer Institute	173
Structural Genomics Consortium—Oxford	43
NIH Chemical Genomics Center	30
BindingDB	19
Scripps Research Institute	15
Diabetic Complications Screening	14
San Diego Center for Chemical Genomics	11
New Mexico Molecular Libraries Screening Center	8
Penn Center for Molecular Discovery	7
Columbia University Molecular Screening Center	6

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