NCBI News

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

Fall 2003
Winter 2004

Entrez Query Goes “Global”

The Entrez search and retrieval system now offers a cross-database search that allows a single query to span the traditional NCBI-sequence databases; Nucleotide and Protein; the literature databases, such as PubMed®, PMC, Books, OMIM™, Journals, and MeSH; the structurally-oriented databases, Structures, the Conserved Domain Database, 3D-Domains; the NCBI Taxonomy, Gene Expression Omnibus (GEO), Single Nucleotide Polymorphisms (SNPs), Population Sets, Genomes, Sequence Tagged Sites, UniGene, Gene-centered information (Gene), and, finally, the NCBI Web site itself. The cross database search option, labeled “Entrez” on the NCBI homepage search menu, replaces 'GenBank' as the default.

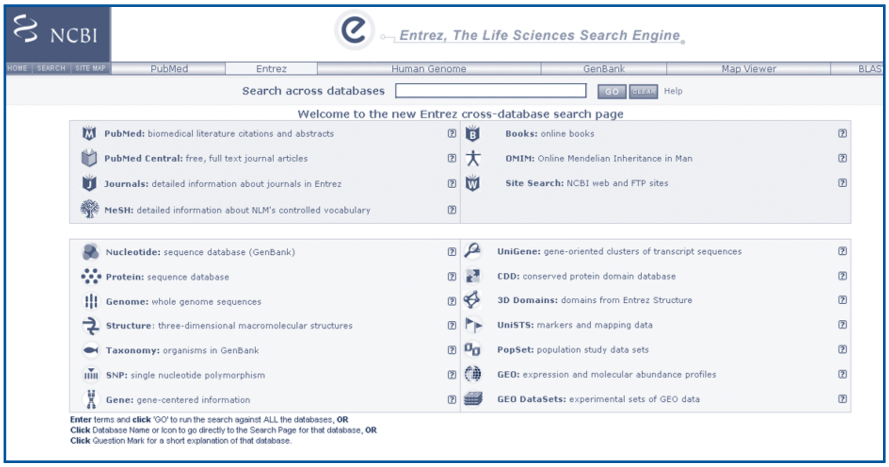
The New Entrez Home Page

The Entrez toolbar link on the NCBI Home Page leads to a new Entrez Home Page which provides a cross-database search box and a listing of the Entrez databases that can be searched in tandem. Question mark icons to the right of each database name lead to descriptions of database content. The database names and icons link to homepages where single-database queries can be constructed using lists of database-specific field restrictions, or tables that can be used to define search limits. In addition, the Entrez homepage toolbar provides links to popular NCBI tools and resources such as the Map Viewer and BLAST®.

Cross-Database Search Results

When a cross-database search is completed, the number of matches for each database is displayed in boxes adjacent to the database name as shown in Figure 2 on the next

continued on page 4



Welcome to the new Entrez cross-database search page

PubMed: biomedical literature citations and abstracts
PubMed Central: free, full text journal articles
Journals: detailed information about journals in Entrez
MeSH: detailed information about NLM's controlled vocabulary
Books: online books
OMIM: Online Mendelian Inheritance in Man
Site Search: NCBI web and FTP sites

Nucleotide: sequence database (GenBank)
Protein: sequence database
Genome: whole genome sequences
Structure: three-dimensional macromolecular structures
Taxonomy: organisms in GenBank
SNP: single nucleotide polymorphism
Gene: gene-centered information
UniGene: gene-oriented clusters of transcript sequences
CDD: conserved protein domain database
3D Domains: domains from Entrez Structure
UniSTS: markers and mapping data
PopSet: population study data sets
GEO: expression and molecular abundance profiles
GEO DataSets: experimental sets of GEO data

Enter terms and **click** 'GO' to run the search against ALL the databases, **OR**
Click Database Name or Icon to go directly to the Search Page for that database, **OR**
Click Question Mark for a short explanation of that database.

Figure 1. Entrez homepage showing the new cross-database search engine with links to the 21 Entrez databases covered.

Register Your Genome Project Online at NCBI

NCBI builds a reference sequence, or RefSeq, for each complete genomic sequence in GenBank® as well as for contigs of incomplete genomes. These RefSeqs are used to integrate the genomic sequence with genomic and other data at NCBI to provide pre-computed analyses that allow researchers to uncover relationships between sequences from different organisms, and between a sequence and a biological function. In addition to the reference human genome, and other major eukaryotic genomes, Entrez Genomes includes sequence for complete viral, microbial, and organellar genomes representing a diverse group of model organisms, pathogens, and organisms of environmental importance. Since the utility of a genomic sequence to the

continued on page 2

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
Medha Bhagwat
Monica Romiti
Eric Sayers

Writers
Vyvy Pham
David Wheeler

Editing and Production
Robert Yates

Graphic 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.

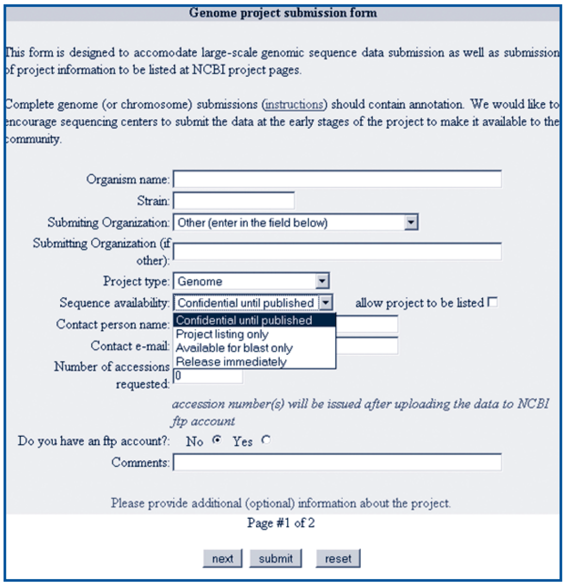
The contents of this newsletter may be reprinted without permission. The mention of trade names, commercial products, or organizations does not imply endorsement by NCBI, NIH, or the U.S. Government.

NIH Publication No. 04-3272

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

Register Your Genome Project

continued from page 1



Genome project submission form

This form is designed to accomodate large-scale genomic sequence data submission as well as submission of project information to be listed at NCBI project pages.

Complete genome (or chromosome) submissions (instructions) should contain annotation. We would like to encourage sequencing centers to submit the data at the early stages of the project to make it available to the community.

Organism name:
Strain:
Submiting Organization: Other (enter in the field below)
Submitting Organization (if other):
Project type: Genome
Sequence availability: Confidential until published   allow project to be listed □
Contact person name:
Contact e-mail:
Number of accessions requested: 0

accession number(s) will be issued after uploading the data to NCBI ftp account

Do you have an ftp account?: No ● Yes ○
Comments:

Please provide additional (optional) information about the project.

Page #1 of 2

Figure 1. Views of the first Web page of the Genome Project submission form. The pulldown shows the four "sequence availability" options.

research community increases dramatically when it is placed in the context of other genomic sequences, NCBI strongly encourages the submission of genomic data ranging from mapping information to complete chromosomal sequences with annotation.

NCBI is working in close collaboration with many sequencing centers to provide new and updated information about ongoing genome sequencing programs for organisms ranging from microbes to multicellular plants and animals. Sequencing centers can register a sequencing project with NCBI prior to the submission of any data registered. Sequencing projects may include those producing finished and unfinished genomic sequence, or Whole Genome Shotgun (WGS) sequence. For each registered project, a sequencing project page on the NCBI Genomes website is created that gives a description of the project, links out to genome-specific resources, and provides a focal point for the addition of links to NCBI resources, such as the Map Viewer and genomic BLAST. Sequencing centers are encouraged to begin submitting mapping and sequence data at the early stages of the project in order to make it available to the research community quickly.

To make the registration of a sequencing project as simple as possible, NCBI has created an online form, shown in Figure 1, for the submission of project information. The form allows for the entry of general information such as the name of the organism to be sequenced, the name of the sequencing center, and the sequencing strategy to be used. In addition, the conditions under which the project data, including sequences, are to be submitted to NCBI can be specified. A project can be listed or unlisted, and may be listed without sequence data. Sequences may be held until publication, the default, released immediately, or made available for BLAST searches only. The form can also be used to set up an FTP site for the upload of data to NCBI, or to specify a URL to be used by NCBI for the download of project or sequence data. The Genome Project submission form along with links to detailed instructions for the submission of genomic sequence and annotations, can be found under "Submitting" in the sidebar of the Entrez Genomes page, or at:

www.ncbi.nih.gov/genomes/mpfsubmission.cgi

Questions about genome project registration or genome submission can be addressed to the NCBI Genomes group at:

genomes@ncbi.nlm.nih.gov

—VP

New Genome Builds and Annotations at NCBI

One Build, Multiple Rounds of Annotation

As reference genome assemblies, such as those for human, mouse, and rat stabilize, a single build is expected to pass through multiple cycles of annotation. For this reason, a "build X version Y" identification system is now being used for many genomes shown in the Map Viewer. An identifier such as, "build 34 version 1", indicates the first version of annotation for genome build 34.

Gene Annotation

NCBI has switched from GenomeScan as its standard method of predicting gene models, to Gnomon, a program developed by NCBI scientists. Gnomon differs from GenomeScan by putting a greater emphasis on coding propensity and matches to existing proteins when predicting genes. Gnomon also checks more rigorously for shifts in reading frame within transcript models that are often indicative of pseudogenes. As a result, the number of genes appearing in the most recent NCBI annotations of the genomes of human, mouse, and rat has decreased significantly, while the number of models identified as pseudogenes has increased. About 20% of the gene models appearing in human genome build 34 version 1 were produced using Gnomon; the remaining 80% were derived from NCBI RefSeq transcript alignments. For more on the Gnomon algorithm, see the shaded box entitled "Gnomon".

Human Genome Build 34 Version 3

The NCBI Map Viewer now shows build 34 version 3 of the human genome reference sequence, which is based on data available as of July 2003, and includes the pseudoautosomal region of the Y chromosome. Supplementing the reference sequence are a separate assembly of chromosome 7 submitted by the The Center for Applied Genomics (TCAG), and the reference sequence for the DR51 haplotype in the Major Histocompatibility Complex region. Transcripts annotated on the TCAG assembly by the TCAG group are shown on a separate track in the Map Viewer. Other new tracks for the human Map Viewer show the alignment to the genome of all human, mouse, rat, pig, and cow ESTs along with mRNAs. A new *ab initio* track replaces the GenomeScan for the display of Gnomon gene predictions.

Rat Genome Build 2 Version 1

NCBI build 2 version 1 of the rat genome reference sequence uses the Rat Genome Sequencing Consortium version 3.1 assembly, which covers 2.8 billion bases of the genome in Whole Genome Shotgun (WGS) contigs. Shown alongside this assembly in the Map Viewer is the NCBI "NT" assembly, which covers 25 million bases of sequence and uses contigs assembled by NCBI from finished BAC sequences in GenBank. New in the Map Viewer for this build, are tracks showing the alignment of all human, mouse, and rat ESTs to the rat genome, as well as the *ab initio* track, which replaces the GenomeScan map for the display of gene models.

Mouse Build 32 Version 1, *Anopheles gambiae* Build 2 Version 1

Mouse build 32 version 1, based on data available as of September 2003, includes 24,819 mapped genes. Shown with the build 32 version 1 reference assembly in the Map Viewer is the Celera assembly of chromosome 16 with NCBI annotations. Build 2 Version 1 of the *Anopheles gambiae* genome, based on data available as of July 2003, is also available for browsing in the Map Viewer with over 12,000 annotated genes.

—DW

Magnaporthe grisea, Bos taurus, Sus scrofa, Canis familiaris are new in Map Viewer

NCBI has recently created Map Viewer displays for four more organisms. The display for *Magnaporthe grisea*, a pathogen of rice that is a close relative of *Neurospora crassa*, includes contig, gene, and transcript tracks. For cow[1] and pig[2], the Map Viewer displays Meat Animal Research Center (MARC) linkage maps while for the dog the Map Viewer shows the Canine 1Mb Radiation Hybrid map (RHDF5000).[3] New Genome Guide pages, created by NCBI in cooperation with the genomic research communities to provide links to an array of genome-specific resources, are available for cow, pig, and dog. These pages can be seen at:

www.ncbi.nlm.nih.gov/genome/guide/pig
www.ncbi.nlm.nih.gov/genome/guide/cow
www.ncbi.nlm.nih.gov/genome/guide/dog

[1] Keele, J. A second-generation linkage map of the bovine genome. *Genome Res*. 1997 7(3): 235-49. PMID: 9074927

[2] Rohrer GA, *et al*. A comprehensive map of the porcine genome. *Genome Res*. 1996 6(5): 371-91. PMID: 8743988

[3] Guyon R, *et al*. A 1-Mb resolution radiation hybrid map of the canine genome. *PNAS USA*. 2003: 100(9): 5296-301. PMID: 12700351

Gnomon

To create a gene model, Gnomon finds the best self-consistent set of transcript and protein alignments to a genomic region and uses these alignments as constraints for a Hidden Markov Model (HMM)-based gene prediction. Several steps are involved.

Gnomon evaluates the statistical properties of the transcripts aligned to a genome in order to determine their most probable coding regions. For each gene model, the set of non-overlapping transcript alignments with the best coding propensity is chosen, after which the best matching proteins for the transcript sequences are aligned to the genomic DNA sequence. For HMM-based gene models without supporting transcript evidence, the proteins that are the best matches to the translated genomic sequence are aligned. Gnomon checks that the resulting predicted gene has every exon in a reading frame consistent with the protein alignment, however, the program is free to choose splice sites and to introduce additional exons between segments of the protein alignment.

Entrez Gene Database Debuts

Entrez Gene, the most recent addition to the growing Entrez database collection, provides detailed reports on genes, both sequenced and unsequenced, with links to related resources.

Entrez Gene supports the queries and provides much the same information as the familiar LocusLink resource it will eventually replace. However, the Gene database has a larger scope and currently covers over 350 microbial species, more than a thousand viral genomes, and more than 400 eukaryotes. Being integrated into Entrez also means that the connections between Gene and other databases in Entrez such as PubMed, OMIM, Nucleotide, Protein, GEO, and UniGene are more readily apparent in the Links menus, rather than a step away via LinkOut. Entrez Gene can be queried using the new Cross Database search described in the previous article beginning on Page 1. The results of such a query are shown in Figure 2 on the facing page with the 968 hits to the Gene database indicated in the lower left-hand corner.

The Entrez Gene database will be featured in the Spring NCBI News. If you are interested in being notified of changes in the Gene interface, consider subscribing to the gene-announce mailing list:

www.ncbi.nlm.nih.gov/mailman/listinfo/gene-announce

—DW

Entrez Goes "Global" continued from page 1

page. The Figure shows the results of a cross-database search using the Entrez query "rat[organism] AND kinase" indicating hits to a wide variety of databases, including those for literature, sequence, gene expression, and structure. The four representative Entrez displays indicated by the letters A, B, C, and D are reached by first clicking on the the database name, icon, or number of matches to reach the Entrez document summary of the hits, and then selecting a record and a format for viewing.

Beginning in the upper left hand corner, the figure shows one of the 54,988 hits to the PubMed database, using the "Abstract" display from the "display" pulldown menu. A link to the full text of the article in PubMed Central appears just to the right of the pulldown menu. Moving in a counter-clockwise direction, one of the 2,613 hits to the Nucleotide database is shown using the FASTA display format with other possible formats listed in the "display" pulldown menu. One of the 968 records from the Entrez Gene database that gives a detailed report on a gene. The report contains a gene description, a graphic of the gene's structure, along with links to bibliographical references, GenBank sequences, RefSeqs, computed and curated functional classifications, and sequence viewers such as the Map Viewer. The last example shows a single hit to the Conserved Domain Database (CDD) giving the sequence alignments used to define the domain indicated. A link to a KOG (see "KOGs and COGs are Now Included in CDD" in this issue) appears in the upper right-hand corner.

Advanced Entrez Features

A cross-database search produces an individual History page, giving a list of previously executed searches with links to the results, for each of the Entrez databases. Other familiar Entrez features such as the Clipboard, Details, Limits, and Preview/Index are also retained separately for each database since the databases support specialized arrays of fields used to limit searches. For example, note that in Figure 2, some of the boxes containing the number of hits to the databases are shaded. The shading indicates that the cross-database query contained a field delimiter not supported by the database in question. The query of Figure 2, "rat[organism] AND kinase", uses the "organism" field restriction; the "organism" field is supported within the Nucleotide, CDD and Gene databases, but not within the PubMed literature database. In this case, the search was performed using the term "rat" in "all fields". This can be verified by clicking on PubMed in the cross-database results page, to reach the PubMed database display of the hits, and then clicking on the "Details" link in the PubMed display to see how the cross-database search was executed within the PubMed database.

The Entrez search and retrieval system is robust and flexible. To learn more about searching and navigating between Entrez databases or displaying and downloading results, see:

Geer RC, **Sayers EW**. Entrez: making use of its power. *Briefings in Bioinformatics.* 2003 Jun;4(2):179-84. PMID: 12846398

www.ncbi.nlm.nih.gov/Entrez

—MR, DW

Query Across All Entrez Databases by Script

Entrez cross-database searches can be performed by script using a new E-Utility program called EGQuery. E-Utilities are programs residing on NCBI servers that can be accessed by posting a standard URL containing the script name and a set of parameters. For example, posting the following URL from within a script will return the number of hits to each Entrez database for the query "stem cells":

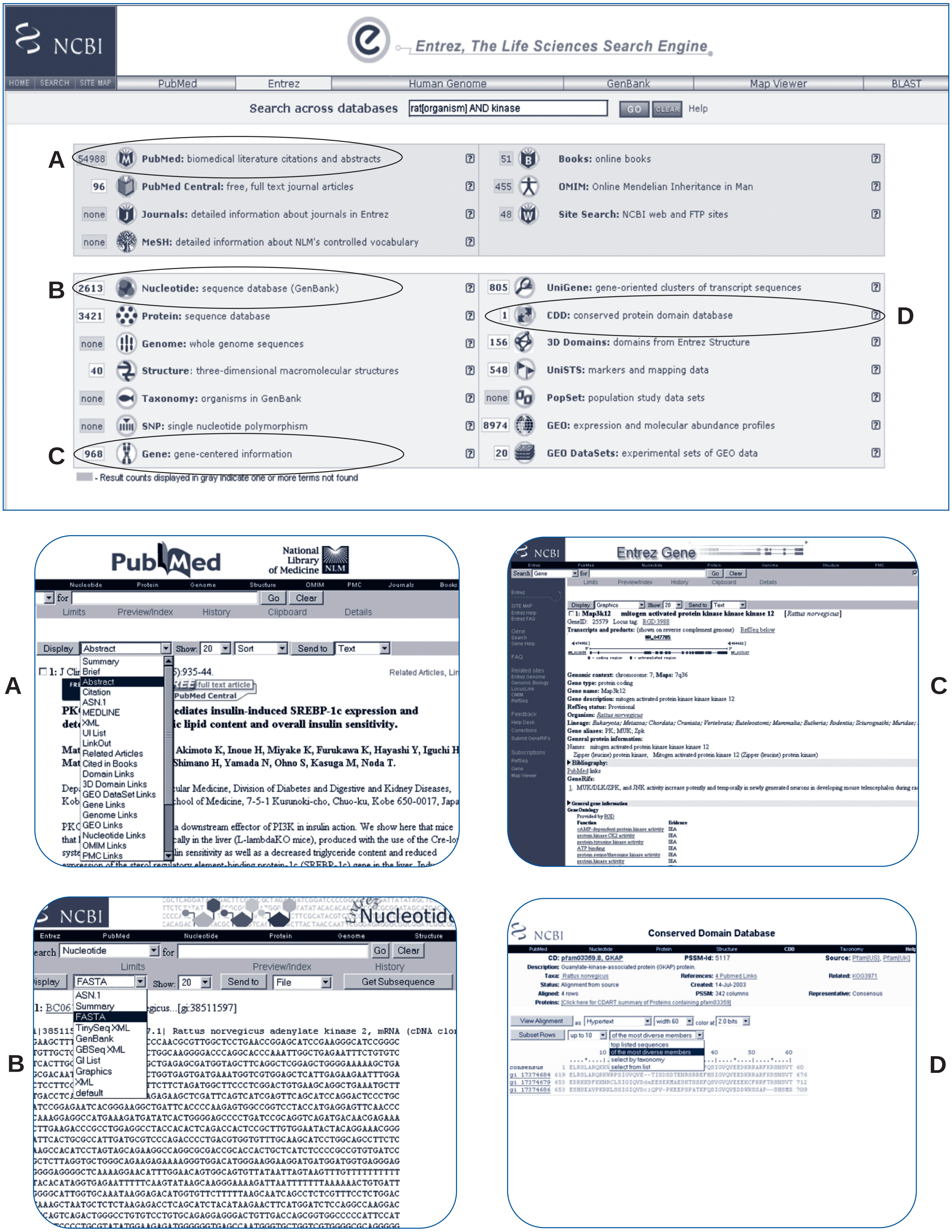
eutils.ncbi.nlm.nih.gov/entrez/eutils/egquery.fcgi?term=stem+cells

Results are returned in XML format. For details on the use of EGQuery or other E-Utilities, see:

eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

To receive announcements of new E-Utility features, subscribe to the E-"utilities-announce" listserve at:

www.ncbi.nlm.nih.gov/mailman/listinfo/utilities-announce



NCBI

Entrez, The Life Sciences Search Engine

HOME | SEARCH | SITE MAP | PubMed | Entrez | Human Genome | GenBank | Map Viewer | BLAST

Search across databases: rat[organism] AND kinase GO CLEAR Help

Hits	Database	Hits	Database
54988	**PubMed:** biomedical literature citations and abstracts	51	**Books:** online books
96	**PubMed Central:** free, full text journal articles	455	**OMIM:** Online Mendelian Inheritance in Man
none	**Journals:** detailed information about journals in Entrez	48	**Site Search:** NCBI web and FTP sites
none	**MeSH:** detailed information about NLM's controlled vocabulary		
2613	**Nucleotide:** sequence database (GenBank)	805	**UniGene:** gene-oriented clusters of transcript sequences
3421	**Protein:** sequence database	1	**CDD:** conserved protein domain database
none	**Genome:** whole genome sequences	156	**3D Domains:** domains from Entrez Structure
40	**Structure:** three-dimensional macromolecular structures	548	**UniSTS:** markers and mapping data
none	**Taxonomy:** organisms in GenBank	none	**PopSet:** population study data sets
none	**SNP:** single nucleotide polymorphism	8974	**GEO:** expression and molecular abundance profiles
968	**Gene:** gene-centered information	20	**GEO DataSets:** experimental sets of GEO data

▨ - Result counts displayed in gray indicate one or more terms not found

PK… mediates insulin-induced SREBP-1c expression and hepatic lipid content and overall insulin sensitivity.

1: Map3k12 mitogen activated protein kinase kinase kinase 12 [Rattus norvegicus]

CD: pfam03359.8, GKAP — Description: Guanylate-kinase-associated protein (GKAP) protein.

Figure 2. Cross-database search results for the Entrez query “rat[organism] AND kinase”. The number of hits returned within each Entrez database is given in boxes to the left of the database name. Shaded boxes indicate that the database in question did not support the field restriction “organism” used to limit the term “rat” in the query. Portions of Entrez displays for a representative record returned from six of the Entrez databases are also shown.

Selected Recent Publications by NCBI Staff

To view the citation for any article listed below, click on the PubMed link on the navigation bar at the top of the NCBI Home Page, enter the PubMed ID number in the search query box, and click Go.

Schriml LM, Hill DP, Blake JA, Bono H, Wynshaw-Boris A, Pavan WJ, Ring BZ, Beisel K, Setou M, Okazaki Y. Human disease genes and their cloned mouse orthologs: exploration of the FANTOM2 cDNA sequence data set. *Genome Research*. 2003 Jun;13(6B):1496-500. PMID: 12819148

Szak ST, Pickeral OK, **Landsman D**, Boeke JD. Identifying related L1 retrotransposons by analyzing 3' transduced sequences. *Genome Biology*. 2003;4(5):R30. Epub 2003 Apr 16. PMID: 12734010

Daraselia N, **Dernovoy D**, Tian Y, Borodovsky M, **Tatusov R**, **Tatusova T**. Reannotation of *Shewanella oneidensis* genome. *Omics : a journal of integrative biology*. 2003 Summer;7(2):171-5. PMID: 14506846

Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V, Duprat S, **Galperin MY**, **Koonin EV**, Le Gall F, **Makarova KS**, Ostrowski M, Oztas S, Robert C, **Rogozin IB**, Scanlan DJ, Tandeau de Marsac N, Weissenbach J, Wincker P, **Wolf YI**, Hess WR. Genome sequence of the *cyanobacterium Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proceedings of the National Academy of Sciences, USA*. 2003 Aug 19;100(17):10020-5. Epub 2003 Aug 13. PMID: 12917486

Mu J, Ferdig MT, Feng X, Joy DA, Duan J, Furuya T, Subramanian G, **Aravind L**, Cooper RA, **Wootton JC**, Xiong M, Su XZ. Multiple transporters associated with malaria parasite responses to chloroquine and quinine. *Molecular Microbiology*. 2003 Aug;49(4):977-89. PMID: 12890022

Shabalina SA, **Ogurtsov AY**, **Lipman DJ**, **Kondrashov AS**. Patterns in interspecies similarity correlate with nucleotide composition in mammalian 3'UTRs. *Nucleic Acids Research*. 2003 Sep 15;31(18):5433-9. PMID: 12954780

Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, **Rozanov M**, Spaan WJ, Gorbalenya AE. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *Journal of Molecular Biology*. 2003 Aug 29;331(5):991-1004. PMID: 12927536

New Microbial Genomes in GenBank

Organism	GenBank | RefSeq Accession Numbers
Photorhabdus luminescens subsp. laumondii TTO1	BX470251 | NC_005126
Gloeobacter violaceus	BA000045 | NC_005125
Chromobacterium violaceum ATCC 12472	AE016825 | NC_005085
Porphyromonas gingivalis W83	AE015924 | NC_002950
Wolinella succinogenes	BX571656 | NC_005090
Geobacter sulfurreducens PCA	AE017180 | NC_002939
Rhodopseudomonas palustris str CGA009	BX571963 | NC_005296
Onion yellows phytoplasma	AP006628 | NC_005303

For more detailed information, see the online version of the Fall 2003/Winter 2004 NCBI News, or use the GenBank or RefSeq Accession Number to query Entrez “Genome” database using the query box on the NCBI Home Page.

KOGs and COGs Are Now Included In CDD

As part of the latest release (v1.63) of the Conserved Domain Database (CDD), the alignment sets of the KOG[1] database (clusters of euKaryotic Orthologous Groups) have been merged into CDD . The KOG database is essentially a eukaryotic version of the COG database (Clusters of Orthologous Groups) that was integrated into CDD in late 2002 (v1.60). KOGs and COGs cluster eukaryotic and prokaryotic proteins respectively into groups containing sequences that are mutual best hits in sequence similarity searches between different species. The KOG database includes proteins from *H. sapiens*, *D. melanogaster*, *C. elegans*, *A. thaliana*, *S. cerevisiae*, *S. pombe*, and *E. cuniculi*. With RPS-BLAST searches available for KOGs and COGs in CDD, users can now classify query sequences by similarity to these pre-determined sets alongside the alignments from Pfam, SMART, and the curated NCBI Conserved Domains. Because CDD data is also incorporated into Entrez as the Domains database, KOGs and COGs can be found using standard Entrez queries by fields such as title, organism, or text words.

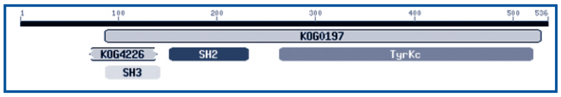


Figure 1. Graphical overview of Conserved Domain Search results for human SRC protein, RefSeq accession NP_005408, showing hits to KOG0197 and a PFAM-based conserved domain for tyrosine kinases, as well as hits to SH2 and SH3 domains.

With KOGs and COGs now included in CDD, the displays of pre-computed RPS-BLAST results have been updated to reflect the different clustering schemes underlying the several datasets within CDD. CDD now contains datasets that cluster proteins based on overall sequence similarity (COGs and KOGs) along with those that cluster based on the presence of defined functional domains (Pfam, SMART, curated CDs). Multiple domain proteins will therefore often have two sets of hits in CDD: hits from COGs and KOGs to large portions of the sequence, and hits to Pfam, SMART, and/or CDD for each functional domain. In order to show both sets of hits in a simple display, each CDD record is now classified as either a "single" or "multiple" domain record, and the best hits from each set are shown when the Domains link is clicked for a record in Entrez Protein. Moreover, the Conserved Domain Architecture Retrieval Tool (CDART) only uses single domain records to group protein sequences by domain architecture. In the example shown above for NP_005408, the human SRC protein, hits are shown to both the multiple domain KOG0197 (tyrosine kinases) and to single domains pfam00018 (SH3), pfam00017 (SH2), and cd00192 (TyrKc, tyrosine kinase catalytic domain). *—ES*

[1] **Tatusov RL**, **Fedorova ND**, **Jackson JJ**, **Jacobs AR**, **Kiryutin B**, **Koonin EV**, **Krylov DM**, Mazumder R, **Mekhedov SL**, Nikolskaya AN, **Rao BS**, **Smirnov S**, **Sverdlov AV**, **Vasudevan S**, **Wolf YI**, **Yin JJ**, Natale DA. The COG database: an updated version includes eukaryotes. *BioMed Central Bioinformatics*. 2003 Sep 11 [Epub ahead of print] PMID: 12969510

Submissions Corner

Submitting a Population or Phylogenetic Sequence Set

The Entrez PopSet database accommodates four varieties of sequence set, to represent versions of a gene or sequence region derived from varying sources. The sources may be isolates of a single organism, comprising a “population set”, an ensemble of organisms, comprising a “phylogenetic set”, individuals from a population of unclassified or unknown organisms, comprising an “environmental set”, or various mutational forms, comprising a “mutation set”. Figure 1 shows a PopSet, consisting of GenBank records AF474412-AF474791 for Adelie penguin mitochondrial DNA. PopSets such as this one can be submitted to GenBank in four easy steps using NCBI’s sequence-submission tool, Sequin.

Step 1—Generate an alignment or FASTA set

Sequin can import alignments in any of the FASTA+GAP, PHYLIP, or NEXUS formats illustrated at:

www.ncbi.nlm.nih.gov/Sequin/QuickGuide/sequin.htm#before

The source information, such as isolates and specimen vouchers, can be included in the definition lines of the alignment to be imported by Sequin. The definition line begins with a “>” sign and is included at the bottom of the alignment for the PHYLIP or NEXUS format or just above the sequence for the FASTA+GAP format. An example of a definition line for a population set in NEXUS format from *Escherichia coli* strain ECOR10 is:

```
>[organism= Escherichia coli] [strain=ECOR10] [clone=1]
```

The modifier information must be included in square brackets with no spaces on either side of the “=” sign. There is no limit to the number of modifiers you can add from the list found at:

www.ncbi.nlm.nih.gov/Sequin/sequin.hlp.html

If the sequences are included in the form of an aligned or unaligned multiple FASTA sequence set, include the sequence identifier, in this case, “seqid”, after the “>” sign followed by the modifiers, as in:

```
>seqid1 [organism= Escherichia coli] [strain=ECOR10] [clone=1]
```

Step 2—Import the Sequences

After entering the submission and contact information in Sequin, choose “Population study”, “Phylogenetic study”, “Mutation study” or “Environmental samples” in the “Sequence Format” panel. Next, import the set of nucleotide sequences in the desired alignment format by clicking on “Import Nucleotide”.

Step 3—Add and propagate features

If you imported the sequences as an alignment, as opposed to an unaligned multiple FASTA set, you have an easy way of propagating features from one member to all members of the set. Select the first entry under “Target Sequence” and add the appropriate features to this entry using the Annotate panel. Then, you may propagate all or the selected features to the remaining entries using the Edit—Feature propagate option.

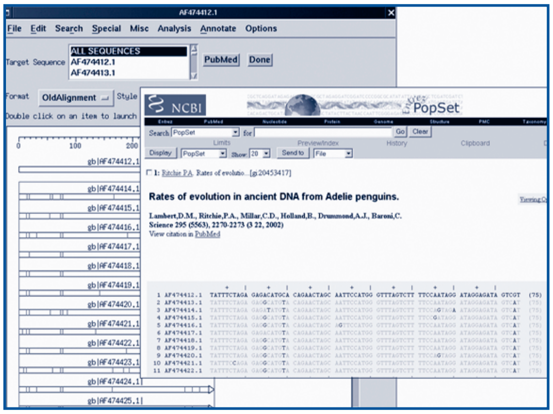
Step 4—Adding Distinguishing information

NCBI strongly encourages distinguishing information for the individual sequences in PopSets. This information can include strains for cultured bacteria, algae, fungi and laboratory animals; clones for sequences obtained by direct PCR-amplification and cloning of an environmental bulk DNA sample; and specimen vouchers for sets of multicellular organisms.

Strain identifiers help distinguish specific cultures from other isolates of the same taxon. A strain may be designated in a variety of ways, such as by the name of an individual, by a culture collection number or locality, by an arbitrary identifier, or by a label used within the submitting laboratory.

A specimen voucher is the remainder from which a sequence has been obtained, or, where coidentity is assured, a representative of the sequence source specimen. Vouchers should be deposited in respositories accessible to the public, such as herbaria or museum collections. Specimen vouchers allow verification of the identity of a taxon and serve as a source for additional molecular analyses. A common format for vouchers includes the collector's name and a unique number, plus the repository or its abbreviation. For example:

C.S. Shen 2459 (HMAS)
A.J. Smith 12.iii.2002 (AMNH)
H. Perrier s.n. (P)



□ 1: Ritchie PA. Rates of evolutio...[gi:20453417]

Rates of evolution in ancient DNA from Adelie penguins.

Lambert,D.M., Ritchie,P.A., Millar,C.D., Holland,B., Drummond,A.J., Baroni,C.
Science 295 (5563), 2270-2273 (3 22, 2002)
View citation in PubMed

```
                     +    |     +    |     +    |     +    |     +    |     +    |     +    |
 1 AF474412.1   TATTTCTAGA GAGACATGCA CAGAACTAGC AATTCCATGG GTTTAGTCTT TTCCAATAGG ATAGGAGATA GTCGT  (75)
 2 AF474413.1   TATTTCTAGA GAGGCATGTA CAGAACTAGC AATTCCATGG GTTTAGTCTT TTCCAATAGG ATAGGAGATA GTCAT  (75)
 3 AF474414.1   TATTTCTAGA GAGATATGTA CAGAACTAGC AATTCCATGG GTTTAGTCTT TTCCAGTAGA ATAGGAGATA GTCAT  (75)
 4 AF474415.1   TATTTCTAGA GAGGCATGTA CAGAACTAGC AATTCCATGG GTTTAGTCTT TTCCGATAGG ATAGGAGATA GTCAT  (75)
 5 AF474416.1   TATTTCTAGA GAGGCATGTA CAGAACTAGC AGTTCCATGG GTTTAGTCTT TTCCAATAGG ATAGGAGATA GTCAT  (75)
 6 AF474417.1   TATTTCTAGA GAGACATGTA CAGAACTAGC AATTCCATGG GTTTAGTCTT TTCCAATAGG ATAGGAGATA GTCAT  (75)
 7 AF474418.1   TATTTCTAGA GAGGCATGTA CAGAACTAGC AATTCCATGG GTTTAGTCTT TTCCAATAGG ATAGGAGATA GTCAT  (75)
 8 AF474419.1   TATTTCTAGA GAGGCATGTA CAGAACTAGC AATTCCATGG GTTTAGTCTT TTCCAATAGG ATAGGAGATA GTCAT  (75)
 9 AF474420.1   TATTTCTAGA GAGGCATGTA CAGAACTAGC AATTCCATGG GTTTAGTCTT TTCCAGTAGG ATAGGAGATA GTCAT  (75)
10 AF474421.1   TATTTCCAGA GAGGCATGTA CAGAACTAGC AATTCCATGG GTTTAGTCTT TTCCAATAGG ATAGGAGATA GTCAT  (75)
11 AF474422.1   TATTTCTAGA GAGGCATGTA CAGAACTAGC AATTCCATGG GTTTAGTCTT TTCCAATAGG ATAGGAGATA GTCAT  (75)
```

Figure1. Views from Sequin and Entrez, respectively, of a population set of Adelie penguin mitochondrial DNA sequences taken from ancient bone and fresh blood for a study of rates of evolution.

In the absence of specimen vouchers, the following source modifiers are helpful:

- cultivar, strain, isolate, breed, ecotype, or genotype name
- germplasm, seed, or stock center accession number
- collection locality, date, and/or collection number.

Along with the specimen voucher information, you may provide online images of the specimens that will be made available in Entrez through LinkOut. For an example, retrieve the entry AY090229 in Entrez and click on LinkOut through "Links" on the right hand side of the page to view an image of an insect specimen.

When your submission is complete, save the entries in the native Sequin format and e-mail to:

gb-sub@ncbi.nlm.nih.gov

You will receive confirmation of your submission along with your accession numbers within approximately 2 business days.

—MB

GenBank® Release 139

GenBank Release 139, made available in December 2003, contains over 30 million sequence entries totaling more than 36 billion base pairs. Release 140 is expected in February. 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 139 flatfiles consume about 122 gigabytes while the ASN.1 version consumes about 108 gigabytes. The data can also be downloaded at two mirror sites:

genbank.sdsc.edu/pub
bio-mirror.net/biomirror/genbank

UniGene Adds Four

Four new organisms are now represented in the UniGene collection of gene-oriented transcript sequences; *Toxoplasma gondii* with 3,106 clusters, *Saccharum officinarum* (noble cane)with 5,197 clusters, *Schistosoma mansoni* (the blood fluke) with 1,170 clusters, *Strongylocentrotus purpuratus* (the purple sea urchin) with 2,592 clusters, and *Physcomitrella patens* (physcomitrella moss) with 6,927 clusters. UniGene clusters are created at NCBI for all organisms represented in GenBank with more than 70,000 Expressed Sequence Tag sequences. UniGene now covers over 30 animals and plants and can be searched using the Entrez search system where it is linked to nucleotide records.

RefSeq Version 3 Released

RefSeq Release 3, covering 2,218 organisms, and containing over 844,000 protein sequences and annotations, includes the RefSeq data as of January 13, 2004.

Download the release at:

ftp.ncbi.nih.gov/refseq/release

Daily updates between releases are made available at:

ftp.ncbi.nih.gov/refseq/daily/new

To receive announcements of future RefSeq releases and incremental large updates, subscribe to NCBI's refseq-announce mail list:

refseq-announce@ncbi.nlm.nih.gov