

NCBI News, September 2010

Peter Cooper, Ph.D.¹ and Dawn Lipshultz, M.S.²

Created: September 27, 2010.

New Databases and Tools

Find-in-Sequence now available in nucleotide and protein databases

The new Find-in-Sequence search finds and highlights subsequences, patterns, or motifs in nucleotide or protein sequences displayed in the Entrez system. In nucleotide sequences, sub-sequences or patterns of interest may include restriction enzyme recognition sites, transcription factor binding sites and other promoter elements, polyadenylation signals, start and stop codons, and many others. In proteins, these may include potential cleavage locations, sites of potential post-translational modification, motifs representing active sites, cofactor binding pockets, or other structural or functional signatures. The link to open the Find-in-Sequence search box appears in the Analysis Tools portlet in the right-hand Discovery column of nucleotide and protein record views. Find-in-sequence works with single and multiple sequence displays with any format that shows the sequence (GenBank, GenPept, FASTA). The search box that opens at the bottom of the display has the familiar look and feel of the find in page function of the Web browser showing all matches and providing means to jump to the next or previous match. However, unlike the simple find function, Find-in-Sequence allows matching across spaces and line breaks in the formatted sequence and can use standard nucleotide and protein ambiguity codes as well as [Prosites patterns](#) for protein sequences. Find-in-Sequence provides rapid mapping of custom features onto nucleotide and protein sequences and is a powerful addition to the suite of analysis tools now available directly from sequence records in the Entrez system.

A [video](#) demonstrating this feature is now available on the NCBI YouTube channel.

¹ NCBI; Email: cooper@ncbi.nlm.nih.gov. ² NCBI; Email: lipshult@ncbi.nlm.nih.gov.

Display Settings: GenBank Send: Change region shown

Homo sapiens prolactin (PRL), transcript variant 1, mRNA

NCBI Reference Sequence: NM_000948.4
FASTA Graphics

Go to: Analyze this sequence
Run BLAST
Pick Primers
Find in this Sequence

LOCUS NM_000948 1359 bp mRNA linear PRI 20-SEP-2010
DEFINITION Homo sapiens prolactin (PRL), transcript variant 1, mRNA.
ACCESSION NM_000948
VERSION NM_000948.4 GI:254281326

1. brown, isoform A [Drosophila melanogaster]
NCBI Reference Sequence: NP_523824.1
GenPept Graphics Related Sequences Identical Proteins

```
>gi|17647219|ref|NP_523824.1| brown, isoform A [Drosophila melanogaster]
MQESGGSGGGPSLCLWKLNYVPPDQFQSNYSFWNECRKRLRLIQDASGHMKTGDLIAIIGSQA
GSLTLAATSQRLEGNLTDGVDLVNGMAMERHQMTRISSFLPQFQEVNKTPTAYEHLVFMHFRRTTK
AEKQRVADLLLAUGLRDAATHRIQQLSGGKRKLSLAELITDPIFLCDEPFTGLDSFSAVSVITLR
...
DSFETPSGESSASGSGSKS IEMEVVAESHESLQTRRELPAUGLVGNLSPNG
LFTTHILMDGGRIYVQGRTEQAQKFFDGLGYELPLNCPADPYLKLTLADKE
LYSGSWLLARSYSGDYLVKQVNFKKIRMIYQVYLLMVFRTEDLRNIRSGL
IGGLTQRTVQDVGGISIFLMSNEMIFPFSYGVYIIFPAALPIIRREVGEVGY
MGYVFLSVIYASIIYTRGFLLYLSMGLMSLAVAAVGVYVFLSSLFESQK
YMNVDVTPGLKYLISLFFYSNEALMYKFWIDIDINDCPVNEHPHCITGVEV
LVVVAVIPHIVSFLVRRYIHRSGYY
```

1 atccttattc tatatctctt ggtatttagt gtaaaaaatt taaaatcttt acctagcaat
61 cttgaggaag aaacttgata aetgataata catgagattg ttacctaatg gaataataat
121 cctatatatt caacaactt tagagaata agataaattt taaagtaaat gactctctga
181 gttttataga toctccaacc caactctatg toagatctca cctctcatcat ttctctcatt
241 tccttttggc ctaattaatc aaaaatcttc ctagaattgt cattctggc cagtattgtct
301 tcctgaatat gaataagaaa taaaatacca ttgatgttt gaaattatgg gggtaatctc
361 aatgacggaa atagatgacc aggaaaaggg aaacgaatgc ctgattcatt atattcaatga
421 agatatcaaa gttttataaa gccaatatct gggaaaagaa aaaccgtgag acttccagat
481 cttctctggt gaagtgtggt tctgcaaac atcaagaac caaacatcaa aggatcgccc
541 tggaaaagggt cctctctgct gctgctggtg tcaaacctgc tctgtgcca gagcgtggcc
601 cctctgccc cctctctgct gctgctggtg tcaaacctgc tctgtgcca gagcgtggcc
661 gaccgcgcc tcgtctctgct ccaactacat cataaactct cctcagaagt gttcagcgaa
721 ttcgataaac ggtatacccc tggccggggg ttcattacca aggccatcaa cagctgccac
781 acttctctcc ttgccacccc cgaagacaag gagcaagccc aacagatgaa tcaaaaagac
841 tttctgagcc tgatagtcag catattcgca tctctggaatg agcctctgta tcatctggtc
901 acggaagtae gttgatgca agaagcccc gaggtatctc tatccaagc tgtagagatt
961 gaggagcaaa ccaaacgctt tctagagggc atggagctga tagtcagcca ggtctatcct
1021 gaaaccaaag aaaatgagat ctaccctgct tggctgggac tccatcctct gcagatggct
1081 gatgaagagt ctgctctctc tgccttattt aacctgctcc actgctcaac cagggatcca
1141 cataaaatcg acaattatct caagctctg agtgccgaa tcaatccaca caacaactgc
1201 taagcccaca tccattatct ctatttctga gaaggtcctt aatgatcctg tccattgcaa
1261 gctctcttta gttgatctc ttttgaatcc atgcttgggt gtaacaggtc tctctttaa
1321 aataaaaaa tgactcetta gagacatcaa aatctaaaa

atg Find 12 of 25 NM_000948: 520-522

[AG]-x(4)-G-K-[ST] Find 1 of 3 NP_523824: 66-73

Recent NCBI Publication on dbVar

The NCBI Database of Genomic Structural Variation (dbVar) announced in the February 2010 issue of the NCBI News was described in a recent correspondence article in *Nature Genetics*.

Church DM, Lappalainen I, Sneddon TP, Hinton J, Maguire M, Lopez J, Garner J, Paschall J, DiCuccio M, Yaschenko E, Scherer SW, Feuk L, Flicek P. **Public data archives for genomic structural variation.** *Nat Genet.* 42, 813-814 (2010). PMID: 20877315

The publication and resource are also announced in a recent [NIH press release](#).

dbVAR provides information on large scale (> 1Kb) variations in genome sequences and includes copy number variants and other deletions and insertions.

Epigenomics Resource

A new Epigenomics resource (www.ncbi.nlm.nih.gov/epigenomics) is part of the Entrez search and retrieval system. Epigenomics is a new field of research with the goal of understanding how different cell types and lineages acquire distinct patterns of gene expression. The Epigenomics database contains results of genome-wide studies on modifications of chromatin (histone modification, DNA methylation, DNAase footprinting) in various cell types. These studies assay programmable changes that affect gene expression (epigenetics). A Sample Browser provides access to experimental data. Data can be filtered based on biological attributes such as cell type, differentiation stage, health status, and others. Data may be displayed on the genome using either the UCSC genome browser or NCBI graphical sequence viewer as shown below.

Term Filter Containing word(s):

Attribute Filter

| Species | Cell Type | Lab |
|-------------------------|-----------------------------|--|
| All | centroblast | All |
| Arabidopsis thaliana | centrocyte | Biotech Research and Innovation Centre/Center for Genomic Regulation |
| Caenorhabditis elegans | embryonic kidney cell | Cold Spring Harbor Laboratory |
| Drosophila melanogaster | embryonic stem cell | Dana-Farber Cancer Institute |
| Homo sapiens | epithelial cell | Laboratory of Molecular Immunology/National Institute of Health |
| Mus musculus | erythrocyte progenitor cell | University of Michigan |
| | fibroblast | |

Samples

| Sample ID | Species | Cell Type | Tissue Type | Cell Line | Cell Population | Differentiation State |
|---|--------------|-----------------|---------------|------------|-----------------|-----------------------|
| <input checked="" type="checkbox"/> ESM000113 | Homo sapiens | epithelial cell | mammary gland | MDA-MB-231 | | |
| <input type="checkbox"/> ESM000262 | Homo sapiens | epithelial cell | | LNCaP | | |
| <input type="checkbox"/> ESM000263 | Homo sapiens | epithelial cell | | LNCaP | | |
| <input type="checkbox"/> ESM000264 | Homo sapiens | epithelial cell | | LNCaP | | |
| <input type="checkbox"/> ESM000330 | Homo sapiens | epithelial cell | | SH-SY5Y | | |
| <input type="checkbox"/> ESM000331 | Homo sapiens | epithelial cell | | HeLa S3 | | |
| <input type="checkbox"/> ESM000366 | Homo sapiens | epithelial cell | | VCaP | | |
| <input type="checkbox"/> ESM000367 | Homo sapiens | epithelial cell | | VCaP | | |
| <input type="checkbox"/> ESM000368 | Homo sapiens | epithelial cell | | LNCaP | | |
| <input type="checkbox"/> ESM000369 | Homo sapiens | epithelial cell | | LNCaP | | |
| <input type="checkbox"/> ESM000370 | Homo sapiens | epithelial cell | | VCaP | | |
| <input type="checkbox"/> ESM000371 | Homo sapiens | epithelial cell | | LNCaP | | |
| <input type="checkbox"/> ESM000372 | Homo sapiens | epithelial cell | | LNCaP | | |

NC_000006.11 (171,115,067 bases)

Sequence | Set Origin | Views & Tools | Markers | Search...

1 10 M 20 M 30 M 40 M 50 M 60 M 70 M 80 M 90 M 100 M 110 M 120 M 130 M 140 M 150 M 160 M 170 M 171,115,067

151,089,060 : 151,445,269 (356,210 bases shown, positive strand)

Sequence | Flip Strands | Tools | Markers | Default | Options

151,100 K 151,150 K 151,200 K 151,250 K 151,300 K 151,350 K 151,400 K 151,450 K

y - Sequence NC_000006.11: Homo sapiens chromosome 6, GRCh37 pr

a: H3K4me2, Max:188 / Min:0 - Human breast cancer epithelial cell, MDA-MB-231 cell line, mammary gland

a: H3K36me3, Max:145 / Min:0 - Human breast cancer epithelial cell, MDA-MB-231 cell line, mammary gland

- Genes

LOC644850 | MTHFD1L

150465793..152068537
 name: NA000000603.1
 Title: Human breast cancer epithelial cell, MDA-MB-231 cell line, mammary gland, adenocarcinoma : H3K36me3
 Comment: High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing Submitted By: Cold Spring Harbor Laboratory
 Global DNA methylation is examined by bisulfite-seq in both human primary and cancer cell lines. H3K4me2 and H3K36me3 are also examined.

You are here: NCBI > DNA & RNA > Nucleotide Database

Microbial Genomes

Eleven finished microbial genomes were released during September 2010. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site:

<ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

GenBank News

GenBank release 179.0 is available via web and FTP. The current release includes information available as of August 16, 2010. Release notes are available on the on the NCBI ftp site: <ftp.ncbi.nih.gov/genbank/gbrel.txt>

Updates and Enhancements

dbGaP

The Entrez dbGaP resource has migrated to the streamlined discovery-oriented design that has been in service in PubMed for nearly a year, described fully in the [November 2009](#) issue of the NCBI News. Included in the re-design is a new home page (www.ncbi.nlm.nih.gov/gap), Limits and Advanced Search page. The image of the human face on the homepage, shown immediately below, is a composite of faces from approximately 10,000 people.



The screenshot shows the dbGaP homepage. At the top, there is a navigation bar with "NCBI Resources" and "How To" menus, and "My NCBI Sign In" links. Below this is a search bar with "dbGaP" entered, and "Limits", "Advanced search", and "Help" links. The main content area features a large banner with a composite human face image and the text: "dbGaP The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype." Below the banner are three columns of links: "Getting Started" (dbGaP Tutorial, Overview, FAQ, How to Submit), "Access dbGaP Data" (Apply for Controlled Access Data, Public Data via ftp Download, Browse All Studies), and "Important Links" (dbGaP RSS Feed, Code of Conduct, Security Procedures, Contact Us).

PubMed DTDs

PubMed E-Utility 2011 DTDs will be updated in mid-December, approximately on December 13. The new DTDs can be found:

http://eutils.ncbi.nlm.nih.gov/corehtml/query/DTD/pubmed_110101.dtd

RefSeq

RefSeq Release 43 is now available through the Entrez system and can be downloaded from the FTP site (<ftp.ncbi.nlm.nih.gov/refseq/release>). This full release incorporates genomic, transcript, and protein data available as of September 16. It includes 11,223,078 records from 10,854 different species and strains. Changes since the last release are described in the release notes (<ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/RefSeq-release43.txt>). More information on the RefSeq project is available on the RefSeq Homepage (www.ncbi.nlm.nih.gov/RefSeq/).

Entrez Gene Record Status

Entrez Gene (www.ncbi.nlm.nih.gov/gene) has two new properties field terms to identify the status of records that are no longer current (alive). Replaced records, retrieved with the search term “replaced”[Properties], are those that have been made secondary to (merged with) another gene record while discontinued records, retrieved with “discontinued”[Properties], are no longer current but not identified with another record. As before, current records can be retrieved with the term “alive”[properties] or by using “current only”[Filter].

NCBI Workshop at ASHG Annual Meeting

NCBI scientists will present a special workshop at the American Society for Human Genetics meeting on November 4 at 7:15 p.m. The workshop entitled, “[A Practical Guide to Genome-Scale Data at NCBI](#)”, will provide information on genome-scale resources available at NCBI including finding and downloading data, analysis, and management of data sets. NCBI will also staff an exhibit booth at the meeting.

Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html. To receive updates on the *NCBI News*, please see: www.ncbi.nlm.nih.gov/About/news/announce_submit.html.

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: www.ncbi.nlm.nih.gov/feed/.

Users can also stay updated on NCBI's resources on Facebook and Twitter: twitter.com/NCBI.

Send comments and questions about NCBI resources to: info@ncbi.nlm.nih.gov, or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.