NCBI FieldGuide

NCBI Molecular Biology Resources

September 2006

NCBI FieldGuide

NCBI Resources

- About NCBI
- NCBI Databases
- Tools
- Entrez
- Examples

NCBI FieldGuide

The National Center for Biotechnology Information

Created in 1988 as a part of the National Library of Medicine at NIH

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

NCBI FieldGuide

Molecular Databases

- Sequence
- Structure
- Expression

Types of Databases

NCBI FieldGuide

- Primary Databases
  - Original submissions by experimentalists
  - Content controlled by the submitter
    - Examples: GenBank, SNP, GEO
- Derivative Databases
  - Built from primary data
  - Content controlled by third party (NCBI)
    - Examples: Refseq, TPA, RefSNP, UniGene, NCBI Protein, Structure, Conserved Domain

GenBank: NCBI's Primary Sequence Database

NCBI FieldGuide

Release 155	**August 2006**
79,093,266	**Records**
145,739,069,776	**Total Bases**
240 Gigabytes (non-WGS)	1010 files (non-WGS)

- full release every two months
- incremental updates daily
- available only via ftp

ftp://ftp.ncbi.nih.gov/genbank/

Traditional GenBank Divisions

NCBI FieldGuide

- Direct Submissions (Sequin and BankIt)
- Accurate
- Well characterized

BCT	Bacterial and Archeal
INV	Invertebrate
MAM	Mammalian (ex. ROD and PRI)
PHG	Phage
PLN	Plant and Fungal
PRI	Primate
ROD	Rodent
SYN	Synthetic (vectors, synth. genes)
VRL	Viral
VRT	Other Vertebrate

Bulk Divisions

NCBI FieldGuide

- Batch Submission and htg (email and ftp)
- Inaccurate
- Poorly Characterized
- Largest component of GenBank (non-WGS)

- Expressed Sequence Tag
  - 1st pass single read cDNA
- Genome Survey Sequence
  - 1st pass single read gDNA
- High Throughput Genomic
  - incomplete sequences of genomic clones
- Sequence Tagged Site
  - PCR-based mapping reagents

GenBank Records

NCBI FieldGuide

```
LOCUS       AY198415                 4675 bp    mRNA    linear   PRI 11-AUG-2003
DEFINITION  Homo sapiens PR domain-containing protein 1 beta (PRDM1) mRNA,
            complete cds.
ACCESSION   AY198415
VERSION     AY198415.1  GI:28630982
```

Primate division (traditional)

```
LOCUS       AI588321                  482 bp    mRNA    linear   EST 07-JUN-2001
DEFINITION  fb99c12.x1 Zebrafish WashU MPIMG EST Danio rerio cDNA clone
            IMAGE:3720022 3', mRNA sequence.
ACCESSION   AI588321
VERSION     AI588321.1  GI:4597368
KEYWORDS    EST.
```

EST division (bulk)

```
LOCUS       AC027700              229988 bp    DNA     linear   HTG 19-MAR-2004
DEFINITION  Mus musculus chromosome 10 clone RP23-145G7 map 10, *** SEQUENCING
            IN PROGRESS ***, 3 unordered pieces.
ACCESSION   AC027700
VERSION     AC027700.7  GI:45593019
KEYWORDS    HTG; HTGS_PHASE1; HTGS_FULLTOP; HTGS_ACTIVEFIN.
SOURCE      Mus musculus (house mouse)
```

HTG division (bulk)

```
LOCUS       AADN01000358           53718 bp    DNA     linear   VRT 29-FEB-2004
DEFINITION  Gallus gallus chromosome 3 Cont0.358, whole genome shotgun
            sequence.
ACCESSION   AADN01000358 AADN01000000
VERSION     AADN01000358.1  GI:44803930
KEYWORDS    WGS.
SOURCE      Gallus gallus (chicken)
```

Vertebrate division (chicken WGS)

RefSeq: NCBI's Derivative Sequence Database

NCBI FieldGuide

- **Curated transcripts and proteins [NM_, NP_]**
  - reviewed
  - human, mouse, rat, cow, fruit fly, zebrafish, arabidopsis *C.elegans*
- **Model transcripts and proteins [XM_, XP_]**
- **Assembled Genomic Regions [NT_, NW_]**
  - draft human genome
  - mouse genome
- **Chromosome records [NC_]**
  - microbial
  - organelle

`srcdb_refseq[Properties]`

ftp://ftp.ncbi.nih.gov/refseq/release/

RefSeq Records

NCBI FieldGuide

```
LOCUS       NM_007548                5281 bp    mRNA    linear   ROD 25-AUG-2004
DEFINITION  Mus musculus PR domain containing 1, with ZNF domain (Prdm1), mRNA.
ACCESSION   NM_007548
VERSION     NM_007548.1  GI:6680783
```

Curated transcript

```
LOCUS       NT_039492            23191165 bp    DNA     linear   CON 31-AUG-2004
DEFINITION  Mus musculus chromosome 10 genomic contig, strain C57BL/6J.
ACCESSION   NT_039492
VERSION     NT_039492.3  GI:51765603
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
```

RefSeq contig

```
LOCUS       XP_228320             1022 aa            linear   ROD 23-OCT-2003
DEFINITION  similar to Blimp-1 [Rattus norvegicus].
ACCESSION   XP_228320
VERSION     XP_228320.2  GI:34853200
DBSOURCE    REFSEQ: accession XM_228320.2
KEYWORDS    .
SOURCE      Rattus norvegicus (Norway rat)
```

RefSeq gene model

```
LOCUS       NC_000006           170975699 bp    DNA     linear   CON 24-AUG-2004
DEFINITION  Homo sapiens chromosome 6, complete sequence.
ACCESSION   NC_000006
VERSION     NC_000006.9  GI:51511722
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
```

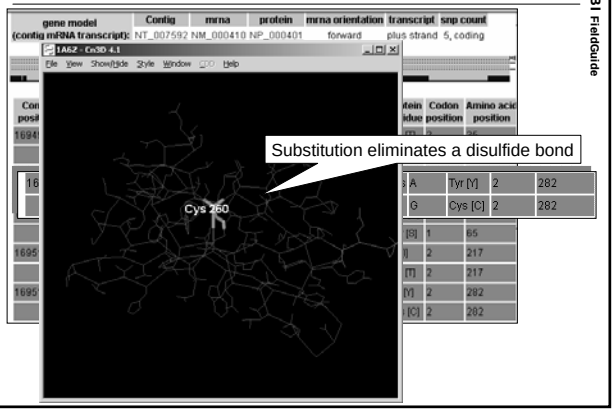
RefSeq chromosome

NCBI's SNP Database

NCBI FieldGuide

- Primary Database and Derivative (RefSNP)
- Single Nucleotide Polymorphism
- Repeat polymorphisms
- Insertion-Deletion Polymorphisms
- 35 species
- Over 87 million submissions
- Over 50 million RefSNP clusters

RefSNP HFE



NCBI FieldGuide

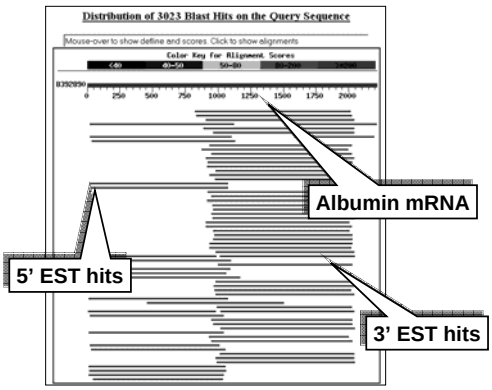
What is UniGene?

A gene-oriented view of sequence entries

- **MegaBlast based automated sequence clustering**
- **Now informed by genome hits *New!***
- **Nonredundant set of gene oriented clusters**
- **Each cluster a unique gene**
- **Information on tissue types and map locations**
- **Includes known genes and uncharaterized ESTs**
- **Useful for gene discovery and selection of mapping reagents**

NCBI FieldGuide

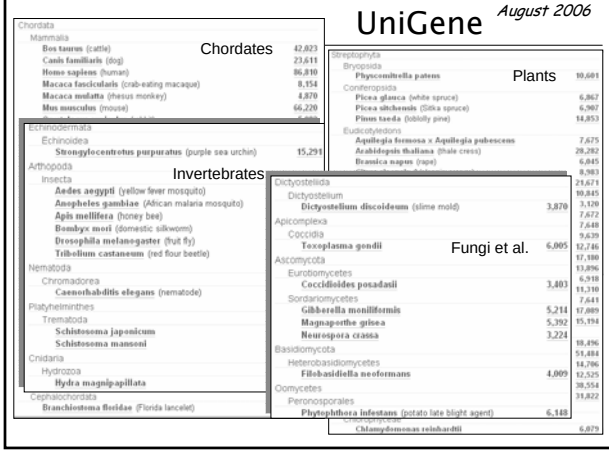
EST hits: Human mRNA



NCBI FieldGuide

UniGene

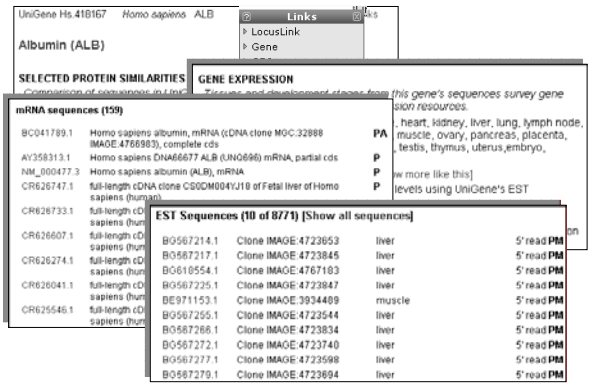
August 2006



NCBI FieldGuide

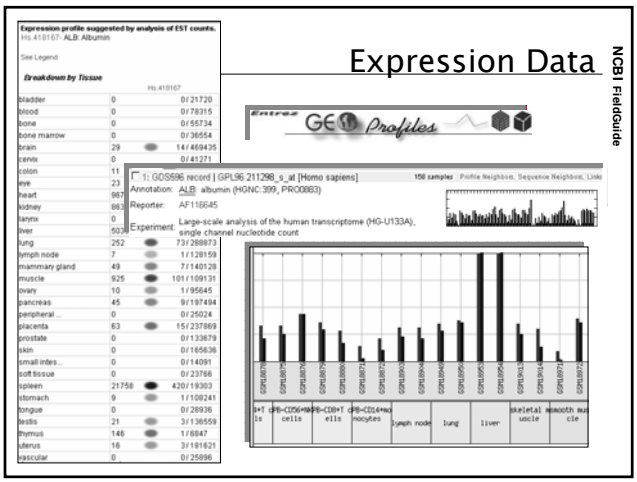
UniGene: Expressed Sequences

NCBI FieldGuide



Expression Data

NCBI FieldGuide



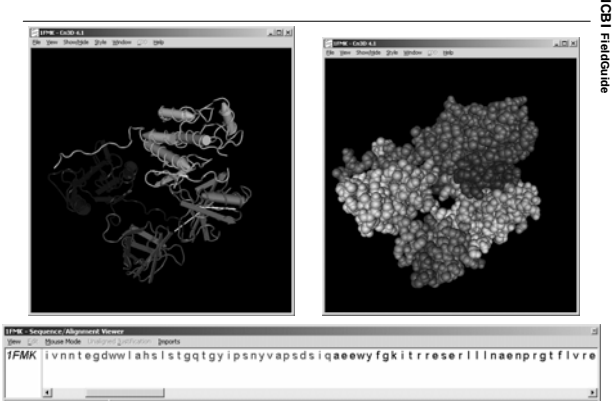
MMDB: Molecular Modeling Data Base

NCBI FieldGuide

- **Derived from experimentally determined PDB records**
- **Value added to PDB records including:**
  - **Addition of explicit chemical graph information**
  - **Validation**
  - **Inclusion of Taxonomy, Citation, and other information**
  - **Conversion to ASN.1 data description language**
- **Structure neighbors determined by Vector Alignment Search Tool (VAST)**

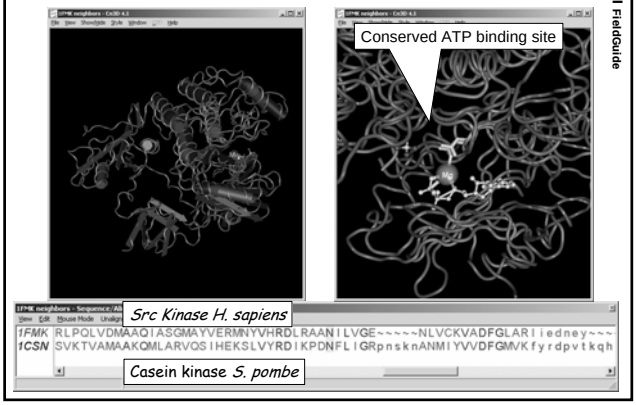
Cn3D 4.1: C-SRC

NCBI FieldGuide



Cn3D 4.1: Structural Alignment

NCBI FieldGuide



Cn3D: Simple Homology Modeling

NCBI FieldGuide

Protein Domains

NCBI FieldGuide

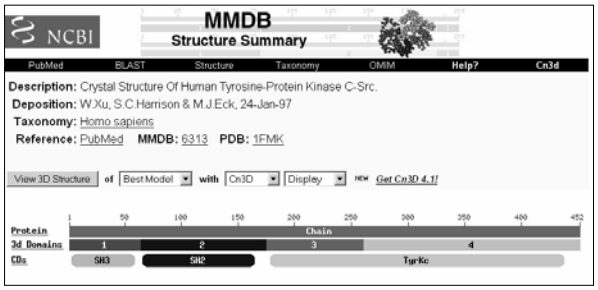
- Structural Domain
  - Discrete independently folding unit of a protein
- Conserved Domain (sequence-based)
  - Protein region with recognizable position specific pattern of sequence conservation
- Sequence-based domains often roughly correspond to structural domains
- Domains often have distinct, identifiable functions

NCBI's Conserved Domain Database

NCBI FieldGuide

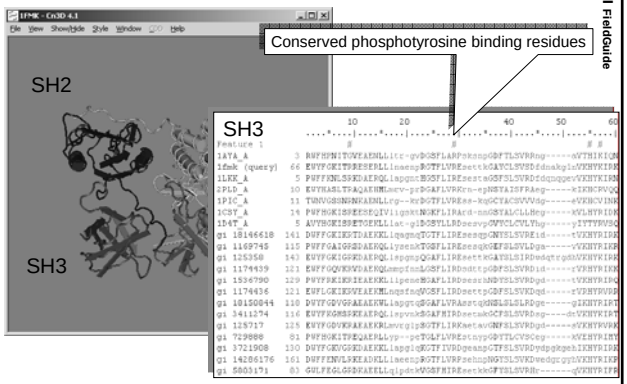
- PSI-BLAST –based score matrices
- Searchable with RPS-BLAST
- Sources
  - SMART
  - PFAM
  - COGs
  - NCBI curated domains
    - structure informed alignments

Src Domains



NCBI FieldGuide

Structure vs Conserved Domain



NCBI FieldGuide

NCBI FieldGuide

NCBI Tools

BLAST

VAST

Basic Local Alignment Search Tool

NCBI FieldGuide

- Calculates similarity for biological sequences
- Finds best local alignments
- Heuristic approach based on Smith-Waterman algorithm
- Searches for matching “words” rather than individual residues
- Uses statistical theory to determine if a match might have occurred by chance

BLAST and BLAST-like programs

NCBI FieldGuide

- Traditional BLAST (blastall) nucleotide, protein, translations
  - blastn nucleotide query vs. nucleotide database
  - blastp protein query vs. protein database
  - blastx nucleotide query vs. protein database
  - tblastn protein query vs. translated nucleotide database
  - tblastx translated query vs. translated database
- Megablast nucleotide only
  - Contiguous megablast
    - **Nearly identical sequences**
  - Discontiguous megablast
    - **Cross-species comparison**
- Position Specific BLAST Programs protein only
  - Position Specific Iterative BLAST (PSI-BLAST)
    - **Automatically generates a position specific score matrix (PSSM)**
  - Reverse PSI-BLAST (RPS-BLAST)
    - **Searches a database of PSI-BLAST PSSMs**

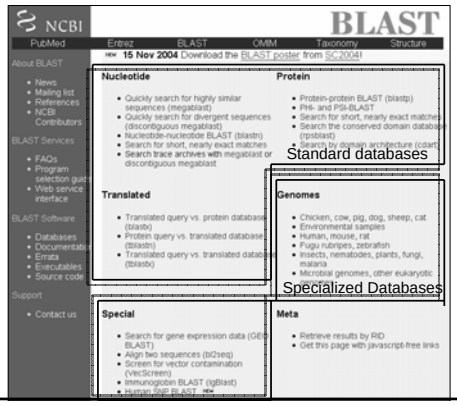
The Flavors of BLAST

NCBI FieldGuide

- Position independent scoring
  - Standard BLAST
    - traditional contiguous word hit
    - nucleotide, protein and translations
  - Megablast
    - can use discontiguous words
    - nucleotide only
    - optimized for large batch searches
- Position dependent scoring
  - PSI-BLAST
    - constructs PSSMs automatically
    - searches protein database with PSSMs
  - RPS BLAST
    - searches a database of PSSMs
    - basis of conserved domain database

The BLAST homepage

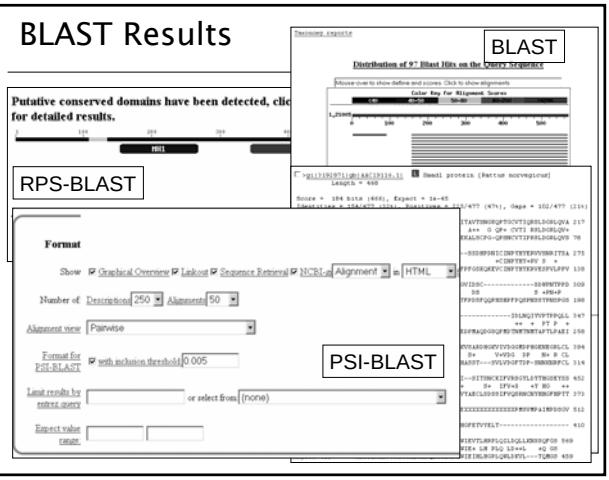
NCBI FieldGuide



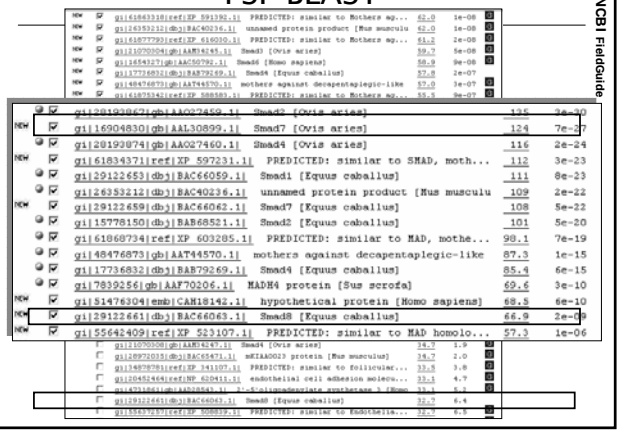
BLAST Search

NCBI FieldGuide

BLAST Results



PSI-BLAST



NCBI FieldGuide

VAST: Structure Neighbors

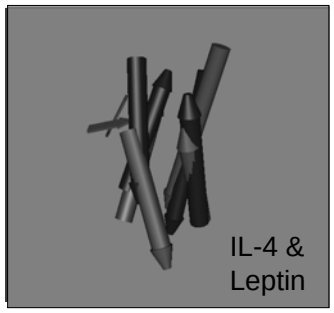
Vector Alignment Search Tool

For each protein chain,

locate SSEs (secondary structure elements),

and represent them as individual vectors.

align the vectors



Human IL-4

NCBI FieldGuide

Finding a human PR domain RefSeq

- Finding PRDM1
- Links and Neighbors with PRDM1

NCBI FieldGuide

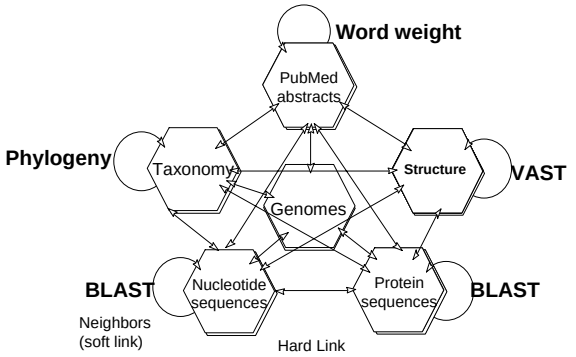
NCBI FieldGuide

WWW Access

Entrez & BLAST

Entrez: Database Integration

NCBI FieldGuide

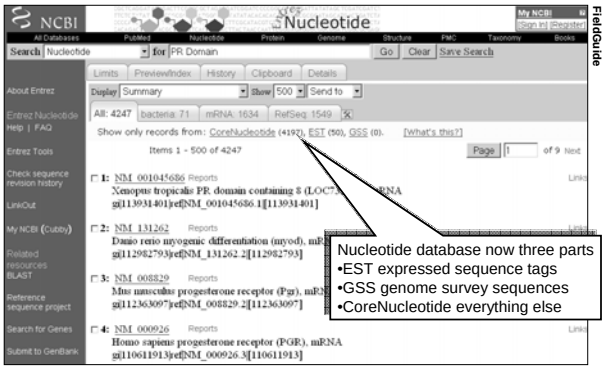


PR Domain Global Search

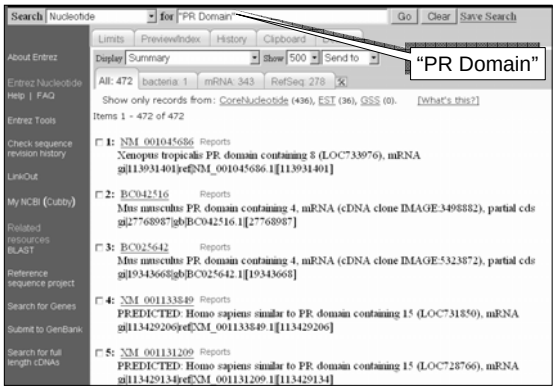
NCBI FieldGuide

PR Domain search results

NCBI FieldGuide

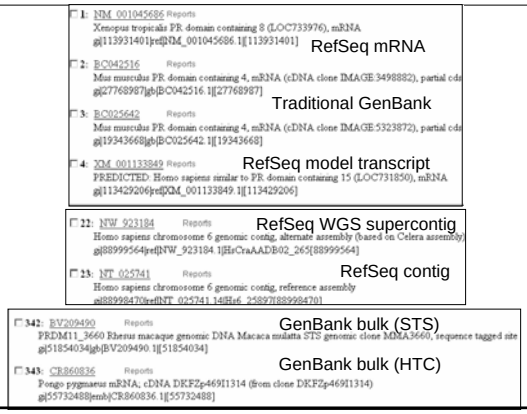


Sometimes Quotes Help



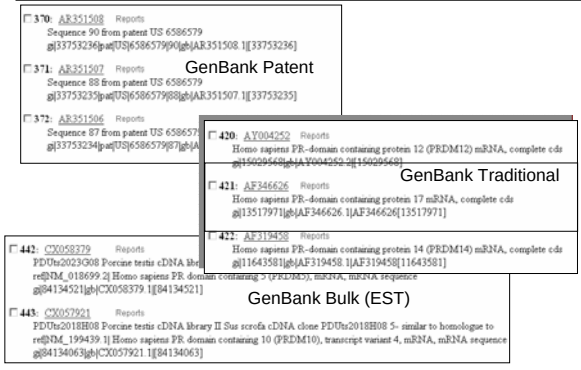
NCBI FieldGuide

"PR Domain" search results



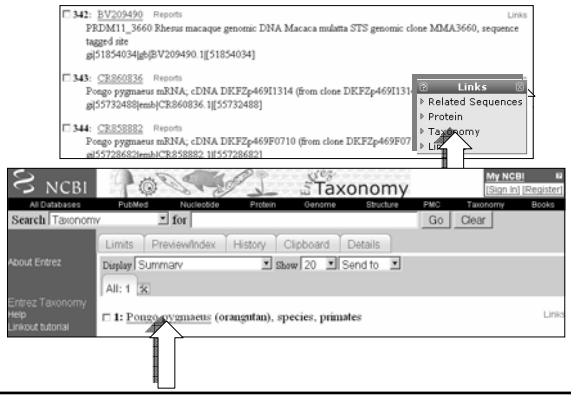
NCBI FieldGuide

PR Domain search results (continued)



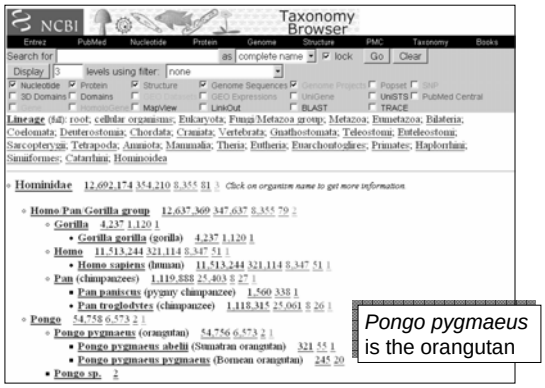
NCBI FieldGuide

What is *Pongo pygmaeus*?



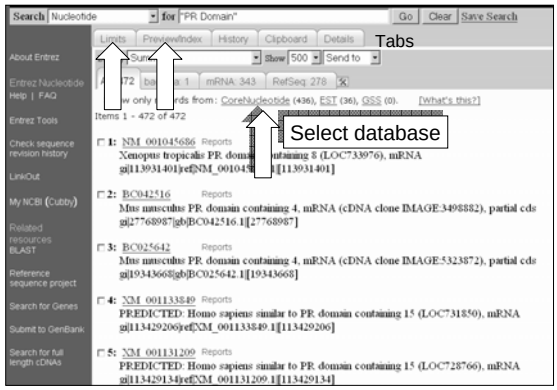
NCBI FieldGuide

Pongo pygmaeus: Taxonomy Browser



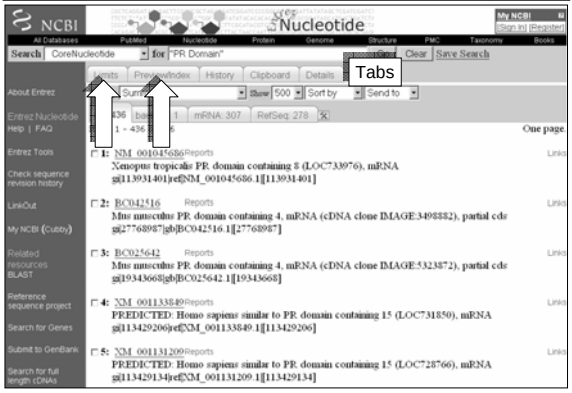
NCBI FieldGuide

Refining Search



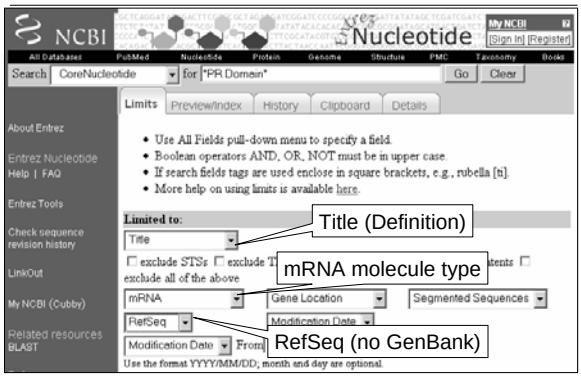
NCBI FieldGuide

Refining Search: CoreNucleotide



NCBI FieldGuide

Limits



NCBI FieldGuide

Limited Search Results

NCBI FieldGuide

Limits | Preview/Index | History | Clipboard | Details

Field: Title, Limits: mRNA, RefSeq

Display Summary Show 500 Sort by Send to

All: 212 | bacteria: 0 | mRNA: 212 | RefSeq: 212

Items 1 - 212 of 212

1: NM_001045686 Reports
Xenopus tropicalis PR domain containing 8 (LOC733976), mRNA
gi|113931401|ref|NM_001045686.1|[113931401]

2: XM_001133849 Reports
PREDICTED: Homo sapiens similar to PR domain containing 15 (LOC731850), mRNA
gi|113429206|ref|XM_001133849.1|[113429206]

3: XM_001131209 Reports
PREDICTED: Homo sapiens similar to PR domain containing 15 (LOC728766), mRNA
gi|113429134|ref|XM_001131209.1|[113429134]

4: XM_937753 Reports
PREDICTED: Homo sapiens PR domain containing 6 (PRDM6), mRNA
gi|113417194|ref|XM_937753.2|[113417194]

5: XM_927647 Reports
PREDICTED: Homo sapiens PR domain containing 6 (PRDM6), mRNA
gi|113416741|ref|XM_927647.2|[113416741]

Preview / Index: Adding Terms

NCBI FieldGuide

Add Term(s) to Query or View Index:

- Enter a term in the text box; use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within a field.
- Multiple terms selected from Index will be ORed; click AND to add to search.

Organism | human | Preview | Index

Click AND | OR | NOT to add terms selected from Index to the query box.

Up

human(2647826)
human adenovirus 1 isolate from a cat(3)
human adenovirus 12(28)
human adenovirus 14(16)
human adenovirus 19a(2)
human adenovirus 19p(2)
human adenovirus 2(75)
human adenovirus 3(125)
human adenovirus 3 7(1)
human adenovirus 35(24)

Down

Organism Field index is NCBI's Taxonomy

Human PR domain RefSeqs

NCBI FieldGuide

All: 31 | bacteria: 0 | mRNA: 31 | RefSeq: 31

Items 1 - 31 of 31

1: XM_001133849 Reports
PREDICTED: Homo sapiens similar to PR domain cont
gi|113429206|ref|XM_001133849.1|[113429206]

2: XM_001131209 Reports
PREDICTED: Homo sapiens similar to PR domain cont

13: NM_182907 Reports
Homo sapiens PR domain containing 1, with ZNF domain (PRDM1), transcript variant 2, mRNA
gi|33946273|ref|NM_182907.1|[33946273]

14: NM_001198 Reports
Homo sapiens PR domain containing 1, with ZNF domain (PRDM1), transcript variant 1, mRNA
gi|33946272|ref|NM_001198.2|[33946272]

gi|113411897|ref|XM_001126711.1|[113411897]

6: XM_936932 Reports
PREDICTED: Homo sapiens similar to PR domain zinc
(Transcription factor MEL1) (LOC647868), mRNA
gi|113411893|ref|XM_936932.2|[113411893]

7: NM_015866 Reports
Homo sapiens PR domain containing 2, with ZNF doma
gi|55953108|ref|NM_015866.3|[55953108]

8: NM_012231 Reports
Homo sapiens PR domain containing 2, with ZNF domain

24: XM_929341 Reports
PREDICTED: Homo sapiens similar to PR-domain zinc finger protein 9 (LOC653738)
gi|88961756|ref|XM_929341.1|[88961756]

25: NM_052996 Reports
Homo sapiens PR domain containing 7 (PRDM7), mRNA
gi|41349477|ref|NM_052996.2|[41349477]

29: NM_024504 Reports
Homo sapiens PR domain containing 14 (PRDM14), mRNA
gi|41349468|ref|NM_024504.2|[41349468]

30: NM_021620 Reports
Homo sapiens PR domain containing 13 (PRDM13), mRNA
gi|41349467|ref|NM_021620.2|[41349467]

31: NM_020229 Reports
Homo sapiens PR domain containing 11 (PRDM11), mRNA
gi|41349465|ref|NM_020229.2|[41349465]

PRDM Reviewed RefSeq

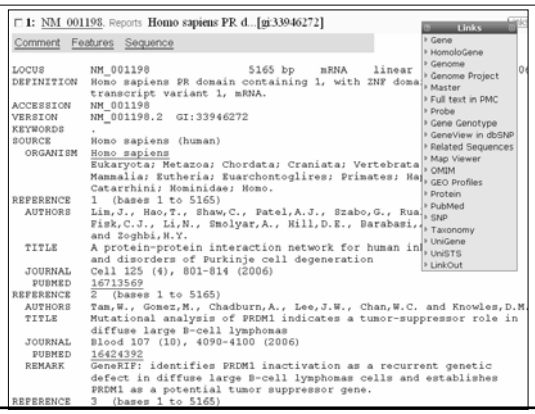
NCBI FieldGuide

1: NM_001198 Reports **Homo sapiens PR d...[gi:33946272]** Links

Comment | Features | Sequence

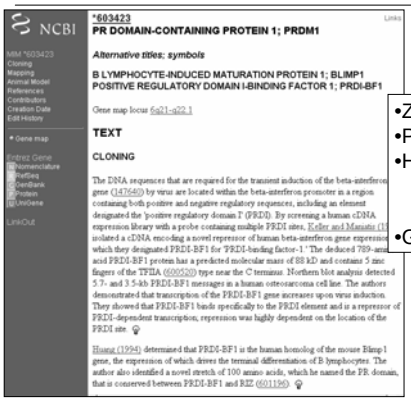
```
LOCUS       NM_001198               5165 bp    mRNA    linear   PRI 20-AUG-2006
DEFINITION  Homo sapiens PR domain containing 1, with ZNF domain (PRDM1),
            transcript variant 1, mRNA.
ACCESSION   NM_001198
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from AY198414.1.
            On Aug 20, 2003 this sequence version replaced gi:4557362.

            Summary: This gene encodes a protein that acts as a repressor of
            beta-interferon gene expression. The protein binds specifically to
            the PRDI (positive regulatory domain I element) of the beta-IFN
            gene promoter. Transcription of this gene increases upon virus
            induction. Two alternatively spliced transcript variants that
            encode different isoforms have been reported.

            Transcript Variant: This variant (1), also known as PRDM1alpha,
            encodes the longer isoform (1). The protein encoded by this
            transcript has also been named PRDI-BF1alpha.
            COMPLETENESS: complete on the 3' end.
  TITLE     Mutational analysis of PRDM1 indicates a tumor-suppressor role in
            diffuse large B-cell lymphomas
  JOURNAL   Blood 107 (10), 4090-4100 (2006)
   PUBMED   16424392
  REMARK    GeneRIF: identifies PRDM1 inactivation as a recurrent genetic
            defect in diffuse large B-cell lymphomas cells and establishes
            PRDM1 as a potential tumor suppressor gene.
REFERENCE   3  (bases 1 to 5165)
```

Links Menu



NCBI FieldGuide

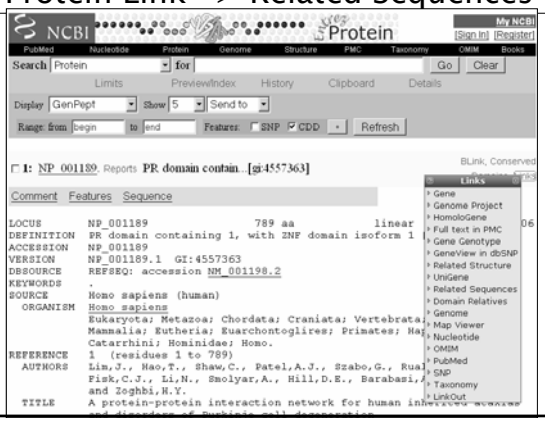
Online Mendelian Inheritance in Man



- Zn Finger domains
- PR Domain
- Homologs
  - BLIMP1 Mouse
  - U-boot Zebrafish
- Genomic context

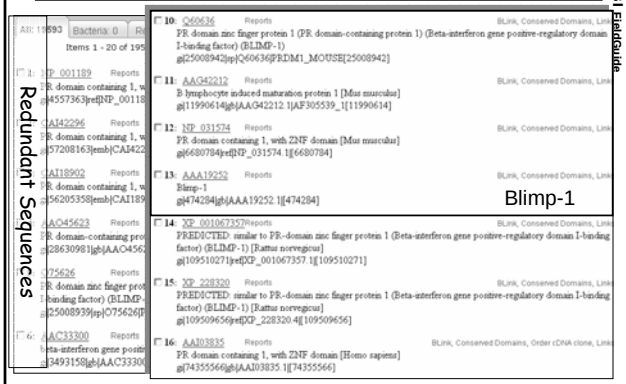
NCBI FieldGuide

Protein Link -> Related Sequences



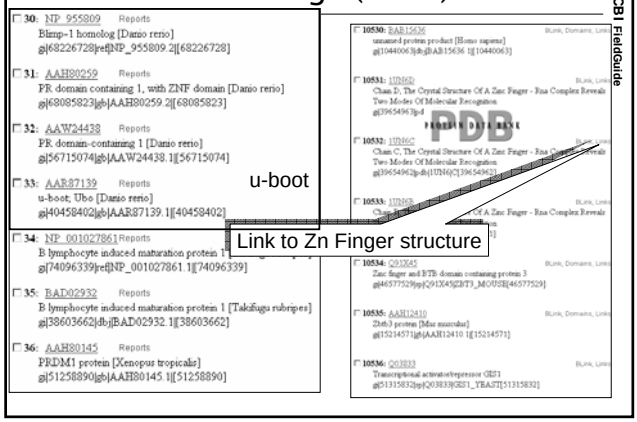
NCBI FieldGuide

Protein Neighbors: Redundancy and homologs



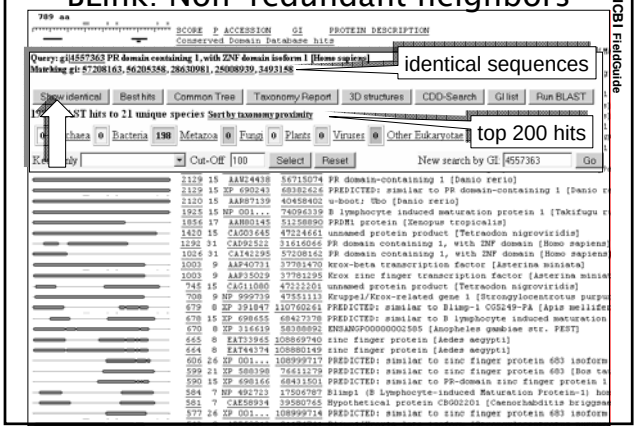
NCBI FieldGuide

Homologs (cont.)



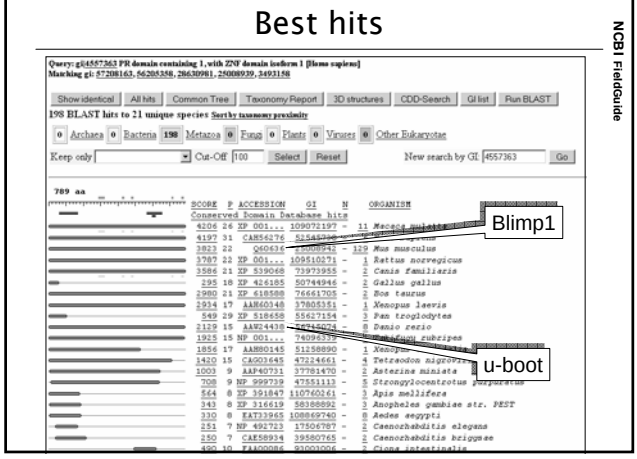
NCBI FieldGuide

BLink: Non-redundant neighbors



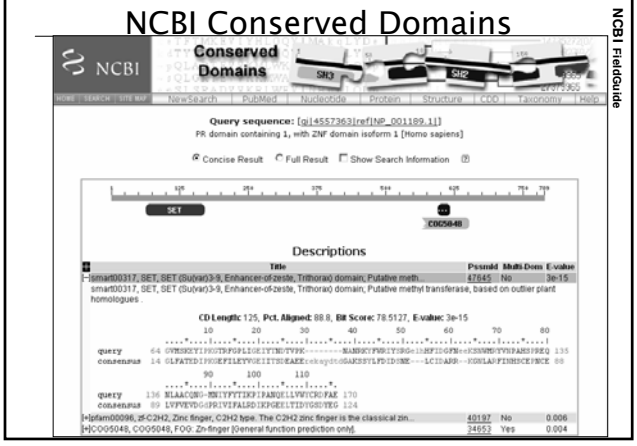
NCBI FieldGuide

Best hits



NCBI FieldGuide

NCBI Conserved Domains

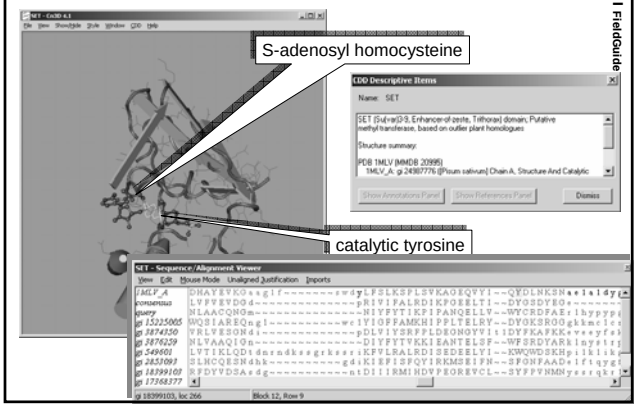


NCBI FieldGuide

The SET Domain

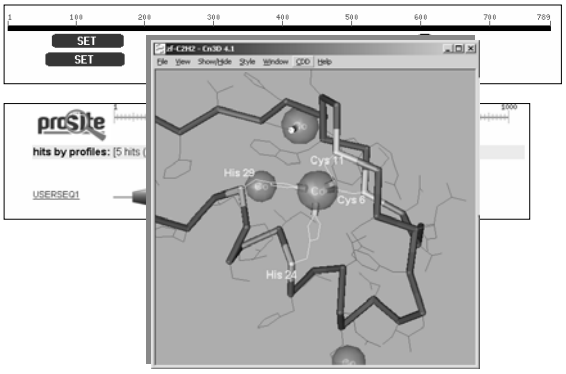
NCBI FieldGuide

SET Domain Structural Model



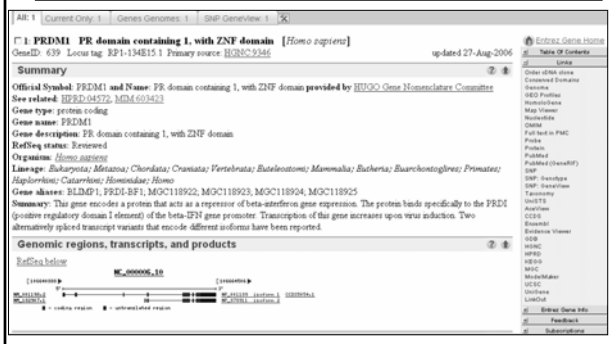
NCBI FieldGuide

Zn finger: a DNA binding motif



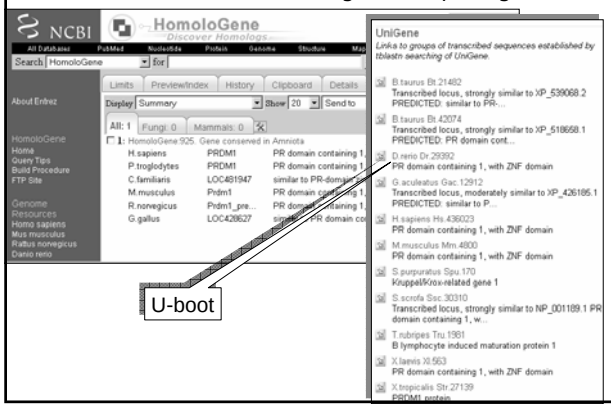
NCBI FieldGuide

Entrez Gene: One stop shopping



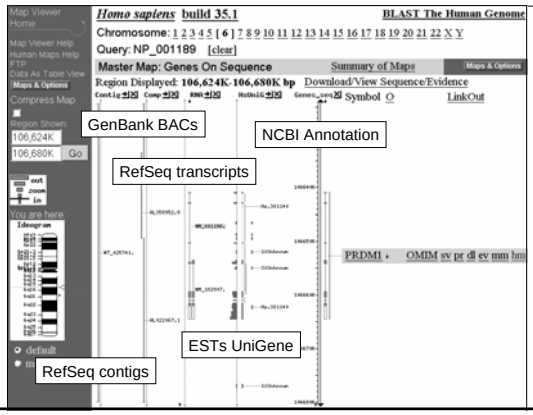
NCBI FieldGuide

HomoloGene: homologs in complete genomes



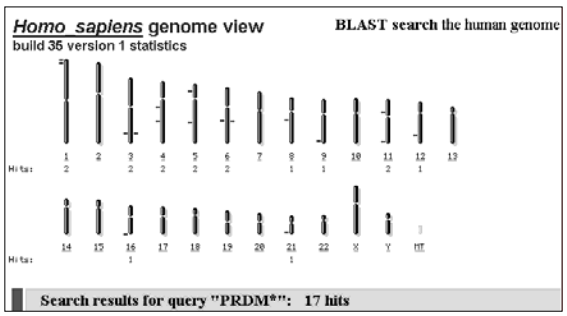
NCBI FieldGuide

Map Viewer: PRDM1



NCBI FieldGuide

PRDM1: Potential Paralogs



NCBI FieldGuide