NCBI FieldGuide

# NCBI Molecular Biology Resources

*September 2005*

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 149 | August 2005 |
|---|---|
| 46,947,388 | Records |
| 51,674,486,881 | Nucleotides (non-WGS) |
| 53,346,605,784 | Shotgun (WGS) |

GenBank now exceeds 100 GigaBases ! (with WGS)

| | |
|---|---|
| >140,000 | Species |
| *160 Gigabytes* | *658 files* |

- full release every two months
- incremental and cumulative updates daily
- available only through internet

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

- 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**
  - reviewed
  - human, mouse, rat, fruit fly, zebrafish, arabidopsis microbial genomes (proteins)
- **Model transcripts and proteins**
- **Assembled Genomic Regions (contigs)**
  - human genome
  - mouse genome
- **Chromosome records**
  - Human genome
  - 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

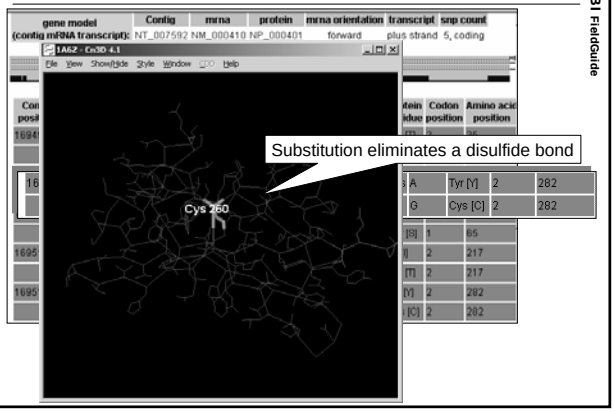
## NCBI's SNP Database

NCBI FieldGuide

- Primary Database and Derivative (RefSNP)
- Single Nucleotide Polymorphism
- Repeat polymorphisms
- Insertion-Deletion Polymorphisms
- 19 Species
- Over 11 million submissions

## RefSNP HFE

NCBI FieldGuide




## What is UniGene?

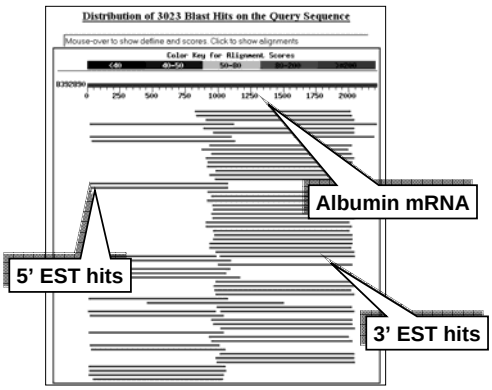
NCBI FieldGuide

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**

## EST hits: Human mRNA

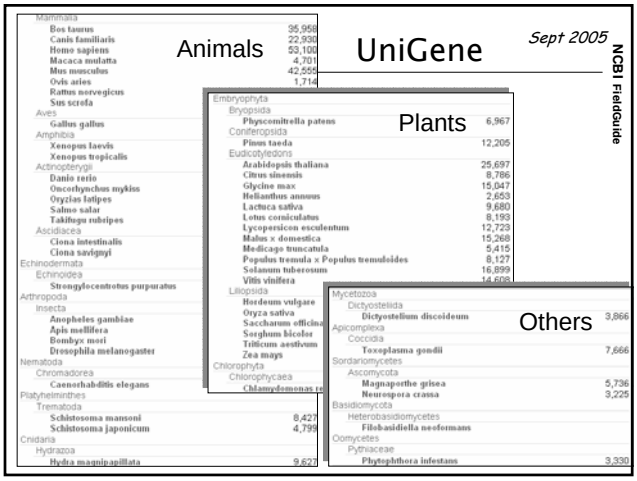
NCBI FieldGuide




## UniGene

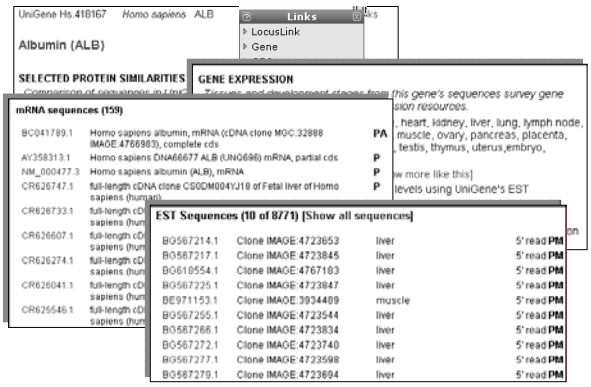
*Sept 2005*

NCBI FieldGuide



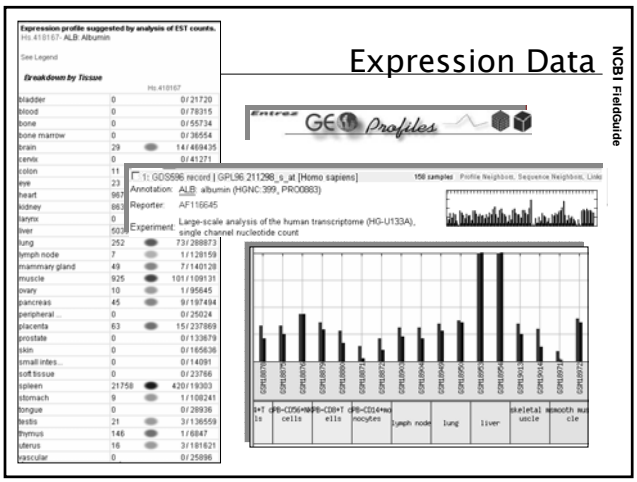
## UniGene: Expressed Sequences

NCBI FieldGuide




## Expression Data

NCBI FieldGuide



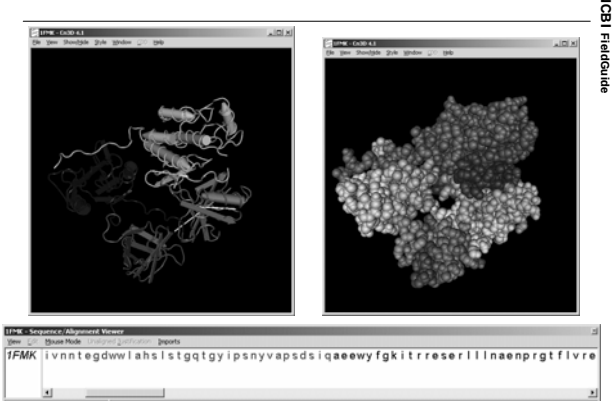

## MMDB: Molecular Modeling Data Base

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)**

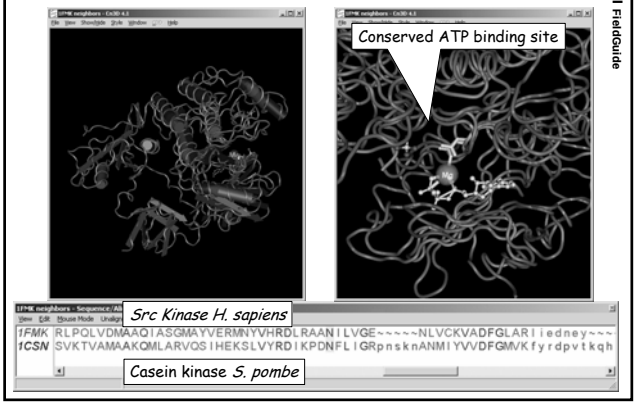
## Cn3D 4.1: C-SRC

NCBI FieldGuide



## Cn3D 4.1: Structural Alignment

NCBI FieldGuide




## Cn3D: Simple Homology Modeling

NCBI FieldGuide

## Protein Domains

NCBI FieldGuide

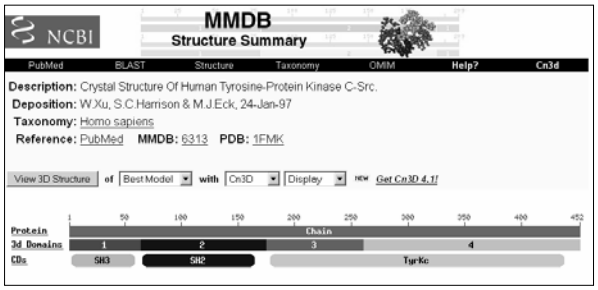
- Structural Domain
  - Discrete independently folding unit of a protein
- Conserved Domain (sequence-based)
  - Protein region with recognizable position specific pattern of sequence conservation
- Sequence-based domains often roughly correspond to structural domains
- Domains often have distinct, identifiable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

## NCBI's Conserved Domain Database

NCBI FieldGuide

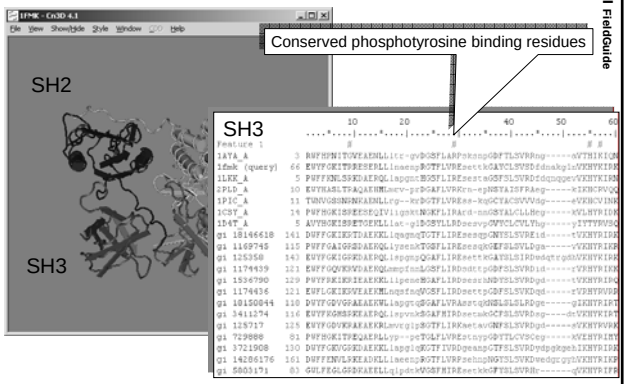
- Multiple sequence alignments
- PSI-BLAST –based score matrices
- Sources SMART, PFAM, COGs
- New NCBI curated domains
  - structure informed alignments
- Stats:
  - COGS 4,873
  - Pfam 5,193
  - Smart 653
  - NCBI CDD 316

## 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

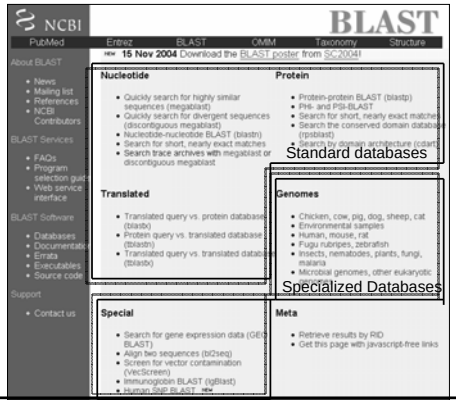
## The Flavors of BLAST

NCBI FieldGuide

- Standard BLAST
  - traditional contiguous word hit
  - position independent scoring
  - nucleotide, protein and translations
- Megablast
  - optimized for large batch searches
  - can use discontiguous words
- 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

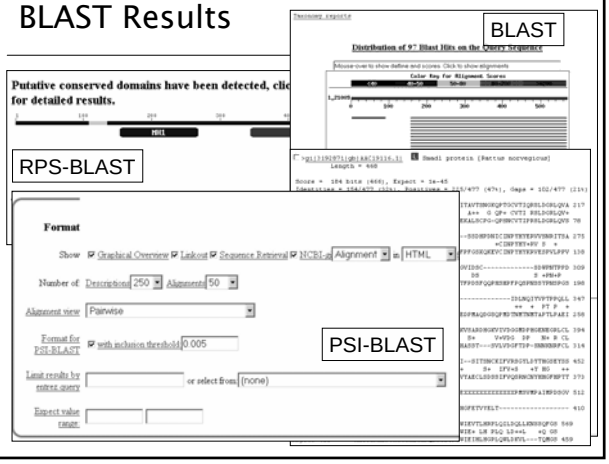
NCBI FieldGuide



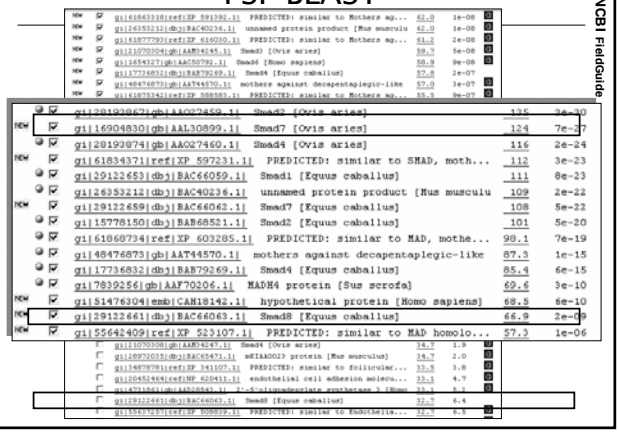

## BLAST Search

NCBI FieldGuide

## BLAST Results



## PSI-BLAST




NCBI FieldGuide

## VAST: Structure Neighbors

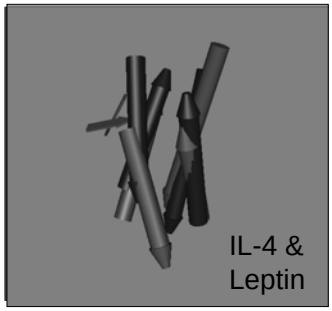
*Vector Alignment Search Tool*

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

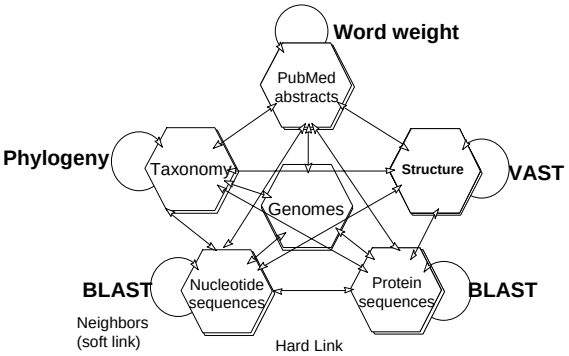
WWW Access

Entrez & BLAST

NCBI FieldGuide

## Entrez: Database Integration

NCBI FieldGuide

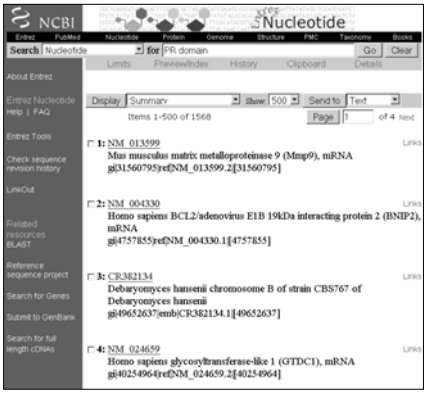



## PR Domain Global Search

NCBI FieldGuide

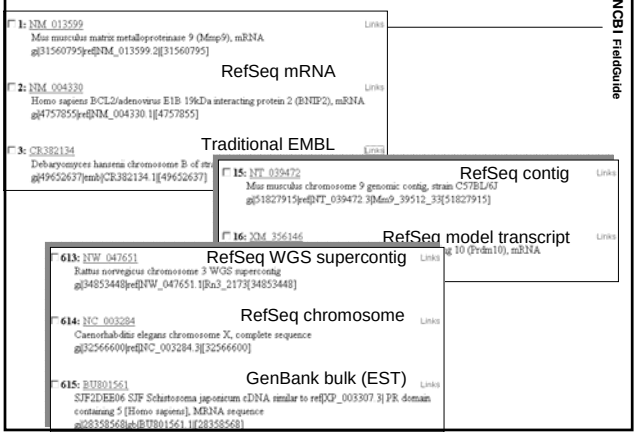
## PR Domain search results

NCBI FieldGuide



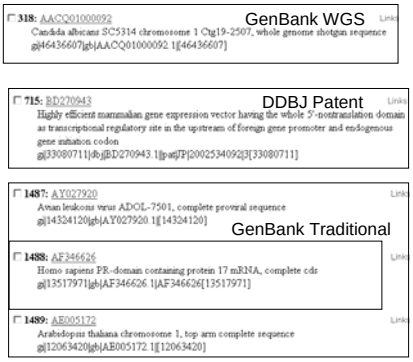

## PR Domain search results

NCBI FieldGuide



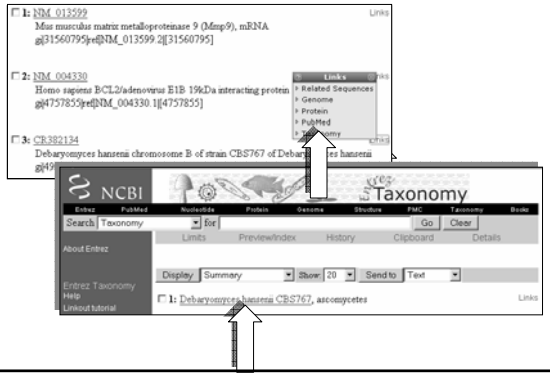
## PR Domain search results (continued)

NCBI FieldGuide



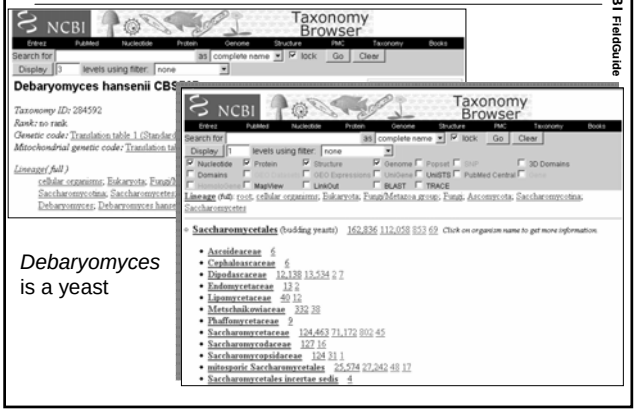

## What is *Debaryomyces*?

NCBI FieldGuide




## *Debaryomyces:* Taxonomy Browser

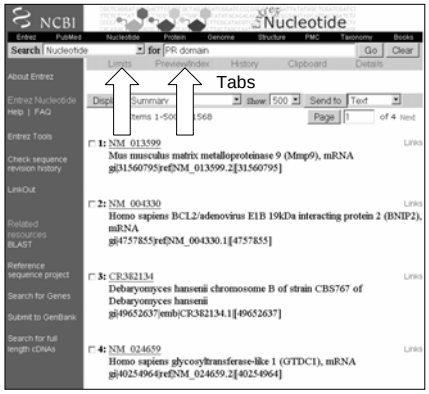
NCBI FieldGuide



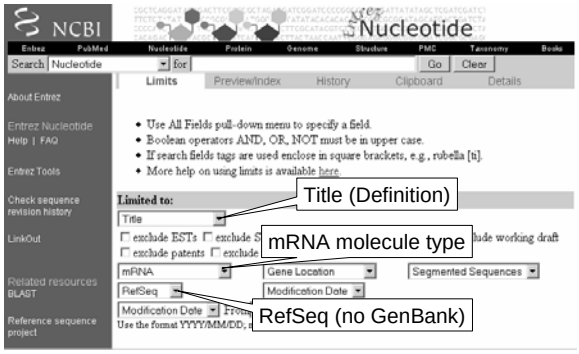

*Debaryomyces* is a yeast

## Refining Search

NCBI FieldGuide

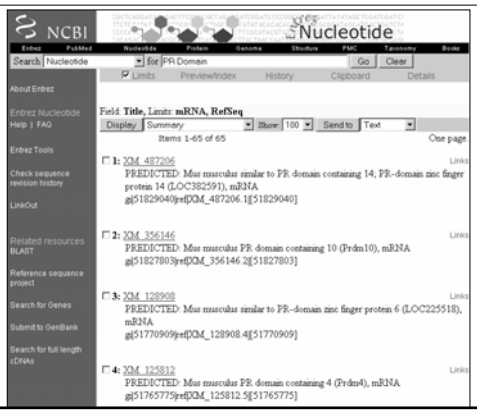


## Limits



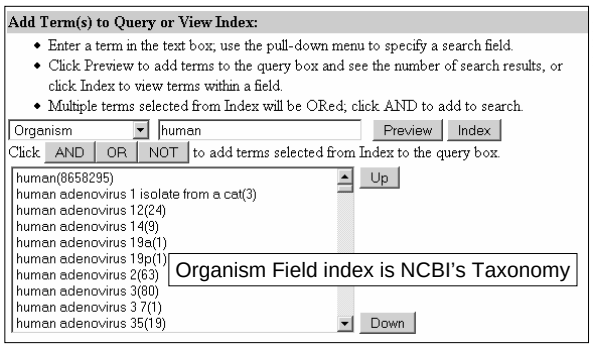

NCBI FieldGuide

## Limited Search Results



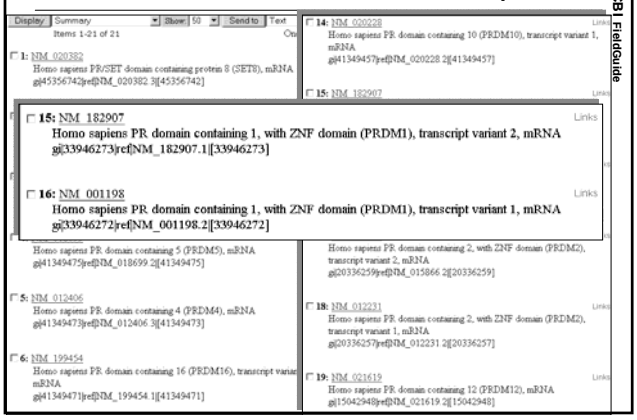

NCBI FieldGuide

## Preview / Index: Adding Terms




NCBI FieldGuide

## Human PR domain RefSeqs




NCBI FieldGuide

## PRDM Reviewed RefSeq

NCBI FieldGuide

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

LOCUS       NM_001198                5165 bp    mRNA    linear   PRI 23-AUG-2004
DEFINITION  Homo sapiens PR domain containing 1, with ZNF domain (PRDM1),
            transcript variant 1, mRNA.
ACCESSION   NM_001198
VERSION     NM_001198.2  GI:33946272
KEYWORDS
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.
              lines
  JOURNAL   J. Immunol. 170 (6), 3125-3133 (2003)
   PUBMED   12626569
   REMARK   GeneRIF: PRDI-BF1 beta is abundantly expressed in myeloma cell
            lines through alternative transcription initiation, has a disrupted
            PR domain and a significantly impaired transcription repressor
            function on multiple target genes.
REFERENCE   3  (bases 1 to 5165)
```

## Links Menu

NCBI FieldGuide

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

LOCUS       NM_001198                5165 bp    mRNA    linear
DEFINITION  Homo sapiens PR domain containing 1, with ZNF doma
            transcript variant 1, mRNA.
ACCESSION   NM_001198
VERSION     NM_001198.2  GI:33946272
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata
            Mammalia; Eutheria; Primates; Catarrhini; Hominida
REFERENCE   1  (bases 1 to 5165)
  AUTHORS   Gyory,I., Wu,J., Fejer,G., Seto,E. and Wright,K.L.
  TITLE     PRDI-BF1 recruits the histone H3 methyltransferase
            transcriptional silencing
  JOURNAL   Nat. Immunol. 5 (3), 299-308 (2004)
   PUBMED   14985713
   REMARK   GeneRIF: PRDI-BF1 mediates silencing of interferon-beta gene
            transcription via recruitment of the histone H3 methyltransferase
            G9a.
REFERENCE   2  (bases 1 to 5165)
  AUTHORS   Gyory,I., Fejer,G., Ghosh,N., Seto,E. and Wright,K.L.
  TITLE     Identification of a functionally impaired positive regulatory
            domain I binding factor 1 transcription repressor in myeloma cell
            lines
  JOURNAL   J. Immunol. 170 (6), 3125-3133 (2003)
   PUBMED   12626569
   REMARK   GeneRIF: PRDI-BF1 beta is abundantly expressed in myeloma cell
            lines through alternative transcription initiation, has a disrupted
            PR domain and a significantly impaired transcription repressor
            function on multiple target genes.
REFERENCE   3  (bases 1 to 5165)
```

Links
- Gene
- HomoloGene
- Master
- Related Sequences
- Map Viewer
- OMIM
- Protein
- PubMed
- SNP
- Taxonomy
- UniGene
- UniSTS
- LinkOut

## Online Mendelian Inheritance in Man

NCBI FieldGuide

*603423
**PR DOMAIN-CONTAINING PROTEIN 1; PRDM1**

*Alternative titles; symbols*

**B LYMPHOCYTE-INDUCED MATURATION PROTEIN 1; BLIMP1**
**POSITIVE REGULATORY DOMAIN I-BINDING FACTOR 1; PRDI-BF1**

Gene map locus 6q21-q22.1

**TEXT**

**CLONING**

The DNA sequences that are required for the transient induction of the beta-interferon gene (147640) by virus are located within the beta-interferon promoter in a region containing both positive and negative regulatory sequences, including an element designated the 'positive regulatory domain I' (PRDI). By screening a human cDNA expression library with a probe containing multiple PRDI sites, Keller and Maniatis (1 isolated a cDNA encoding a novel repressor of human beta-interferon gene expression which they designated PRDI-BF1 for 'PRDI-binding factor-1.' The deduced 789-amino acid PRDI-BF1 protein has a predicted molecular mass of 88 kD and contains 5 zinc fingers of the TFIIA (600520) type near the C terminus. Northern blot analysis detected 5.7- and 3.5-kb PRDI-BF1 messages in a human osteosarcoma cell line. The authors demonstrated that transcription of the PRDI-BF1 gene increases upon virus induction. They showed that PRDI-BF1 binds specifically to the PRDI element and is a repressor of PRDI-dependent transcription, repression was highly dependent on the location of the PRDI site.

Huang (1994) determined that PRDI-BF1 is the human homolog of the mouse Blimp1 gene, the expression of which drives the terminal differentiation of B lymphocytes. The author also identified a novel stretch of 100 amino acids, which he named the PR domain, that is conserved between PRDI-BF1 and RIZ (601156).

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

## Protein Link -> Related Sequences

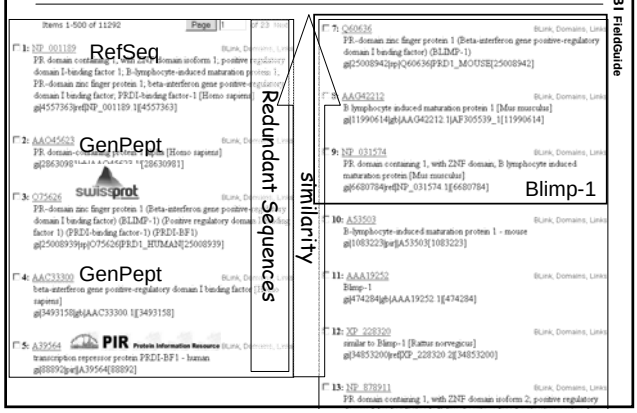
NCBI FieldGuide

```
1: NP_001189. PR domain contain...[gi:4557363]

LOCUS       NP_001189                789 aa            linear   PRI 23-
DEFINITION  PR domain containing 1, with ZNF domain isoform 1; positive
            regulatory domain I-binding factor 1; B-lymphocyte-induced
            maturation protein 1; PR-domain zinc finger protein 1;
            beta-interferon gene positive-regulatory domain I binding f
            PRDI-binding factor-1 [Homo sapiens].
ACCESSION   NP_001189
VERSION     NP_001189.1  GI:4557363
DBSOURCE    REFSEQ: accession NM_001198.2
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleo
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 789)
  AUTHORS   Gyory,I., Wu,J., Fejer,G., Seto,E. and Wright,K.L.
  TITLE     PRDI-BF1 recruits the histone H3 methyltransferase G9a in
            transcriptional silencing
  JOURNAL   Nat. Immunol. 5 (3), 299-308 (2004)
   PUBMED   14985713
   REMARK   GeneRIF: PRDI-BF1 mediates silencing of interferon-beta gene
```

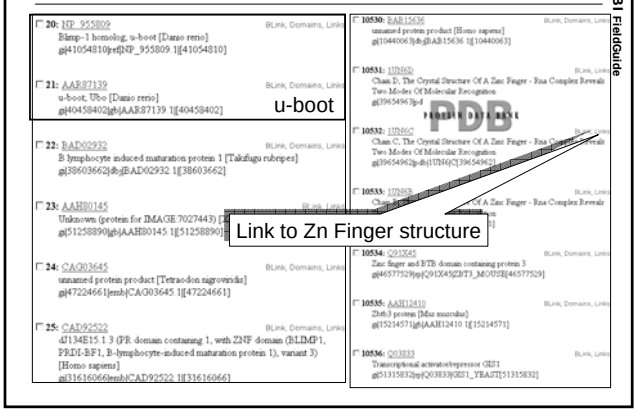
Links
- Gene
- HomoloGene
- UniGene
- Related Sequences
- Domain Relatives
- Map Viewer
- Nucleotide
- OMIM
- PubMed
- SNP
- Taxonomy
- LinkOut

## Protein Neighbors: Redundancy and homologs



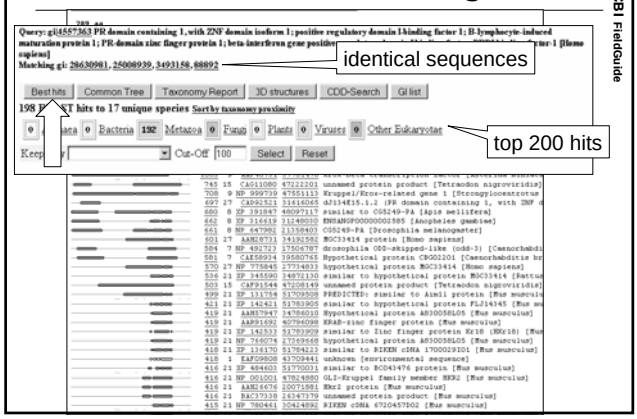

NCBI FieldGuide

## Homologs (cont.)



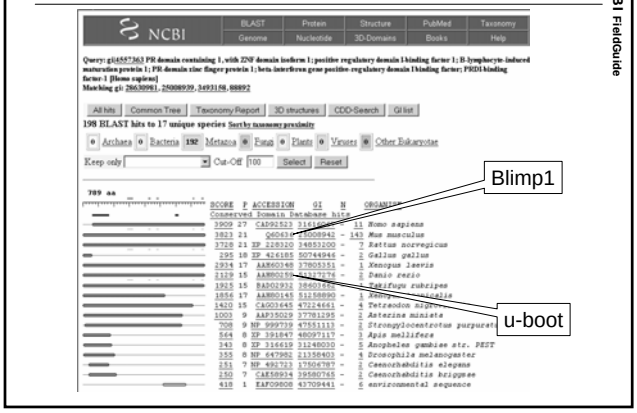

NCBI FieldGuide

## BLink: Non-redundant neighbors



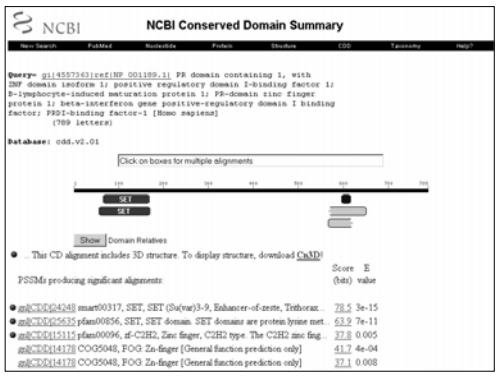

NCBI FieldGuide

## Best hits




NCBI FieldGuide

## NCBI Conserved Domains

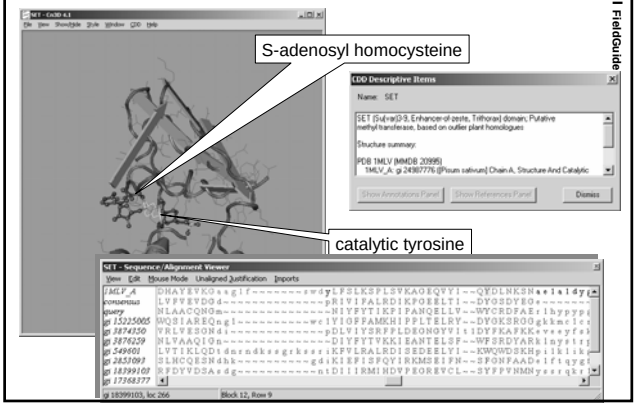



NCBI FieldGuide

## The SET Domain

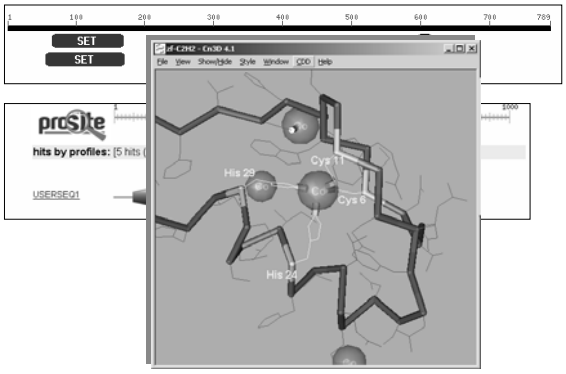
NCBI FieldGuide

## SET Domain Structural Model



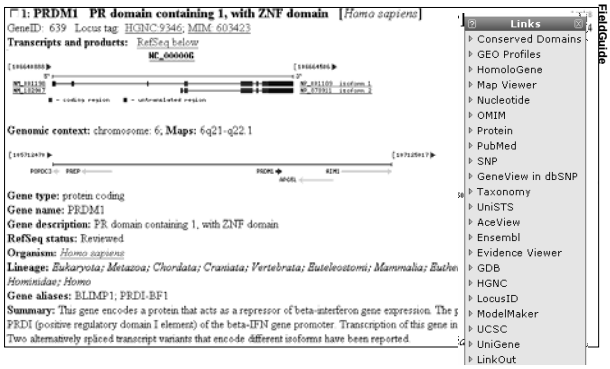

NCBI FieldGuide

## Zn finger: a DNA binding motif



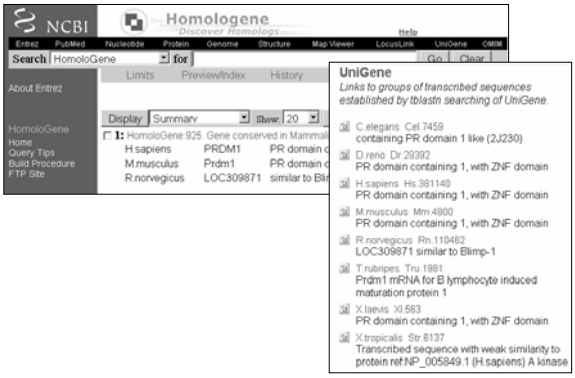

NCBI FieldGuide

## Entrez Gene: One stop shopping



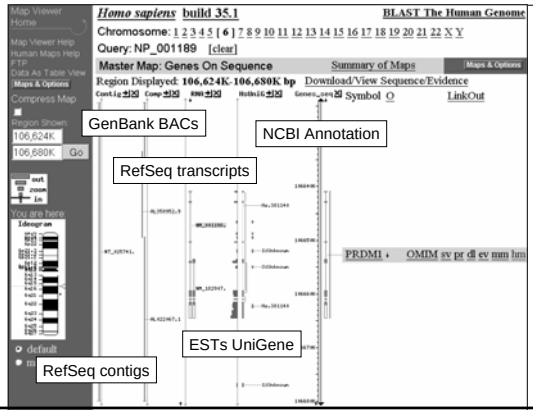

NCBI FieldGuide

## HomoloGene: homologs in complete genomes



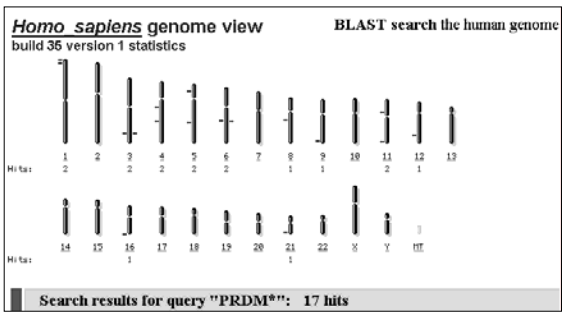

NCBI FieldGuide

## Map Viewer: PRDM1




NCBI FieldGuide

## PRDM1: Potential Paralogs




NCBI FieldGuide