



NCBI Molecular Biology Resources

NCBI Databases

March 2007

NCBI Field Guide

## 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 Field Guide

### Web Access: [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)



NCBI Field Guide

### NCBI Databases and Services

- GenBank largest sequence database
- Free public access to biomedical literature
  - PubMed free Medline
  - PubMed Central full text online access
- Entrez integrated molecular and literature databases
- BLAST highest volume sequence search service
- VAST structure similarity searches
- Software and Databases

NCBI Field Guide

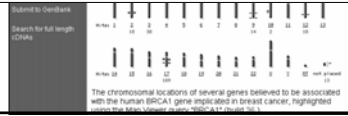
## Types of Databases

- 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

NCBI Field Guide

## Entrez Nucleotides

Primary	
• GenBank / EMBL / DDBJ	<b>86,766,287</b>
Derivative	
• RefSeq	<b>1,715,255</b>
• Third Party Annotation	<b>5,312</b>
• PDB	<b>7,334</b>
<b>Total</b>	<b>88,494,392</b>



NCBI Field Guide

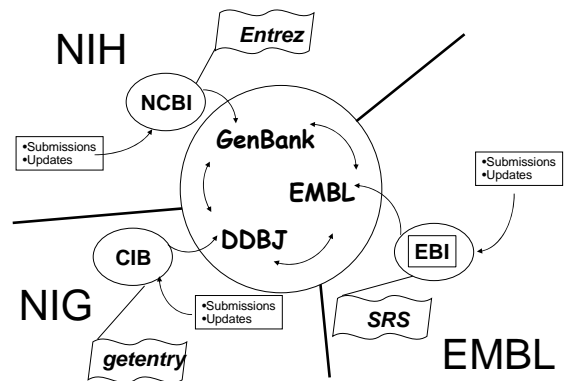
## What is GenBank?

NCBI's Primary Sequence Database

- Nucleotide only sequence database
- Archival in nature
  - Historical
  - Reflective of submitter point of view (subjective)
  - Redundant
- GenBank Data
  - Direct submissions (traditional records)
  - Batch submissions (EST, GSS, STS)
  - ftp accounts (genome data)
- Three collaborating databases
  - GenBank
  - DNA Database of Japan (DDBJ)
  - European Molecular Biology Laboratory (EMBL) Database

NCBI Field Guide

## International Sequence Database Collaboration



NCBI Field Guide

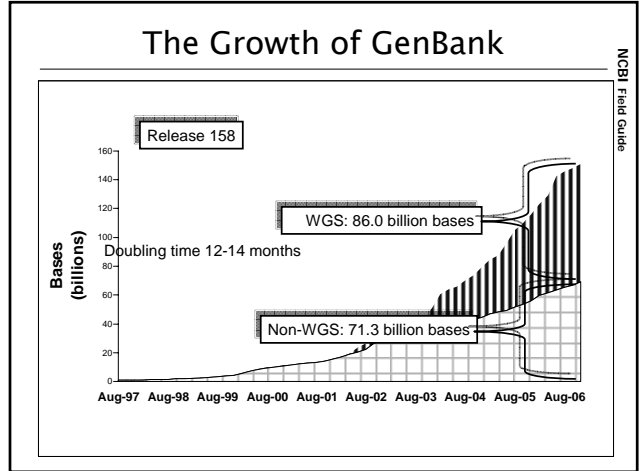
### GenBank: NCBI's Primary Sequence Database

<b>Release 158</b>	<b>February 2007</b>
<b>86,639,920</b>	<b>Records</b>
<b>157,335,689,977</b>	<b>Total Bases</b>
263 Gigabytes (non-WGS)	1115 files (non-WGS)

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

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

NCBI Field Guide



### Organization of GenBank: Traditional Divisions

Records are divided into 18 Divisions.

- 12 Traditional
- 6 Bulk

**Traditional Divisions:**

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

PRI Primate  
 PLN Plant and Fungal  
 BCT Bacterial and Archeal  
 INV Invertebrate  
 ROD Rodent  
 VRL Viral  
 VRT Other Vertebrate  
 MAM Mammalian  
 PHG Phage  
 SYN Synthetic (cloning vectors)  
 ENV Environmental Samples  
 UNA Unannotated

Entrez query: `gbdiv_XXX[Properties]`

NCBI Field Guide

### Organization of GenBank: Bulk Divisions

Records are divided into 18 Divisions.

- 12 Traditional
- 6 Bulk

**BULK Divisions:**

- Batch Submission (Email and FTP)
- Inaccurate
- Poorly characterized

EST Expressed Sequence Tag  
 GSS Genome Survey Sequence  
 HTG High Throughput Genomic  
 STS Sequence Tagged Site  
 HTC High Throughput cDNA  
 PAT Patent

Entrez query: `gbdiv_XXX[Properties]`

NCBI Field Guide



## ESTs in Entrez

All: 260886 bacteria: 0 mRNA: 260886

Items 1 - 20 of 260886 Page 1 of 13045 Next

Item	Accession	Organism	Sequence Type
1	CX724902	1332502 NOCCWA 09RT	Coccolithus mykiss cDNA clone 09RT08P21 3-, mRNA sequence
2	CX724901	1332500 NOCCWA 09RT	Coccolithus mykiss cDNA clone 09RT08P21 3-, mRNA sequence
3	CX724900	1332499 NOCCWA 09RT	Coccolithus mykiss cDNA clone 09RT08P21 3-, mRNA sequence
4	CX724799	1332498 NOCCWA 09RT	Coccolithus mykiss cDNA clone 09RT08P21 3-, mRNA sequence
5	CX724798	1332497 NOCCWA 09RT	Coccolithus mykiss cDNA clone 09RT08P21 3-, mRNA sequence
6	CX724797	1332496 NOCCWA 09RT	Coccolithus mykiss cDNA clone 09RT08P21 3-, mRNA sequence
7	CX724796	1332495 NOCCWA 09RT	Coccolithus mykiss cDNA clone 09RT08P21 3-, mRNA sequence

Category	Count
<b>Total</b>	<b>41 million records</b>
Human	7.9 million
Mouse	4.7 million
Cow	1.3 million
Rice	1.2 million
Zebrafish	1.2 million
Maize	1.2 million
Xenopus tropicalis	1.0 million
Rat	0.9 million
Wheat	0.9 million
Chicken	0.6 million
Barley	0.4 million

NCBI Field Guide

## Derivative Databases

NCBI Field Guide

## Entrez Protein: Derivative Database

Data Source	Sequences
GenPept	6,937,176
RefSeq	3,359,561
Third Party Annotation	5,136
Swiss Prot	255,159
PIR	29,996
PRF	12,079
PDB	91,116
PAT Division	669,035
<b>Total</b>	<b>10,690,223</b>
<b>BLAST nr total (no patents or env)</b>	<b>4,545,310</b>

NCBI Field Guide

## GenPept: GenBank CDS translations

```

FEATURES             Location/Qualifiers
     source            1..2484
                        /organism="Homo sapiens"
                        /mol_type="mRNA"
                        /db_xref="taxon:9606"
                        /chromosome="3"
                        /map="3p22-p23"
     gene              1..2484
     CDS                22..2292
                        /gene="DGL463989|gb|AAC50285.1| DNA mismatch repair prote...
                        /note="homolog of S. cerevisiae PMS1 (Swiss-Prot Accession
                        Number P14242), E. coli MUTL (GenBank Accession
                        Number U07187), S. typhimurium MUTL (Swiss-Prot Accession
                        Number P23367), Salmonella typhimurium MUTL (Swiss-Prot Accession
                        Number P14161), Streptococcus pneumoniae (Swiss-Prot
                        Accession Number P14160)"
                        /codon_start=1
                        /product="DNA mismatch repair protein homolog"
                        /protein_id="AAC50285.1"
                        /db_xref="GI:463989"
                        /translation="MSFVAGVIRRLDETIVNRIAAGEVIQRPANAIKEMINCLDAKS
                        TSIQIVKEGGLKLIQIQDNGTGIRKEDLDIVCFRTTSLKQSFEDLASISTYGRGFE
                        ALASISHVAHVITTTADGKCAVRSYSDGKLGKAPKPCAGNQTITVEDLFYNI
                        TRRKALKNPSEYGGKILEVGVRSVHNAGISFVKKQGETVADVRTLPNASTVDNIRS
    
```

NCBI Field Guide

## Redundant Proteins

```
>gi|463989|gb|AAC50285.1| DNA mismatch repair prote...
MSFVAGVIRRLDET VVNR IAAGEVIQR PANAIKEMIENCLDAKSTSIQIV...
EDLDIVCERFTTSK LQSFEDLAS ISTYGRGEALAS ISHVAHVTTTKTAD...

>gi|13905126|gb|AAH06850.1| MutL protein homolog 1 ...
MSFVAGVIRRLDET VVNR IAAGEVIQR PANAIKEMIENCLDAKSTSIQIV...
EDLDIVCERFTTSK LQSFEDLAS ISTYGRGEALAS ISHVAHVTTTKTAD...

>gi|1079787|gb|AAA82079.1| DNA mismatch repair prote...
MSFVAGVIRRLDET VVNR IAAGEVIQR PANAIKEMIENCLDAKSTSIQIV...
EDLDIVCERFTTSK LQSFEDLAS ISTYGRGEALAS ISHVAHVTTTKTAD...

>gi|4557757|ref|NP_000240.1| MutL protein homolog 1...
MSFVAGVIRRLDET VVNR IAAGEVIQR PANAIKEMIENCLDAKSTSIQIV...
EDLDIVCERFTTSK LQSFEDLAS ISTYGRGEALAS ISHVAHVTTTKTAD...

>gi|730028|sp|P40692|MLH1_HUMAN DNA mismatch repair...
MSFVAGVIRRLDET VVNR IAAGEVIQR PANAIKEMIENCLDAKSTSIQIV...
EDLDIVCERFTTSK LQSFEDLAS ISTYGRGEALAS ISHVAHVTTTKTAD...

>gi|741682|prf|2007430A DNA mismatch repair protei...
MSFVAGVIRRLDET VVNR IAAGEVIQR PANAIKEMIENCLDAKSTSIQIV...
EDLDIVCERFTTSK LQSFEDLAS ISTYGRGEALAS ISHVAHVTTTKTAD...
```

NCBI Field Guide

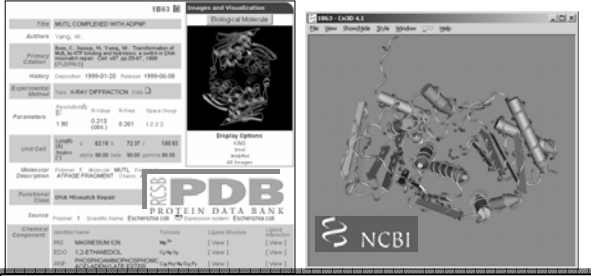
} GenPept

} NCBI RefSeq

} Swiss-Prot

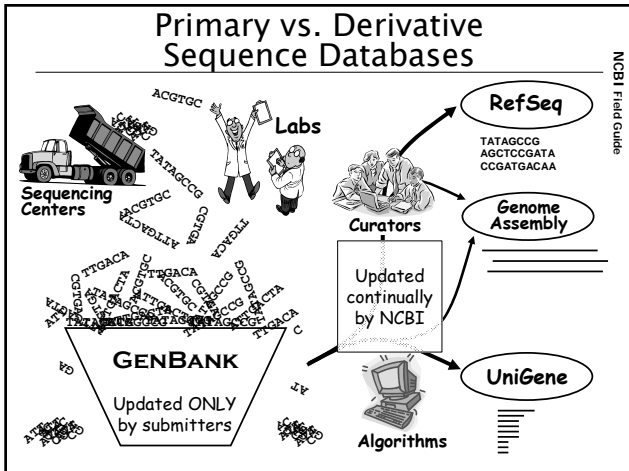
} PRF

## Protein Sequences from Structures



```
>gi|5542073|pdb|1B63|A Chain A, MutL Complexed With Adpnp
SHMP I QV L P P Q L A N Q I A A G E V V E R P A S V V K E L V E N S L D A G A T R I D I D I E R G G A K L I R I R D N G C G I K K D E L
A L A L A R H A T S K I A S L D D L E A I S L G F R G E A L A S I S S V S R L T L T S R T A E Q Q E A W Q A Y A E G R D M N V T K F P A A
H P V G T L E V L D L F Y N T P A R R K F L R T E K T E F N H I D E I I R R I A L A R F D V T I N L S H N G K I V R Y R A V P E G G Q K
E R R L G A I C G T A F L E Q A L A E W Q H G D L T L R G W A D P N H T P A L A E I Q Y C V N G R M M R D R L I N H A I R Q A C E D
K L G A D Q Q P A F V L Y L E I D P H Q V D V N V H P A K H E V R F H Q S R L V H D F I Y Q G V L S V L Q
```

NCBI Field Guide



## RefSeq: NCBI's Derivative Sequence Database

- **Curated transcripts and proteins**
  - reviewed
  - human, mouse, rat, fruit fly, zebrafish, arabidopsis
  - microbial genomes (proteins), and more
- **Model transcripts and proteins**
- **Assembled Genomic Regions (contigs)**
  - human genome
  - mouse genome
  - rat genome
  - chicken
  - honeybee
  - sea urchin
- **Chromosome records**
  - Human genome
  - microbial
  - organelle

srcdb\_refseq[Properties]

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

## Selected RefSeq Accession Numbers

NCBI Field Guide

### mRNAs and Proteins

NM_123456	Curated mRNA
NP_123456	Curated Protein
NR_123456	Curated non-coding RNA
XM_123456	Predicted mRNA
XP_123456	Predicted Protein
XR_123456	Predicted non-coding RNA
NG_123456	Reference Genomic Sequence
NC_123455	Microbial replicons, organelle
NT_123456	Contig
NW_123456	WGS Supercontig

### Gene Records

### Chromosome

### Assemblies

### Assemblies

### Assemblies

### Assemblies

## GenBank to RefSeq

NCBI Field Guide

```

1. A123456 RefSeq
   Homo sapiens hypothetical ribonucleoprotein A123456 mRNA, partial cds
   gb|AF047373.1|gb|AF047373.1|

2. A123456 RefSeq
   Homo sapiens CDNA, noncoding, partial cds
   gb|AF047373.1|gb|AF047373.1|

3. A123456 RefSeq
   Homo sapiens CDNA, noncoding, partial cds
   gb|AF047373.1|gb|AF047373.1|

4. A123456 RefSeq
   Homo sapiens CDNA, noncoding, partial cds
   gb|AF047373.1|gb|AF047373.1|

5. A123456 RefSeq
   Homo sapiens CDNA, noncoding, partial cds
   gb|AF047373.1|gb|AF047373.1|
    
```

□ 1: [NM\\_000249](#) Reports  
Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), mRNA  
[gi|28559089|ref|NM\\_000249.2||28559089](#)

```

6. A123456 RefSeq
   Homo sapiens CDNA, noncoding, partial cds
   gb|AF047373.1|gb|AF047373.1|

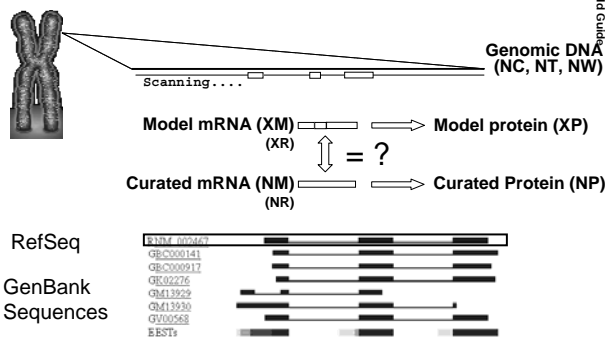
7. A123456 RefSeq
   Homo sapiens CDNA, noncoding, partial cds
   gb|AF047373.1|gb|AF047373.1|

8. A123456 RefSeq
   Homo sapiens CDNA, noncoding, partial cds
   gb|AF047373.1|gb|AF047373.1|

9. A123456 RefSeq
   Homo sapiens CDNA, noncoding, partial cds
   gb|AF047373.1|gb|AF047373.1|
    
```

## RefSeqs: Annotation Reagents

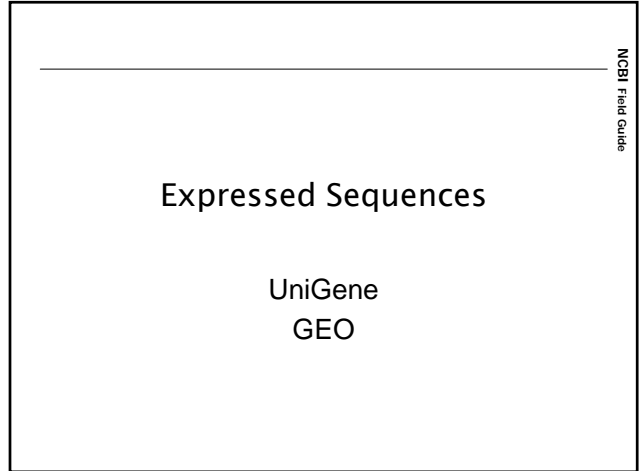
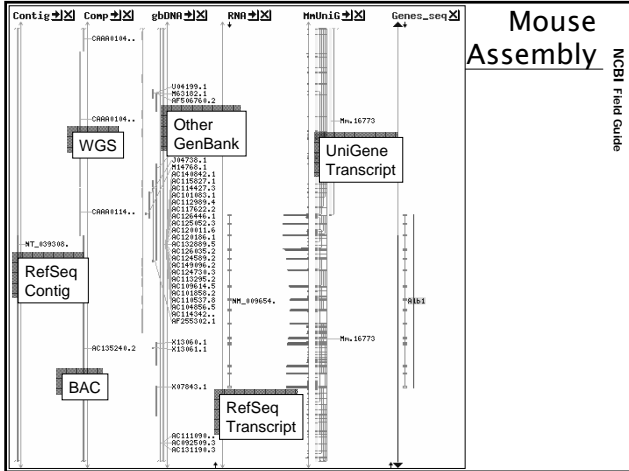
NCBI Field Guide



## RefSeq Benefits

NCBI Field Guide

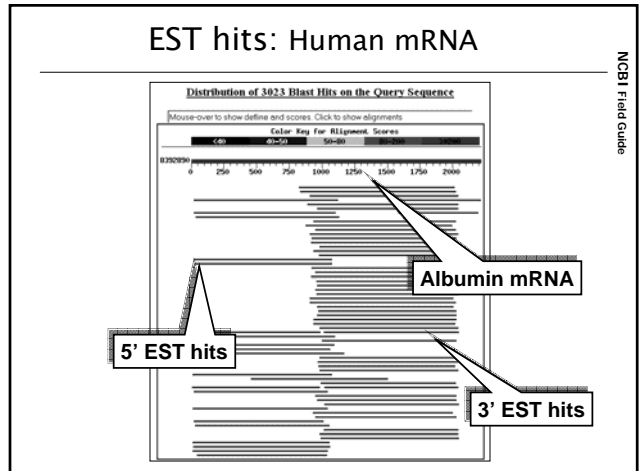
- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current sequence data and biology
- data validation
- format consistency
- distinct accession series
- stewardship by NCBI staff and collaborators



### 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 uncharacterized ESTs
- Useful for gene discovery and selection of mapping reagents



## UniGene

**Chordates**

- Bos taurus (cattle) 42,893
- Canis familiaris (dog) 27,649
- Homo sapiens (human) 86,310
- Macaca fascicularis (crah-eating macaque) 8,154
- Macaca mulatta (rhesus monkey) 4,826

**Invertebrates**

- Strongylocentrotus purpuratus (purple sea urchin) 15,291
- Aedes aegypti (yellow fever mosquito) 16,081
- Anopheles gambiae (African malaria mosquito) 15,291
- Apis mellifera (honey bee) 15,291
- Bombus morio (domestic bumblebee) 15,291
- Drosophila melanogaster (fruit fly) 15,291
- Tribolium castaneum (red flour beetle) 15,291

**Fungi et al.**

- Neurospora crassa 3,224
- Aspergillus nidulans 3,224
- Aspergillus fumigatus 3,224
- Aspergillus niger 3,224
- Aspergillus oryzae 3,224
- Aspergillus terreus 3,224
- Aspergillus versicolor 3,224
- Aspergillus nidulans 3,224
- Aspergillus fumigatus 3,224
- Aspergillus niger 3,224
- Aspergillus oryzae 3,224
- Aspergillus terreus 3,224
- Aspergillus versicolor 3,224

**Plants**

- Physcomitrella patens 19,691
- Picea glauca (white spruce) 6,867
- Picea sitchensis (Sitka spruce) 6,967
- Picea taeda (loblolly pine) 14,953
- Arabidopsis thaliana (thale cress) 28,282
- Arabidopsis lyrata (thale cress) 28,282
- Brassica napus (rape) 8,983
- Brassica oleracea (broccoli) 8,983
- Brassica rapa (turnip) 8,983
- Brassica campestris (turnip) 8,983
- Brassica napus (rape) 8,983
- Brassica oleracea (broccoli) 8,983
- Brassica rapa (turnip) 8,983
- Brassica campestris (turnip) 8,983

NCBI Field Guide

## Xenopus laevis MLH1 Cluster

**SELECTED PROTEIN SIMILARITIES**

Comparison of sequences in UniGene with proteins supported by a complete genome. The alignment can suggest function of a gene.


A. thaliana	AT5G15120 - T51620 DNA mismatch repair protein MLH1A [Imported] - Arabidopsis thaliana	34.62 % / 202 aa (see ProteoEST)
C. elegans	CE039736.1 - CNA mismatch repair protein [Caenorhabditis elegans]	28.85 % / 196 aa (see ProteoEST)
D. melanogaster	dmel_472022.1 - Mh1-LP1 [Drosophila melanogaster]	41.79 % / 201 aa (see ProteoEST)
H. sapiens	NP_002245.1 - mutL homolog 1, mutL	78.11 % / 199 aa (see ProteoEST)
M. musculus	U02820.1 - MLH1, MISMATCH REPAIR PROTEIN Mh1	77.61 % / 199 aa (see ProteoEST)
R. norvegicus	NP_122315.1 - mismatch repair protein [Rattus norvegicus]	75.62 % / 199 aa (see ProteoEST)
S. cerevisiae	YJ052525 - YJ042525 mismatch repair protein MLH1 - yeast	34.93 % / 197 aa (see ProteoEST)

**SEQUENCES**

Sequences representing this gene, mRNAs, ESTs, and gene predictions supported by a complete genome.

**EST Sequences (15)**

- BE691219.1 Clone IMAGE:3748204 whole 5' read P
- AW542932.1 Clone PEX0123C03 whole 5' read P
- AW543236.1 Clone PEX0126E06 whole 5' read P
- AW546546.1 Clone PEX0165B01 whole 5' read P
- BM190937.1 Clone IMAGE:5078219 body 3' read PA
- BQ385912.1 Clone IMAGE:5073362 ovary 3' read PA
- BQ385913.1 Clone IMAGE:5073362 ovary 5' read P
- CD307263.1 Clone IMAGE:6954962 whole 5' read P
- B615909.1 Clone XL174022 body 3' read P
- B632520.1 Clone XL174022 whole 3' read P
- BX551911.1 Clone IMAGE:3748204 body 5' read PA
- DY553264.1 Clone IMAGE:8329313 ovary 3' read A
- DY556751.1 Clone IMAGE:8329180 ovary 5' read A
- DY558112.1 Clone IMAGE:8329313 ovary 5' read
- DY556724.1 Clone IMAGE:8329180 ovary 5' read



NCBI Field Guide

## Human ALB Cluster

**SELECTED PROTEIN SIMILARITIES**

Comparison of sequences in UniGene with proteins supported by a complete genome.

BQ041789.1	Homo sapiens albumin, mRNA (cDNA clone MGC:32888, IMAGE:4768983), complete cds	PA
AY358313.1	Homo sapiens DNA66677 ALB (UNQ696) mRNA, partial cds	P
NM_000477.3	Homo sapiens albumin (ALB), mRNA	P

**GENE EXPRESSION**

Expression and developmental data from this gene's sequences survey gene (show resources).

**mRNA sequences (159)**

- BC041789.1 Homo sapiens albumin, mRNA (cDNA clone MGC:32888, IMAGE:4768983), complete cds PA
- AY358313.1 Homo sapiens DNA66677 ALB (UNQ696) mRNA, partial cds P
- NM\_000477.3 Homo sapiens albumin (ALB), mRNA P

**EST sequences (10 of 15944) [Show all sequences]**

CR626747.1	R16512.1	Clone IMAGE:126673	mixed	5' read P
CR628733.1	R16809.1	Clone IMAGE:126673	mixed	3' read P
CR626607.1	AA988197.1	Clone IMAGE:1436838	liver	3' read PA
CR626274.1	A032512.1	Clone IMAGE:1654762	mixed	3' read P
CR626041.1	A097221.1	Clone IMAGE:1707169	heart	3' read PA
CR625446.1	A3849088.1	Clone IMAGE:1508806	mixed	3' read PA
	AI127447.1	Clone IMAGE:1705923	heart	3' read PA
	AI139674.1	Clone IMAGE:1710130	heart	3' read PA
	AI140200.1	Clone IMAGE:1739542	lung	3' read PA
	AI140227.1	Clone IMAGE:1739555	lung	3' read PA

NCBI Field Guide

## Expression Data

**Expression profile, suggested by analysis of EST counts.**

Hs.418167: ALB Albumin

See Legend

Breakdown by Tissue	Hs.418167	Hs.418167
bladder	0	0/21720
blood	0	0/78315
bone	0	0/55724
bone marrow	0	0/38554
brain	29	14/488435
breast	0	0/41271
colon	11	2/23
eye	23	0/117404
heart	967	0/117404
kidney	863	0/117404
liver	0	0/117404
liver	523	0/117404
lung	252	73/28873
lymph node	7	1/128159
mammary gland	49	2/40028
muscle	925	101/109121
ovary	10	1/95645
pancreas	45	0/117404
peripheral blood leukocytes	0	0/29024
placenta	63	15/237869
prostate	0	0/133879
skin	0	0/166636
small intestine	0	0/140911
soft tissue	0	0/23766
spleen	21758	420/19303
stomach	8	1/116241
testis	0	0/23926
testis	21	3/136559
thymus	146	1/6847
uterus	16	3/116124
vascular	0	0/25886

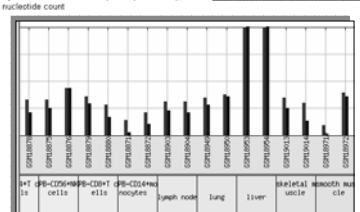
**GE Profiles**

Annotation: ALB albumin (HGNC:399, PRO0063)

Reporter: A1155245

Experiment: Large-scale analysis of the human transcriptome (HG-U133A).

150 samples | Profile Neighbors, Sequence Neighbors, Links



NCBI Field Guide

## Other NCBI Databases

- **Structure:** imported structures (PDB)  
Cn3D viewer, NCBI curation
- **CDD:** conserved domain database  
Protein families (COGs and KOGs)  
Single domains (PFAM, SMART, CD)
- **dbSNP:** nucleotide polymorphism
- **Gene:** gene records  
Unifies LocusLink and Microbial Genomes

NCBI Field Guide

PROTEIN DATA BANK →

Structure



## NCBI Structures and Domains



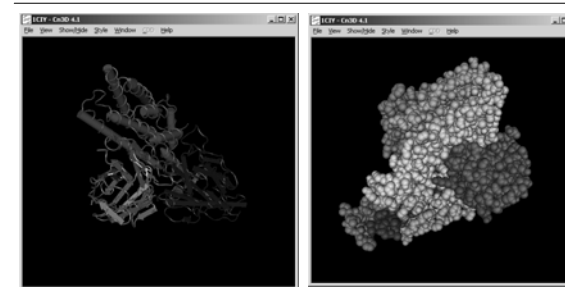
NCBI Field Guide

## MMDB: Molecular Modeling Data Base

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

NCBI Field Guide

## Cn3D 4.1: *Bacillus thuringiensis* Toxin

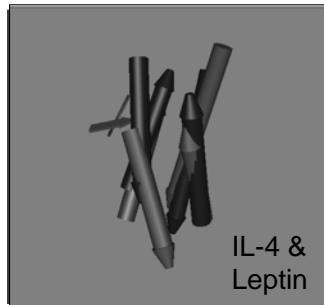


NCBI Field Guide

## 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 Field Guide

## Protein Domains

- 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

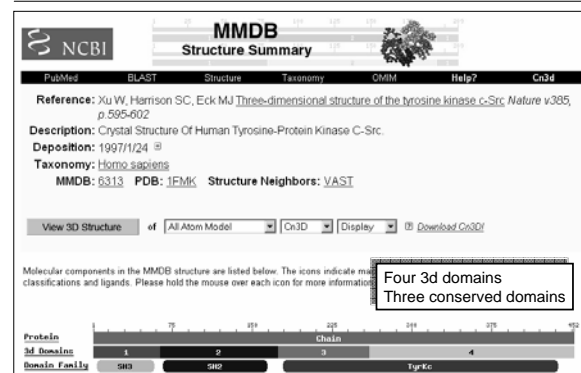
NCBI Field Guide

## NCBI's Conserved Domain Database

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

NCBI Field Guide

## Src Domains



NCBI MMDb Structure Summary

PubMed BLAST Structure Taxonomy OMM Help? Cn3D

**Reference:** Xu W, Harrison SC, Eck MJ Three-dimensional structure of the tyrosine kinase c-Src. *Nature* v.385, p.595-602

**Description:** Crystal Structure Of Human Tyrosine-Protein Kinase C-Src.

**Deposition:** 1997/1/24

**Taxonomy:** [Homo sapiens](#)

**MMDb:** 6313 **PDB:** 1FMK **Structure Neighbors:** VAST

View 3D Structure of All Atom Model Cn3D Display Download Co3D

Molecular components in the MMDb structure are listed below. The icons indicate molecular classifications and ligands. Please hold the mouse over each icon for more information.

Four 3d domains  
Three conserved domains

Protein Chain 1 2 3 4

3d Domains 1 2 3 4

Domain Family SH3 SH2 TyKc

NCBI Field Guide

