NCBI National Center for Biotechnology Information

NCBI FieldGuide

A Field Guide to GenBank

and NCBI's Molecular Biology Resources

April 11, 2007 *Rutgers*

ftp://ftp.ncbi.nih.gov/pub/FieldGuide/Slides/Archive/2007/Rutgers.04.11.07

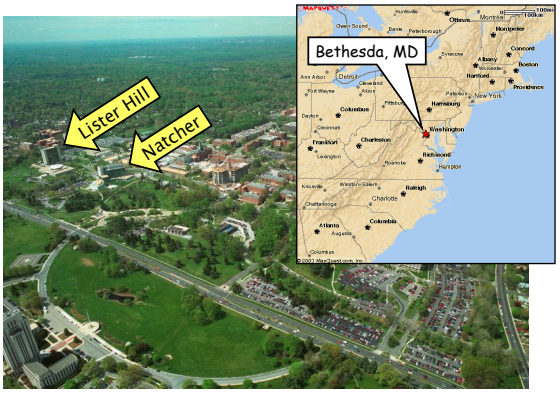
Topics

- About NCBI
- GenBank overview
- Primary vs derivative databases
  - The Reference Sequence (RefSeq) project
- The Entrez engine and databases

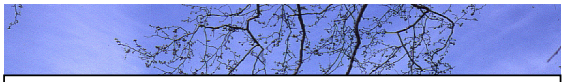
-break-

- Entrez text searching
- Genomic resources
- Sequence similarity – BLAST
- An integrated example

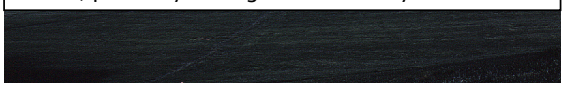
The National Institutes of Health



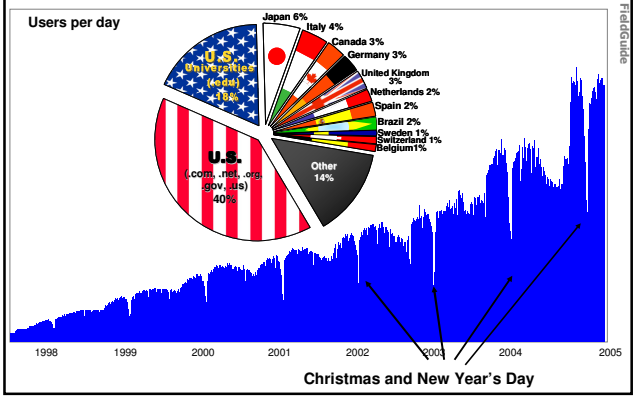
The National Center for Biotechnology Information



- Accepts submissions of primary data
- Develops tools to analyze these data
- Creates derivative databases based on the primary data
- Provides free search, link, and retrieval of these data, primarily through the Entrez system

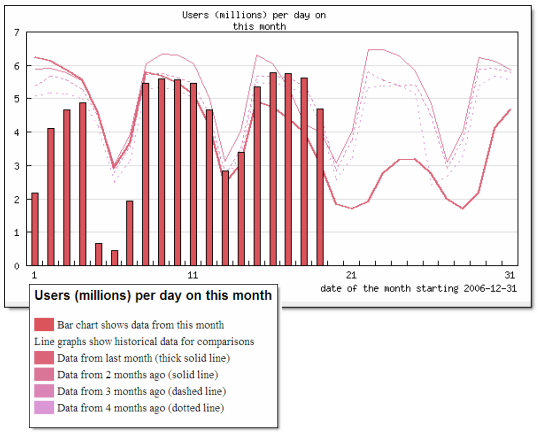


NCBI Web Traffic



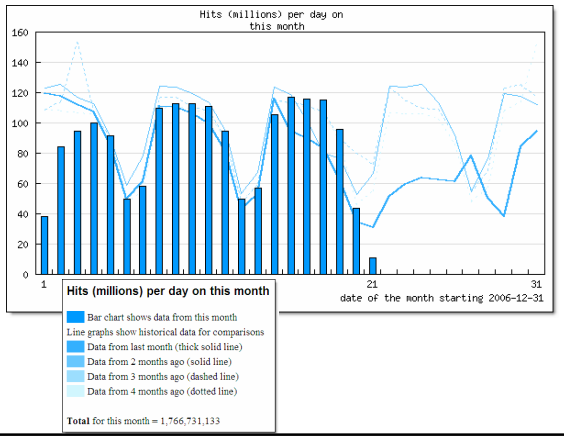
NCBI FieldGuide

NCBI Web Usage – Users per Day, January 2007

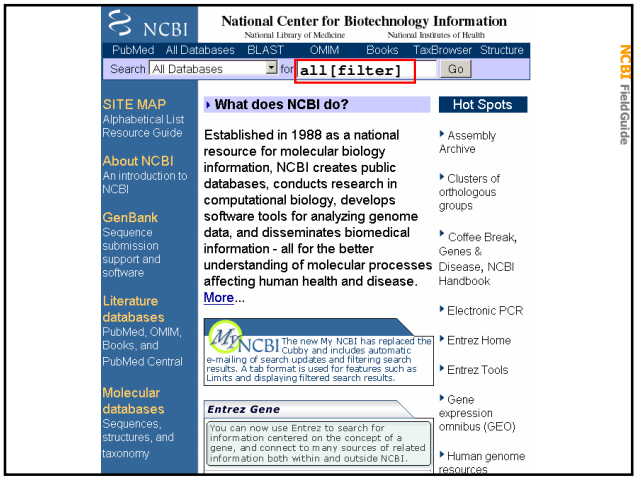


NCBI FieldGuide

NCBI Web Usage – Hits per Day, January 2007

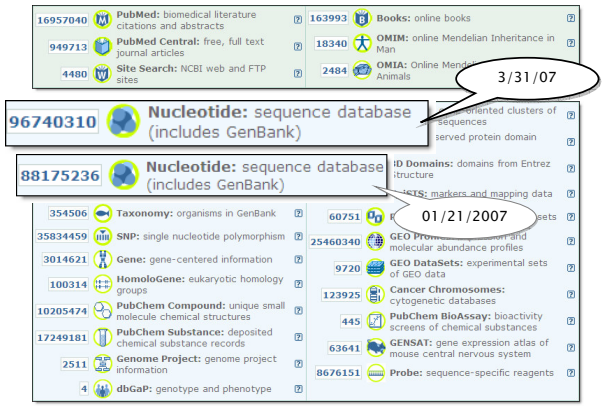


NCBI FieldGuide



NCBI FieldGuide

Global Query Page



NCBI FieldGuide

Types of Databases

- Primary Databases
  - Original data
  - 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 FieldGuide

GenBank

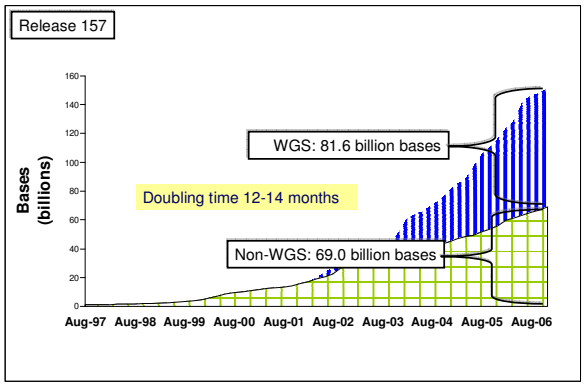
Release 158	February 2007
87 x 10^6	Records
157 x 10^9	Nucleotides
263 Gb *(non-WGS)*	1115 files *(non-WGS)*

- full release every two months
- incremental and cumulative updates daily
- available only via ftp
- release notes: gbrel.txt

ftp://ftp.ncbi.nih.gov/genbank/
ftp://genbank.sdsc.edu/pub
ftp://bio-mirror.net/biomirror/genbank

NCBI FieldGuide

The Growth of GenBank



NCBI FieldGuide

What is GenBank?

- 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 FieldGuide

GenBank Divisions

PRI (28)	Primate
ROD (15)	Rodent
PLN (20)	Plant and Fungal
BCT (18)	Bacterial/Archeal
INV (7)	Invertebrate
VRT (7)	Other Vertebrate
VRL (4)	Viral
MAM (2)	Mammalian
PHG (1)	Phage
SYN (1)	Synthetic
ENV (4)	Envir. samples
UNA (1)	Unannotated

EST (570)	Expressed Sequence Tag
GSS (197)	Genome Survey Sequence
HTG (88)	High Throughput Genomic
PAT (27)	Patent
STS (9)	Sequence Tagged Site
CON (1)	Contigs, virtual

"Organismal"
(Traditional)

- Organized by taxonomy (sort of)
- Direct submissions (Sequin/Bankit)
- Accurate (~1 error per 10,000 bp)
- Well characterized

"Functional"
(Bulk)

- Organized by sequence type
- Batch submissions (ftp/email)
- Less accurate
- Poorly characterized

NCBI FieldGuide

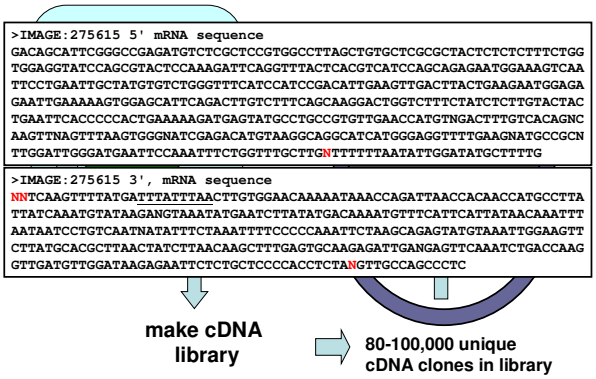
GenBank Functional (Bulk) Divisions

- **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

Whole Genome Shotgun

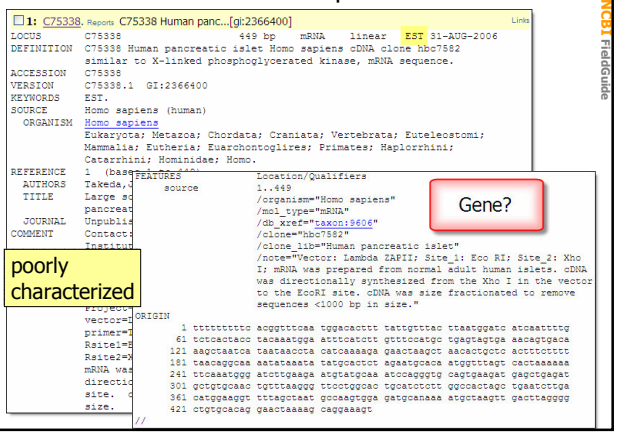
NCBI FieldGuide

EST Division: Expressed Sequence Tags



NCBI FieldGuide

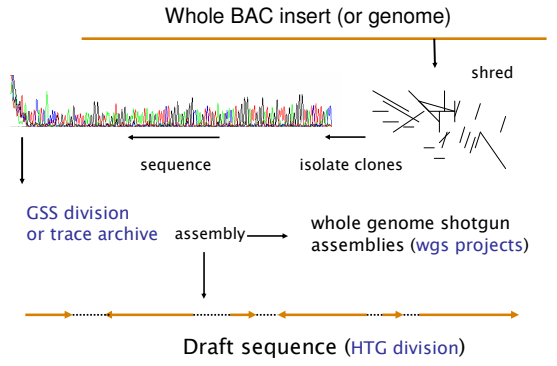
GenBank Bulk Sequence: EST



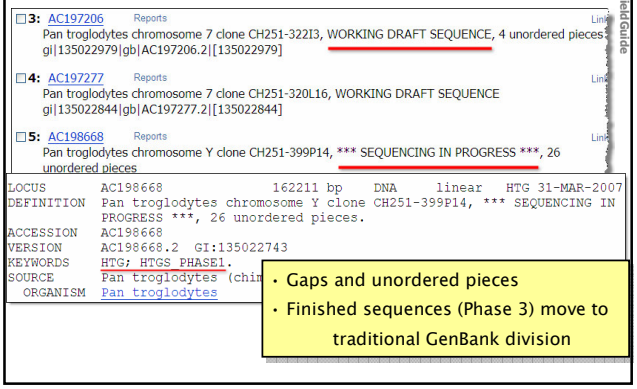
Gene?

poorly characterized

GSS, HTG, WGS

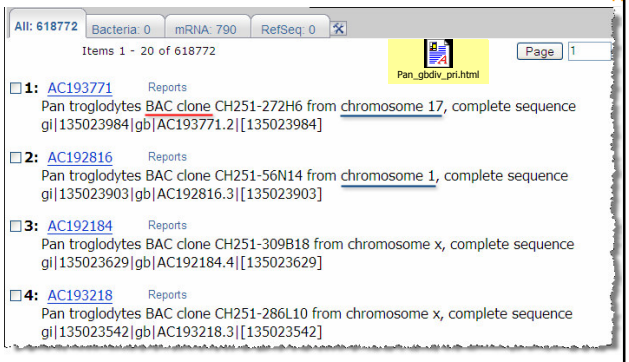


HTG Example: Chimpanzee



- Gaps and unordered pieces
- Finished sequences (Phase 3) move to traditional GenBank division

Pan troglodytes[orgn] AND gbdiv pri[Properties]

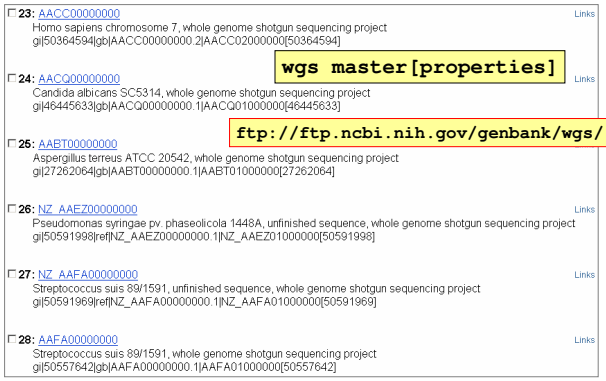


Whole Genome Shotgun Projects

- 685 projects
  - Bacteria (320)
  - Environmental sequences (14)
  - Archaea (8)
  - Eukaryotes (140), including:
    - Chicken, Rat, Mouse, Dog (2), Chimpanzee, Human
    - Pufferfish (2)
    - Honeybee, Anopheles, Fruit Flies (3), Silkworm
    - Nematode (2)
    - Yeasts (8), Aspergillus (2)
    - Rice (2)

NCBI FieldGuide

Whole Genome Shotgun (WGS) Projects



NCBI FieldGuide

Topics

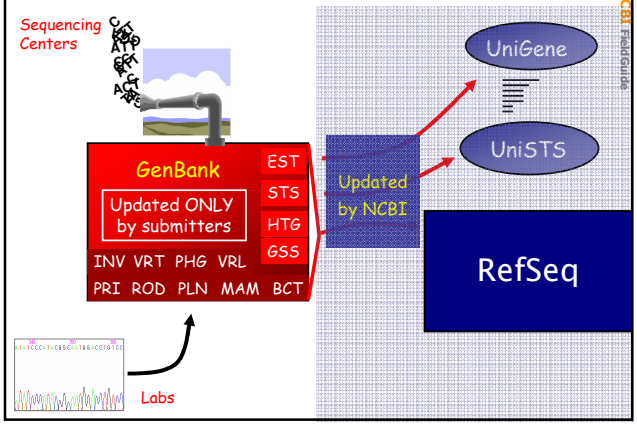
- About NCBI
- GenBank overview
- Primary vs derivative databases
  - The Reference Sequence (RefSeq) project
- The Entrez engine and databases

–break–

- Entrez text searching
- Genomic resources
- Sequence similarity – BLAST
- An integrated example

NCBI FieldGuide

Derivative Databases



NCBI FieldGuide

Why Make Reference Sequences?

Entrez Nucleotide query:

human[organism] AND lipase[title]

NCBI FieldGuide

Entrez Nucleotide query: human[organism] AND lipase[title]

All: 1661 | bacteria: 0 | mRNA: 1546 | RefSeq: 18

Items 1 - 20 of 1661 Page 1 of 84 Next

1: X68111 Reports Links
H.sapiens 5'-flanking region of gene for lipoprotein lipase
gi|34389|emb|X68111.1|HSLIPLP[34389]

2: BC070041 Reports Links
Homo sapiens lipase, hormone-sensitive, mRNA (cDNA clone MGC:87080 IMAGE:5296155), complete cds
gi|47124455|gb|BC070041.1|[47124455]

3: BC060825 Reports Links
Homo sapiens lipase, endothelial, mRNA (cDNA clone MGC:71687 IMAGE:30338950), complete cds
gi|38174525|gb|BC060825.1|[38174525]

4: NM_004190 Reports Links
Homo sapiens lipase, gastric (LIPF), mRNA
gi|4758675|ref|NM_004190.1|[4758675]

5: NM_005357 Reports Links
Homo sapiens lipase, hormone-sensitive (LIPE), mRNA
gi|21328445|ref|NM_005357.2|[21328445]

human[organism] AND lipase[title] AND endothelial[title]

Items 1 - 5 of 5 One page.

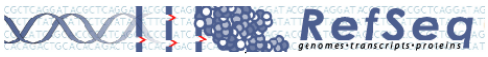
1: NM_006033 Reports 3927 bp Links
Homo sapiens lipase, endothelial (LIPG), mRNA
gi|5174496|ref|NM_006033.1|[5174496]

2: BC060825 Reports 4150 bp Links
Homo sapiens lipase, endothelial, mRNA (cDNA clone MGC:71687 IMAGE:30338950), complete cds
gi|38174525|gb|BC060825.1|[38174525]

3: AK125344 Reports 2323 bp Links
Homo sapiens cDNA FLJ43354 fis, clone NT2RP7010599, highly similar to Homo sapiens endothelial lipase mRNA
gi|34531414|dbj|AK125344.1|[34531414]

4: AF118767 Reports 3927 bp Links
Homo sapiens endothelial lipase mRNA, complete cds
gi|4836418|gb|AF118767.1|AF118767[4836418]

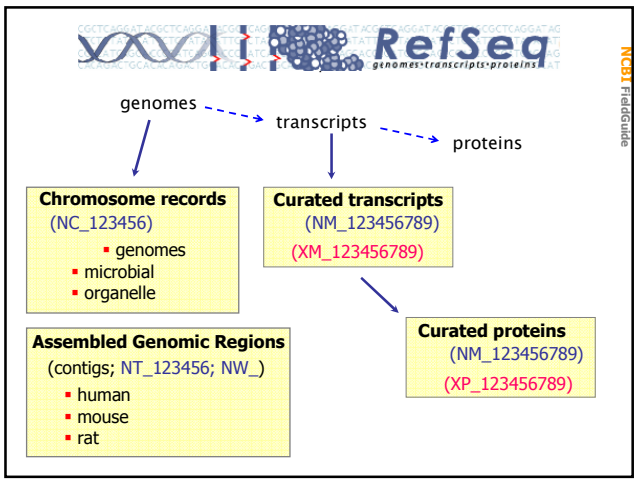
5: AA303576 Reports 261 bp Links
EST16216 Aorta endothelial cells, TNF alpha-treated Homo sapiens cDNA 5' end similar to cholesteryl ester hydrolase/ lysosomal acid lipase A, MRNA sequence
gi|1955909|gb|AA303576.1|[1955909]



genomes → transcripts → proteins

NCBI FieldGuide

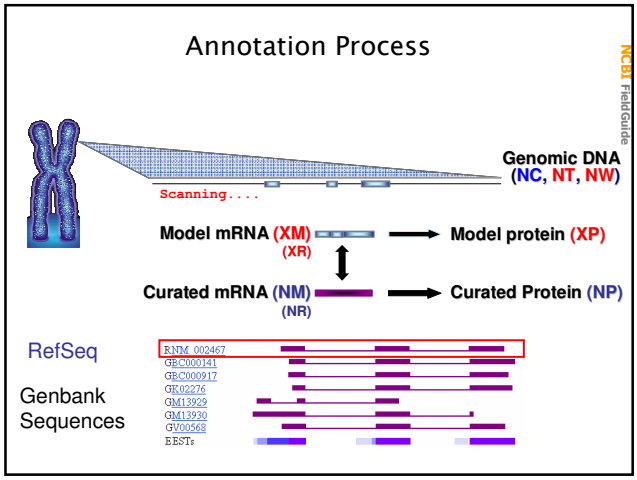
- non-redundant; best representative
- updates to reflect current sequence data and biology
- distinct, stable accession series

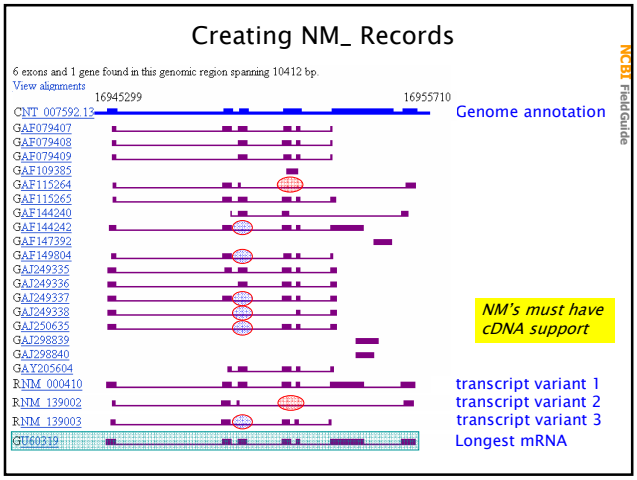


Reference Sequence: RefSeq

Accession	Sequence Type
NM_123456789	mRNA
NP_123456789	protein, from NM_
NR_123456	non-coding RNA
XM_123456	predicted mRNA
XP_123456	predicted protein
XR_123456	predicted non-coding RNA
ZP_12345678	predicted from NZ_
NC_123456	genomic, *e.g.*, chromosomes
NG_123455	genomic, incomplete region
NT_123456	genomic, BAC assembly
NW_123456	genomic, WGS assembly
NZ_ABCD12345678	genomic, WGS collection

NCBI FieldGuide





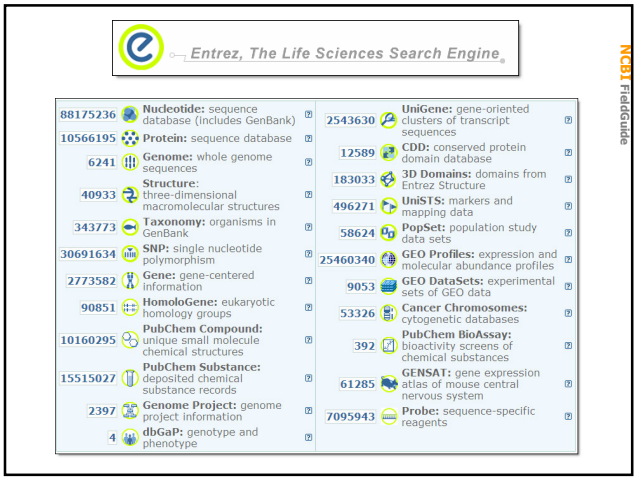
Topics

- About NCBI
- GenBank overview
- Primary vs derivative databases
  - The Reference Sequence (RefSeq) project
- The Entrez engine and databases

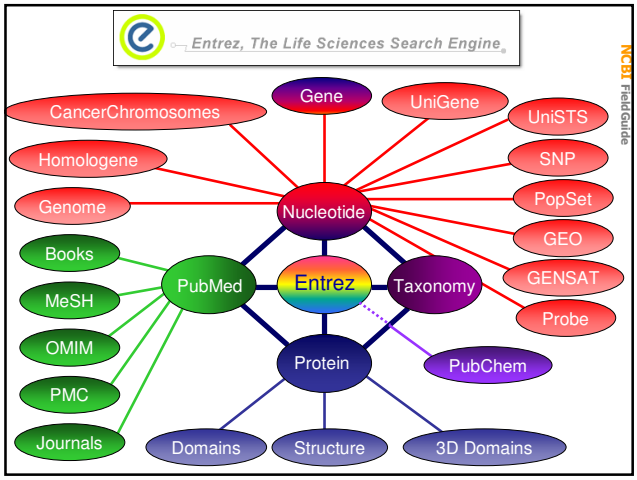
–break–

- Entrez text searching
- Genomic resources
- Sequence similarity – BLAST
- An integrated example

NCBI FieldGuide



NCBI FieldGuide



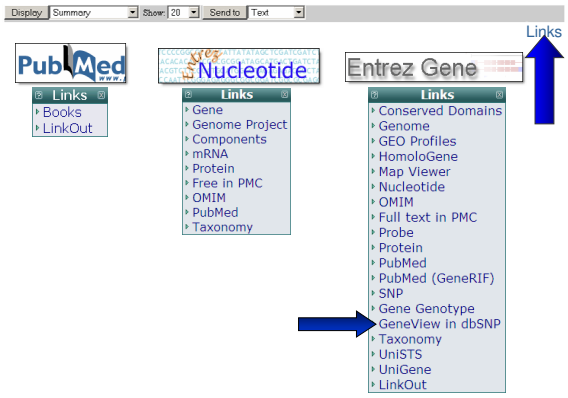
NCBI FieldGuide

Entrez Databases

- All Molecular Database entries are organized by organism ***(Taxonomy Database)***.
- Each record is assigned a UID.
  - A "unique integer identifier" for internal tracking
- Each record is indexed by data fields.
  - [author], [title], [organism], and many others
- Each record is given a Document Summary.
  - a summary of the record's content (DocSum)
- Each record is manually or computationally assigned links to biologically related UIDs in and across databases.

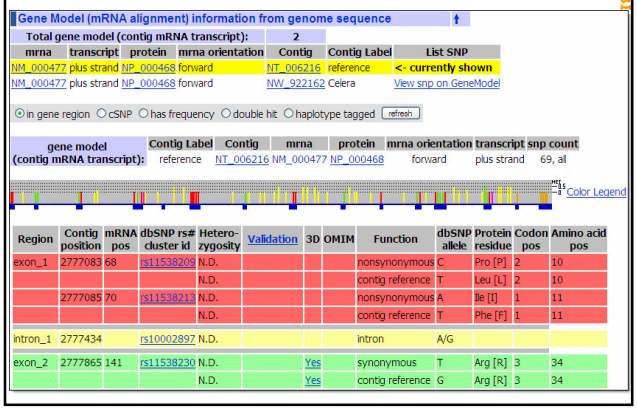
NCBI FieldGuide

Entrez Links



NCBI FieldGuide

GeneView in dbSNP



NCBI FieldGuide

Entrez Databases

UniGene	Clusters of ESTs, mRNAs
dbSNP	Single Nucleotide Polymorphisms
	…and more
CDD	Conserved Domain Database
	protein families (COGs and KOGs)
	single domains (PFAM, SMART, CD)

NCBI FieldGuide

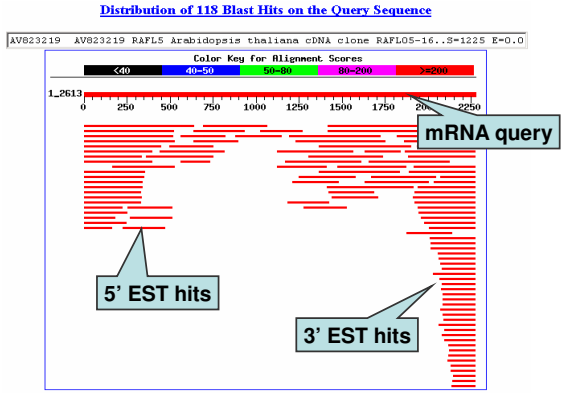
UniGene
ORGANIZED VIEW OF THE TRANSCRIPTOME

Gene-oriented clusters of expressed sequences

- Automatic clustering using MegaBlast
- Each cluster represents a unique gene
- Informed by genome hits
- Information on tissue types and map locations
- Useful for gene discovery and selection of mapping reagents

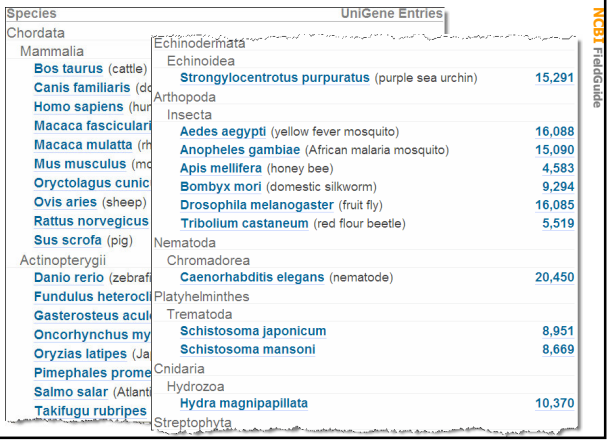
NCBI FieldGuide

A Cluster of ESTs



NCBI FieldGuide

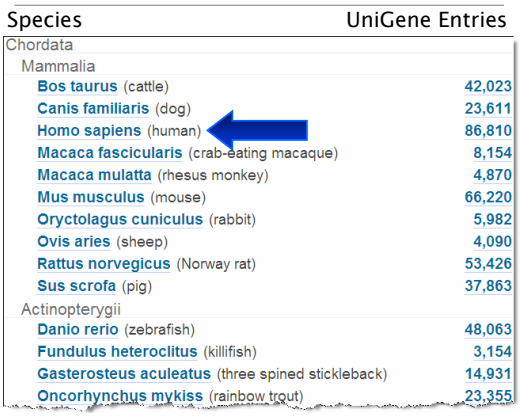
UniGene Collections



Species	UniGene Entries
Chordata	
Mammalia	
Bos taurus (cattle)	
Canis familiaris (dog)	
Homo sapiens (human)	
Macaca fascicularis (crab-eating macaque)	
Macaca mulatta (rhesus monkey)	
Mus musculus (mouse)	
Oryctolagus cuniculus (rabbit)	
Ovis aries (sheep)	
Rattus norvegicus (Norway rat)	
Sus scrofa (pig)	
Actinopterygii	
Danio rerio (zebrafish)	
Fundulus heteroclitus (killifish)	
Gasterosteus aculeatus (three spined stickleback)	
Oncorhynchus mykiss (rainbow trout)	
Oryzias latipes (Japanese medaka)	
Pimephales promelas (fathead minnow)	
Salmo salar (Atlantic salmon)	
Takifugu rubripes (fugu)	
Echinodermata	
Echinoidea	
Strongylocentrotus purpuratus (purple sea urchin)	15,291
Arthopoda	
Insecta	
Aedes aegypti (yellow fever mosquito)	16,088
Anopheles gambiae (African malaria mosquito)	15,090
Apis mellifera (honey bee)	4,583
Bombyx mori (domestic silkworm)	9,294
Drosophila melanogaster (fruit fly)	16,085
Tribolium castaneum (red flour beetle)	5,519
Nematoda	
Chromadorea	
Caenorhabditis elegans (nematode)	20,450
Platyhelminthes	
Trematoda	
Schistosoma japonicum	8,951
Schistosoma mansoni	8,669
Cnidaria	
Hydrozoa	
Hydra magnipapillata	10,370
Streptophyta	

NCBI FieldGuide

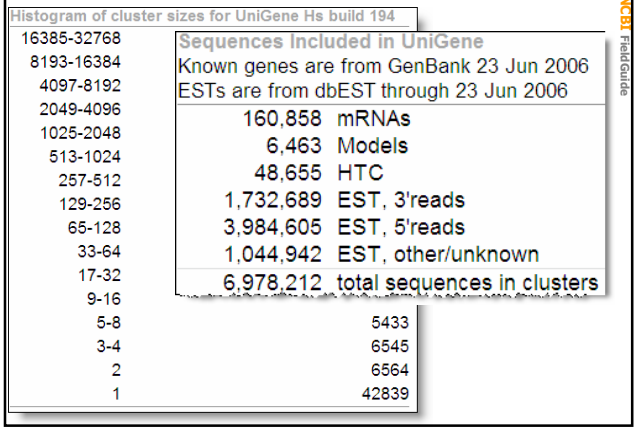
UniGene Collections



Species	UniGene Entries
Chordata	
Mammalia	
Bos taurus (cattle)	42,023
Canis familiaris (dog)	23,611
Homo sapiens (human)	86,810
Macaca fascicularis (crab-eating macaque)	8,154
Macaca mulatta (rhesus monkey)	4,870
Mus musculus (mouse)	66,220
Oryctolagus cuniculus (rabbit)	5,982
Ovis aries (sheep)	4,090
Rattus norvegicus (Norway rat)	53,426
Sus scrofa (pig)	37,863
Actinopterygii	
Danio rerio (zebrafish)	48,063
Fundulus heteroclitus (killifish)	3,154
Gasterosteus aculeatus (three spined stickleback)	14,931
Oncorhynchus mykiss (rainbow trout)	23,355

NCBI FieldGuide

UniGene Hs build 194 *(now at #201)*



Histogram of cluster sizes for UniGene Hs build 194

16385-32768	
8193-16384	
4097-8192	
2049-4096	
1025-2048	
513-1024	
257-512	
129-256	
65-128	
33-64	
17-32	
9-16	
5-8	5433
3-4	6545
2	6564
1	42839

Sequences Included in UniGene

Known genes are from GenBank 23 Jun 2006

ESTs are from dbEST through 23 Jun 2006

160,858	mRNAs
6,463	Models
48,655	HTC
1,732,689	EST, 3'reads
3,984,605	EST, 5'reads
1,044,942	EST, other/unknown
6,978,212	total sequences in clusters

NCBI FieldGuide

UniGene Cluster Hs.656980
Lipase, hormone-sensitive (LIPE)

NCBI FieldGuide

SELECTED PROTEIN SIMILARITIES

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

A. thaliana	pir:T46214 - T46214 hypothetical protein T8P19.210 - Arabidopsis thaliana	33.65 % / 94 aa (see ProtEST)
C. elegans	ref:NP_508794.1 - hormone sensitive lipase [Caenorhabditis elegans]	32.83 % / 527 aa (see ProtEST)
E. coli	ref:NP_415009.1 - putative lipase	36.46 % / 96 aa (see ProtEST)
H. sapiens	sp:Q05469 - LIPS_HUMAN Hormone sensitive lipase	100.00 % / 775 aa (see ProtEST)
M. musculus	ref:NP_034849.1 - lipase, hormone sensitive [Mus musculus]	85.13 % / 754 aa (see ProtEST)

UniGene Cluster Hs.656980

NCBI FieldGuide

GENE EXPRESSION

Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

Expression Profile: View expression levels using UniGene's EST ProfileViewer

cDNA Sources: brain; mammary gland; mixed; testis; pancreas; uncharacterized tissue; kidney; thymus; embryonic tissue; colon; ovary; adrenal gland; small intestine; blood; eye; cranial nerve; adipose tissue; bone; lymph node; lung; vascular; cervix

Expression profile suggested by analysis of EST counts.
Hs.656980- LIPE: Lipase, hormone-sensitive

NCBI FieldGuide

Breakdown by Tissue

	Hs.656980	
adipose tissue	391	5/12777
adrenal gland	31	1/32215
ascites	0	0/40022
bladder	0	0/29175
blood	8	1/119874
bone	41	3/71667
bone marrow	0	0/47392
brain	39	35/890811
cervix	21	1/47558
cochlea	0	0/16098
colon	16	3/181250
connective tissue	0	0/145437
cranial nerve	220	4/18109
embryonic tissue	15	3/194985
esophagus	0	0/18916
eye	30	6/199696
heart	0	0/87149
kidney	9	2/206123

Expression profile suggested by analysis of EST counts.
Hs.656980- LIPE: Lipase, hormone-sensitive

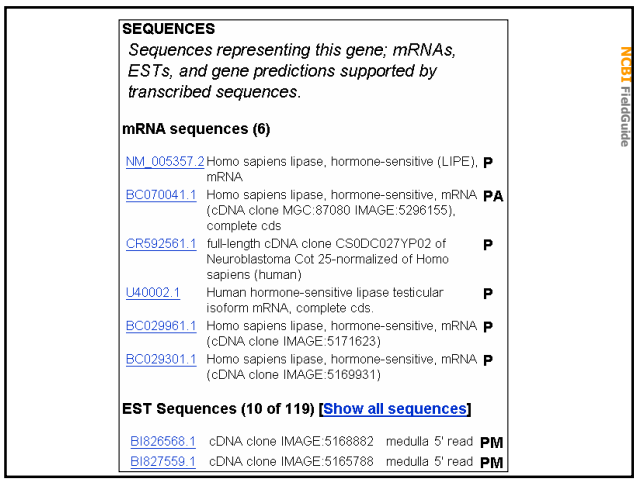
NCBI FieldGuide

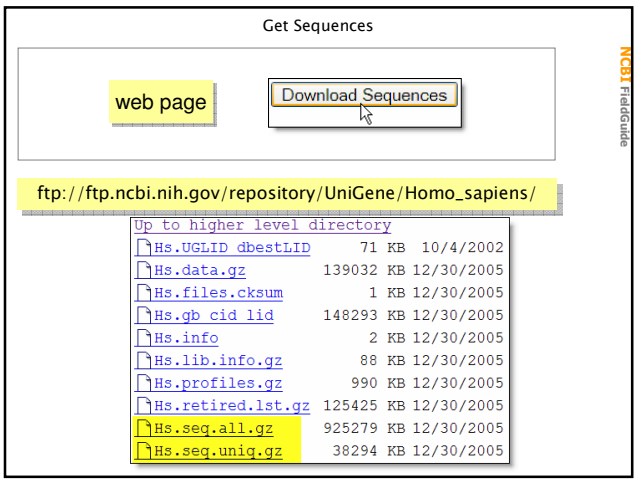
Breakdown by Developmental Stage

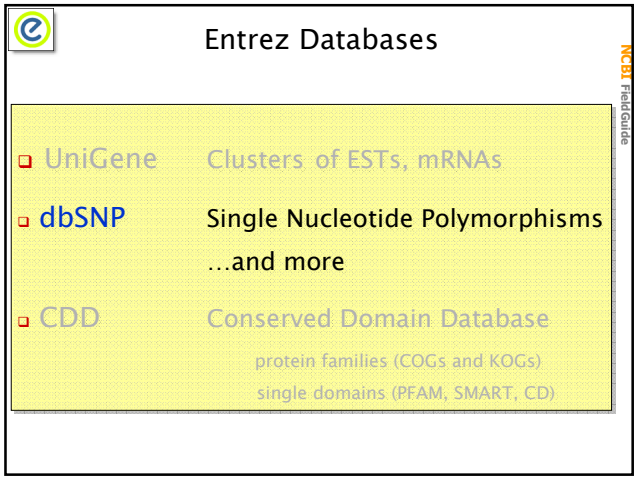
	Hs.656980	
embryo	16	3/179566
embryoid body	0	0/70535
fetus	19	11/553710
neonate (less than 4 weeks old)	0	0/26593
infant (less than 3 years old)	45	1/21845

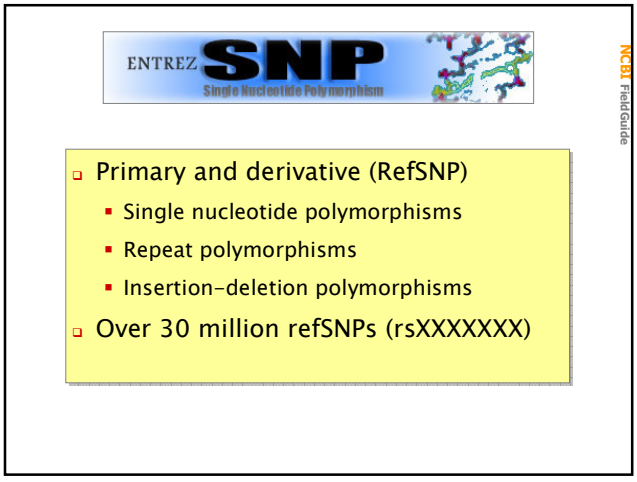
Breakdown by Health State

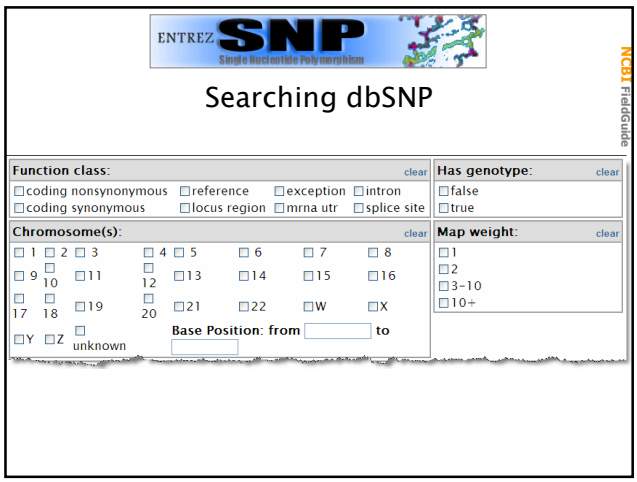
	Hs.656980	
adrenal tumor	78	1/12703
bone tumor	0	0/99675
breast (mammary gland) cancer	10	1/93020
cervical tumor	29	1/33938
colorectal cancer	17	2/112359
esophageal tumor	0	0/16386
gastrointestinal tumor	8	1/119030
germ cell tumor	7	2/254527

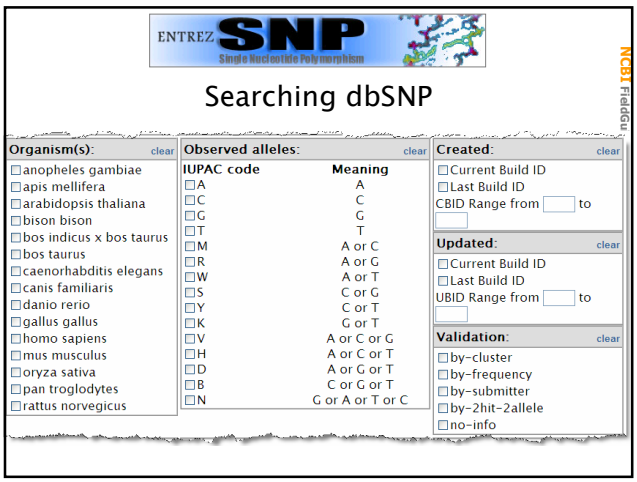

SEQUENCES
Sequences representing this gene; mRNAs, ESTs, and gene predictions supported by transcribed sequences.
mRNA sequences (6)
NM_005357.2 Homo sapiens lipase, hormone-sensitive (LIPE), mRNA P
BC070041.1 Homo sapiens lipase, hormone-sensitive, mRNA (cDNA clone MGC:87080 IMAGE:5296155), complete cds PA
CR592561.1 full-length cDNA clone CS0DC027YP02 of Neuroblastoma Cot 25-normalized of Homo sapiens (human) P
U40002.1 Human hormone-sensitive lipase testicular isoform mRNA, complete cds. P
BC029961.1 Homo sapiens lipase, hormone-sensitive, mRNA (cDNA clone IMAGE:5171623) P
BC029301.1 Homo sapiens lipase, hormone-sensitive, mRNA (cDNA clone IMAGE:5169931) P
EST Sequences (10 of 119) [Show all sequences]
BI826568.1 cDNA clone IMAGE:5168882 medulla 5' read PM
BI827559.1 cDNA clone IMAGE:5165788 medulla 5' read PM
NCBI FieldGuide

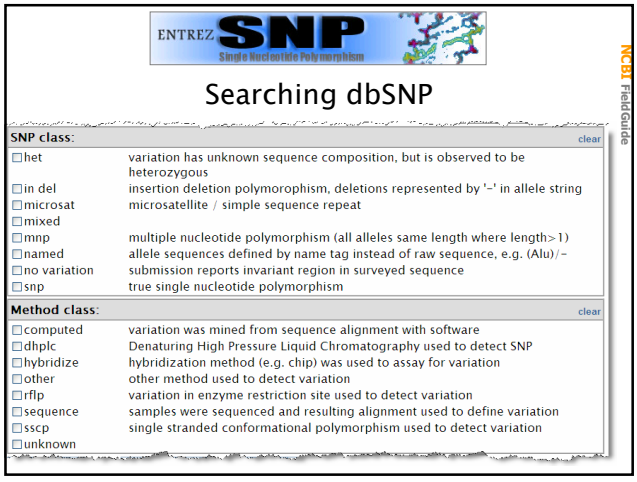

Get Sequences
web page
Download Sequences
ftp://ftp.ncbi.nih.gov/repository/UniGene/Homo_sapiens/
Up to higher level directory
Hs.UGLID dbestLID 71 KB 10/4/2002
Hs.data.gz 139032 KB 12/30/2005
Hs.files.cksum 1 KB 12/30/2005
Hs.gb_cid_lid 148293 KB 12/30/2005
Hs.info 2 KB 12/30/2005
Hs.lib.info.gz 88 KB 12/30/2005
Hs.profiles.gz 990 KB 12/30/2005
Hs.retired.lst.gz 125425 KB 12/30/2005
Hs.seq.all.gz 925279 KB 12/30/2005
Hs.seq.unig.gz 38294 KB 12/30/2005
NCBI FieldGuide

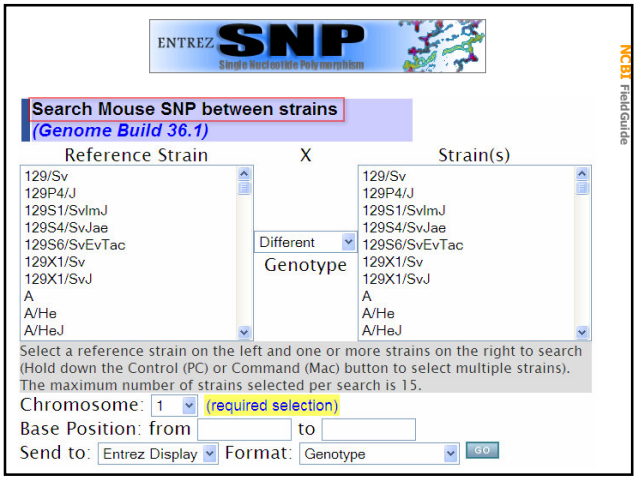

Entrez Databases
UniGene Clusters of ESTs, mRNAs
dbSNP Single Nucleotide Polymorphisms
...and more
CDD Conserved Domain Database
protein families (COGs and KOGs)
single domains (PFAM, SMART, CD)
NCBI FieldGuide


ENTREZ SNP
Single Nucleotide Polymorphism
Primary and derivative (RefSNP)
Single nucleotide polymorphisms
Repeat polymorphisms
Insertion-deletion polymorphisms
Over 30 million refSNPs (rsXXXXXXX)
NCBI FieldGuide


ENTREZ SNP
Single Nucleotide Polymorphism
NCBI FieldGuide
Searching dbSNP
Function class: clear
coding nonsynonymous
coding synonymous
reference
locus region
exception
mrna utr
intron
splice site
Has genotype: clear
false
true
Chromosome(s): clear
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 W X Y Z unknown
Base Position: from to
Map weight: clear
1
2
3-10
10+


ENTREZ SNP
Single Nucleotide Polymorphism
NCBI FieldGu
Searching dbSNP
Organism(s): clear
anopheles gambiae
apis mellifera
arabidopsis thaliana
bison bison
bos indicus x bos taurus
bos taurus
caenorhabditis elegans
canis familiaris
danio rerio
gallus gallus
homo sapiens
mus musculus
oryza sativa
pan troglodytes
rattus norvegicus
Observed alleles: clear
IUPAC code Meaning
A A
C C
G G
T T
M A or C
R A or G
W A or T
S C or G
Y C or T
K G or T
V A or C or G
H A or C or T
D A or G or T
B C or G or T
N G or A or T or C
Created: clear
Current Build ID
Last Build ID
CBID Range from to
Updated: clear
Current Build ID
Last Build ID
UBID Range from to
Validation: clear
by-cluster
by-frequency
by-submitter
by-2hit-2allele
no-info

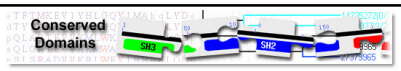

ENTREZ SNP
Single Nucleotide Polymorphism
NCBI FieldGuide
Searching dbSNP
SNP class: clear
het variation has unknown sequence composition, but is observed to be heterozygous
in del insertion deletion polymorphism, deletions represented by '-' in allele string
microsat microsatellite / simple sequence repeat
mixed
mnp multiple nucleotide polymorphism (all alleles same length where length>1)
named allele sequences defined by name tag instead of raw sequence, e.g. (Alu)/-
no variation submission reports invariant region in surveyed sequence
snp true single nucleotide polymorphism
Method class: clear
computed variation was mined from sequence alignment with software
dhplc Denaturing High Pressure Liquid Chromatography used to detect SNP
hybridize hybridization method (e.g. chip) was used to assay for variation
other other method used to detect variation
rflp variation in enzyme restriction site used to detect variation
sequence samples were sequenced and resulting alignment used to define variation
sscp single stranded conformational polymorphism used to detect variation
unknown


ENTREZ SNP
Single Nucleotide Polymorphism
NCBI FieldGuide
Search Mouse SNP between strains
(Genome Build 36.1)
Reference Strain
X
Strain(s)
129/Sv
129P4/J
129S1/SvImJ
129S4/SvJae
129S6/SvEvTac
129X1/Sv
129X1/SvJ
A
A/He
A/HeJ
Different
Genotype
Select a reference strain on the left and one or more strains on the right to search
(Hold down the Control (PC) or Command (Mac) button to select multiple strains).
The maximum number of strains selected per search is 15.
Chromosome: 1 (required selection)
Base Position: from to
Send to: Entrez Display Format: Genotype
GO

Entrez Databases

- UniGene — Clusters of ESTs, mRNAs
- dbSNP — Single Nucleotide Polymorphisms
  …and more
- CDD — Conserved Domain Database
  protein families (COGs and KOGs)
  single domains (PFAM, SMART, CD)

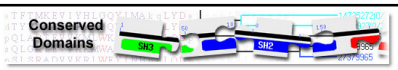
NCBI FieldGuide



Conserved Domain Database

- Multiple sequence alignments
- Position-specific scoring matrices (PSSM)
- Sources SMART, PFAM, COGs, KOGs, and NCBI curated domains (structure-informed alignments)

NCBI FieldGuide



A Conserved Domain Database and Search Service, v2.08

Proteins often contain several modules or domains, each with a distinct evolutionary origin and function. NCBI's Conserved Domain Database is a collection of multiple sequence alignments for ancient domains and full-length proteins. The CD-Search service may be used to identify the conserved domains present in a protein query sequence:

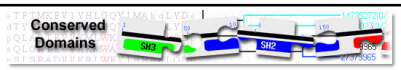
Submit Query — Search Database: CDD v2.08 - 12147 PSSMs

Enter a **Protein** query as Accession, GI, or Sequence in FASTA f

- SMART v4.0 - 663 PSSMs
- Pfam v11.0 - 7255 PSSMs
- COG v1.00 - 4873 PSSMs
- KOG v1.00 - 4825 PSSMs
- CDD v2.08 - 12147 PSSMs

Read about the FASTA format description. Click here for advanced options.

NCBI FieldGuide



Search Entrez CDD

All: 467 | archaeal: 8 | bacterial: 96 | curated: 192 | eukaryotic: 153 | uncurated: 275

Items 1 - 20 of 467 — Page 1 of 24 Next

1: cd01919 — Links

PEPCK: Phosphoenolpyruvate carboxykinase (PEPCK), a critical gluconeogenic enzyme, catalyzes the first committed step in the diversion of tricarboxylic acid cycle intermediates toward gluconeogenesis. It catalyzes the reversible decarboxylation and phosphorylation of oxaloacetate to yield phosphoenolpyruvate and carbon dioxide, using a nucleotide molecule (ATP or GTP) for the phosphoryl transfer, and has a strict requirement for divalent metal ions for activity. PEPCK's separate into two phylogenetic groups based on their nucleotide substrate specificity (the ATP-, and GTP-dependent groups). [cd01919|29834]

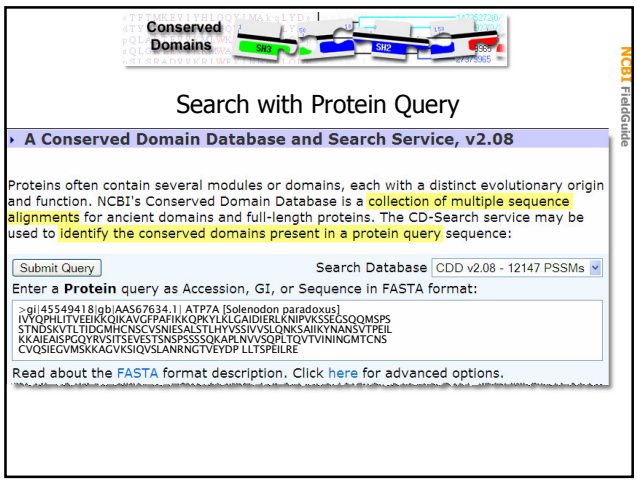
2: COG3340 — Links

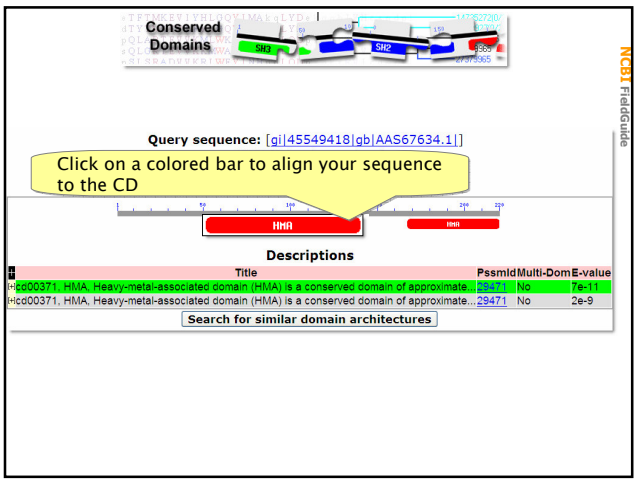
PepE: Peptidase E [Amino acid transport and metabolism] [COG3340|33149]

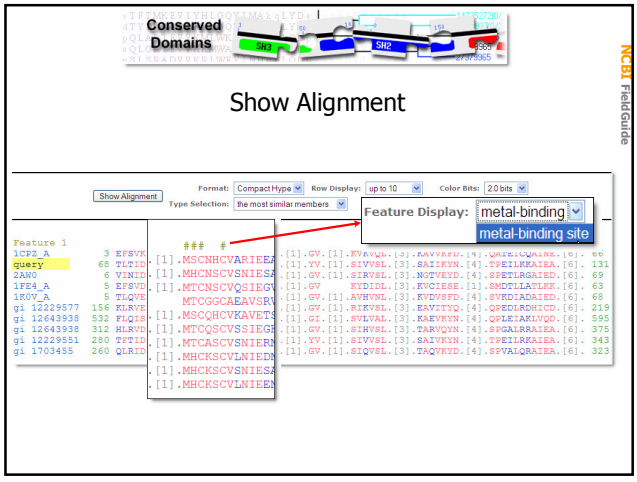
Cn3D

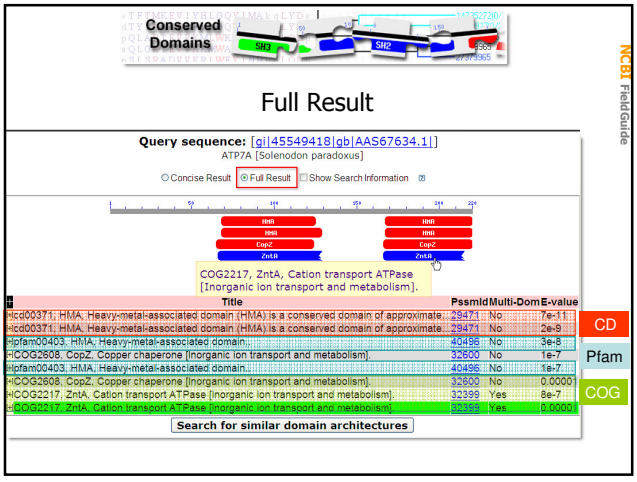
VAST

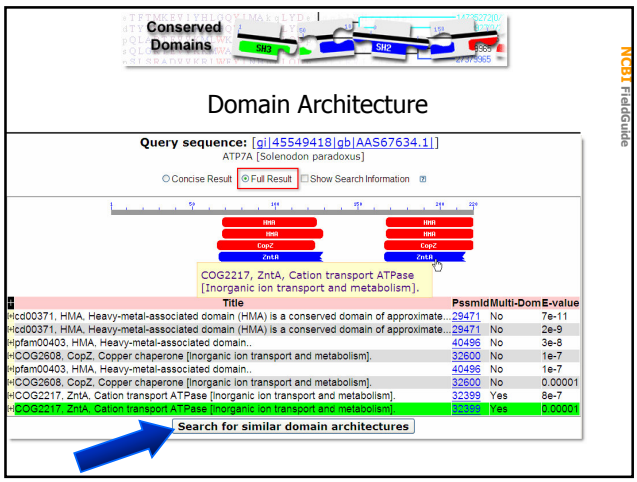
NCBI FieldGuide

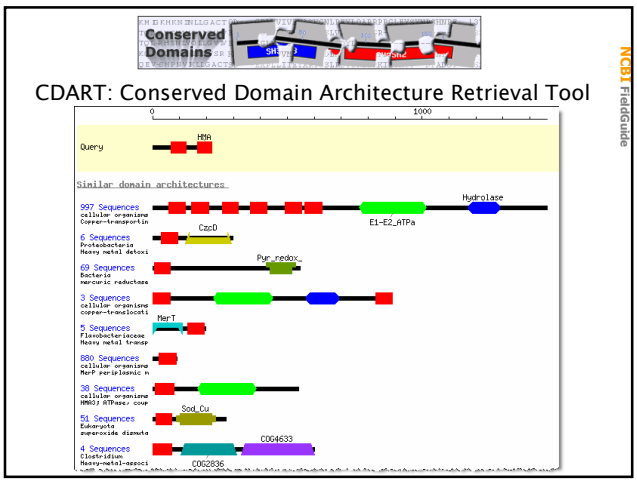

Conserved Domains
Search with Protein Query
A Conserved Domain Database and Search Service, v2.08
Proteins often contain several modules or domains, each with a distinct evolutionary origin and function. NCBI's Conserved Domain Database is a collection of multiple sequence alignments for ancient domains and full-length proteins. The CD-Search service may be used to identify the conserved domains present in a protein query sequence:
Submit Query
Search Database CDD v2.08 - 12147 PSSMs
Enter a Protein query as Accession, GI, or Sequence in FASTA format:
>gi|45549418|gb|AAS67634.1| ATP7A [Solenodon paradoxus]
IVYQPHLITVEEIKKQIKAVGFPAFIKKQPKYLKLGAIDIERLKNIPVKSSECSQQMSPS
STNDSKVTLTIDGMHCNSCVSNIESALSTLHYVSSIVVSLQNKSAIIKYNANSVTPEIL
KKAIEAISPCQYRVSITSEVESTSNSPSSSSQKAPLNVVSQPLTQVTVININGMTCNS
CVQSIEGVMSKKAGVKSIQVSLANRNGTVEYDP LLTSPEILRE
Read about the FASTA format description. Click here for advanced options.
NCBI FieldGuide

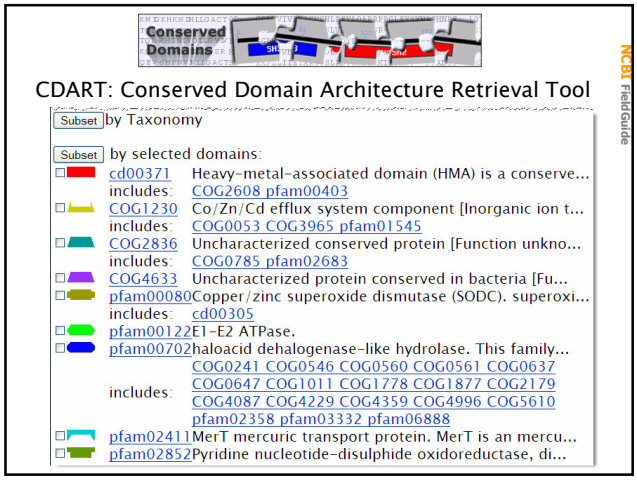

Conserved Domains
Query sequence: [gi|45549418|gb|AAS67634.1|]
Click on a colored bar to align your sequence to the CD
HMA
HMA
Descriptions
Title PssmId Multi-Dom E-value
cd00371, HMA, Heavy-metal-associated domain (HMA) is a conserved domain of approximate... 29471 No 7e-11
cd00371, HMA, Heavy-metal-associated domain (HMA) is a conserved domain of approximate... 29471 No 2e-9
Search for similar domain architectures
NCBI FieldGuide

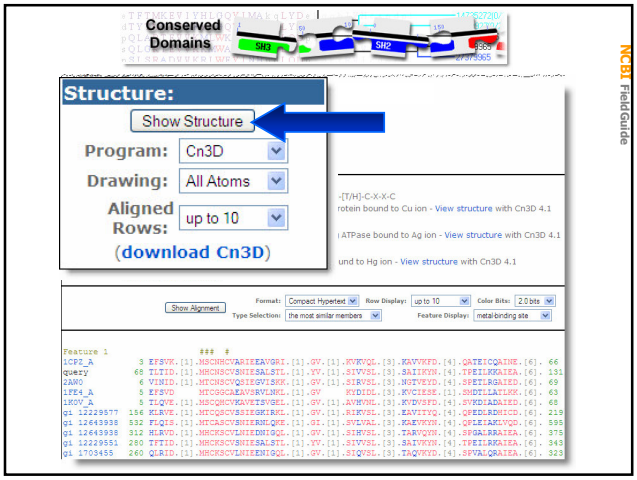

Conserved Domains
Show Alignment
Show Alignment
Format: Compact Hype
Row Display: up to 10
Color Bits: 2.0 bits
Type Selection: the most similar members
Feature Display: metal-binding
metal-binding site
Feature 1
1CPZ_A 3 EFSVE
query 68 TLTID
2AW0 6 VINID
1FE4_A 5 EFSVD
1KOV_A 5 TLQVE
gi 12229577 156 KLRVE
gi 12643938 532 FLQIS
gi 12643938 312 HLRVD
gi 12229551 280 TPTID
gi 1703455 260 QLRID
NCBI FieldGuide

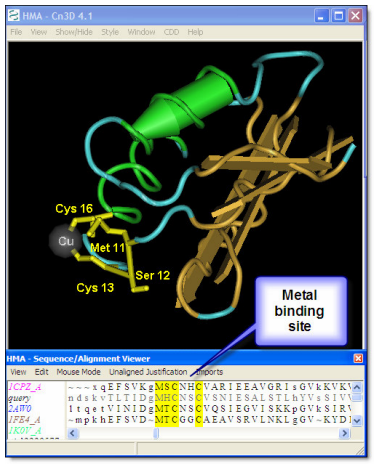

Conserved Domains
Full Result
Query sequence: [gi|45549418|gb|AAS67634.1|]
ATP7A [Solenodon paradoxus]
Concise Result
Full Result
Show Search Information
COG2217, ZntA, Cation transport ATPase [Inorganic ion transport and metabolism].
Title PssmId Multi-Dom E-value
cd00371, HMA, Heavy-metal-associated domain (HMA) is a conserved domain of approximate... 29471 No 7e-11
cd00371, HMA, Heavy-metal-associated domain (HMA) is a conserved domain of approximate... 29471 No 2e-9
pfam00403, HMA, Heavy-metal-associated domain... 40496 No 3e-8
COG2608, CopZ, Copper chaperone [Inorganic ion transport and metabolism]. 32600 No 1e-7
pfam00403, HMA, Heavy-metal-associated domain... 40496 No 1e-7
COG2608, CopZ, Copper chaperone [Inorganic ion transport and metabolism]. 32600 No 0.0000
COG2217, ZntA, Cation transport ATPase [Inorganic ion transport and metabolism]. 32399 Yes 8e-7
COG2217, ZntA, Cation transport ATPase [Inorganic ion transport and metabolism] 72359 Yes 0.0000
Search for similar domain architectures
CD
Pfam
COG
NCBI FieldGuide


Conserved Domains
Domain Architecture
Query sequence: [gi|45549418|gb|AAS67634.1|]
ATP7A [Solenodon paradoxus]
Concise Result
Full Result
Show Search Information
COG2217, ZntA, Cation transport ATPase [Inorganic ion transport and metabolism].
Title
PssmId
Multi-Dom
E-value
cd00371, HMA, Heavy-metal-associated domain (HMA) is a conserved domain of approximate...
29471
No
7e-11
cd00371, HMA, Heavy-metal-associated domain (HMA) is a conserved domain of approximate...
29471
No
2e-9
pfam00403, HMA, Heavy-metal-associated domain..
40496
No
3e-8
COG2608, CopZ, Copper chaperone [Inorganic ion transport and metabolism].
32600
No
1e-7
pfam00403, HMA, Heavy-metal-associated domain..
40496
No
1e-7
COG2608, CopZ, Copper chaperone [Inorganic ion transport and metabolism].
32600
No
0.00001
COG2217, ZntA, Cation transport ATPase [Inorganic ion transport and metabolism].
32399
Yes
8e-7
COG2217, ZntA, Cation transport ATPase [Inorganic ion transport and metabolism].
32399
Yes
0.00001
Search for similar domain architectures
NCBI FieldGuide


Conserved Domains
CDART: Conserved Domain Architecture Retrieval Tool
Query
Similar domain architectures
997 Sequences
Hydrolase
E1-E2_ATPa
6 Sequences
CzcD
69 Sequences
Pyr_redox_
3 Sequences
5 Sequences
MerT
880 Sequences
38 Sequences
51 Sequences
Sod_Cu
4 Sequences
COG4633
COG2836
NCBI FieldGuide


Conserved Domains
CDART: Conserved Domain Architecture Retrieval Tool
Subset by Taxonomy
Subset by selected domains:
cd00371 Heavy-metal-associated domain (HMA) is a conserve...
includes: COG2608 pfam00403
COG1230 Co/Zn/Cd efflux system component [Inorganic ion t...
includes: COG0053 COG3965 pfam01545
COG2836 Uncharacterized conserved protein [Function unkno...
includes: COG0785 pfam02683
COG4633 Uncharacterized protein conserved in bacteria [Fu...
pfam00080 Copper/zinc superoxide dismutase (SODC). superoxi...
includes: cd00305
pfam00122 E1-E2 ATPase.
pfam00702 haloacid dehalogenase-like hydrolase. This family...
includes: COG0241 COG0546 COG0560 COG0561 COG0637 COG0647 COG1011 COG1778 COG1877 COG2179 COG4087 COG4229 COG4359 COG4996 COG5610 pfam02358 pfam03332 pfam06888
pfam02411 MerT mercuric transport protein. MerT is an mercu...
pfam02852 Pyridine nucleotide-disulphide oxidoreductase, di...
NCBI FieldGuide


Conserved Domains
Structure:
Show Structure
Program: Cn3D
Drawing: All Atoms
Aligned Rows: up to 10
(download Cn3D)
Show Alignment
NCBI FieldGuide



NCBI FieldGuide

Topics

- About NCBI
- GenBank overview
- Primary vs derivative databases
  - The Reference Sequence (RefSeq) project
- Selected Entrez databases
- Bookshelf

-break-

- Entrez text searching
- Selected genomic resources
- Sequence similarity – BLAST
- An integrated example

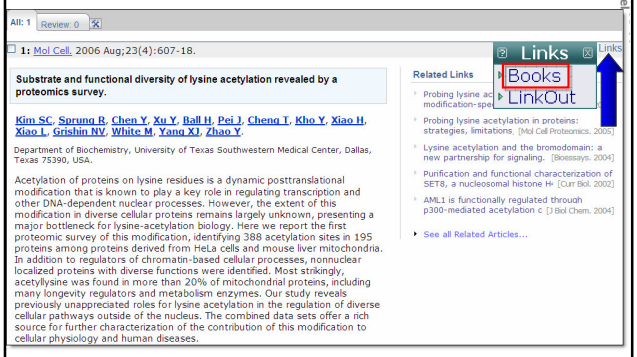
NCBI FieldGuide

Literature Links

Bookshelf

NCBI FieldGuide

Bookshelf



NCBI Field

BOOKS Database: Hyperlinked Terms

NCBI FieldGuide

1: Mol Cell. 2006 Aug;23(4):607-18. Related Articles, Links

Cell Press

Substrate and functional diversity of lysine acetylation revealed by a proteomics survey.

Kim SC, Sprung R, Chen Y, Xu Y, Ball H, Pei J, Cheng T, Kho Y, Xiao H, Xiao L, Grishin NV, White M, Yang XJ, Zhao Y.

Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA.

Acetylation of proteins on lysine residues is a dynamic posttranslational modification that is known to play a key role in regulating transcription and other DNA-dependent nuclear processes. However, the extent of this modification in diverse cellular proteins remains largely unknown, presenting a major bottleneck in lysine-acetylation biology. Here we report the first proteomic survey of this modification, identifying 388 acetylation sites in 195 proteins among proteins derived from HeLa cells and liver mitochondria. In addition to regulators of chromatin-based cellular processes, nonnuclear localized proteins with diverse functions were identified. Most strikingly, acetyllysine was found in more than 20% of mitochondrial proteins, including many longevity regulators and metabolism enzymes. Our study reveals previously unappreciated roles for lysine acetylation in the regulation of diverse cellular pathways outside of the nucleus. The combined data sets offer a rich source for further characterization of the contribution of this modification to cellular physiology and human diseases.

Bookshelf

NCBI FieldGuide

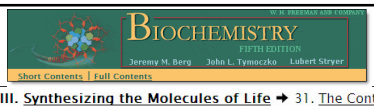
All: 3 Figures: 1

Items 1 - 3 of 3 One page

1: Chromatin Structure Is Modulated Through Covalent Modifications of Histone Tails
Biochemistry -> III. Synthesizing the Molecules of Life -> 31. The Control of Gene Expression -> 31.3. Transcriptional Activation and Repression Are Mediated by Protein-Protein Interactions

2: Key Terms
Biochemistry -> III. Synthesizing the Molecules of Life -> 31. The Control of Gene Expression -> Summary

3: Structure of a Bromodomain.
Biochemistry -> III. Synthesizing the Molecules of Life -> 31. The Control of Gene Expression -> 31.3. Transcriptional Activation and Repression Are Mediated by Protein-Protein Interactions



NCBI FieldGuide

Biochemistry ➜ III. Synthesizing the Molecules of Life ➜ 31. The Control of Gene Expression ➜ 31.3. Transcriptional Activation and Repression Are Mediated by Protein-Protein Interactions

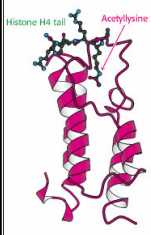


Figure 31.29. Structure of a Bromodomain. This four-helix-bundle domain binds peptides containing acetyllysine. An acetylated peptide of histone H4 is bound in the structure shown.

For More Information...

NCBI FieldGuide

The NCBI Handbook The National Library of Medicine

Short Contents | Full Contents Other books @ NCBI

The NCBI Handbook

Search GO This book All books PubMed

Part 1. The Databases

1. GenBank: The Nucleotide Sequence Database
Ilene Mizrachi.
Created: October 9, 2002, Updated: July 27, 2004

2. PubMed: The Bibliographic Database
Kathi Canese, Jennifer Jentsch, and Carol Myers.
Created: October 9, 2002, Updated: August 13, 2003

3. Macromolecular Structure Databases
Eric Sayers and Steve Bryant.
Created: October 9, 2002, Updated: August 13, 2003

4. The Taxonomy Project
Scott Federhen.
Created: October 9, 2002, Updated: August 13, 2003

5. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation
Adrienne Kitts and Stephen Sherry.
Created: October 09, 2002, Updated: June 16, 2006

6. The Gene Expression Omnibus (GEO): A Gene Expression and Hybridization Repository
Ron Edgar and Alex Lash.

Intermission

NCBI FieldGuide