
NCBI National Center for Biotechnology Information

**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>

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

The National Institutes of Health




NCBI FieldGuide

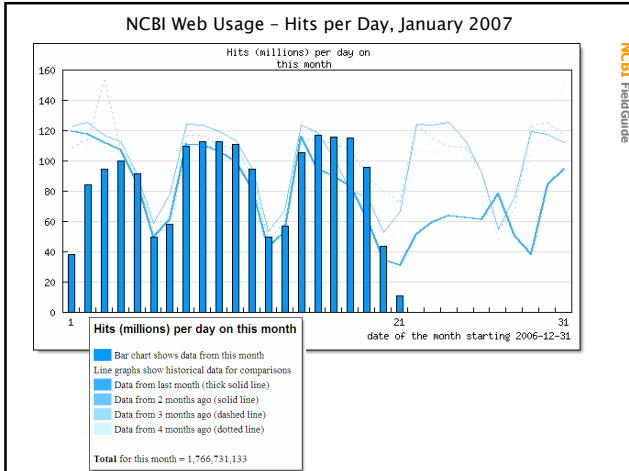
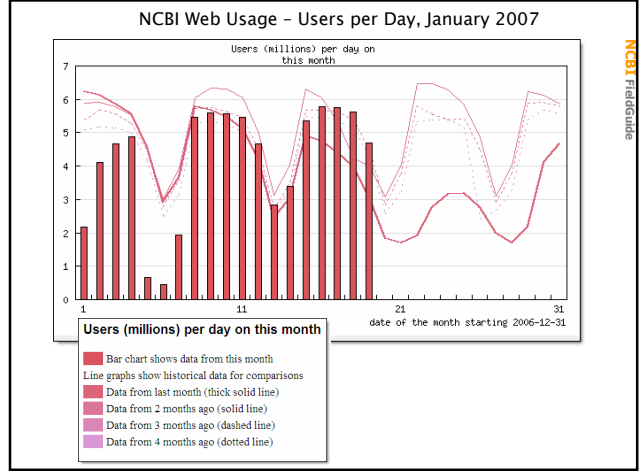
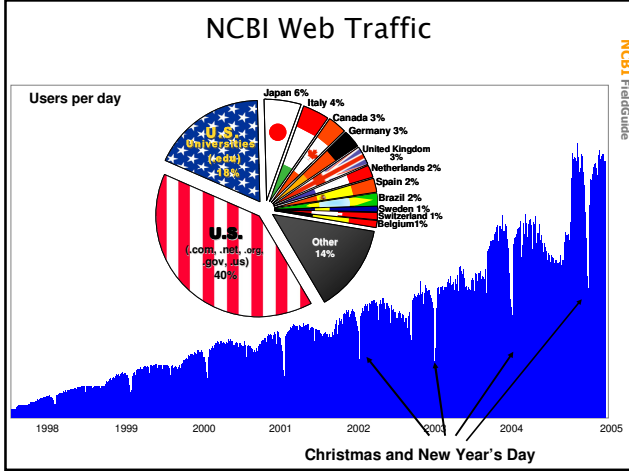
The National Center for Biotechnology Information



- 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 FieldGuide



National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search All Databases for **all [filter]** Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.

Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources

Entrez Gene
You can now use Entrez to search for information centered on the context of a gene, and connect to many sources of related information both within and outside NCBI.

Global Query Page

16957040 PubMed: biomedical literature citations and abstracts

163993 Books: online books

949713 PubMed Central: free, full text journal articles

18340 OMIM: online Mendelian Inheritance in Man

4480 Site Search: NCBI web and FTP sites

2484 OMIA: Online Mendelian Inheritance in Animals

96740310 Nucleotide: sequence database (includes GenBank) 3/31/07

88175236 Nucleotide: sequence database (includes GenBank) 01/21/2007

354506 Taxonomy: organisms in GenBank

35834459 SNP: single nucleotide polymorphism

3014621 Gene: gene-centered information

100314 Homologene: eukaryotic homology groups

10205474 PubChem Compound: unique small molecule chemical structures

17249181 PubChem Substance: deposited chemical substance records

2511 Genome Project: genome project information

4 dbGaP: genotype and phenotype

60751 ...

25460340 GEO Profiles: gene expression and molecular abundance profiles

9720 GEO DataSets: experimental sets of GEO data

123925 Cancer Chromosomes: cytogenetic databases

445 PubChem BioAssay: bioactivity screens of chemical substances

63641 GENSAT: gene expression atlas of mouse central nervous system

8676151 Probe: sequence-specific reagents

Types of Databases

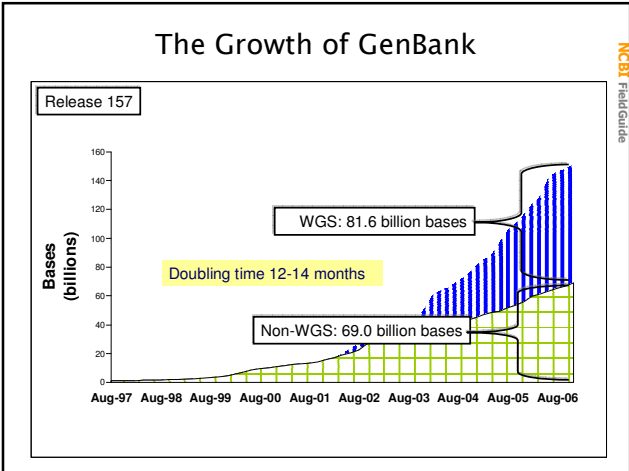
- **Primary Databases**
 - Original data
 - Content controlled by the submitter
 - Examples: GenBank, SNP, GEO
- **Derivative Databases**
 - Built from primary data
 - Content controlled by third party (NCBI)
 - Examples: Refseq, TPA, RefSNP, UniGene, NCBI Protein, Structure, Conserved Domain

GenBank

Release 158	February 2007
87 x 10 ⁶ Records	157 x 10 ⁹ Nucleotides
263 Gb (<i>non-WGS</i>)	1115 files (<i>non-WGS</i>)

- 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>



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

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

“Organismal”
(Traditional)

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

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

“Functional”
(Bulk)

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

GenBank Functional (Bulk) Divisions

GenBank

EST

GSS

HTG

STS

- **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

EST Division: Expressed Sequence Tags

```

>IMAGE:275615 5' mRNA sequence
GACAGCATTCCGGCCGAGATGTCTCGCTCCGTGGCCTTAGCTGTGCTCGCGCTACTCTCTCTTTCTGG
TGGAGGTATCCAGCGTACTCCAAGATTCCAGTTTACTCAGCTATCCAGCAGAGAATGGAAAAGTCAA
TTCTGAAATTGCTATGTGTCTGGGTTTCATCCATCCGACATTGAAGTTGACTTACTGAAGAATGGAGA
GAATTGAAAAAGTGGAGCATTGAGTGTCTTTTCAGCAAGGACTGGTCTTCTATCTCTTTGTACTAC
TGAATTCACCCCACTGAAAAAGATGAGTATGCTGCGCTGTTGAACCATGTNGACTTTGTACAGNC
AAGTTNAGTTTAAAGTGGNATCGAGACATGTAAGGCAGGCATCATGGGAGGTTTTGAAGNATGCCGCN
TTGGATTGGGATGAATCCAAATTTCTGGTTGCTTGNNTTTTAAATATGGATATGCTTTTG

>IMAGE:275615 3', mRNA sequence
NNTCAAGTTTATGATTATTTAACTGTGGAACAAAAATAAACCCAGATTAACCACAACCATGCCTTA
TTATCAAATGTATAAGANGTAAATATGAATCTTATATGACAAAAATGTTTCATTCAATTAACAAATTT
AATAATCTCTCAATNATATTTCTAAATTTCCCCCAATTTCTAAGCAGAGATGTAATTTGAAAGTT
CTTATGCACGCTTAACTATCTTAACAAGCTTTGAGTGCAAGAGATTGANGAGTTCAAATCTGACCAAG
GTTGATGTTGGATAAGAGAATTTCTGCTCCCACTTANNTTGGCAGCCCTC
  
```

↓

make cDNA library

→ **80-100,000 unique cDNA clones in library**

GenBank Bulk Sequence: EST

1: C75338 Reports C75338 Human panc...[gi2366400] Lines

LOCUS C75338 449 bp mRNA linear EST 31-AUG-2006

DEFINITION C75338 Human pancreatic islet Homo sapiens cDNA clone hbc7582 similar to X-linked phosphoglycerated kinase, mRNA sequence.

ACCESSION C75338

VERSION C75338.1 GI:2366400

KEYWORDS EST.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominoidea; Homo.

REFERENCE 1 (base pair) source Location/Qualifiers

1..449

source /organism="Homo sapiens"

title Large s4

pancreas /mol_type="mRNA"

journal Unpubl

contact /db_xref="taxon:9606"

comment /clone_lib="Human pancreatic islet"

note="Vector: Lambda ZAPII; Site_1: Eco RI; Site_2: Xho I; mRNA was prepared from normal adult human islets. cDNA was directionally synthesized from the Xho I in the vector to the EcoRI site. cDNA was size fractionated to remove sequences <1000 bp in size."

ORIGIN

vector= 1 tttttttttt accgttttca tggacaattt tatgttttac ttaattgata atcaattttg

primer= 61 tttcaactac tacaatttga attcaatttt gtttccatg tgaatagta acaattgaca

Ra1cc1= 121 aagctaacca taataaacca catcaaaaga gaactaagct acaactcttc actttttttt

Ra1cc2= 181 taataggaca aatacaaca tatgaattct agaatgaca atgttttagc caacaaaaaa

mRNA wa 241 tcaaatggg actttgaaga agtatgaa atcaatggtg cagtgaagat gaatgaagat

directi 301 gctgtgcaac ttttttaagg tttctggaac tgaatttttt gccaactatg tgaatttga

size. c 361 atgtgaagat ttttagtaat gcaaatgga gatcaaaaa atgataaatt gactttaggg

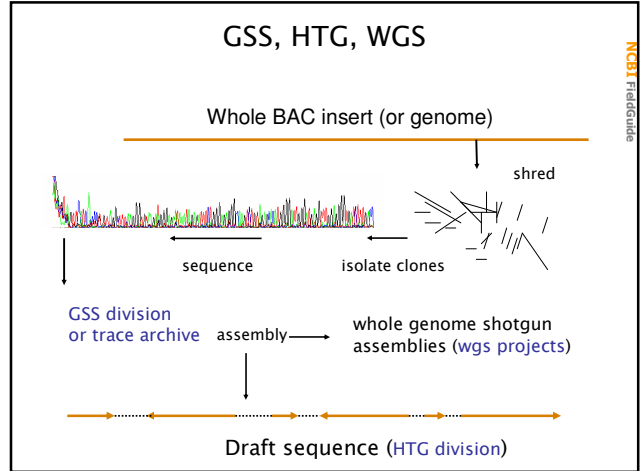
size. 421 cttgtgcaag gaactaaag caggaatg

//

Gene?

poorly characterized

NCBI FieldGuide



HTG Example: Chimpanzee

3: AC197206 Reports
Pan troglodytes chromosome 7 clone CH251-32213, WORKING DRAFT SEQUENCE, 4 unordered pieces
gi|135022979|gb|AC197206.2|[135022979]

4: AC197277 Reports
Pan troglodytes chromosome 7 clone CH251-320L16, WORKING DRAFT SEQUENCE
gi|135022844|gb|AC197277.2|[135022844]

5: AC198668 Reports
Pan troglodytes chromosome Y clone CH251-399P14, *** SEQUENCING IN PROGRESS ***, 26 unordered pieces

LOCUS AC198668 162211 bp DNA linear HTG 31-MAR-2007

DEFINITION Pan troglodytes chromosome Y clone CH251-399P14, *** SEQUENCING IN PROGRESS ***, 26 unordered pieces.

ACCESSION AC198668

VERSION AC198668.2 GI:135022743

KEYWORDS HTG; HTGS PHASE1.

SOURCE Pan troglodytes (chimpanzee)

ORGANISM [Pan troglodytes](#)

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

NCBI FieldGuide

Pan troglodytes[orgn] AND gbdiv pri[Properties]

All: 618772 Bacteria: 0 mRNA: 790 RefSeq: 0

Items 1 - 20 of 618772

Page 1

1: AC193771 Reports
Pan troglodytes BAC clone CH251-272H6 from chromosome 17, complete sequence
gi|135023984|gb|AC193771.2|[135023984]

2: AC192816 Reports
Pan troglodytes BAC clone CH251-56N14 from chromosome 1, complete sequence
gi|135023903|gb|AC192816.3|[135023903]

3: AC192184 Reports
Pan troglodytes BAC clone CH251-309B18 from chromosome x, complete sequence
gi|135023629|gb|AC192184.4|[135023629]

4: AC193218 Reports
Pan troglodytes BAC clone CH251-286L10 from chromosome x, complete sequence
gi|135023542|gb|AC193218.3|[135023542]

NCBI FieldGuide

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)

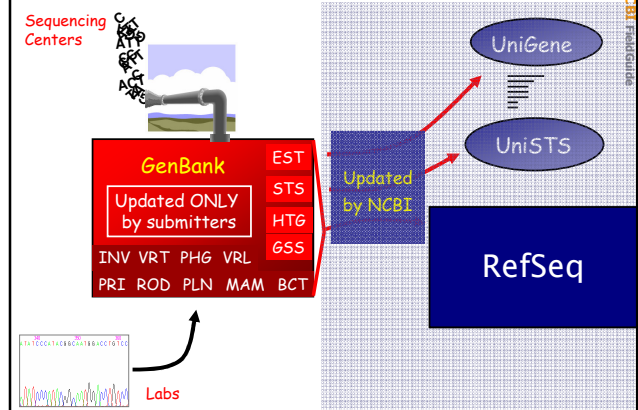
Whole Genome Shotgun (WGS) Projects

- 23: [AACG000000000](#) Links
Homo sapiens chromosome 7, whole genome shotgun sequencing project
gi|50364594|gb|AACG00000000.2|AACG02000000{50364594}
 - 24: [AACG000000000](#) Links
Candida albicans SC5314, whole genome shotgun sequencing project
gi|46445833|gb|AACG00000000.1|AACG01000000{46445833}
 - 25: [AABT000000000](#) Links
Aspergillus terreus ATCC 20542, whole genome shotgun sequencing project
gi|27262064|gb|AABT00000000.1|AABT01000000{27262064}
 - 26: [NZ_AAEZ000000000](#) Links
Pseudomonas syringae pv. phaseolicola 1448A, unfinished sequence, whole genome shotgun sequencing project
gi|50591998|ref|NZ_AAEZ00000000.1|NZ_AAEZ01000000{50591998}
 - 27: [NZ_AAFA000000000](#) Links
Streptococcus suis 89/1591, unfinished sequence, whole genome shotgun sequencing project
gi|50591969|ref|NZ_AAFA00000000.1|NZ_AAFA01000000{50591969}
 - 28: [AAFA000000000](#) Links
Streptococcus suis 89/1591, whole genome shotgun sequencing project
gi|50557642|gb|AAFA00000000.1|AAFA01000000{50557642}
- wgs master[properties]**
- ftp://ftp.ncbi.nih.gov/genbank/wgs/**

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

Derivative Databases



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 X

Items 1 - 20 of 1661 Page 1 of 84 Next

- [X68111](#) Reports Links
H.sapiens 5'-flanking region of gene for lipoprotein lipase
gi|34389|emb|X68111.1|HSLPLP[34389]
- [BC070041](#) Reports Links
Homo sapiens lipase, hormone-sensitive, mRNA (cDNA clone MGC:87080 IMAGE:5296155), complete cds
gi|47124455|gb|BC070041.1|[47124455]
- [BC060825](#) Reports Links
Homo sapiens lipase, endothelial, mRNA (cDNA clone MGC:71687 IMAGE:30338950), complete cds
gi|38174525|gb|BC060825.1|[38174525]
- [NM_004190](#) Reports Links
Homo sapiens lipase, gastric (LIPF), mRNA
gi|4758675|ref|NM_004190.1|[4758675]
- [NM_005357](#) Reports Links
Homo sapiens lipase, hormone-sensitive (LIPE), mRNA
gi|21328445|ref|NM_005357.2|[21328445]

NCBI

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

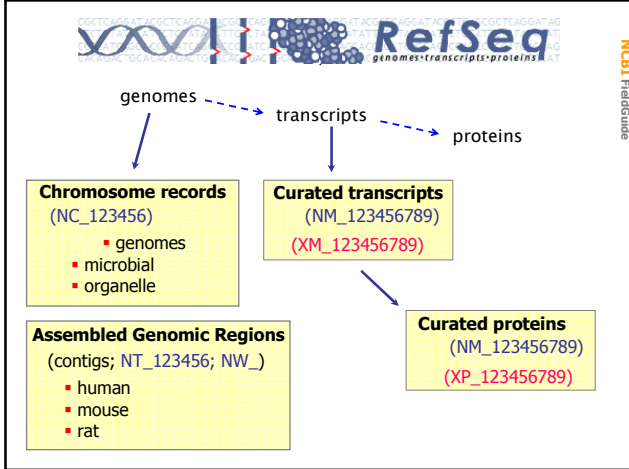
Items 1 - 5 of 5 One page.

- [NM_006033](#) Reports **3927 bp** Links
Homo sapiens lipase, endothelial (LIPG), mRNA
gi|5174496|ref|NM_006033.1|[5174496]
- [BC060825](#) Reports **4150 bp** Links
Homo sapiens lipase, endothelial, mRNA (cDNA clone MGC:71687 IMAGE:30338950), complete cds
gi|38174525|gb|BC060825.1|[38174525]
- [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]
- [AF118767](#) Reports **3927 bp** Links
Homo sapiens endothelial lipase mRNA, complete cds
gi|4836418|gb|AF118767.1|AF118767[4836418]
- [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]



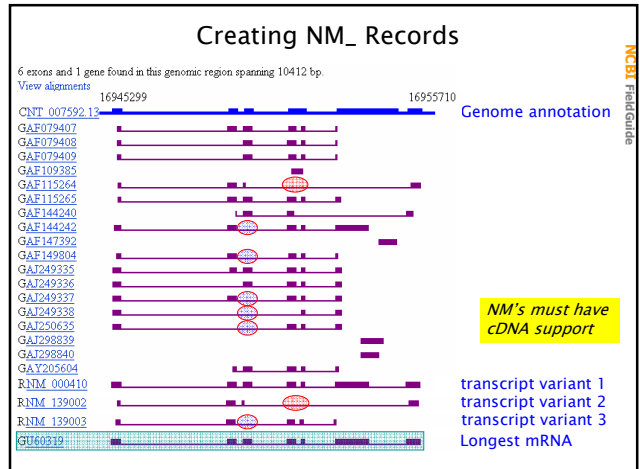
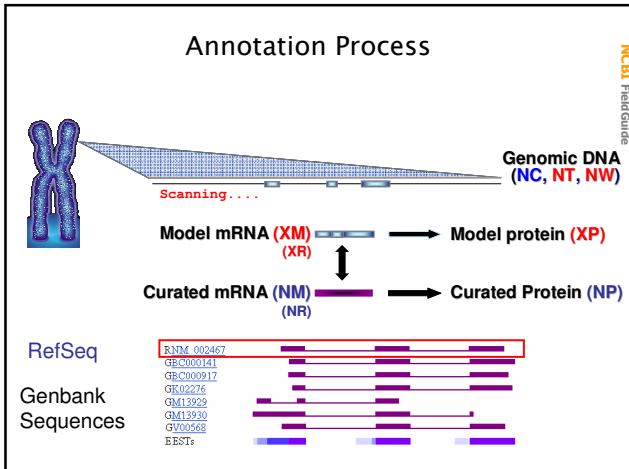
- non-redundant; best representative
- updates to reflect current sequence data and biology
- distinct, stable accession series

NCBI Fieldguide



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



Topics

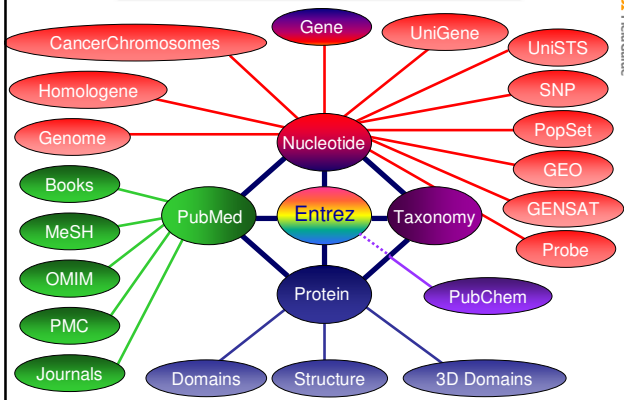
- About NCBI
 - GenBank overview
 - Primary vs derivative databases
 - The Reference Sequence (RefSeq) project
 - The Entrez engine and databases
- break-
- Entrez text searching
 - Genomic resources
 - Sequence similarity – BLAST
 - An integrated example

NCBI Fieldguide



88175236	Nucleotide: sequence database (includes GenBank)	2543630	UniGene: gene-oriented clusters of transcript sequences
10566195	Protein: sequence database	12589	CDD: conserved protein domain database
6241	Genome: whole genome sequences	183033	3D Domains: domains from Entrez Structure
40933	Structure: three-dimensional macromolecular structures	496271	UniSTS: markers and mapping data
343773	Taxonomy: organisms in GenBank	58624	PopSet: population study data sets
30691634	SNP: single nucleotide polymorphism	25460340	GEO Profiles: expression and molecular abundance profiles
2773582	Gene: gene-centered information	9053	GEO DataSets: experimental sets of GEO data
90851	HomoloGene: eukaryotic homology groups	53326	Cancer Chromosomes: cytogenetic databases
10160295	PubChem Compound: unique small molecule chemical structures	392	PubChem BioAssay: bioactivity screens of chemical substances
15515027	PubChem Substance: deposited chemical substance records	61285	GENSAT: gene expression atlas of mouse central nervous system
2397	Genome Project: genome project information	7095943	Probe: sequence-specific reagents
4	dbGap: genotype and phenotype		

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

Display: Summary Show: 20 Send to: Text

PubMed

- Links
- Books
- LinkOut

Nucleotide

- Links
- Gene
- Genome Project
- Components
- mRNA
- Protein
- Free in PMC
- OMIM
- PubMed
- Taxonomy

Entrez Gene

- Links
- Conserved Domains
- Genome
- GEO Profiles
- HomoloGene
- Map Viewer
- Nucleotide
- OMIM
- Full text in PMC
- Probe
- Protein
- PubMed
- PubMed (GeneRIF)
- SNP
- Gene Genotype
- GeneView in dbSNP
- Taxonomy
- UniSTS
- UniGene
- LinkOut

NCBI FieldGuide

GeneView in dbSNP

Gene Model (mRNA alignment) information from genome sequence

Total gene model (contig mRNA transcript): 2

mRNA	transcript	protein	mRNA orientation	Contig	Contig Label	List SNP
NM_000477	plus strand	NP_000468	forward	NT_006216	reference	<- currently shown
NM_000477	plus strand	NP_000468	forward	WV_922162	Celera	View snp on GeneModel

in gene region cSNP has frequency double hit haplotype tagged

gene model (contig mRNA transcript): reference NT_006216 NM_000477 NP_000468 forward plus strand 69, al

Region	Contig position	mRNA pos	dbSNP rs# cluster id	Heterozygosity	Validation	SD	OMIM	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos
exon_1	2777083	68	rs11538209	N.D.				nonsynonymous	C	Pro [P]	2	10
				N.D.				contig reference	T	Leu [L]	2	10
	2777085	70	rs11538213	N.D.				nonsynonymous	A	Ile [I]	1	11
				N.D.				contig reference	T	Phe [F]	1	11
intron_1	2777434		rs10002897	N.D.				intron	A/G			
exon_2	2777865	141	rs11538230	N.D.	Yes			synonymous	T	Arg [R]	3	34
				N.D.	Yes			contig reference	G	Arg [R]	3	34

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

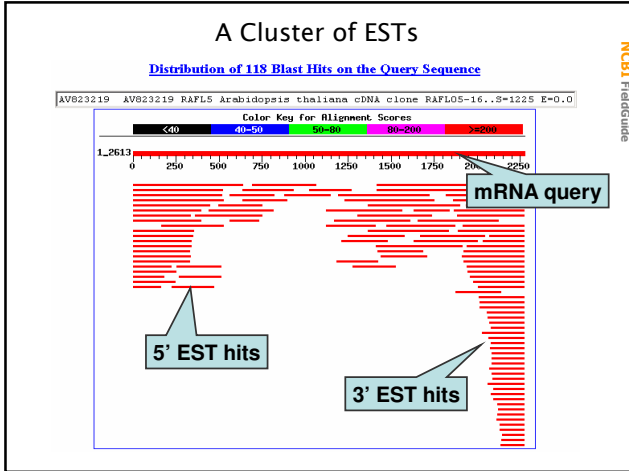
UniGene

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



UniGene Collections

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

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

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

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

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

<i>A. thaliana</i>	pir:T46214 - T46214 hypothetical protein T8P19.210 - Arabidopsis thaliana	33.85 % / 94 aa (see ProtEST)
<i>C. elegans</i>	ref.NP_508794.1 - hormone sensitive lipase [Caenorhabditis elegans]	32.83 % / 527 aa (see ProtEST)
<i>E. coli</i>	ref.NP_415009.1 - putative lipase	36.46 % / 96 aa (see ProtEST)
<i>H. sapiens</i>	sp:Q05469 - LIPS_HUMAN Hormone sensitive lipase	100.00 % / 775 aa (see ProtEST)
<i>M. musculus</i>	ref.NP_034849.1 - lipase, hormone sensitive [Mus musculus]	85.13 % / 754 aa (see ProtEST)

NCBI FieldGuide

UniGene Cluster Hs.656980

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

NCBI FieldGuide

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

Breakdown by Tissue

Tissue	Hs.656980	Total
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

NCBI FieldGuide

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

Breakdown by Developmental Stage

Stage	Hs.656980	Total
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

Health State	Hs.656980	Total
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

NCBI FieldGuide

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 (rsXXXXXX)**

NCBI FieldGuide

ENTREZ SNP
Single Nucleotide Polymorphism

Searching dbSNP

NCBI FieldGuide

Function class: clear <input type="checkbox"/> coding nonsynonymous <input type="checkbox"/> reference <input type="checkbox"/> exception <input type="checkbox"/> intron <input type="checkbox"/> coding synonymous <input type="checkbox"/> locus region <input type="checkbox"/> mrna utr <input type="checkbox"/> splice site	Has genotype: clear <input type="checkbox"/> false <input type="checkbox"/> true
Chromosome(s): clear <input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> 6 <input type="checkbox"/> 7 <input type="checkbox"/> 8 <input type="checkbox"/> 9 <input type="checkbox"/> 10 <input type="checkbox"/> 11 <input type="checkbox"/> 12 <input type="checkbox"/> 13 <input type="checkbox"/> 14 <input type="checkbox"/> 15 <input type="checkbox"/> 16 <input type="checkbox"/> 17 <input type="checkbox"/> 18 <input type="checkbox"/> 19 <input type="checkbox"/> 20 <input type="checkbox"/> 21 <input type="checkbox"/> 22 <input type="checkbox"/> W <input type="checkbox"/> X <input type="checkbox"/> Y <input type="checkbox"/> Z <input type="checkbox"/> unknown	Map weight: clear <input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3-10 <input type="checkbox"/> 10+
Base Position: from <input type="text"/> to <input type="text"/>	

ENTREZ SNP
Single Nucleotide Polymorphism

Searching dbSNP

NCBI FieldGuide

Organism(s): clear <input type="checkbox"/> anopheles gambiae <input type="checkbox"/> apis mellifera <input type="checkbox"/> arabidopsis thaliana <input type="checkbox"/> bison bison <input type="checkbox"/> bos indicus x bos taurus <input type="checkbox"/> bos taurus <input type="checkbox"/> caenorhabditis elegans <input type="checkbox"/> canis familiaris <input type="checkbox"/> danio rerio <input type="checkbox"/> gallus gallus <input type="checkbox"/> homo sapiens <input type="checkbox"/> mus musculus <input type="checkbox"/> oryza sativa <input type="checkbox"/> pan troglodytes <input type="checkbox"/> rattus norvegicus	Observed alleles: clear <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">IUPAC code</th> <th style="text-align: left;">Meaning</th> </tr> </thead> <tbody> <tr><td><input type="checkbox"/> A</td><td>A</td></tr> <tr><td><input type="checkbox"/> C</td><td>C</td></tr> <tr><td><input type="checkbox"/> G</td><td>G</td></tr> <tr><td><input type="checkbox"/> T</td><td>T</td></tr> <tr><td><input type="checkbox"/> M</td><td>A or C</td></tr> <tr><td><input type="checkbox"/> R</td><td>A or G</td></tr> <tr><td><input type="checkbox"/> W</td><td>A or T</td></tr> <tr><td><input type="checkbox"/> S</td><td>C or G</td></tr> <tr><td><input type="checkbox"/> Y</td><td>C or T</td></tr> <tr><td><input type="checkbox"/> K</td><td>G or T</td></tr> <tr><td><input type="checkbox"/> V</td><td>A or C or G</td></tr> <tr><td><input type="checkbox"/> H</td><td>A or C or T</td></tr> <tr><td><input type="checkbox"/> D</td><td>A or G or T</td></tr> <tr><td><input type="checkbox"/> B</td><td>C or G or T</td></tr> <tr><td><input type="checkbox"/> N</td><td>G or A or T or C</td></tr> </tbody> </table>	IUPAC code	Meaning	<input type="checkbox"/> A	A	<input type="checkbox"/> C	C	<input type="checkbox"/> G	G	<input type="checkbox"/> T	T	<input type="checkbox"/> M	A or C	<input type="checkbox"/> R	A or G	<input type="checkbox"/> W	A or T	<input type="checkbox"/> S	C or G	<input type="checkbox"/> Y	C or T	<input type="checkbox"/> K	G or T	<input type="checkbox"/> V	A or C or G	<input type="checkbox"/> H	A or C or T	<input type="checkbox"/> D	A or G or T	<input type="checkbox"/> B	C or G or T	<input type="checkbox"/> N	G or A or T or C	Created: clear <input type="checkbox"/> Current Build ID <input type="checkbox"/> Last Build ID CBID Range from <input type="text"/> to <input type="text"/> Updated: clear <input type="checkbox"/> Current Build ID <input type="checkbox"/> Last Build ID UBID Range from <input type="text"/> to <input type="text"/> Validation: clear <input type="checkbox"/> by-cluster <input type="checkbox"/> by-frequency <input type="checkbox"/> by-submitter <input type="checkbox"/> by-2hit-2allele <input type="checkbox"/> no-info
IUPAC code	Meaning																																	
<input type="checkbox"/> A	A																																	
<input type="checkbox"/> C	C																																	
<input type="checkbox"/> G	G																																	
<input type="checkbox"/> T	T																																	
<input type="checkbox"/> M	A or C																																	
<input type="checkbox"/> R	A or G																																	
<input type="checkbox"/> W	A or T																																	
<input type="checkbox"/> S	C or G																																	
<input type="checkbox"/> Y	C or T																																	
<input type="checkbox"/> K	G or T																																	
<input type="checkbox"/> V	A or C or G																																	
<input type="checkbox"/> H	A or C or T																																	
<input type="checkbox"/> D	A or G or T																																	
<input type="checkbox"/> B	C or G or T																																	
<input type="checkbox"/> N	G or A or T or C																																	

ENTREZ SNP
Single Nucleotide Polymorphism

Searching dbSNP

NCBI FieldGuide

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


ENTREZ SNP
Single Nucleotide Polymorphism

Search Mouse SNP between strains

(Genome Build 36.1)


NCBI FieldGuide

Reference Strain 129/Sv 129P4/J 129S1/SvImJ 129S4/SvJae 129S6/SvEvTac 129X1/Sv 129X1/SvJ A A/He A/HeJ	X Different Genotype	Strain(s) 129/Sv 129P4/J 129S1/SvImJ 129S4/SvJae 129S6/SvEvTac 129X1/Sv 129X1/SvJ A A/He A/HeJ
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: <input type="text" value="1"/> (required selection)		
Base Position: from <input type="text"/> to <input type="text"/>		
Send to: <input type="text" value="Entrez Display"/> Format: <input type="text" value="Genotype"/> <input type="button" value="GO"/>		



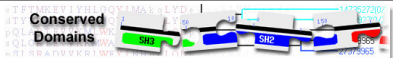
Entrez Databases

- ❑ 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)



Conserved Domain Database

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



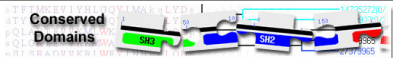
A Conserved Domain Database and Search Service, v2.08

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: 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.



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

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|AA567634.1| ATP7A [Solenodon paradoxus]
IVYQPHLITVEIKKQIKAVGFPPIKPKYKLGADIERLKNIPVKSSECSQQMSPS
STNDSKVTLLDGAIHCNSCSNIESALTIHYYSIVVSLGNSAIKIYNAVYTFEIL
KKAIEAISPGQVRVTSIVTESVTSNPSSSQKAPLNVSQPLTQVTVINMGTCNS
CVQSIQGVMSKKAGVKSIQVSLANRNGTVEYDP LLTSPLEIR
```

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

Query sequence: [gi|45549418|gb|AA567634.1|]

Click on a colored bar to align your sequence to the CD

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

Show Alignment

Format: CompactType 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	3	EFSTY	[1]	.MHCNHCVARIEE	[1]	.GV	[1]	.RYAVV	[3]	.AVVAV	[4]	.VAREVQVIRN	[6]	.66
ICP2_A	69	TIKID	[1] <td>.MHCNHCVARIEE</td> <td>[1] <td>.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.SAIVVY</td> <td>[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.151</td> </td></td></td></td></td>	.MHCNHCVARIEE	[1] <td>.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.SAIVVY</td> <td>[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.151</td> </td></td></td></td>	.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.SAIVVY</td> <td>[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.151</td> </td></td></td>	[1] <td>.SIYVSL</td> <td>[3] <td>.SAIVVY</td> <td>[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.151</td> </td></td>	.SIYVSL	[3] <td>.SAIVVY</td> <td>[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.151</td> </td>	.SAIVVY	[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.151</td>	.PPEILRAIEA	[6]	.151
2AW0	6	VINID	[1] <td>.MHCNHCVARIEE</td> <td>[1] <td>.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.NOTVEYD</td> <td>[4] <td>.SPEILRAIED</td> <td>[6]</td> <td>.69</td> </td></td></td></td></td>	.MHCNHCVARIEE	[1] <td>.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.NOTVEYD</td> <td>[4] <td>.SPEILRAIED</td> <td>[6]</td> <td>.69</td> </td></td></td></td>	.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.NOTVEYD</td> <td>[4] <td>.SPEILRAIED</td> <td>[6]</td> <td>.69</td> </td></td></td>	[1] <td>.SIYVSL</td> <td>[3] <td>.NOTVEYD</td> <td>[4] <td>.SPEILRAIED</td> <td>[6]</td> <td>.69</td> </td></td>	.SIYVSL	[3] <td>.NOTVEYD</td> <td>[4] <td>.SPEILRAIED</td> <td>[6]</td> <td>.69</td> </td>	.NOTVEYD	[4] <td>.SPEILRAIED</td> <td>[6]</td> <td>.69</td>	.SPEILRAIED	[6]	.69
1FE4_A	5	EFSTY	[1] <td>.MTCNSVCQSIQV</td> <td>[1] <td>.GV</td> <td>[1] <td>.YVDIL</td> <td>[3] <td>.RVCTESE</td> <td>[1] <td>.SMDTLATLER</td> <td>[6]</td> <td>.63</td> </td></td></td></td>	.MTCNSVCQSIQV	[1] <td>.GV</td> <td>[1] <td>.YVDIL</td> <td>[3] <td>.RVCTESE</td> <td>[1] <td>.SMDTLATLER</td> <td>[6]</td> <td>.63</td> </td></td></td>	.GV	[1] <td>.YVDIL</td> <td>[3] <td>.RVCTESE</td> <td>[1] <td>.SMDTLATLER</td> <td>[6]</td> <td>.63</td> </td></td>	.YVDIL	[3] <td>.RVCTESE</td> <td>[1] <td>.SMDTLATLER</td> <td>[6]</td> <td>.63</td> </td>	.RVCTESE	[1] <td>.SMDTLATLER</td> <td>[6]</td> <td>.63</td>	.SMDTLATLER	[6]	.63
1K0V_A	5	TIQGE	[1] <td>.MTCNSVCQSIQV</td> <td>[1] <td>.GV <td>[1] <td>.AVRVSL</td> <td>[3] <td>.KVDVSD</td> <td>[4] <td>.SVEIDLADIED</td> <td>[6]</td> <td>.68</td> </td></td></td></td></td>	.MTCNSVCQSIQV	[1] <td>.GV <td>[1] <td>.AVRVSL</td> <td>[3] <td>.KVDVSD</td> <td>[4] <td>.SVEIDLADIED</td> <td>[6]</td> <td>.68</td> </td></td></td></td>	.GV <td>[1] <td>.AVRVSL</td> <td>[3] <td>.KVDVSD</td> <td>[4] <td>.SVEIDLADIED</td> <td>[6]</td> <td>.68</td> </td></td></td>	[1] <td>.AVRVSL</td> <td>[3] <td>.KVDVSD</td> <td>[4] <td>.SVEIDLADIED</td> <td>[6]</td> <td>.68</td> </td></td>	.AVRVSL	[3] <td>.KVDVSD</td> <td>[4] <td>.SVEIDLADIED</td> <td>[6]</td> <td>.68</td> </td>	.KVDVSD	[4] <td>.SVEIDLADIED</td> <td>[6]</td> <td>.68</td>	.SVEIDLADIED	[6]	.68
gi 12229577 156	FLRVE	[1] <td>.MTCNSVCQSIQV</td> <td>[1] <td>.GV <td>[1] <td>.RIYVSL</td> <td>[3] <td>.EAVITVQ</td> <td>[4] <td>.QPEDLRDIED</td> <td>[6]</td> <td>.219</td> </td></td></td></td></td>	.MTCNSVCQSIQV	[1] <td>.GV <td>[1] <td>.RIYVSL</td> <td>[3] <td>.EAVITVQ</td> <td>[4] <td>.QPEDLRDIED</td> <td>[6]</td> <td>.219</td> </td></td></td></td>	.GV <td>[1] <td>.RIYVSL</td> <td>[3] <td>.EAVITVQ</td> <td>[4] <td>.QPEDLRDIED</td> <td>[6]</td> <td>.219</td> </td></td></td>	[1] <td>.RIYVSL</td> <td>[3] <td>.EAVITVQ</td> <td>[4] <td>.QPEDLRDIED</td> <td>[6]</td> <td>.219</td> </td></td>	.RIYVSL	[3] <td>.EAVITVQ</td> <td>[4] <td>.QPEDLRDIED</td> <td>[6]</td> <td>.219</td> </td>	.EAVITVQ	[4] <td>.QPEDLRDIED</td> <td>[6]</td> <td>.219</td>	.QPEDLRDIED	[6]	.219	
gi 12643938 532	FLQIS	[1] <td>.MTCNSVCQSIQV</td> <td>[1] <td>.GV <td>[1] <td>.SVLVAL</td> <td>[3] <td>.KAEVYV</td> <td>[4] <td>.QPLEYAKIVQD</td> <td>[6]</td> <td>.595</td> </td></td></td></td></td>	.MTCNSVCQSIQV	[1] <td>.GV <td>[1] <td>.SVLVAL</td> <td>[3] <td>.KAEVYV</td> <td>[4] <td>.QPLEYAKIVQD</td> <td>[6]</td> <td>.595</td> </td></td></td></td>	.GV <td>[1] <td>.SVLVAL</td> <td>[3] <td>.KAEVYV</td> <td>[4] <td>.QPLEYAKIVQD</td> <td>[6]</td> <td>.595</td> </td></td></td>	[1] <td>.SVLVAL</td> <td>[3] <td>.KAEVYV</td> <td>[4] <td>.QPLEYAKIVQD</td> <td>[6]</td> <td>.595</td> </td></td>	.SVLVAL	[3] <td>.KAEVYV</td> <td>[4] <td>.QPLEYAKIVQD</td> <td>[6]</td> <td>.595</td> </td>	.KAEVYV	[4] <td>.QPLEYAKIVQD</td> <td>[6]</td> <td>.595</td>	.QPLEYAKIVQD	[6]	.595	
gi 12643938 312	HLRVD	[1] <td>.MTCNSVCQSIQV</td> <td>[1] <td>.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.TAVYVY</td> <td>[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.375</td> </td></td></td></td></td>	.MTCNSVCQSIQV	[1] <td>.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.TAVYVY</td> <td>[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.375</td> </td></td></td></td>	.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.TAVYVY</td> <td>[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.375</td> </td></td></td>	[1] <td>.SIYVSL</td> <td>[3] <td>.TAVYVY</td> <td>[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.375</td> </td></td>	.SIYVSL	[3] <td>.TAVYVY</td> <td>[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.375</td> </td>	.TAVYVY	[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.375</td>	.SFGALRAIEA	[6]	.375	
gi 12229551 280	TFPID	[1] <td>.MTCNSVCQSIQV</td> <td>[1] <td>.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.SAIVYV</td> <td>[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.343</td> </td></td></td></td></td>	.MTCNSVCQSIQV	[1] <td>.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.SAIVYV</td> <td>[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.343</td> </td></td></td></td>	.GV <td>[1] <td>.SIYVSL</td> <td>[3] <td>.SAIVYV</td> <td>[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.343</td> </td></td></td>	[1] <td>.SIYVSL</td> <td>[3] <td>.SAIVYV</td> <td>[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.343</td> </td></td>	.SIYVSL	[3] <td>.SAIVYV</td> <td>[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.343</td> </td>	.SAIVYV	[4] <td>.PPEILRAIEA</td> <td>[6]</td> <td>.343</td>	.PPEILRAIEA	[6]	.343	
gi 1703455 260	QLRID	[1] <td>.MHCNHCVARIEE</td> <td>[1] <td>.GV <td>[1] <td>.SIQVSL</td> <td>[3] <td>.TAQVYVY</td> <td>[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.323</td> </td></td></td></td></td>	.MHCNHCVARIEE	[1] <td>.GV <td>[1] <td>.SIQVSL</td> <td>[3] <td>.TAQVYVY</td> <td>[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.323</td> </td></td></td></td>	.GV <td>[1] <td>.SIQVSL</td> <td>[3] <td>.TAQVYVY</td> <td>[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.323</td> </td></td></td>	[1] <td>.SIQVSL</td> <td>[3] <td>.TAQVYVY</td> <td>[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.323</td> </td></td>	.SIQVSL	[3] <td>.TAQVYVY</td> <td>[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.323</td> </td>	.TAQVYVY	[4] <td>.SFGALRAIEA</td> <td>[6]</td> <td>.323</td>	.SFGALRAIEA	[6]	.323	

Full Result

Query sequence: [gi|45549418|gb|AA567634.1|]

ATP7A [Solenodon paradoxus]

Concise Result Full Result Show Search Information

CD Pfam COG

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
pfam0403.HMA, Heavy-metal-associated domain	40498	No	3e-8
COG2808.CopZ, Copper chaperone [Inorganic ion transport and metabolism]	32500	No	1e-7
pfam00403.HMA, Heavy-metal-associated domain	40498	No	1e-7
COG2808.CopZ, Copper chaperone [Inorganic ion transport and metabolism]	32500	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]	32399	Yes	1.0000

Search for similar domain architectures

Conserved Domains

Domain Architecture

Query sequence: [q|45549418|gb|AAS67634.1|]
ATP7A [Solenodon paradoxus]

COG2217, ZnTA, Cation transport ATPase [Inorganic ion transport and metabolism].

Title	Pssmid	Multi-Dom	E-value
hcd00371, HMA, Heavy-metal-associated domain (HMA) is a conserved domain of approximate...	29471	No	7e-11
hcd00371, HMA, Heavy-metal-associated domain (HMA) is a conserved domain of approximate...	29471	No	2e-9
hpfam0403, HMA, Heavy-metal-associated domain.	40496	No	3e-8
hCOG2608, CopZ, Copper chaperone [Inorganic ion transport and metabolism].	32600	No	1e-7
hpfam0403, HMA, Heavy-metal-associated domain.	40496	No	1e-7
hCOG2608, CopZ, Copper chaperone [Inorganic ion transport and metabolism].	32600	No	0.00001
hCOG2217, ZnTA, Cation transport ATPase [Inorganic ion transport and metabolism].	32399	Yes	8e-7
hCOG2217, ZnTA, Cation transport ATPase [Inorganic ion transport and metabolism].	32399	Yes	6.95000e-7

Search for similar domain architectures

Conserved Domains

CDART: Conserved Domain Architecture Retrieval Tool

Query: HMA

Similar domain architectures:

- 997 Sequences: cellulin-organime, Copper-transportin, CuzD, E1-E2_ATPase, Hydrolase
- 6 Sequences: Proteobacteria, Heavy metal detoxi, Pgm_redox
- 69 Sequences: Bacteria, peroxide dismutase
- 3 Sequences: cellulin-organime, copper-transportin, MerT
- 5 Sequences: Flavobacteriaceae, Heavy metal transp
- 880 Sequences: cellulin-organime, HMO1_ATPase, Sot_Du
- 38 Sequences: cellulin-organime, HMO1_ATPase, COG4633
- 51 Sequences: Eubacteria, peroxide dismuta, COG2836
- 4 Sequences: Chloroflexi, Heavy-metal-associ, COG2836

Conserved Domains

CDART: Conserved Domain Architecture Retrieval Tool

Subset by Taxonomy

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... includes: cd00305
- 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... includes: pfam02852
- Pyridine nucleotide-disulphide oxidoreductase, di...

Conserved Domains

Structure:

Show Structure

Program: Cn3D

Drawing: All Atoms

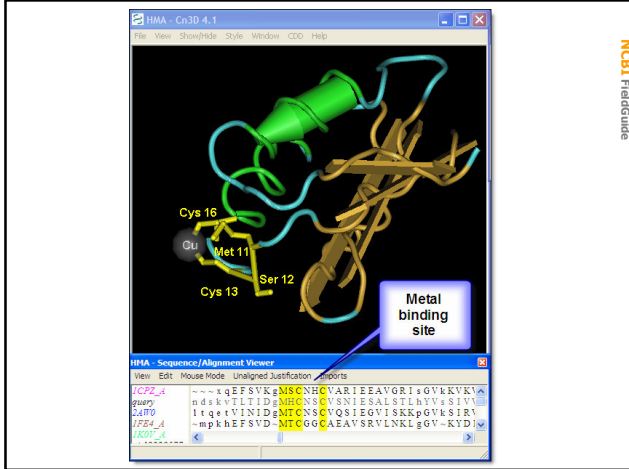
Aligned Rows: up to 10 (download Cn3D)

Format: Compact HyperText Row Display: up to 10 Color Bits: 2.0 bits

Type Selection: the most similar members Feature Display: metal-binding site

```

Feature 1
1027_A 3 EFWK [1] .NCRGQVARIIEA0RI [1] .GV [1] .STPVL [3] .KAVTVD [4] .GATEYQALNE [6] .66
1027_A 68 TLTD [1] .NCRGQVARIIEA0RI [1] .GV [1] .STPVL [3] .KAVTVD [4] .TEELRQALNE [6] .131
1027_A 6 VIND [1] .NCRGQVARIIEA0RI [1] .GV [1] .STPVL [3] .KAVTVD [4] .SPTLRQALNE [6] .69
1027_A 5 EYFD [1] .NCRGQVARIIEA0RI [1] .GV [1] .STPVL [3] .KAVTVD [4] .SPTLRQALNE [6] .65
1027_A 5 TLQV [1] .NCRGQVARIIEA0RI [1] .GV [1] .STPVL [3] .KAVTVD [4] .SPTLRQALNE [6] .68
ql_12229877 136 RLVK [1] .NCRGQVARIIEA0RI [1] .GV [1] .STPVL [3] .KAVTVD [4] .SPTLRQALNE [6] .219
ql_12443938 312 PGRS [1] .NCRGQVARIIEA0RI [1] .GV [1] .STPVL [3] .KAVTVD [4] .SPTLRQALNE [6] .393
ql_12443938 312 RLVK [1] .NCRGQVARIIEA0RI [1] .GV [1] .STPVL [3] .KAVTVD [4] .SPTLRQALNE [6] .378
ql_12229877 280 TLTD [1] .NCRGQVARIIEA0RI [1] .GV [1] .STPVL [3] .KAVTVD [4] .TEELRQALNE [6] .343
ql_1705455 260 QRID [1] .NCRGQVARIIEA0RI [1] .GV [1] .STPVL [3] .KAVTVD [4] .SPTLRQALNE [6] .323
  
```



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

NCBI FieldGuide

Bookshelf

All: 1 | Review: 0

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

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

[Kim SC, Sprung B, 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 for 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 mouse liver mitochondria. In addition to regulators of chromatin-based cellular processes, nonnuclear localized proteins with diverse functions were identified. Most strikingly, acetylysine 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.

Related Links

- Probing lysine acetylation in proteins: strategies, limitations. [Mol Cell Proteomics. 2005]
- Probing lysine acetylation in proteins: a new partnership for signaling. [Bioessays. 2004]
- Purification and functional characterization of SET8, a nucleosomal histone H4. [Curr Biol. 2002]
- AML1 is functionally regulated through p300-mediated acetylation. [J Biol Chem. 2004]
- See all Related Articles...

Links

- Books
- LinkOut

NCBI FieldGuide

BOOKS Database: Hyperlinked Terms

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, **acetylysine** 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

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

BIOCHEMISTRY
 FIFTH EDITION
 Jeremy M. Berg John L. Tymoczko Lubert Stryer
[Short Contents](#) | [Full Contents](#)

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 acetylsine. An acetylated peptide of histone H4 is bound in the structure shown.

For More Information...

The NCBI Handbook The National Library of Medicine
[Short Contents](#) | [Full Contents](#) [Other books @ NCBI](#)

The NCBI Handbook

Part 1. The Databases

1. [GenBank: The Nucleotide Sequence Database](#)
 Ilene Mizrahi.
 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

Search

This book All books PubMed



Intermission