NCBI Mouse Genome Resources

Peter Cooper
National Center for Biotechnology Information

NCBI

Talk Outline

- About NCBI
- Molecular Databases
  - GenBank
  - RefSeq
- Web Access
  - Entrez
  - BLAST
  - Mouse Genome Resources

NCBI

The National Center for Biotechnology Information (NCBI)

- Created as a part of NLM in 1988
  - Establish public databases
  - Research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information
- Tools: BLAST(1990), Entrez (1992)
- GenBank (1992)
- Free MEDLINE (PubMed, 1997)
- Human genome (2001)

NCBI

Molecular Databases

- Primary Databases
  - Original submissions by experimentalists
  - Database staff organize but don't add additional information
    - **Example: GenBank**
- Derivative Databases
  - Human curated
    - compilation and correction of data
    - **Example: SWISS-PROT, NCBI RefSeq mRNA**
  - Computationally Derived
    - **Example: UniGene**
  - Combinations
    - **Example: NCBI Genome Assembly**

NCBI

GenBank: NCBI's Primary Sequence Database

Release 130	June 2002
17,471,130	Records
20,648,748,345	Nucleotides
140,000 +	Species

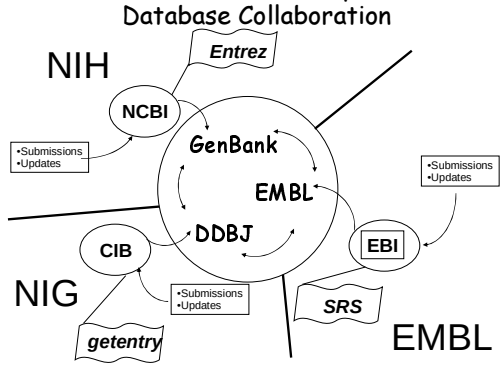
- full release every two months
- incremental and cumulative updates daily
- available only through internet

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

78.75 Gigabytes of data

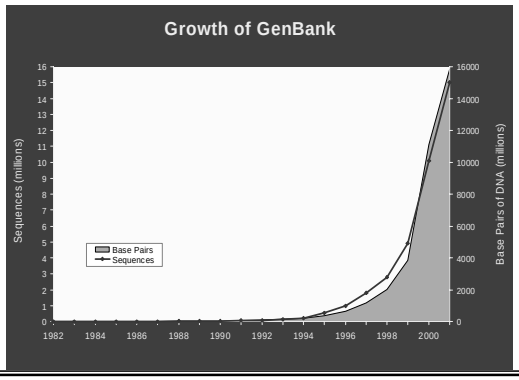
NCBI

The International Sequence Database Collaboration



NCBI

Growth of GenBank



NCBI

GenBank Divisions

Bulk Sequence Divisions

PAT	Patent
EST	Expressed Sequence Tags (142 files)
STS	Sequence Tagged Sites
GSS	Genome Survey Sequences (48 files)
HTG	High Throughput Genome (26 files)
HTC	High Throughput cDNA
CON	Contig

Traditional Divisions

BCT INV MAM PHG PLN PRI
ROD SYN UNA VRL VRT

NCBI

Traditional GenBank Divisions

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

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

NCBI

Bulk GenBank Divisions

- Batch Submission and htg (email and ftp)
- Inaccurate
- Poorly Characterized

EST	Expressed Sequence Tag
STS	Sequence Tagged Site
GSS	Genome Survey Sequence
HTG	High Throughput Genomic

NCBI

EST Division: Expressed Sequence Tags

```
>IMAGE:275615 5' mRNA sequence
GACAGCATTCGGGCCGAGATGTCTCGCTCCGTGGCCTTAGCTGTGCTCGCGCTACTCTCTCTTTCTGGCC
TGGAGGTATCCAGCGTACTCCAAAGATTCAGGTTTACTCACGTCATCCAGCAGAGAATGGAAAGTCAAAT
TTCCTGAATTGCTATGTGTCTGGGTTTCATCCATCCGACATTGAAGTTGACTTACTGAAGAATGGAGAGA
GAATTGAAAAAGTGGAGCATTCAGACTTGTCTTTCAGCAAGGACTGGTCTTTCTATCTCTTGTACTACAC
TGAATTCACCCCCACTGAAAAAGATGAGTATGCCTGCCGTGTTGAACCATGTNGACTTTGTCACAGNCCC
AAGTTNAGTTTAAGTGGGNATCGAGACATGTAAGGCAGGCATCATGGGAGGTTTTGAAGNATGCCGCNTT
TTGGATTGGGATGAATTCCAAATTTCTGGTTTGCTTGNTTTTTTAATATTGGATATGCTTTTG
```

```
>IMAGE:275615 3', mRNA sequence
NNTCAAGTTTTATGATTTATTTAACTTGTGGAACAAAAATAAACCAGATTAACCACAACCATGCCTTACT
TTATCAAATGTATAAGANGTAAATATGAATCTTATATGACAAAATGTTTCATTCATTATAACAAATTTCC
AATAATCCTGTCAATNATATTTCTAAATTTTCCCCCAAATTCTAAGCAGAGTATGTAAATTGGAAGTTAA
CTTATGCACGCTTAACTATCTTAACAAGCTTTGAGTGCAAGAGATTGANGAGTTCAAATCTGACCAAGAT
GTTGATGTTGGATAAGAGAATTCTCTGCTCCCCACCTCTANGTTGCCAGCCCTC
```

NCBI

make cDNA library

80-100,000 unique cDNA clones in library

What is UniGene?

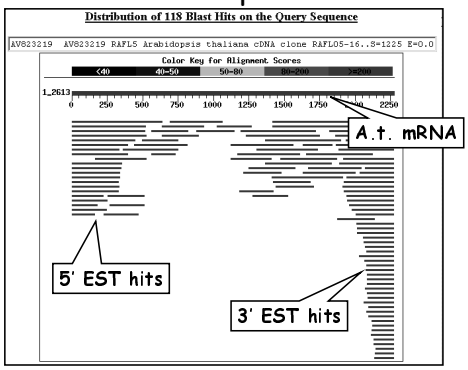
A gene-oriented view of sequence entries

- MegaBlast based automated sequence clustering
- Nonredundant set of gene oriented clusters
- Each cluster a unique gene
- Information on tissue types and map locations
- Includes well-characterized genes and novel ESTs
- Useful for gene discovery and selection of mapping reagents

http://www.ncbi.nlm.nih.gov/UniGene/

NCBI

EST hits A.t. serine protease mRNA



NCBI

Arabidopsis UniGene Statistics

UniGene Build 14
Apr. 9th, 2002

```
    39,855     mRNAs + gene CDSs
    87,006     EST, 3'reads
    42,137     EST, 5'reads
+   32,571     EST, other/unknown
----------
   201,569    total sequences in clusters

Final Number of Clusters (sets)
===============================
26,808   sets total
```

115,000,000 bp
25,498 expected genes
5% uncharacterized transcripts

```
25,474   s                    t one known gene
17,654   s                    t one EST
16,326   sets contain both genes and ESTs
```

NCBI

Mm UniGene Statistics

UniGene Build 112
Jul. 25th, 2002

```
    46,745     mRNAs + gene CDSs
 1,339,208     EST, 3'reads
    89,928     EST, 5'reads
+   51,328     EST, other/unknown
----------
 2,332,864    total sequences in clusters

Final Number of Clusters (sets)
===============================
84,247   sets total
```

3,000,000 base pairs
30 K expected genes
80% uncharacterized transcripts

```
17,723   s                    one known gene
82,866   s                    one EST
16,342   sets contain both genes and ESTs
```

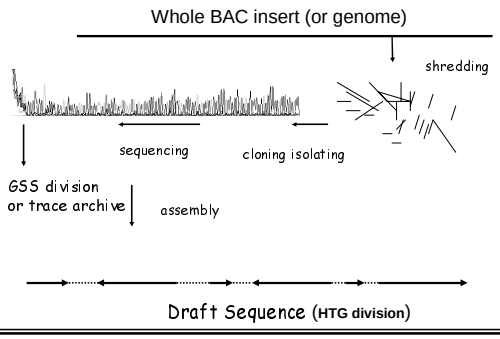
NCBI

UniGene Collections *Jul, 2002*

		Sequences	Clusters
Homo sapiens	human	3,569,546	101,602
Mus musculus	**mouse**	**2,332, 864**	**84,247**
Rattus norvegicus	rat	334,582	62,220
Danio rerio	zebrafish	197,266	15,404
Bos taurus	cow	128,914	10,295
Xenopus laevis	frog	162,269	18,984
D.melanogaster	fruit fly	250,655	11,115
Anopholes gambiae	mosquito	43,126	2,556
Plants			
Arabidopsis thaliana	thale cress	210,693	26,875
Oryzia sativa	rice	78,632	15,802
Triticum aestivum	wheat	139,447	12,575
Hordeum vulgare	barley	160,518	7,324
Zea mays	maize (corn)	131,668	10,301

NCBI

Genome Sequencing



NCBI

HTG Division: High Throughput Genome

```
LOCUS       AC109609               154774 bp    DNA     linear   HTG 06-FEB-2002
DEFINITION  Mus musculus clone rp23-167h1, WORKING DRAFT SEQUENCE, 19 unordered
            pieces.
ACCESSION   AC109609
VERSION     AC109609.1  GI:18543009
KEYWORDS    HTG; HTGS_PHASE1; HTGS_DRAFT.
```

```
LOCUS       AC109609               167074 bp    DNA     linear   HTG 07-MAY-2002
DEFINITION  Mus musculus clone rp23-167h1 strain C57BL/6J, WORKING DRAFT
            SEQUENCE, 3 ordered pieces.
ACCESSION   AC109609
VERSION     AC109609.6  GI:20087170
KEYWORDS    HTG; HTGS_PHASE2; HTGS_DRAFT.
```

```
LOCUS       AC109609               167617 bp    DNA     linear   ROD 10-AUG-2002
DEFINITION  Mus musculus chromosome 17 clone rp23-167h1 strain C57BL/6J,
            complete sequence.
ACCESSION   AC109609
VERSION     AC109609.10  GI:22203259
KEYWORDS    HTG.
```

NCBI

40,000 to > 350,000 bp

Mouse Whole Genome Shotgun

```
LOCUS       CAAA01000000          224713 rc    DNA     linear   ROD 15-MAY-2002
DEFINITION  Mus musculus whole genome shotgun sequencing project.
ACCESSION   CAAA00000000
VERSION     CAAA00000000.1  GI:20800445
KEYWORDS    WGS.
SOURCE      house mouse.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1
  AUTHORS   The Mouse Genome Sequencing,Consortium.
  TITLE     Direct Submission
  JOURNAL   Submitted (01-MAY-2002) The Mouse Genome Sequencing Consortium,
            http://mouse.ensembl.org/
COMMENT     The Mus musculus whole genome shotgun (WGS) project has the project
            accession CAAA00000000.  This version of the project (01) has the
            accession number CAAA01000000, and consists of sequences
            CAAA01000001-CAAA01224713.
FEATURES             Location/Qualifiers
     source          1..224713
                     /organism="Mus musculus"
                     /db_xref="taxon:10090"
WGS         CAAA01000001-CAAA01224713
//
```

NCBI

Celera Whole Genome Shotgun

```
LOCUS       AAAD01000000             20 rc    DNA     linear   ROD 31-MAY-2002
DEFINITION  Mus musculus chromosome 16 whole genome shotgun sequencing project.
ACCESSION   AAAD00000000
VERSION     AAAD00000000.1  GI:21281728
KEYWORDS    WGS.
SOURCE      house mouse.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 20)
  AUTHORS   Mural,R.J.
            Wides,R.,
            Salzberg,S
            Evangelist
            Milshina,N
            Wang,X., Y
            Biddick,K.
            Flanigan,M
            Reinert,K.
            Lai,Z., Lo
            Wang,Z.Y.,
            Wortman,J.
            Johnson,J.
            Chaturvedi,K., Gabriellan,A.E., Ke,Z., Sun,J., Subramanian,G. and
            Venter,J.C.
  TITLE     A Comparison of Whole-Genome Shotgun-Derived Mouse Chromosome 16
            and the Human Genome
  JOURNAL   Science 296 (5573), 1661-1671 (2002)
   PUBMED   12040188
REFERENCE   2  (bases 1 to 20)
  AUTHORS   Mural,R.J., Adams,M.D., Myers,E.W., Smith,H.O. and Venter,C.J.
  TITLE     Direct Submission
  JOURNAL   Submitted (24-MAY-2002) Celera Genomics, 45 W. Gude Dr., Rockville,
            MD 20850, USA
COMMENT     The Mus musculus whole genome shotgun (WGS) project has the project
            accession AAAD00000000.  This version of the project (01) has the
            accession number AAAD01000000, and consists of sequences
            AAAD01000001-AAAD01000020.
            The strings of n's in a record repcesent gaps between contigs, and
            the length of each string corresponds to the length of the gap.
FEATURES             Location/Qualifiers
     source          1..20
                     /organism="Mus musculus"
                     /db_xref="taxon:10090"
                     /chromosome="16"
WGS         AAAD01000001-AAAD01000020
//
```

NCBI

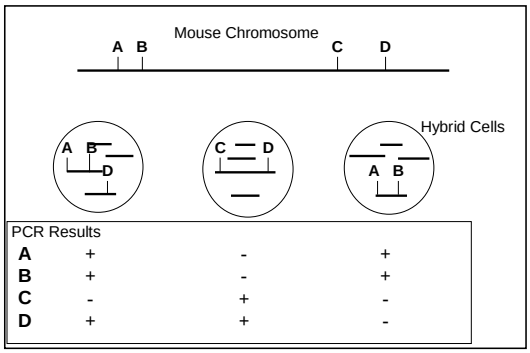
STS Division : Sequence Tagged Sites

- Segment of gene, EST , mRNA or genomic DNA of known position (microsatellite)
- PCR with STS primers gives one product per genome
- Basis of Radiation Hybrid Mapping
  - UniGene
  - Genome Assembly
- Related resource: Electronic PCR

  http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi

NCBI

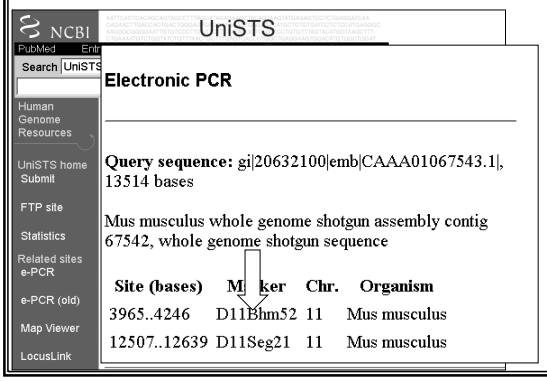
RH mapping using STSs



PCR Results			
A	+	-	+
B	+	-	+
C	-	+	-
D	+	+	-

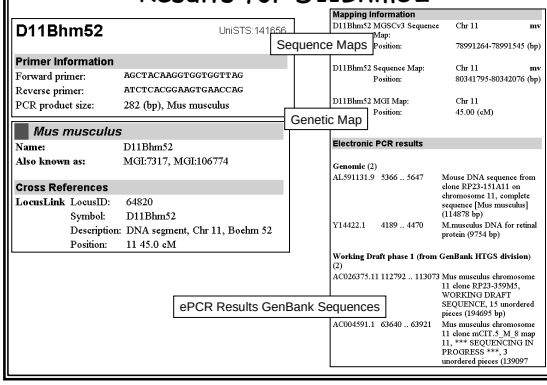
NCBI

Electronic PCR



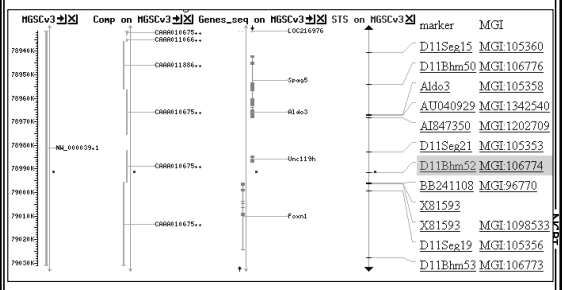
NCBI

Results for D11Bhm52



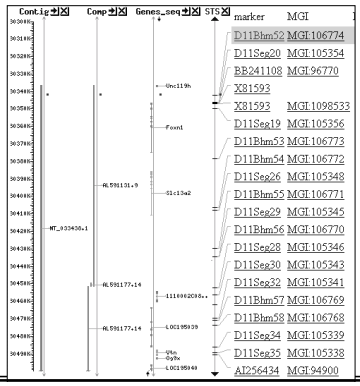
NCBI

Markers on MGSCv3



NCBI

Markers on Finished Contigs



NCBI

RefSeq: NCBI's Derivative Sequence Database

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

NCBI

The RefSeq Accession Numbers

NCBI Reference Sequences

mRNAs and Proteins

NM_123456 Curated mRNA
NP_123456 Curated Protein
XM_123456 Predicted Transcript (human, mouse)
XP_123456 Predicted Protein (human, mouse)

human
mouse
rat
fruit fly
zebrafish
Arabidopsis
Microbial

Gene Records

NG_123456 Reference Genomic Sequence (human)

Assemblies

NT_123456 Contig (Mouse and Human)
NW_123456 WGS Supercontig (Mouse)
NC_123455 Chromosome (Microbial, Arabidopsis)

NCBI

GenBank Sequences: Human Lipoprotein Lipase

Show: 50 Items 1-38 of 38 One page.

1: AK092286
Homo sapiens cDNA FLJ349
LIPOPROTEIN LIPASE PR
gi|21750843|dbj|AK092286.1|

2: AY092404
Homo sapiens lipoprotein lipa
gi|21309830|gb|AY092404.1|

3: AY092405
Homo sapiens lipoprotein lipa
gi|21309828|gb|AY092405.1|

4: BC011353
Homo sapiens, Similar to lipo
mRNA, complete cds
gi|15030192|gb|BC011353.1|

5: S75964
lipoprotein lipase {exon 9}
gi|242490|gb|S75964.1|S75964

6: D83550
Human DNA for lipoprotein l
gi|1695684|dbj|D83550.1|

7: D83549
Human DNA for lipoprotein l

29: S71706 OMIM, Protein, PubMed, Taxonomy
lipoprotein lipase {exon 5} [human, Genomic Mutant, 18 nt]
gi|240930|gb|S71706.1|S71706[240930]

30: S71710 OMIM, Protein, PubMed, Taxonomy
lipoprotein lipase {exon 6} [human, Genomic Mutant, 18 nt]
gi|240928|gb|S71710.1|S71710[240928]

31: S58235 MapView, Related Sequences, PubMed, SNP, Taxonomy
lipoprotein lipase {5' flanking region, exon 1} [human, Genomic, 200 nt]
gi|236815|gb|S58235.1|S58235[236815]

32: X14390 MapView, Related Sequences, OMIM, Protein, PubMed, SNP, Taxonomy, UniSTS
Human mRNA for lipoprotein lipase (EC 3.1.1.34)
gi|34404|emb|X14390.1|HSLPLR[34404]

33: X52978 MapView, Related Sequences, PubMed, SNP, Taxonomy, UniSTS
H.sapiens gene for lipoprotein lipase
gi|34386|emb|X52978.1|HSLIPASE[34386]

34: X54516 MapView, Related Sequences, OMIM, Protein, PubMed, SNP, Taxonomy
Human mRNA for lipoprotein lipase (EC 3.1.1.34)
gi|34382|emb|X54516.1|HSLIPAS[34382]

35: M94221 Taxonomy, UniSTS
Human GT microsatellite DNA flanking the lipoprotein lipase (LPL) gene
gi|187217|gb|M94221.1|HUMLPLMUTA[187217]

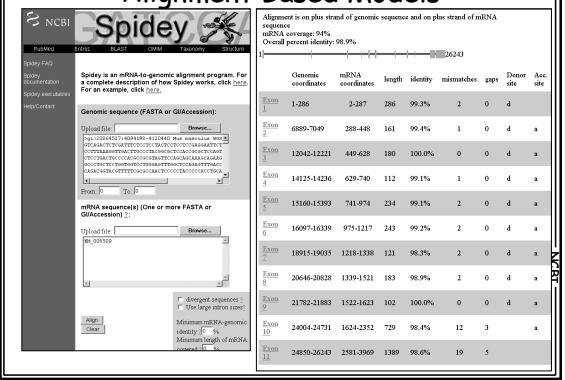
NCBI

Curated RefSeq Records: NM_, NP_

```
LOCUS       Lpl                     3969 bp    mRNA    linear   ROD 07-JAN-2002
DEFINITION  Mus musculus lipoprotein lipase (Lpl), mRNA.
ACCESSION   NM_008509
VERSION     NM_008509.1  GI:6678709
KEYWORDS    .
SOURCE      Mus musculus.
  CDS             200..1624
                  /gene="Lpl"
                  /EC_number="3.1.1.34"
                  /codon_start=1
                  /product="lipoprotein lipase"
                  /protein_id="NP_032535.1"
                  /db_xref="GI:6678710"
                  /db_xref="LocusID:16956"
                  /db_xref="MGD:96820"
                  /translation="MESKALLLVVLGVWLQSLTAFRGGVAAADAGRDFSDIESKFALR
                  TPEDTAEDTCHLIPGLADSVSNCHFNHSSKTFVVIHGWTVTGMYESWVPKLVAALYKR
                  EPDSNVIVVDWLYRAQQHYPVSAGYTKLVGNDVARFINWMEEEFKYPLDNVHLLGYSL
                  GAHAGVAGSLTNKKVNRITGLDPAGPNFEYAEAPSRLSPDDADFVDVLHTFTRGSPG
                  RSIGIQKPVGHVDIYPNGGTFQPGCNIGEAIRVIAERGLGDVDQLVKCSHERSIHLFI
                  DSLLNEENPSKAYRCNSKEAFEKGLCLSCRKNRCNNLGYEINKVRAKRSSKMYLKTRS
                  QMPYKVFHYQVKIHFSGTENGKQHNQAFEISLYGTVAESENIPFTLPEVSTNKTYSFL
                  IYTEVDIGELLMMKLKWMSDSYFSWPDWWSSPSFVIERIRVKAGETQKKVIFCAREKV
                  Severe mutations that cause LPL deficiency result in type I
                  hyperlipoproteinemia, while less extreme mutations in LPL are
                  linked to many disorders of lipoprotein metabolism.
                  analysis
   JOURNAL    Gene 107 (2), 247-258 (1991)
   MEDLINE    92084117
    PUBMED    1748295
REFERENCE   2  (bases 1 to 3969)
```

NCBI

Alignment Based Models



Alignment is on plus strand of genomic sequence and on plus strand of mRNA sequence
mRNA coverage: 94%
Overall percent identity: 98.9%

1 — 26243

	Genomic coordinates	mRNA coordinates	length	identity	mismatches	gaps	Donor site	Acc. site
Exon 1	1-286	2-287	286	99.3%	2	0	d	
Exon 2	6889-7049	288-448	161	99.4%	1	0	d	a
Exon 3	12042-12221	449-628	180	100.0%	0	0	d	a
Exon 4	14125-14236	629-740	112	99.1%	1	0	d	a
Exon 5	15160-15393	741-974	234	99.1%	2	0	d	a
Exon 6	16097-16339	975-1217	243	99.2%	2	0	d	a
Exon 7	18915-19035	1218-1338	121	98.3%	2	0	d	a
Exon 8	20646-20828	1339-1521	183	98.9%	2	0	d	a
Exon 9	21782-21883	1522-1623	102	100.0%	0	0	d	a
Exon 10	24004-24731	1624-2352	729	98.4%	12	3		a
Exon 11	24850-26243	2581-3969	1389	98.6%	19	5		

NCBI

Alignment Based Models

Exon 7: 18915-19035 (genomic); 1218-1338 (mRNA)

```
18915      TTCCCTACAGTGTTCCATTACCAAGTCAAGATTCACTTTTCTGGGACTGA
                        ||||||||||||||||||||||||||||||||||||||||||||
1218                    TGTTCCATTACCAAGTCAAGATTCACTTTTCTGGGACTGA
                        V  F  H  Y  Q  V  K  I  H  F  S  G  T  E

18955      GGATGGCAAGCAACACAACCAGGCCTTCGAGATTTCTCTGTACGGCACAG
           | |||||||||||||||||||||||||||||| |||||||||||||||||||
1258       GAATGGCAAGCAACACAACCAGGCCTTCGAAATTTCTCTGTACGGCACAG
             N  G  K  Q  H  N  Q  A  F  E  I  S  L  Y  G  T
```

AA change

```
19005      TGGCCGAGAGCGAGAACATTCCCTTCACCCTGTGAGTATCT
           |||||||||||||||||||||||||||||||||
1308       TGGCCGAGAGCGAGAACATTCCCTTCACCCT
           V  A  E  S  E  N  I  P  F  T  L
```

NCBI

Alignment Generated Transcripts: XM_, XP_

```
LOCUS       Lpl                     3862 bp    mRNA    linear   ROD 16-MAY-2002
DEFINITION  Mus musculus lipoprotein lipase (Lpl), mRNA.
ACCESSION   XM_134193
VERSION     XM_134193.1  GI:20864361
KEYWORDS    .
SOURCE      Mus musculus.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 3862)
  AUTHORS   NCBI Annotation Project.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-MAY-2002) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
COMMENT     GENOME ANNOTATION REFSEQ:  This model reference sequence was
            predicted from NCBI contig NW_000344 by automated computational
            analysis using gene prediction method: BLAST, supported by mRNA
            evidence.
            Also see:
                Documentation of NCBI's Annotation Process

FEATURES             Location/Qualifiers
     source          1..3862
                     /organism="Mus musculus"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="8"
     gene            1..3862
                     /gene="Lpl"
                     /db_xref="LocusID:16956"
                     /db_xref="MGD:96820"
     CDS             199..1623
                     /gene="Lpl"
                     /codon_start=1
```

NCBI

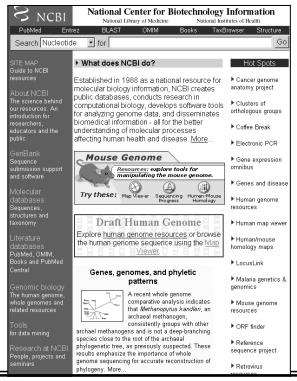
RefSeq Contig: NT_, NW_

```
LOCUS       NT_007933          64386157 bp    DNA     linear   CON 01-AUG-2002
DEFINITION  Homo sapiens chromosome 7 reference genomic contig.
ACCESSION   NT_007933
VERSION     NT_007933.10  GI:22050621
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chor
            Mammalia; Eutheria; Prim
REFERENCE   1  (bases 1 to 64386157)
  AUTHORS   NCBI Annotation Project.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-JUL-2002)
            Information, NIH, Bethes
COMMENT     GENOME ANNOTATION REFSEQ:
            genomic sequence data. T
            sequence.
            On Aug 1, 2002 this sequ
            COMPLETENESS: not full l
FEATURES             Location/Qualifi
     source          1..64386157
                     /organism="Homo
                     /db_xref="taxon
                     /chromosome="7"
     gene            230..24397
                     /gene="MESCR16"
                     /note="DKFZp434
                     /db_xref="Locus
     mRNA            join(230..298,1
                     10510..10683,13
                     /gene="MESCR16"
                     /note="Derived
                     gene prediction
                     evidence includ
                     /transcript_id=
                     /db_xref="GI:22
                     /db_xref="Locus
     CDS             join(240..298,1
```

```
LOCUS       NW_000272          27755676 bp    DNA     linear   CON 04-JUN-2002
DEFINITION  Mus musculus WGS supercontig Mm6_WIFeb01_97.
ACCESSION   NW_000272
VERSION     NW_000272.1  GI:20915425
KEYWORDS    .
SOURCE      house mouse.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 27755676)
  AUTHORS   NCBI Annotation Project.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-MAY-2002) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
COMMENT     GENOME ANNOTATION REFSEQ:  NCBI contigs are derived from assembled
            genomic sequence data. They may include both draft and finished
            sequence.
FEATURES             Location/Qualifiers
     source          1..27755676
                     /organism="Mus musculus"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="6"
     gene            6728..13494
                     /gene="LOC243722"
     mRNA            /db_xref="InterimID:243722"
                     join(6728..6872,7967..8218,11029..11098,11202..11335,
                     11368..11503,11641..11766,12089..12225,12970..13494)
                     /gene="LOC243722"
                     /product="similar to nuclear matrix transcription factor"
                     /note="Derived by automated computational analysis using
                     gene prediction method: GenomeScan.*
                     /transcript_id="XM_145115.1"
                     /db_xref="GI:20915205"
                     /db_xref="InterimID:243722"
     CDS             join(6786..6872,7967..8218,11029..11098,11202..11335,
                     11368..11503,11641..11766,12089..12225,12970..13494)
                     /gene="LOC243722"
                     /codon_start=1
```

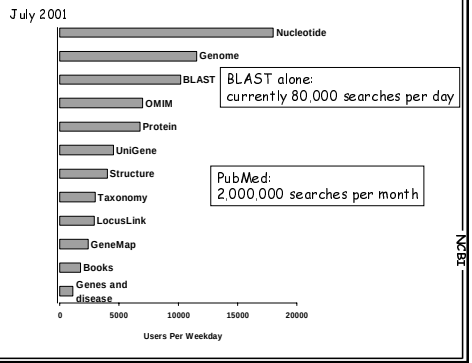
NCBI

Integrated WWW Access: BLAST and Entrez



NCBI

Some Web Statistics

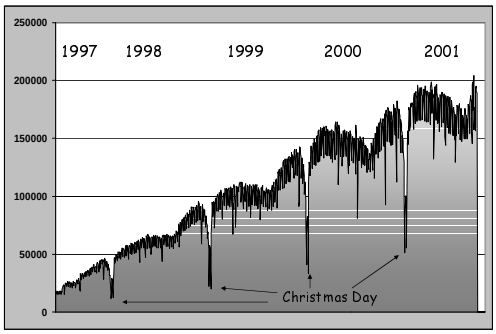


BLAST alone:
currently 80,000 searches per day

PubMed:
2,000,000 searches per month

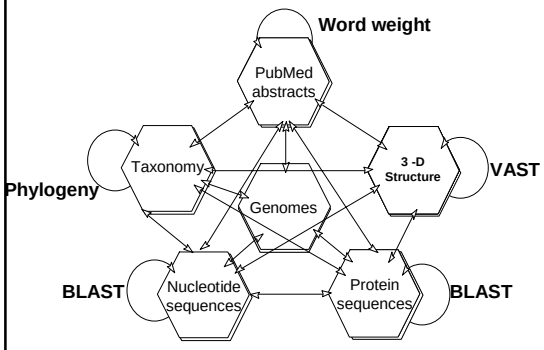
NCBI

Users per day

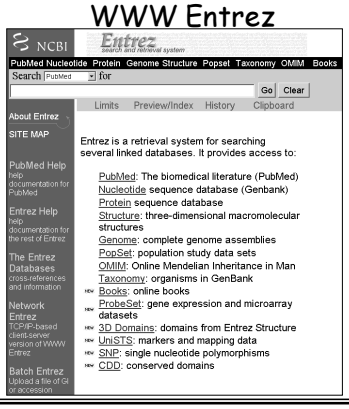


NCBI

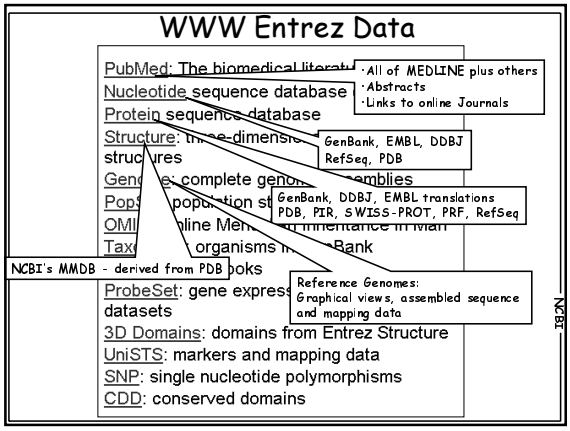
Entrez: Database Integration

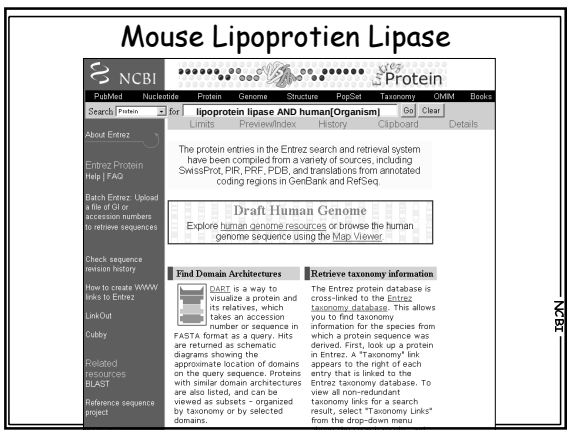


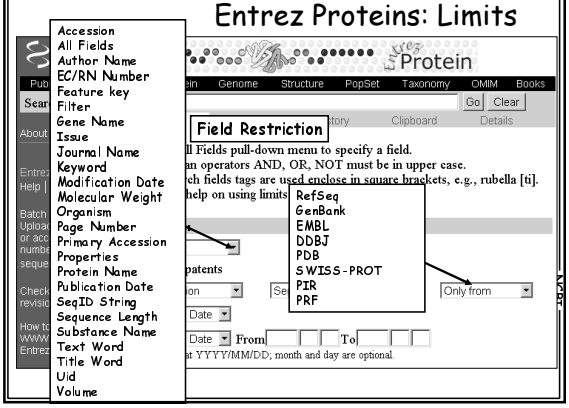
WWW Entrez



NCBI


WWW Entrez Data
PubMed: The biomedical literatu
Nucleotide sequence database
Protein sequence database
Structure: three-dimensio
structures
Genome: complete genom
Pops: population st
OMI: online Men
Taxo: organisms in GenBank
ProbeSet: gene express
datasets
3D Domains: domains from Entrez Structure
UniSTS: markers and mapping data
SNP: single nucleotide polymorphisms
CDD: conserved domains
·All of MEDLINE plus others
·Abstracts
·Links to online Journals
GenBank, EMBL, DDBJ
RefSeq, PDB
GenBank, DDBJ, EMBL translations
PDB, PIR, SWISS-PROT, PRF, RefSeq
NCBI's MMDB - derived from PDB
Reference Genomes:
Graphical views, assembled sequence
and mapping data
NCBI


Mouse Lipoprotien Lipase
NCBI
Protein
lipoprotein lipase AND human[Organism]
Draft Human Genome
Find Domain Architectures
Retrieve taxonomy information
NCBI


Entrez Proteins: Limits
Accession
All Fields
Author Name
EC/RN Number
Feature key
Filter
Gene Name
Issue
Journal Name
Keyword
Modification Date
Molecular Weight
Organism
Page Number
Primary Accession
Properties
Protein Name
Publication Date
SeqID String
Sequence Length
Substance Name
Text Word
Title Word
Uid
Volume
Field Restriction
RefSeq
GenBank
EMBL
DDBJ
PDB
SWISS-PROT
PIR
PRF
Protein

Entrez Nucleotides: Limits

Limited to:

All Fields

Exclude bulk sequences

☐ exclude ESTs ☐ exclude STSs ☐ exclude GSS ☐ exclude working draft
☐ exclude patents ☐ exclude all of the above

Molecule | Gene Location | Segmented Sequences

Only from | Modification Date

Modification Date From To

Use the format YYYY/MM/DD; month and day are optional.

NCBI

Advanced Search:Preview/Index

Add Term(s) to Query or View Index:

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

Organism | mouse | Preview | Index

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

mouse(125052)
mouse/rat hybrid cell lines(3)
mouse/rat neuroblastoma x glioma hybrid ng108 15(3)
mouse/rat ng108 15(3)
mouse adenovirus(65)
mouse adenovirus 1(65)
mouse adenovirus type 1(65)
mouse cytomegalovirus 1(206)
mouse ear cress(94329)
mouse eared bat(9)

Up | Down

NCBI

Lipoprotein lipase records

☐ 1: AAH03305 BLink, Domains, Nucleotide, Related Sequences, Domain Relatives, Taxonomy
lipoprotein lipase [Mus musculus]
gi|13097036|gb|AAH03305.1|[13097036]

☐ 2: P11152 BLink, Domains, OMIM, Related Sequences, Domain Relatives, PubMed, Taxonomy
Lipoprotein lipase precursor (LPL)
gi|417255|sp|P11152|LIPL_MOUSE[417255]

☐ 3: XP_134193 BLink, Domains, Nucleotide, Related Sequences, Domain Relatives, Taxonomy
lipoprotein lipase [Mus musculus]
gi|20864362|ref|XP_134193.1|[20864362]

☐ 4: NP_032535 BLink, Domains, Nucleotide, OMIM, Related Sequences, Domain Relatives, PubMed, Taxonomy
lipoprotein lipase [Mus musculus]
gi|6678710|ref|NP_032535.1|[6678710]

☐ 5: A40570 BLink, Domains, OMIM, Related Sequences, Domain Relatives, PubMed, Taxonomy
lipoprotein lipase (EC 3.1.1.34) precursor - mouse
gi|110647|pir||A40570[110647]

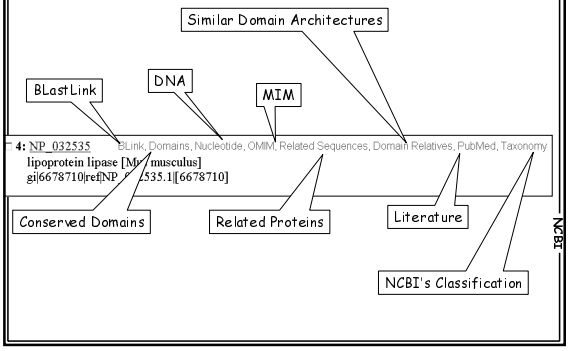
☐ 6: S17503 BLink, Domains, Related Sequences, Domain Relatives, PubMed, Taxonomy
lipoprotein lipase precursor - mouse
gi|110648|pir||S17503[110648]

☐ 7: AAC04464 BLink, Nucleotide, Related Sequences, PubMed, Taxonomy
lipoprotein lipase [Mus musculus]
gi|2911254|gb|AAC04464.1|[2911254]

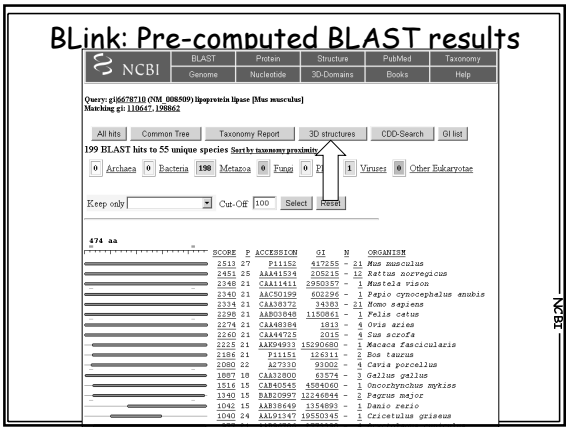
☐ 8: CAA41329 BLink, Domains, Nucleotide, Related Sequences, Domain Relatives, PubMed, Taxonomy
lipoprotein lipase [Mus musculus]
gi|51468|emb|CAA41329.1|[51468]

NCBI

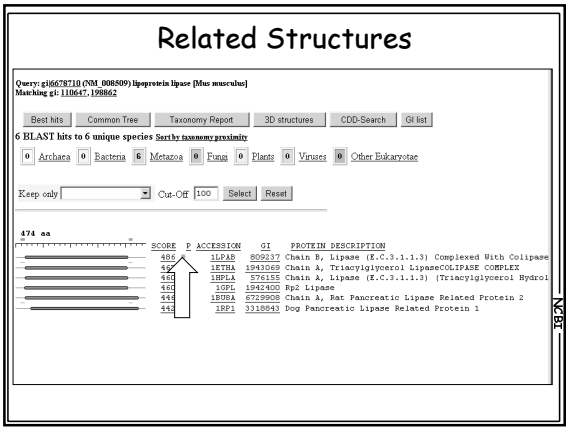
Links and Neighbors lp1



BLink: Pre-computed BLAST results



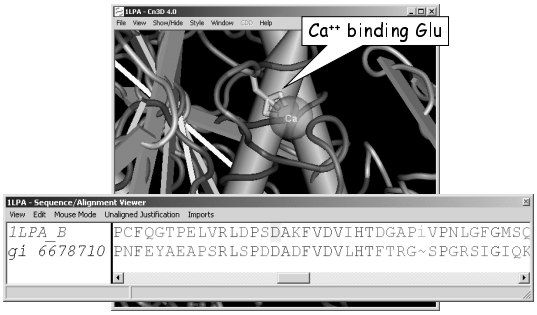
Related Structures



lp1 Modeling Template

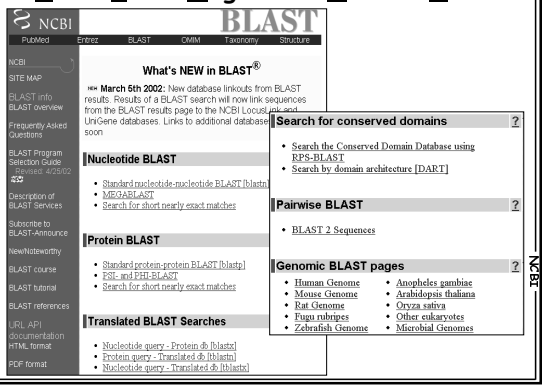
NCBI

lp1 Ca Binding Site



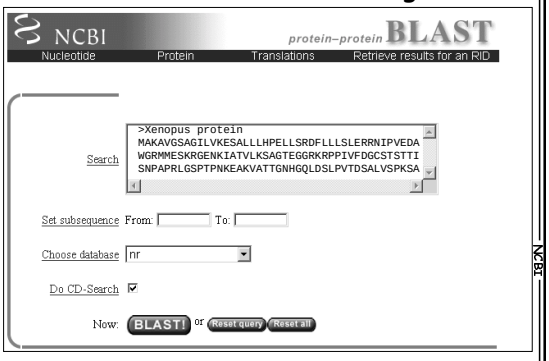
NCBI

Basic Local Alignment Search Tool



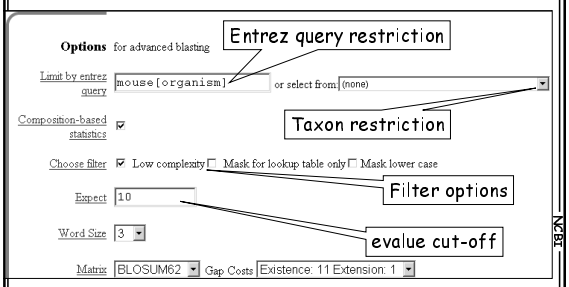
NCBI

The Standard BLAST Pages



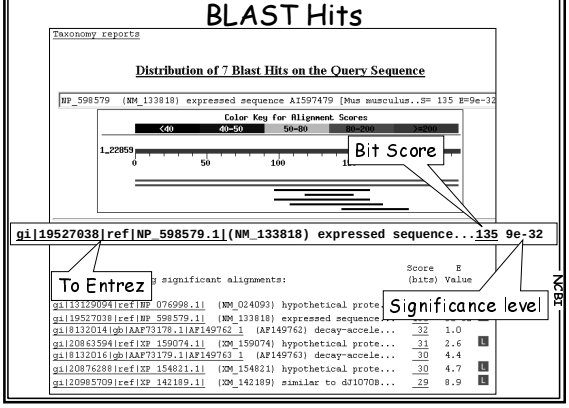
NCBI

Advanced Options



NCBI

BLAST Hits



NCBI

BLAST Alignments

```
>gi|19527038|ref|NP_598579.1|  L (NM_133818) expressed sequence AI597479 [Mus musculus]
 gi|20846094|ref|XP_123625.1|  L (XM_123625) expressed sequence AI597479 [Mus musculus]
 gi|13905264|gb|AAH06931.1|  L (BC006931) Similar to hypothetical protein MGC5509 [Mus
          Length = 232

 Score =  135 bits (340), Expect = 9e-32
 Identities = 100/237 (42%), Positives = 130/237 (54%), Gaps = 16/237 (6%)

Query: 1   MAKAVGSAGILVKESALLLHPXXXXXXXXXXXXXXXXNIPVEDALN-NKEKLTEIFVQHAM 59
           MA  VG  G    ++ LLLHP                NI VE+ +  NK+ LT+++VQHA+
Sbjct: 1   MAGDVG--GRSCTDAELLLHPELLSQEFI
Query: 60  PLPQRDLPRNRWGRMMESKRGENKIATVLK-SAGTEGGRKRPPIVFDGCSTSTTIKVKRT 118
           PLPQRDLP+NRWG+MME KR  +++     K S+  +G RKRP IVFDG STST+IKVKRT
Sbjct: 59  PLPQRDLPKNRWGKMMEKKREHHEVKNDTKK-  -GGLRKRPLIVFDGSSTSTSIKVKRT 118

Query: 119 ENGDASQRLNPSPK---------ESTNSTQ                                169
           ENG A  RL P  +          +S NS+   +P      +   N   K+    N  Q +
Sbjct: 119 ENG-ADDRLKPLP  IGSTSDAFWKSPNSSSRISPLVLFSNLPVNH--KMEHNNNDTQQNH 175

Query: 170 LPVTDSAL                KLQQSMIINAVKLKRNAPKEEFDVANDTKPTEAKKKIQHVTWP 226
             +   +   P  +     +    VKLKR APKEE +  N  KP E K+KIQHVTWP
Sbjct: 176 DLMNRKSPSGPVKSPPLSPVGTTPVKLKRAAPKEEAEATNHLKPPEVKRKIQHVTWP 232
```

Redundant sequences

Filtered low complexity region

Conservative substitutions

Gap

NCBI

BLAST Databases

Protein (nonredundant)

```
nr          Non-redundant GenBank CDS translations
            PDB+SwissProt+SPupdate+PIR
swisspr     Non-redundant SwissProt sequences
pdb                                  quences
```

Entrez Proteins: includes swissprot and PDB

Nucleotide (NO7 nonredundant)

```
nr(nt)   GenBank+EMBL+DDBJ
dbest    Expressed Sequence Tags (EST Division)
htgs     High-Throughput Genome Sequences
         (HTG Division)
chromosome NC_ RefSeqs
```

Entrez Nucleotides: without bulk division sequences

Higher Genomes

NCBI

Note: mouse WGS not here!

Mouse Genome BLAST

NCBI Home » Genomic Biology » Mouse

Search LocusLink for Go

BLAST
overview
FAQs
news
manual
references

Blast the Mouse Genome

BLAST your sequence against mouse specific sequences.

Database: MGSCV3 Program blastn

☑ use MegaBLAST

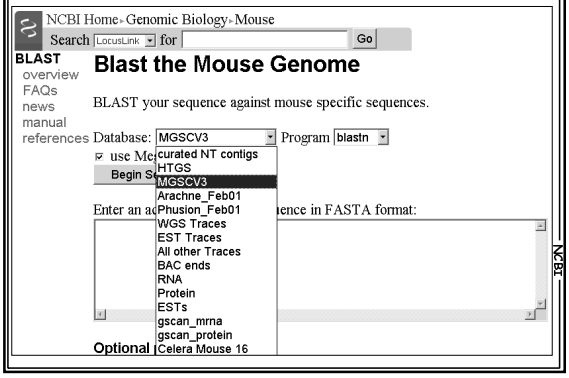
Begin Search

Enter an accession, gi, or a sequence in FASTA format:

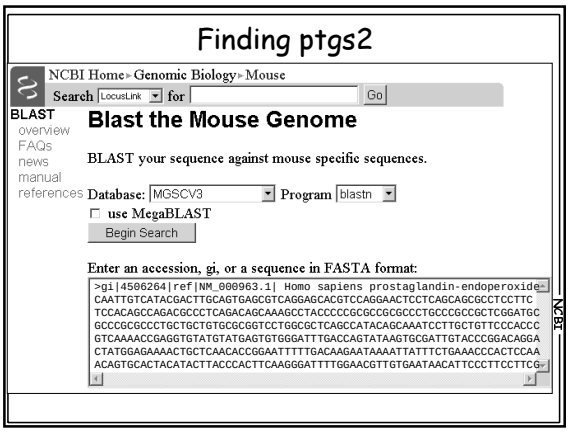
Megablast: same species comparisons

NCBI

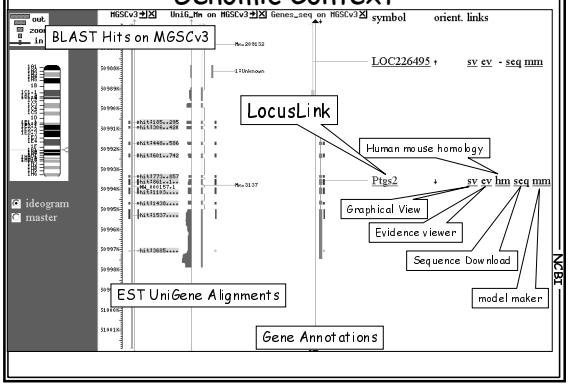
Mouse Genome BLAST: Databases

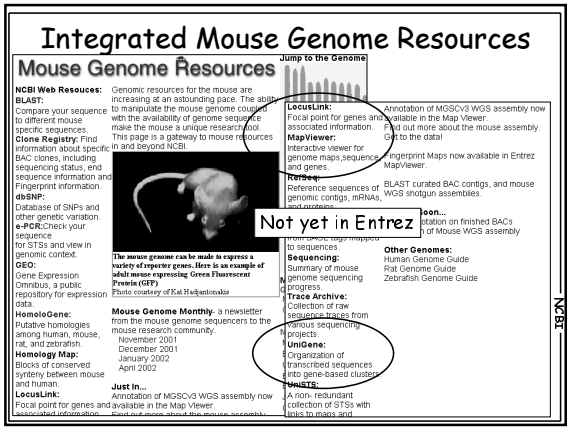


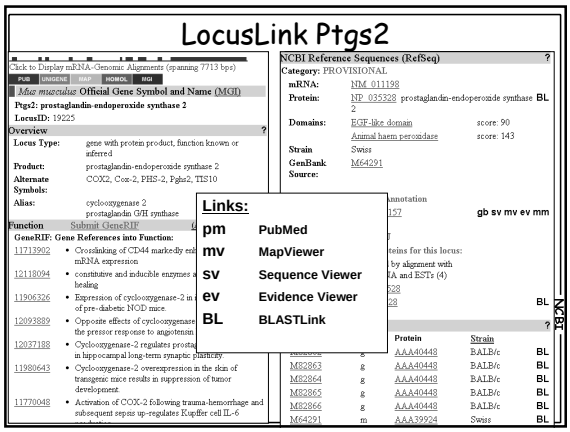
Finding ptgs2

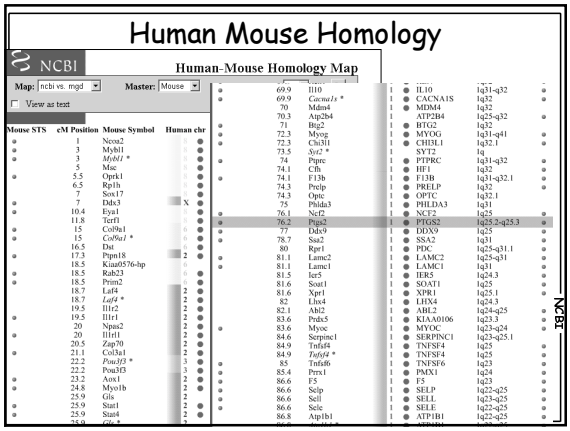


Genomic Context

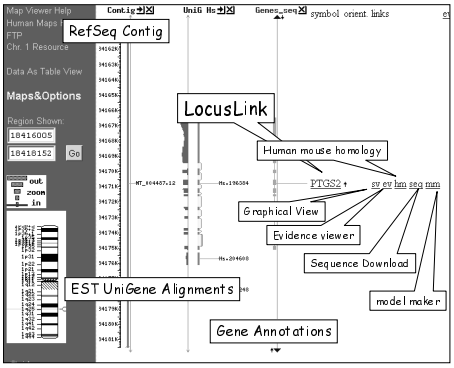



Integrated Mouse Genome Resources
Mouse Genome Resources
Jump to the Genome
NCBI Web Resouces:
BLAST:
Clone Registry:
dbSNP:
e-PCR:
GEO:
HomoloGene:
Homology Map:
LocusLink:
Genomic resources for the mouse are increasing at an astounding pace. The ability to manipulate the mouse genome coupled with the availability of genome sequence make the mouse a unique research tool.
This page is a gateway to mouse resources in and beyond NCBI.
The mouse genome can be made to express a variety of reporter genes. Here is an example of adult mouse expressing Green Fluorescent Protein (GFP)
Photo courtesy of Kat Hadjantonakis
Mouse Genome Monthly - a newsletter from the mouse genome sequencers to the mouse research community.
November 2001
December 2001
January 2002
April 2002
Just In...
Annotation of MGSCv3 WGS assembly now available in the Map Viewer.
LocusLink:
Focal point for genes and associated information.
MapViewer:
Interactive viewer for genome maps,sequence and genes.
RefSeq:
Reference sequences of genomic contigs, mRNAs, and proteins.
Not yet in Entrez
UniGene:
Organization of transcribed sequences into gene-based clusters.
UniSTS:
A non- redundant collection of STSs with links to maps and
Annotation of MGSCv3 WGS assembly now available in the Map Viewer.
Find out more about the mouse assembly.
Get to the data!
Fingerprint Maps now available in Entrez MapViewer.
BLAST curated BAC contigs, and mouse WGS shotgun assemblies.
Sequencing:
Summary of mouse genome sequencing progress.
Trace Archive:
Collection of raw sequence traces from various sequencing projects.
Other Genomes:
Human Genome Guide
Rat Genome Guide
Zebrafish Genome Guide
NCBI


LocusLink Ptgs2
Click to Display mRNA-Genomic Alignments (spanning 7713 bps)
PUB UNIGENE MAP HOMOL MGI
Mus musculus Official Gene Symbol and Name (MGI)
Ptgs2: prostaglandin-endoperoxide synthase 2
LocusID: 19225
Overview
Locus Type: gene with protein product, function known or inferred
Product: prostaglandin-endoperoxide synthase 2
Alternate Symbols: COX2, Cox-2, PHS-2, Pghs2, TIS10
Alias: cyclooxygenase 2, prostaglandin G/H synthase
Function
GeneRIF: Gene References into Function:
11713902
12118094
11906326
12093889
12037188
11980643
11770048
NCBI Reference Sequences (RefSeq)
Category: PROVISIONAL
mRNA: NM_011198
Protein: NP_035328 prostaglandin-endoperoxide synthase 2
Domains: EGF-like domain score: 90
Animal haem peroxidase score: 143
Strain Swiss
GenBank Source: M64291
Links:
pm PubMed
mv MapViewer
sv Sequence Viewer
ev Evidence Viewer
BL BLASTLink
Protein
Strain
M32863 AAA40448 BALB/c
M32864 AAA40448 BALB/c
M32865 AAA40448 BALB/c
M32866 AAA40448 BALB/c
M64291 AAA39924 Swiss
NCBI


Human Mouse Homology
NCBI
Human-Mouse Homology Map
Map: ncbi vs. mgd
Master: Mouse
View as text
Mouse STS
cM Position
Mouse Symbol
Human chr
NCBI

The Human Genome



NCBI