NCBI Molecular Biology Resources

Advances in Cell and Molecular Biology
American Society of Nephrology

Peter Cooper
National Center for Biotechnology Information

NCBI

Talk Outline

- About NCBI
- Molecular Databases
  - GenBank
  - RefSeq
- Web Access
  - Entrez
  - BLAST
- Tour with MLH1
- Live Web Searches

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 131	August 2002
18,197,119	Records
22,616,937,182	Nucleotides
110,000 +	Species

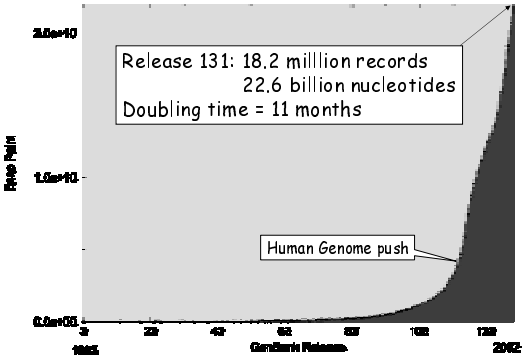
- full release every two months
- incremental and cumulative updates daily
- available only through internet

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

83.65 Gigabytes of data

NCBI

The 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

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

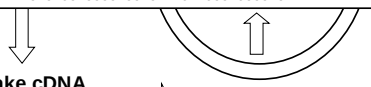
NCBI

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?

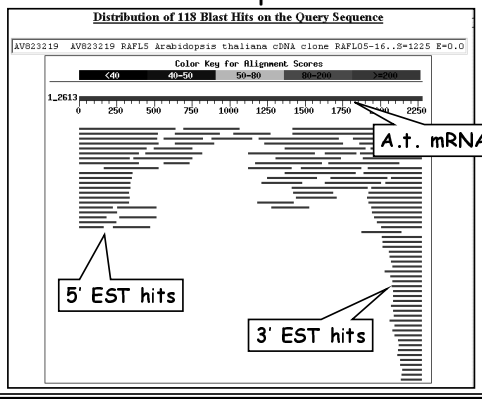
A gene-oriented view of sequence entries

- 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

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

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

NCBI

Hs UniGene Statistics

UniGene Build 148
Apr. 8th, 2002

```
    73,419     mRNAs + gene CDSs
 1,181,855     EST, 3'reads
 1,461,928     EST, 5'reads
+  616,609     EST, other/unknown
----------
 3,333,811    total sequences in clusters

Final Number of Clusters (sets)
================================
98,816    sets total

22,431    s                    one known gene
97,618    s                    one EST
21,233   sets contain both genes and ESTs
```

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

NCBI

UniGene Collections: Animals

		Sequences	Clusters
Homo sapiens	human	3,739,155	104,214
Mus musculus	mouse	2,473,157	88,844
Rattus norvegicus	rat	337,592	62,985
Danio rerio	zebrafish	211,216	15,956
Bos taurus	cow	128,040	10,511
Xenopus laevis	frog	160,336	18,928
Anopheles gambiae	mosquito	43,122	2,568

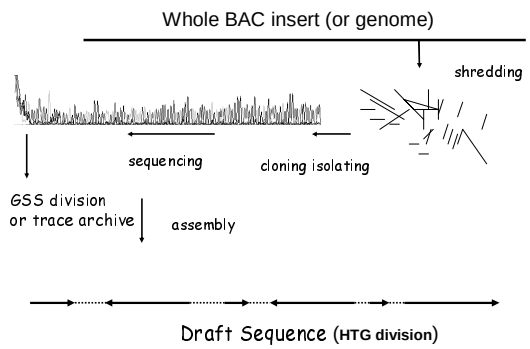
NCBI

UniGene Collections: Plants

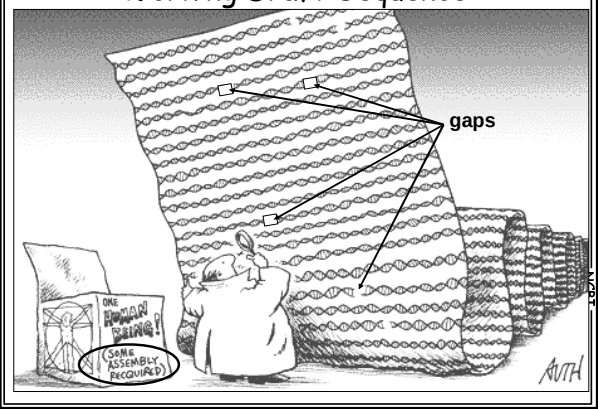
		Sequences	Clusters
Arabidopsis thaliana	thale cress	216,870	27,148
Oryzia sativa	rice	81,977	16,724
Triticum aestivum	wheat	199,738	13,278
Hordeum vulgare	barley	160,069	7,686
Zea mays	maize (corn)	140,279	12,071
Glycine max	soybean	174,243	8,543
Medicago truncatula	barrel medic	92,195	5,721
Lycopersicon esculentum	tomato	70,552	3,779

NCBI

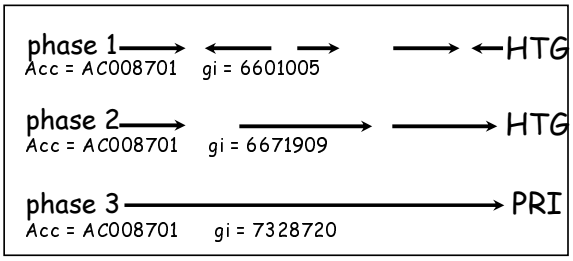
Genome Sequencing



Working Draft Sequence

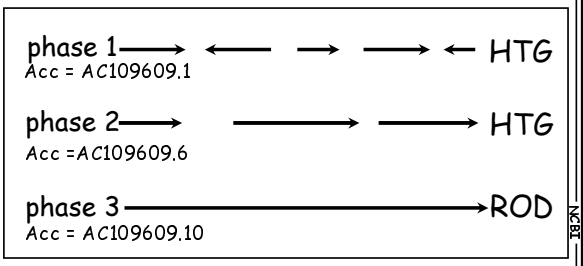


HTG Division: High Throughput Genome



40,000 to > 350,000 bp

HTG Division: High Throughput Genome



40,000 to > 350,000 bp

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.
```

40,000 to > 350,000 bp

NCBI

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

Fugu and Mosquito WGS

```
LOCUS       AAAB01000000           69724 rc    DNA     linear   INV 12-JUN-2002
DEFINITION  Anopheles gambiae str. PEST whole genome shotgun sequencing
            project.
ACCESSION   AAAB00000000
VERSION     AAAB00000000.1  GI:19683831
KEYWORDS
SOURCE
  ORGANIS
REFERENCE
  AUTHORS
  TITLE
  JOURNAL
COMMENT
FEATURES
     sour
WGS         AAAB01000001-AAAB01069724
//
```

```
LOCUS       CAAB01000000           12381 rc    DNA     linear   VRT 22-AUG-2002
DEFINITION  Takifugu rubripes whole genome shotgun sequencing project.
ACCESSION   CAAB00000000
VERSION     CAAB00000000.1  GI:22418063
KEYWORDS    WGS.
SOURCE      Takifugu rubripes.
  ORGANISM  Takifugu rubripes
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Actinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei;
            Acanthomorpha; Acanthopterygii; Percomorpha; Tetraodontiformes; Tetraodontidae; Takifugu.
REFERENCE   1  (bases 1 to 12381)
  AUTHORS   The Fugu Genome Sequencing Consortium.
  TITLE     Direct Submission
  JOURNAL   Submitted (01-JUL-2002) The Fugu Genome Sequencing Consortium,
            http://www.fugubase.org/ http://www.jgi.doe.gov/fugu
COMMENT     The Takifugu rubripes whole genome shotgun (WGS) project has the
            project accession CAAB00000000.  This version of the project (01)
            has the accession number CAAB01000000, and consists of sequences
            CAAB01000001-CAAB01012381.
FEATURES             Location/Qualifiers
     source          1..12381
                     /organism="Takifugu rubripes"
                     /db_xref="taxon:31033"
WGS         CAAB01000001-CAAB01012381
//
```

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

human
mouse
rat
fruit fly
zebrafish
Arabidopsis

NR_123456 Curated non-coding RNA
XM_123456 Predicted Transcript (human, mouse)
XP_123456 Predicted Protein (human, mouse)
XR_123456 Predicted non-coding RNA

Gene Records

NG_123456 Reference Genomic Sequence (human)

Assemblies

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

NCBI

GenBank Sequences: Human Lipoprotein Lipase

Show 500 | Items 1-217 of 217 | One page

- 1: AK092286
  Homo sapiens cDNA FLJ34967 fis, clone NTONG2004690, highl
  LIPASE PRECURSOR (EC 3.1.1.34)
  gi|217508435|dbj|AK092286.1|[217508435]
- 2: AY092404
  Homo sapiens lipoprotein lipase (LPL) gene, promoter region
  gi|21309830|gb|AY092404.1|[21309830]
- 3: AY092405
  Homo sapiens lipoprotein lipase (LPL) gene, exon 7 and partial cd
  gi|21309828|gb|AY092405.1|[21309828]
- 4: BI481389
  IGBPE-1751 Human Retinal Pigment Epithelium (2) Homo sapie
  lipase, mRNA sequence
  gi|18999198|gb|BI481389.1|BI481389[18999198]
- 5: BC011353
  Homo sapiens, Similar to lipoprotein lipase, clone MGC:17090 IM
  cds
  gi|15030192|gb|BC011353.1|BC011353[15030192]
- 6: BG656177
  ab38a05.y1 HR85 islet Homo sapiens cDNA 5' similar to SW:LIP
  LIPOPROTEIN LIPASE PRECURSOR :, mRNA sequence
  gi|13793586|gb|BG656177.1|BG656177[13793586]
- 7: S75964
  lipoprotein lipase (exon 9) [human, Genomic Mutant, 18 nt]
  gi|242490|gb|S75964.1|S75964[242490]
- 8: AU100389
  AU100389 Sugano Homo sapiens cDNA library Homo sapiens cD
  Human lipoprotein lipase mRNA, mRNA sequence
  gi|13551518|dbj|AU100389.1|AU100389[13551518]
- 209: R51310
  Links
  yg72e02.r1 Soares infant brain 1NIB Homo sapiens cDNA clone IMAGE:38573 3' similar to
  gb:M15856 LIPOPROTEIN LIPASE PRECURSOR (HUMAN);, mRNA sequence
  gi|812212|gb|R51310.1|R51310[812212]
- 210: R22123
  Links
  yh25h03.r1 Soares placenta Nb2HP Homo sapiens cDNA clone IMAGE:130805 5' similar to
  gb:M15856 LIPOPROTEIN LIPASE PRECURSOR (HUMAN);, mRNA sequence
  gi|776904|gb|R22123.1|R22123[776904]
- 211: R22066
  Links
  yh25h03.s1 Soares placenta Nb2HP Homo sapiens cDNA clone IMAGE:130805 3' similar to
  gb:M15856 LIPOPROTEIN LIPASE PRECURSOR (HUMAN);, mRNA sequence
  gi|776847|gb|R22066.1|R22066[776847]
- 212: T49529
  Links
  ya76f09.r1 Stratagene placenta (#937225) Homo sapiens cDNA clone IMAGE:67625 5' similar to
  similar to SP:LIPL_CHICK P11602 LIPOPROTEIN LIPASE PRECURSOR, mRNA sequence
  gi|651389|gb|T49529.1|T49529[651389]
- 213: M94221
  Links
  Human GT microsatellite DNA flanking the lipoprotein lipase (LPL) gene
  gi|187217|gb|M94221.1|HUMLPLMUTA[187217]
- 214: M76722
  Links
  H.sapiens lipoprotein lipase (LPL) gene, exons 7,8,and 9, and an Alu repetative element
  gi|187215|gb|M76722.1|HUMLPLF8[187215]
- 215: M29549
  Links
  Human lipoprotein lipase (LPL) gene, 5' flank, clone LPL15/E
  gi|187211|gb|M29549.1|HUMLPLA[187211]
- 216: M26380
  Links
  Human lipoprotein lipase (LPL) gene mRNA
  gi|187212|gb|M26380.1|HUMLPLAA[187212]
- 217: T06274
  Links
  EST04163 Fetal brain, Stratagene (cat#936206) Homo sapiens cDNA clone HFBDR23 similar to
  Lipoprotein lipase, mRNA sequence
  gi|317423|gb|T06274.1|T06274[317423]

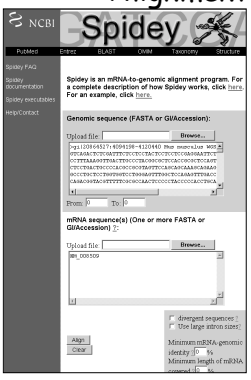
NCBI

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
                  GAHAAGVAGSLTNKKVNRITGLDPAGPNFEYAEAPSRLSPDDADFVDVLHTFTRGSPG
                  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
                     /product="lipoprotein lipase"
                     /protein_id="XP_134193.1"
                     /db_xref="GI:20864362"
```

NCBI

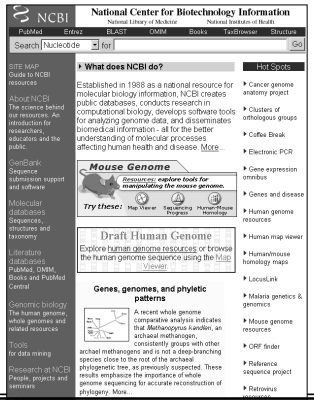
RefSeq Contig: NT_, NW_

```
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:22050862
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chord
            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/Qualif
     source          1..64386157
                     /organism="Homo
                     /db_xref="taxon
                     /chromosome="7"
     gene            230..24397
                     /gene="WBSCR16"
                     /note="DKFZp434
                     /db_xref="Locus
     mRNA            join(230..298,1
                     10510..10683,13
                     /gene="WBSCR16"
                     /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"
                     /db_xref="InterimID:243722"
     mRNA            join(6728..6872,7967..8218,11029..11098,11202..11335,
                     11360..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,
                     11360..11503,11641..11766,12089..12225,12970..13494)
                     /gene="LOC243722"
                     /codon_start=1
```

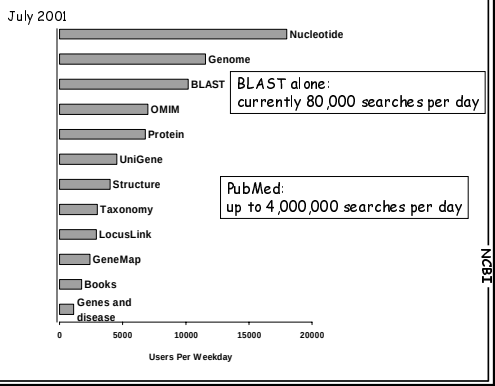
NCBI

Integrated WWW Access: BLAST and Entrez



NCBI

Some Web Statistics

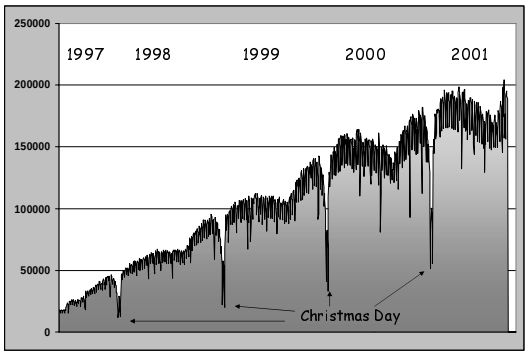


BLAST alone:
currently 80,000 searches per day

PubMed:
up to 4,000,000 searches per day

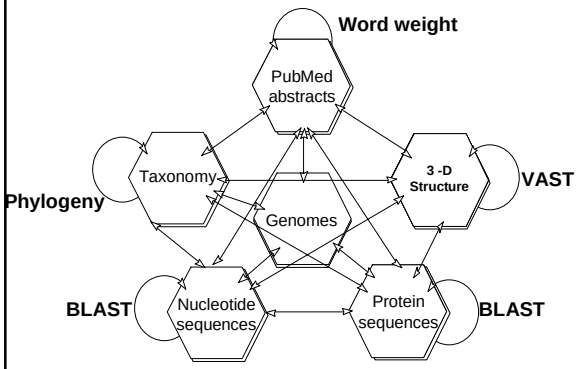
NCBI

Users per day

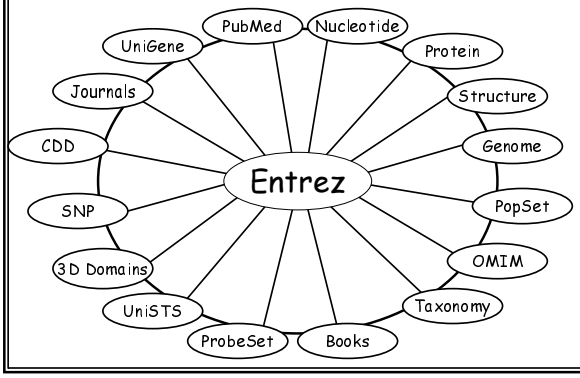


NCBI

Entrez: Database Integration

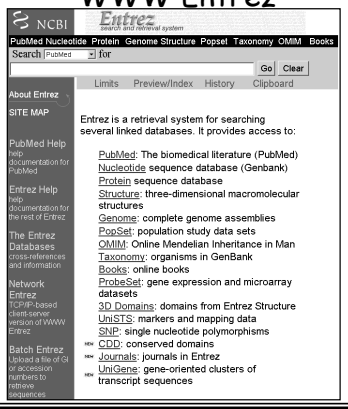


The (ever) Expanding Entrez System



NCBI

WWW Entrez



NCBI

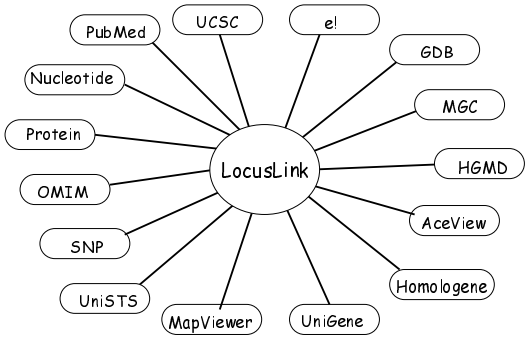
Entrez Databases

PubMed	Biomedical literature
Books	Online textbooks
Nucleotide	GenBank, EMBL, DDBJ, RefSeq, PDB
Protein	[GenBank, EMBL, DDBJ], RefSeq, SWISS-PROT, PIR, PRF, PDB
Genome	Complete genomes
Taxonomy	Organisms in NCBI sequence databases
Structure	MMDB: experimental 3D structures
Domains	CDD: conserved protein domains
3D Domains	Compact 3D protein domains in MMDB
OMIM	Online Mendelian Inheritance in Man
SNP	Single nucleotide polymorphisms (and more!)
UniSTS	Sequence Tagged Site markers
ProbeSet	Gene expression and microarray datasets
PopSet	Population study datasets
UniGene	Nonredundant gene oriented clusters of expressed sequences

NCBI

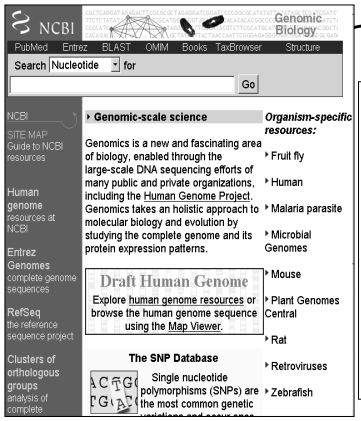
LocusLink

- Entrez
- Other NCBI
- Not NCBI



NCBI

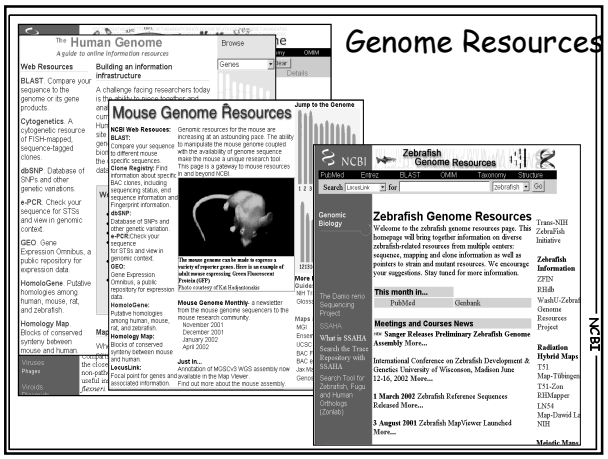
Higher Genomes

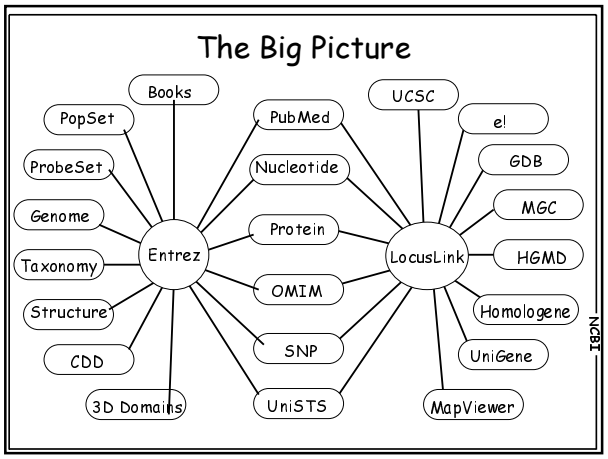


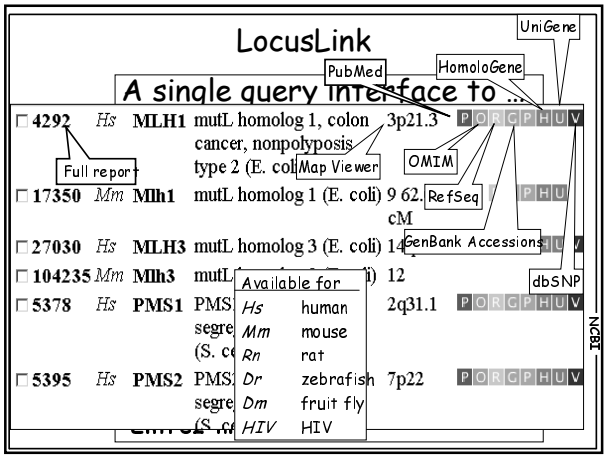
Complete (or nearly)

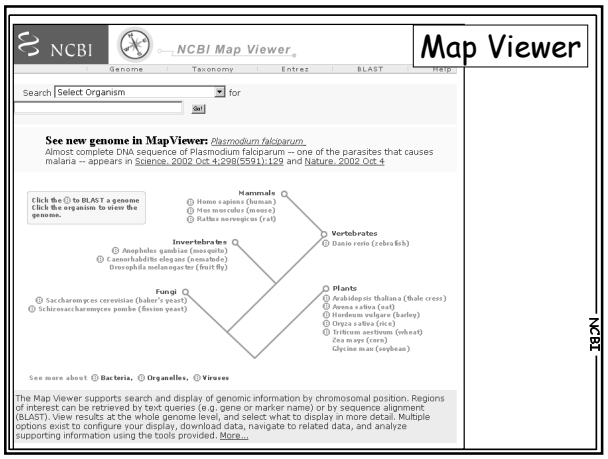
- Human
- Mouse
- Fruit Fly
- Malaria Mosquito
- Fugu
- *C. elegans*
- *Plasmodium falciparum*
- *Arabidopsis*
- Rice

NCBI

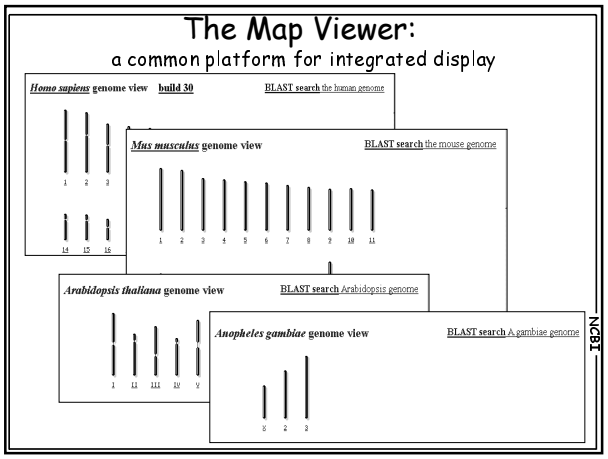

Genome Resources
The Human Genome
Mouse Genome Resources
NCBI
Zebrafish Genome Resources
NCBI


The Big Picture
Books
PopSet
ProbeSet
Genome
Taxonomy
Structure
CDD
3D Domains
Entrez
PubMed
Nucleotide
Protein
OMIM
SNP
UniSTS
UCSC
e!
GDB
MGC
HGMD
LocusLink
Homologene
UniGene
MapViewer
NCBI

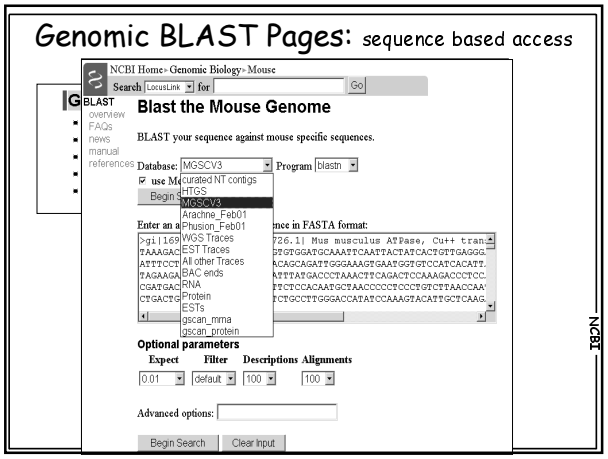

LocusLink
A single query interface to ...
UniGene
PubMed
HomoloGene
Full report
Map Viewer
OMIM
RefSeq
GenBank Accessions
dbSNP
4292 Hs MLH1 mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) 3p21.3
17350 Mm Mlh1 mutL homolog 1 (E. coli) 9 62. cM
27030 Hs MLH3 mutL homolog 3 (E. coli) 14
104235 Mm Mlh3 mutL 12
5378 Hs PMS1 PMS segre (S. co 2q31.1
5395 Hs PMS2 PMS segre (S. co 7p22
Available for
Hs human
Mm mouse
Rn rat
Dr zebrafish
Dm fruit fly
HIV HIV
NCBI



NCBI

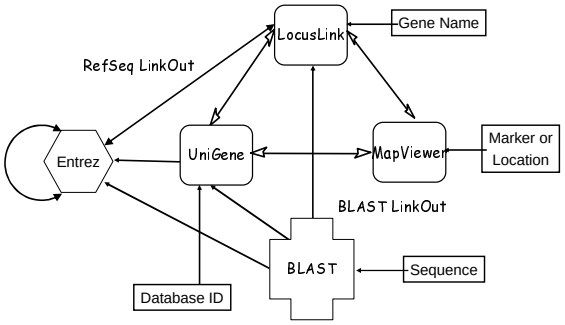


NCBI



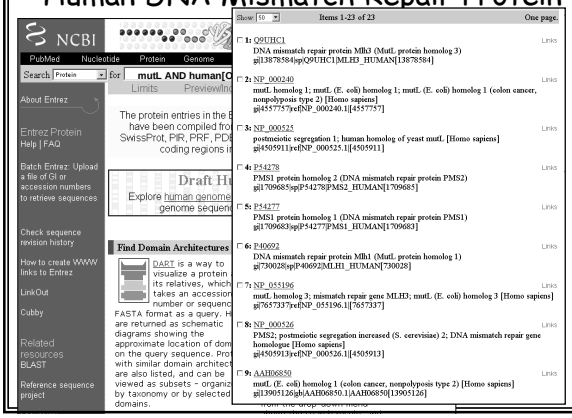
NCBI

Genome Resources Integration



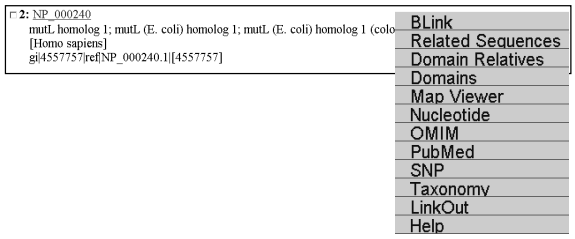
NCBI

Human DNA Mismatch Repair Protein



NCBI

Links and Neighbors Human mutL



NCBI

PubMed

NCBI

Abstract: Full Text, Books

1: Nucleic Acids Res 2001 Nov 1;29(21):4378-86

Related Articles, Links

Full text article at nar.oupjournals.org | PubMed Central access FREE full text articles

Nucleotide
Protein
Free in PMC
Books
LinkOut
Help

Direct association of Bloom's syndrome gene product with the human m
repair protein MLH1.

Pedrazzi G, Perrera C, Blaser H, Kuster P, Marra G, Davies SL, Ryu GH, Freire R, Hickson ID, Jiricny J, Stagljar I.

Institute of Veterinary Biochemistry and Molecular Biology, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.

Bloom's syndrome (BS) is a rare genetic disorder characterised by genomic instability and cancer susceptibility. BLM, the gene mutated in BS, encodes a member of the RecQ family of DNA helicases. Here, we identify hMLH1, which is involved in mismatch repair (MMR) and recombination, as a protein that directly interacts with BLM both in vivo and in vitro, and that the two proteins co-localise to discrete nuclear foci. The interaction between BLM and hMLH1 appears to have been evolutionarily conserved, as Sgs1p, the Saccharomyces cerevisiae homologue of BLM, interacts with yeast Mlh1p. However, cell extracts derived from BS patients show no obvious defects in MMR compared to wild-type- and BLM-complemented BS cell extracts. We conclude that the hMLH1-BLM interaction is not essential for post-replicative MMR, but, more likely, is required for some aspect of genetic recombination.

PMID: 11691925 [PubMed - indexed for MEDLINE]

NCBI

Books

1: Nucleic Acids Res 2001 Nov 1;29(21):4

Full text article at nar.oupjournals.org | PubMed Centr access FREE full text ar

Direct association of Bloom's syn
repair protein MLH1.

Pedrazzi G, Perrera C, Blaser H, Kust
Jiricny J, Stagljar I.

Institute of Veterinary Biochemistry and
190, CH-8057 Zurich, Switzerland.

Bloom's syndrome (BS) is a rare genetic
susceptibility. BLM, the gene mutated in
Here, we identify hMLH1, which is invol
that directly interacts with BLM both in v
discrete nuclear foci. The interaction bet
conserved, as Sgs1p, the Saccharomyces
However, cell extracts derived from BS
wild-type- and BLM-complemented BS
not essential for post-replicative MMR,
recombination.

93 items in **Cancer Medicine.** 5th ed.
Bast, Robert C.; Kufe, Donald W.; Pollock, Raphael E.; Weichselbaum, Ralph R.; Holland, James F.; Frei, Emil, editors.
Canada: BC Decker Inc.; c2000

64 items in **Introduction to Genetic Analysis.** 7th ed.
Griffiths, Anthony J.F.; Miller, Jeffrey H.; Suzuki, David T.; Lewontin, Richard C.; Gelbart, William M.
New York: W.H.Freeman & Co; c1999.

61 items in **Retroviruses.**
Coffin, John M.; Hughes, Stephen H.; Varmus, Harold E.
Plainview (NY): Cold Spring Harbor Laboratory Press; c1997.

55 items in **Molecular Biology of the Cell.** 3rd ed.
Alberts, Bruce; Bray, Dennis; Lewis, Julian; Raff, Martin; Roberts, Keith; Watson, James D.
New York and London: Garland Publishing; c1994

51 items in **C. elegans II.**
Riddle, Donald L.; Blumenthal, Thomas; Meyer, Barbara J.; Priess, James R., editors.
Plainview (NY): Cold Spring Harbor Laboratory Press; c1997.

49 items in **Modern Genetic Analysis.**
Griffiths, Anthony J.F.; Gelbart, William M.; Miller, Jeffrey H.; Lewontin, Richard C.
New York: W.H.Freeman & Co; c1999.

39 items in **Developmental Biology** 6th ed.
Gilbert, Scott F
Sunderland, Massachusetts: Sinauer Associates, Inc.; c2000

34 items in **Medical Microbiology.** 4th ed.
Baron, Samuel, editor
Galveston: University of Texas Medical Branch; c1996.

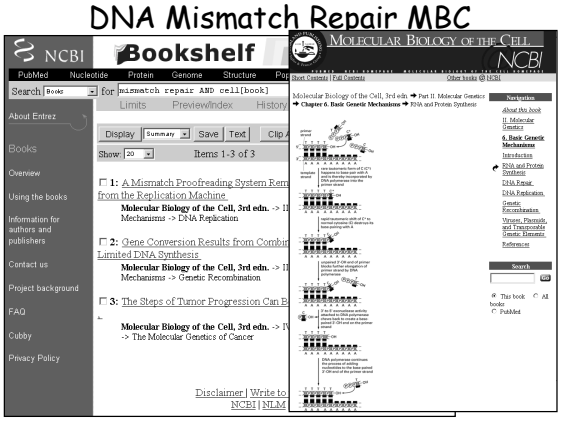
19 items in **Molecular Cell Biology.** 4th ed.
Lodish, Harvey; Berk, Arnold; Zipursky, S. Lawrence; Matsudaira, Paul; Baltimore, David; Darnell, James E.
New York: W.H.Freeman & Co; c2000.

19 items in **Immunobiology.** 5th ed.
Janeway, Charles A.; Travers, Paul; Walport, Mark; Shlomchik, Mark
New York and London: Garland Publishing; c2001.

9 items in **Genes and disease.**
National Institutes of Health.
NCBI; Not under copyright

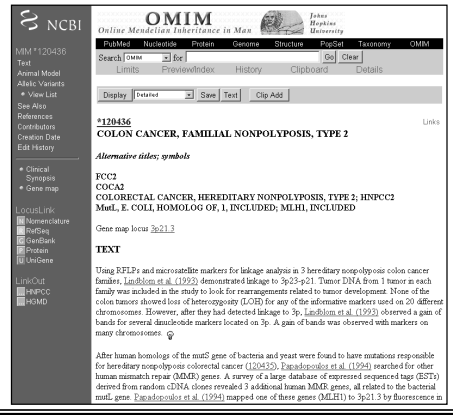
NCBI

DNA Mismatch Repair MBC



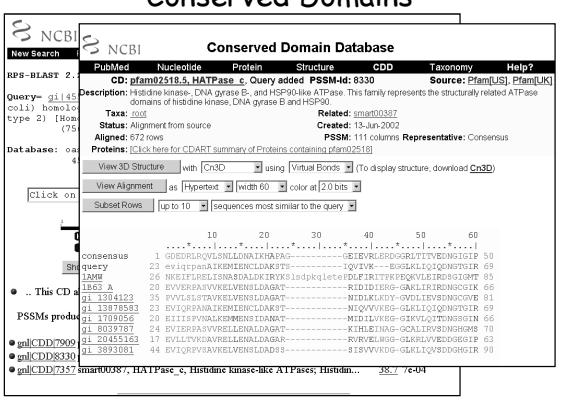
NCBI

OMIM

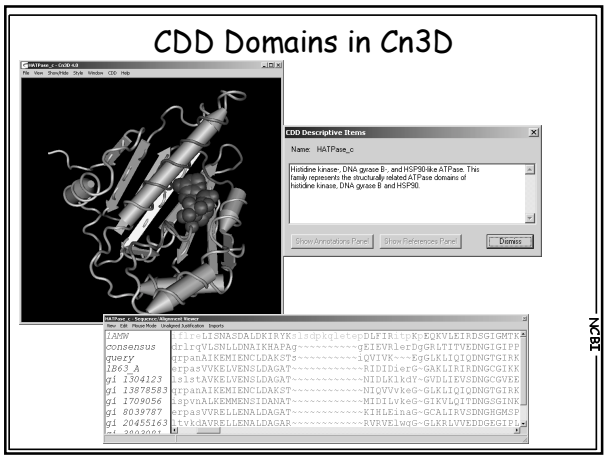


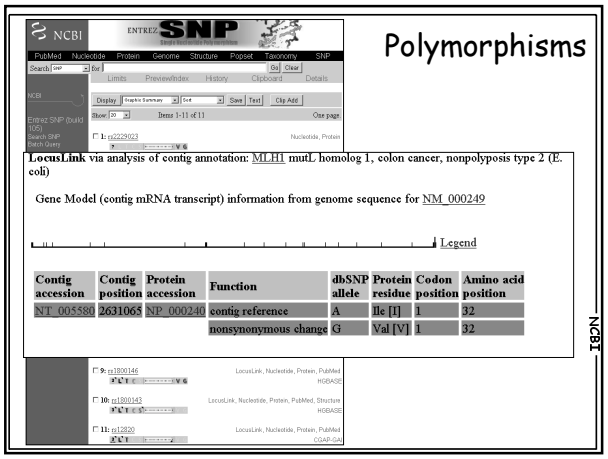
NCBI

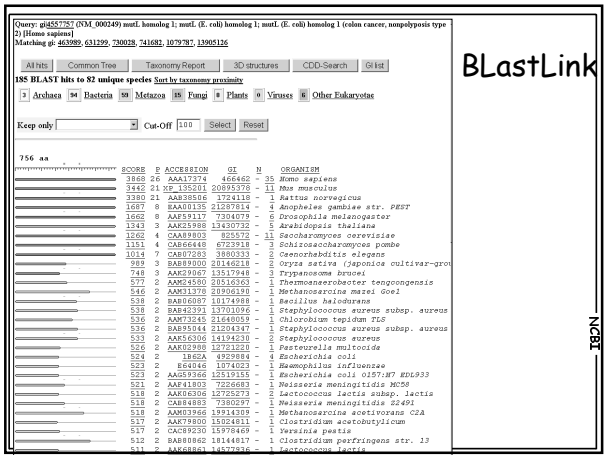
Conserved Domains



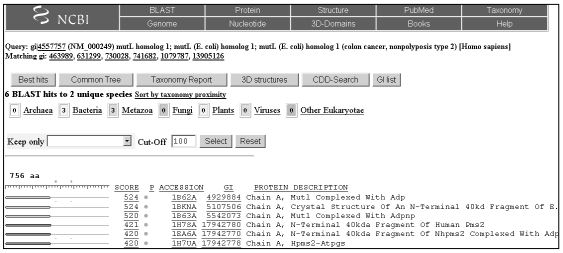
NCBI


CDD Domains in Cn3D
CDD Descriptive Items
Name: HATPase_c
Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase. This family represents the structurally related ATPase domains of histidine kinase, DNA gyrase B and HSP90.
Show Annotations Panel
Show References Panel
Dismiss
NCBI


Polymorphisms
NCBI
ENTREZ SNP
LocusLink via analysis of contig annotation: MLH1 mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
Gene Model (contig mRNA transcript) information from genome sequence for NM_000249
Legend
Contig accession
Contig position
Protein accession
Function
dbSNP allele
Protein residue
Codon position
Amino acid position
NT_005580
2631065
NP_000240
contig reference
A
Ile [I]
1
32
nonsynonymous change
G
Val [V]
1
32
NCBI

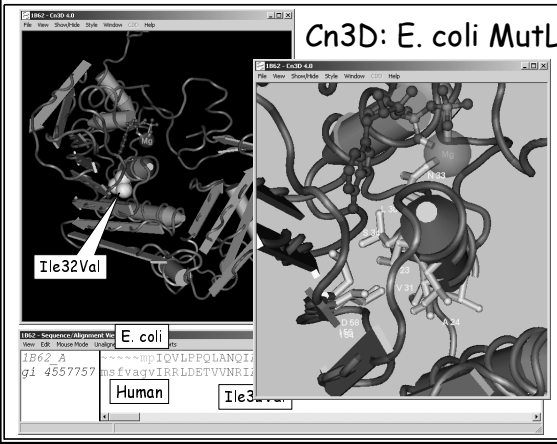

BLastLink
All hits
Common Tree
Taxonomy Report
3D structures
CDD-Search
GI list
185 BLAST hits to 82 unique species
Archaea
Bacteria
Metazoa
Fungi
Plants
Viruses
Other Eukaryotae
Keep only
Cut-Off
Select
Reset
756 aa
SCORE
ACCESSION
GI
ORGANISM
NCBI

Finding Modeling Template



NCBI

Cn3D: E. coli MutL

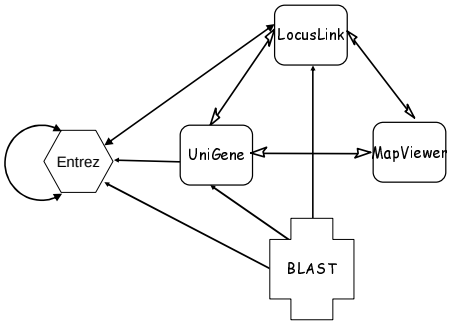


NCBI

Genome Resources with MLH1

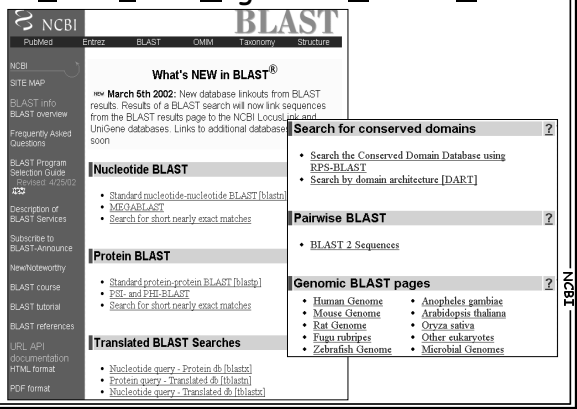
NCBI

Genome Resources Integration



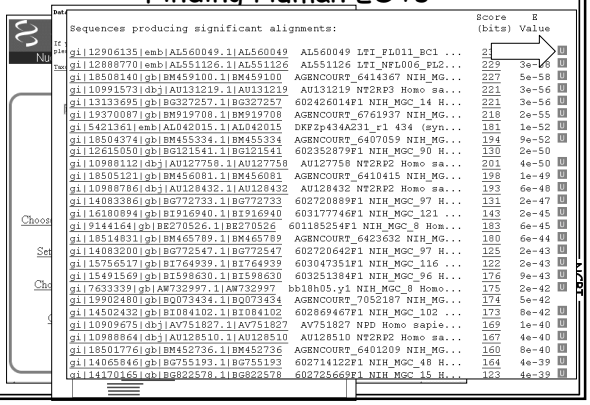
NCBI

Basic Local Alignment Search Tool



NCBI

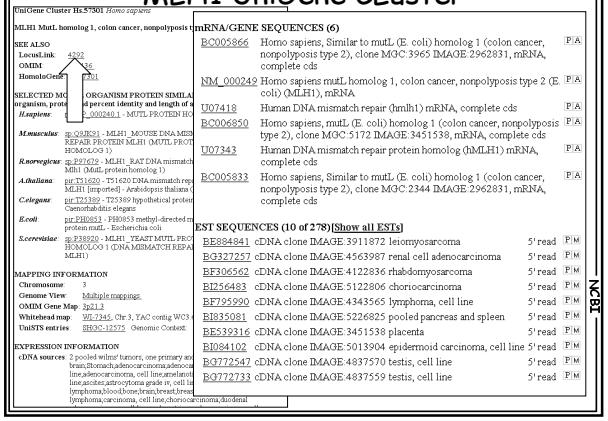
Finding Human ESTs



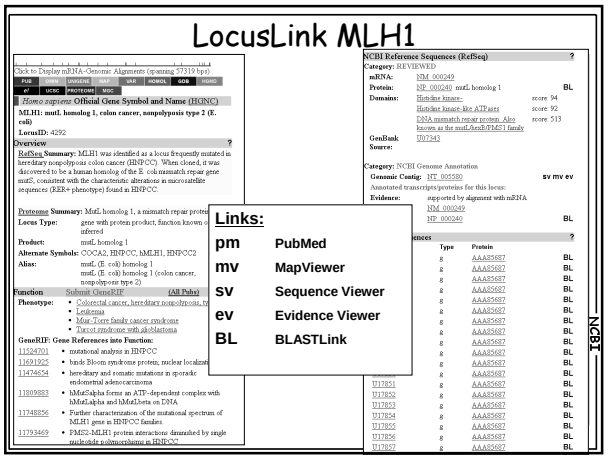
```
                                                          Score    E
Sequences producing significant alignments:               (bits) Value

gi|12906135|emb|AL560049.1|AL560049  AL560049 LTI_FL011_EC1 ...   2
gi|12888770|emb|AL551126.1|AL551126  AL551126 LTI_NFL006_PL2...   229   3e-16
gi|18508140|gb|BM459100.1|BM459100   AGENCOURT_6414367 NIH_MG...  227   5e-58
gi|10991573|dbj|AU131219.1|AU131219  AU131219 NT2RP3 Homo sa...  221   3e-56
gi|13133695|gb|BG327257.1|BG327257   602426014F1 NIH_MGC_14 H...  221   3e-56
gi|19370087|gb|BM919708.1|BM919708   AGENCOURT_6761937 NIH_MG...  218   2e-55
gi|5421361|emb|AL042015.1|AL042015   DKFZp434A231_r1 434 (syn...  181   1e-52
gi|18504374|gb|BM455334.1|BM455334   AGENCOURT_6407059 NIH_MG...  194   9e-52
gi|12615050|gb|BG121541.1|BG121541   602352879F1 NIH_MGC_90 H...  130   2e-50
gi|10988112|dbj|AU127758.1|AU127758  AU127758 NT2RP2 Homo sa...  201   4e-50
gi|18505121|gb|BM456081.1|BM456081   AGENCOURT_6410415 NIH_MG...  198   1e-49
gi|10988786|dbj|AU128432.1|AU128432  AU128432 NT2RP2 Homo sa...  193   6e-48
gi|14083386|gb|BG772733.1|BG772733   602720889F1 NIH_MGC_97 H...  131   2e-47
gi|16180894|gb|BI916940.1|BI916940   603177746F1 NIH_MGC_121 ...  143   2e-45
gi|9144164|gb|BE270526.1|BE270526   601185254F1 NIH_MGC_8 Hom...  183   6e-45
gi|18514831|gb|BM465789.1|BM465789   AGENCOURT_6423632 NIH_MG...  180   6e-44
gi|14083200|gb|BG772547.1|BG772547   602720642F1 NIH_MGC_97 H...  125   2e-43
gi|15756517|gb|BI764939.1|BI764939   603047351F1 NIH_MGC_116 ...  122   2e-43
gi|15491569|gb|BI598630.1|BI598630   603251384F1 NIH_MGC_96 H...  176   9e-43
gi|7633339|gb|AW732997.1|AW732997   bb18h05.y1 NIH_MGC_8 Homo...  175   2e-42
gi|19902480|gb|BQ073434.1|BQ073434   AGENCOURT_7052187 NIH_MG...  174   5e-42
gi|14502432|gb|BI084102.1|BI084102   602869467F1 NIH_MGC_102 ...  173   8e-42
gi|10909675|dbj|AV751827.1|AV751827   AV751827 NPD Homo sapie...  169   1e-40
gi|10988864|dbj|AU128510.1|AU128510   AU128510 NT2RP2 Homo sa...  167   4e-40
gi|18501776|gb|BM452736.1|BM452736   AGENCOURT_6401209 NIH_MG...  160   8e-40
gi|14065846|gb|BG755193.1|BG755193   602714122F1 NIH_MGC_48 H...  164   4e-39
gi|14170165|gb|BG820578.1|BG820578   602725669F1 NIH_MGC_15 H...  123   4e-39
```

NCBI

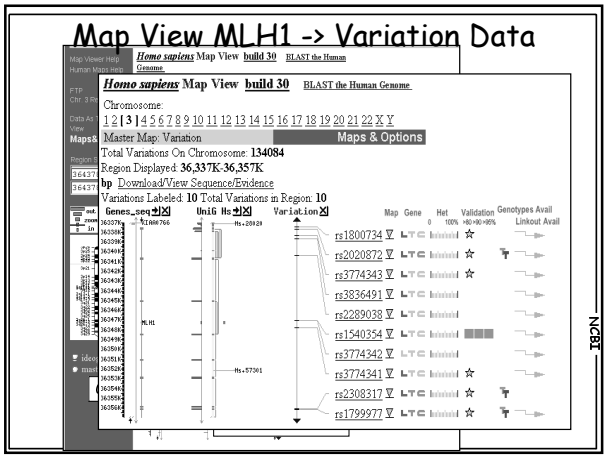
MLH1 UniGene CLuster



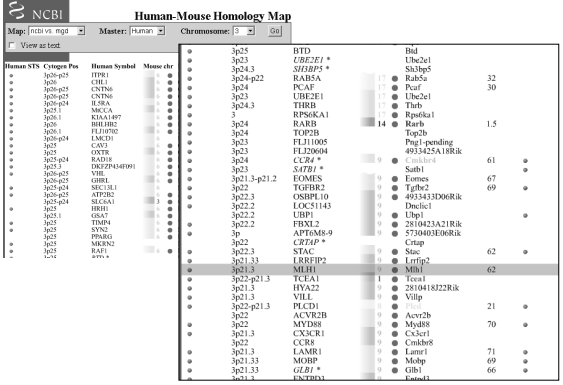
LocusLink MLH1



Map View MLH1 -> Variation Data

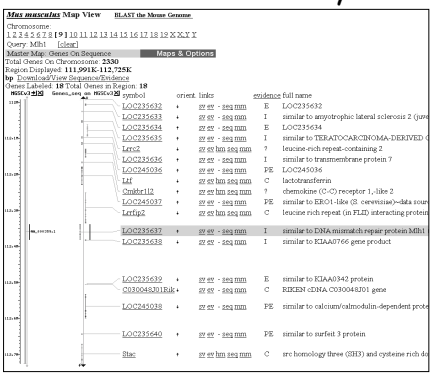


Human Mouse Homology: the mouse assembly



NCBI

The Mouse Assembly



NCBI

Live Web Demos

- Finding a gene by …
  - -Database ID (AI863603)
  - - Gene name (Albumin)
  - - Map location (NPHP4)
    - PMID:11920297 D1S2660 OR D1S2642
  - - BLAST mouse NPHP4
    - NM_015102

NCBI