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
  - Candidate gene by genetic markers
  - Human homologue of rat gene by BLAST

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)

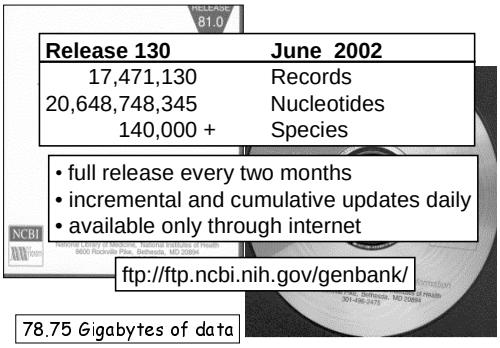
NCBI

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

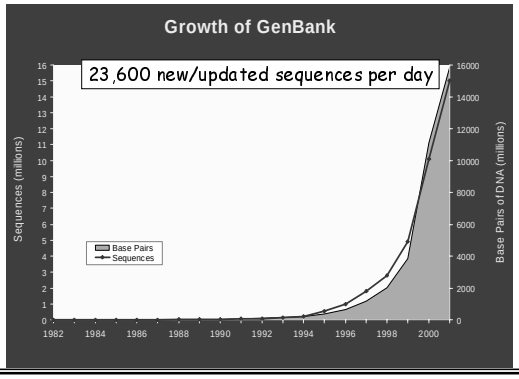
- full release every two months
- incremental and cumulative updates daily
- available only through internet

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

78.75 Gigabytes of data

NCBI

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

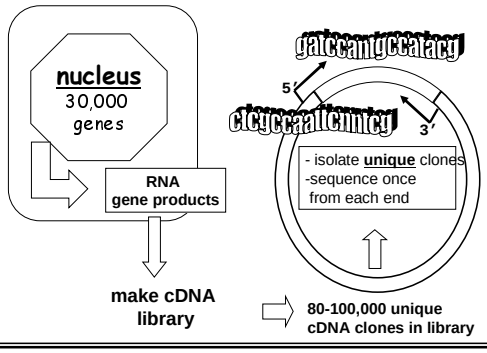
Bulk GenBank Divisions

- 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



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
```

make cDNA library

80-100,000 unique cDNA clones in library

NCBI

What is UniGene?

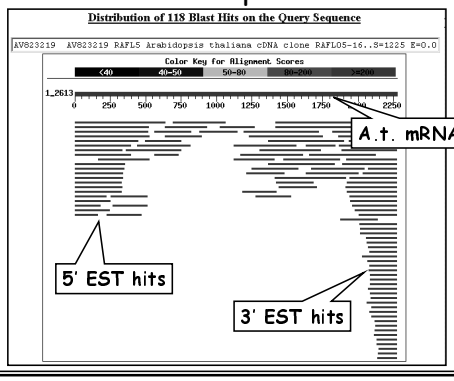
A gene-oriented view of sequence entries

- 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   sets contain at least one known gene
17,654   sets contain at least 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   sets contain at least one known gene
97,618   sets contain at least one EST
21,233  sets contain both genes and ESTs
```

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

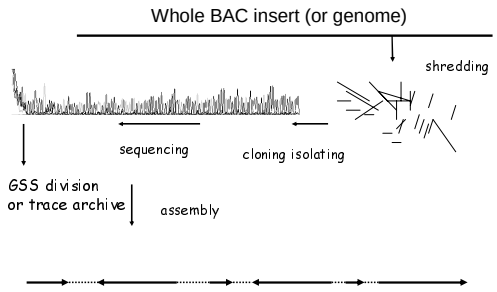
NCBI

UniGene Collections *Apr. 2002*

		Sequences	Clusters
Animals			
Homo sapiens	human	3,333,811	98,816
Mus musculus	mouse	2,274,640	86,897
Rattus norvegicus	rat	308,877	59,882
Danio rerio	zebrafish	159,261	14,893
Bos taurus	cow	122,503	9,303
Xenopus laevis	frog	120,489	16,489
Anopholes gambiae	mosquito	42,590	2,414
Plants			
Arabidopsis thaliana	thale cress	202,099	26,794
Oryzia sativa	rice	77,376	15,283
Triticum aestivum	wheat	35,387	3,091
Hordeum vulgare	barley	108,658	6,984
Zea mays	maize (corn)	108,030	9,889

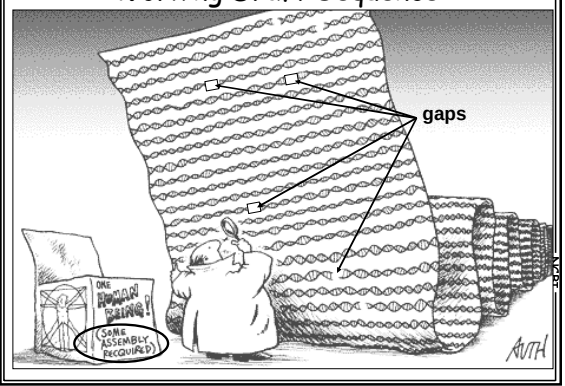
NCBI

Genome Sequencing



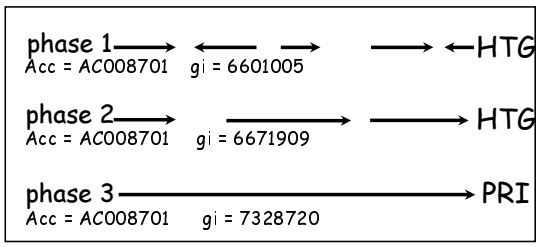
NCBI

Working Draft Sequence



NCBI

HTG Division: High Throughput Genome



NCBI

40,000 to > 350,000 bp

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

RefSeq Resource

- Genome Oriented Resource
  - A sequence for each macromolecule Central Dogma: Chromosome, mRNA, preprotein, mature protein
  - Linked on a residue by residue basis
  - Objectively non-redundant and comprehensive
- Curated Resource
  - Authoritative source by genome
  - Derived from GenBank but corrected, merged, extended
  - Publicly distributed, Entrez Genomes Web site
- Reagents for Genome Annotation and Analysis
- Substrate for Functional Genomics

NCBI

The RefSeq Accession Numbers

NCBI Reference Sequences

mRNAs and Proteins

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

human
mouse
rat
fruit fly
zebrafish
Arabidopsis

Gene Records
NG_123456 Reference Genomic Sequence (human)
Assemblies
NT_123456 Contig (Mouse and Human)
NW_123456 Super Contig (Mouse)
NC_123455 Chromosome (Microbial, Arabidopsis)

NCBI

GenBank Sequences: human CFTR

Show: 100 Items 1-83 of 83 One page.

1: M86631 Related Sequences, PubMed, Taxonomy
Homo sapiens (clone ST-18-5(9/16)) cystic fibrosis transmembrane conductance regulator (CFTR) gene, 3' end intron 17B; complete exon 18; complete intron 18
gi|180296|gb|M86631.1|HUMCFTR[180296]

2: S64699 Related Sequences, OMIM, Protein, PubMed, Taxonomy
Homo sapiens cystic fibrosis transmembrane conductance regulator isoform 36 (CFTR) mRNA, partial cds
gi|408285|gb|S64699.1|S64699[408285]

3: AL121762 Related Sequences, Taxonomy
Human DNA sequence from clone RP4-610C12 on chromosome 20 Contains the 3' end of a novel gene, a putative novel gene, a pseudogene similar to part of the cystic fibrosis transmembrane conductance regulator (CFTR), four CpG islands, ESTs, STSs and GSSs, complete sequence
gi|8574423|emb|AL121762.13|HSJ610C12[8574423]

4: AH006034 OMIM, Protein, PubMed, Taxonomy
Human cystic fibrosis transmembrane conductance regulator (CFTR) gene
gi|306537|gb|AH006034.1|SEG_HUMCFTRA[306537]

5: M55131 Related Sequences, ProbeSet, OMIM, Protein, PubMed, Taxonomy
Human cystic fibrosis transmembrane conductance regulator (CFTR) gene, exon 24
gi|306536|gb|M55131.1|HUMCFTRA26[306536]

6: M55130 Related Sequences, Protein, PubMed, Taxonomy
Human cystic fibrosis transmembrane conductance regulator (CFTR) gene, exon 23

NCBI

Curated RefSeq Records: NM_, NP_

```
LOCUS       NM_000492    6159 bp    mRNA            PRI      26-JUL-1999
DEFINITION  Homo sapiens cystic fibrosis transmembrane conductance
            regulator(CFTR) mRNA.
ACCESSION   NM_000492
```

RefSeq Nucleotide

```
LOCUS       NP_000483    1480 aa                    PRI      26-JUL-1999
DEFINITION  cystic fibrosis transmembrane conductance regulator.
ACCESSION   NP_000483
PID         g4502785
VERSION     NP_000483.1  GI:4502785
DBSOURCE    REFSEQ: accession NM_000492.1
```

RefSeq Protein

NCBI

```
COMMENT     REFSEQ: This reference sequence was derived from M55131.
            PROVISIONAL RefSeq: This is a provisional reference sequence
            record that has not yet been subject to human review. The final
            curated reference sequence record may be somewhat different from
            this one.
```

Curated RefSeq Records: NM_, NP_

```
LOCUS       NM_000492    6159 bp    mRNA            PRI      26-JUL-1999
DEFINITION  Homo sapiens cystic fibrosis transmembrane conductance
            regulator(CFTR) mRNA.
```

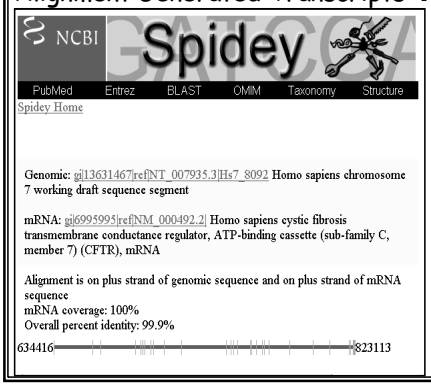
```
REFSEQ: This reference sequence was derived from M28668.1,
M55131.1.
On Feb 17, 2000 this sequence version replaced gi:4502784.
Summary: Cystic fibrosis transmembrane conductance regulator is
member 7 of the ATP-binding cassete sub-family C. The protein
functions as a chloride channel and controls the regulation of
other transport pathways. Mutations in this gene cause the
autosomal recessive disorder, cystic fibrosis (CF) and congenital
bilateral aplasia of the vas deferens (CBAVD). Alternative splice
variants have been described, many of which result from mutations
in the CFTR gene.
COMPLETENESS: full length.
```

Reviewed

NCBI

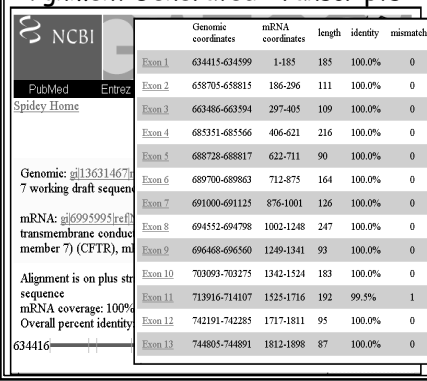
```
COMMENT     REFSEQ: This reference sequence was derived from M55131.
            PROVISIONAL RefSeq: This is a provisional reference sequence
            record that has not yet been subject to human review. The final
            curated reference sequence record may be somewhat different from
            this one.
```

Alignment Generated Transcripts: XM_, XP_



NCBI

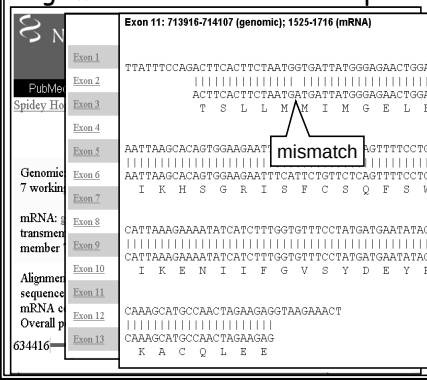
Alignment Generated Transcripts: XM_, XP_



	Genomic coordinates	mRNA coordinates	length	identity	mismatches	gaps	Donor site	Acc. site
Exon 1	634415-634599	1-185	185	100.0%	0	0	d	
Exon 2	658705-658815	186-296	111	100.0%	0	0	d	a
Exon 3	663486-663594	297-405	109	100.0%	0	0	d	a
Exon 4	685351-685566	406-621	216	100.0%	0	0	d	a
Exon 5	688728-688817	622-711	90	100.0%	0	0	d	a
Exon 6	689700-689863	712-875	164	100.0%	0	0	d	a
Exon 7	691000-691125	876-1001	126	100.0%	0	0	d	a
Exon 8	694552-694798	1002-1248	247	100.0%	0	0	d	a
Exon 9	696468-696560	1249-1341	93	100.0%	0	0	d	a
Exon 10	703093-703275	1342-1524	183	100.0%	0	0	d	a
Exon 11	713916-714107	1525-1716	192	99.5%	1	0	d	a
Exon 12	742191-742285	1717-1811	95	100.0%	0	0	d	a
Exon 13	744805-744891	1812-1898	87	100.0%	0	0	d	a

NCBI

Alignment Generated Transcripts: XM_, XP_



NCBI

Alignment Generated Transcripts: **XM_, XP_**

Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Exon 6

Exon 11: 713916-714107 (genomic); 1525-1716 (mRNA)

```
TTATTTCCAGACTTCACTTCTAATGGTGATTATGGGAGAACTGGAGCCTTCAGAGGGTAA
            ||||||||||||||| |||||||||||||||||||||||||||||||||||
           ACTTCACTTCTAATGATGATTATGGGAGAACTGGAGCCTTCAGAGGGTAA
             T  S  L  L  M  M  I  M  G  E  L  E  P  S  E  G  K

AATTAAGCACAGTGGAAGAAT                AGTTTTCCTGGATTATGCCTGGCAC
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCAGTTTTCCTGGATTATGCCTGGCAC
```

mismatch

```
LOCUS       XM_004980    6128 bp    mRNA            PRI       16-NOV-2000
DEFINITION  Homo sapiens cystic fibrosis transmembrane conductance regulator,
            ATP-binding cassette (sub-family C, member 7) (CFTR), mRNA.
ACCESSION   XM_004980
VERSION     XM_004980.3  GI:13631444
```

Exon 11
Exon 12
Exon 13

```
CAAAGCATGCCAACTAGAAGAGGTAAGAAACT
||||||||||||||||||||||
CAAAGCATGCCAACTAGAAGAG
  K  A  C  Q  L  E  E
```

RefSeq Contig: **NT_**

```
LOCUS       NT_007933           22893611 bp    DNA     linear   CON 13-MAY-2002
DEFINITION  Homo sapiens chromosome 7 working draft sequence segment.
ACCESSION   NT_007933
VERSION     NT_007933.9  GI:20543589
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 22893611)
  AUTHORS   NCBI Annotation Project.
  TITLE     Direct Submission
  JOURNAL   Submitted (09-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.
            On May 13, 2002 this sequence version replaced gi:18568843.
            COMPLETENESS: not full length.
```

RefSeq Contig: **NT_**

```
LOCUS       NT_007933           22893611 bp    DNA     linear   CON 13-MAY-2002
DE
AC
VE
KE
SO
RE
CO
```

```
     gene            9640781..9829479
                     /gene="CFTR"
                     /note="CF; MRP7; ABC35; ABCC7"
                     /db_xref="LocusID:1080"
                     /db_xref="MIM:602421"
     mRNA            join(9640781..9640965,9665071..9665181,9669852..9669960,
                     9691717..9691932,9695094..9695183,9696066..9696229,
                     9697366..9697491,9700918..9701164,9702834..9702926,
                     9709459..9709641,9720282..9720473,9748557..9748651,
                     9751171..9751257,9752752..9753475,9755748..9755876,
                     9763644..9763681,9764350..9764600,9767492..9767571,
                     9771337..9771487,9772399..9772626,9775431..9775531,
                     9788340..9788588,9803256..9803411,9813660..9813749,
                     9825506..9825678,9826277..9826382,9827726..9829479)
                     /gene="CFTR"
                     /product="cystic fibrosis transmembrane conductance
                     regulator, ATP-binding cassette (sub-family C, member 7)"
                     /note="Derived by automated computational analysis using
                     gene prediction method: BLAST. Supporting evidence
                     includes similarity to: 2 mRNAs"
                     /transcript_id="XM_004980.4"
                     /db_xref="GI:14753226"
                     /db_xref="LocusID:1080"
                     /db_xref="MIM:602421"
```

RefSeq Contig: **NT_**

```
LOCUS       NT_007933            22893611 bp    DNA     linear   CON 13-MAY-2002
DE
AC
VE
KE
S

R

C
```

```
     gene            9640781..9829479
                     /gene="CFTR"
                     /note="CF; MRP7; ABC35; ABCC7"
```

```
CONTIG      join(AC004917.2:1..110233,AC025728.4:201..36215,
            AC002384.1:201..69248,AC005018.2:201..107527,
            AC079396.7:2001..61359,AC006387.3:2001..138764,
            AC004492.1:201..165608,AC004855.1:201..74816,
            AC002381.1:201..172533,AC004839.1:201..55977,AC078937.1:201..62035,
            AC002467.1:201..97943,AC005046.3:201..219436,AC078853.6:201..4728,
            AC005048.2:201..125290,AC005683.2:201..74673,
            AC007567.2:201..187413,AC005058.1:201..182256,
            AC005487.2:201..54183,AC002487.1:201..181588,AC099338.3:2001..6975,
            AC073272.7:2007..54594,AC004014.2:201..172241,
            AC091766.2:2001..65542,AC073248.5:2001..128294,
            AC093670.4:2001..3994,AC002386.2:1995..131458,
            AC074119.5:201..96106,AC073071.2:201..154216,
            AC073103.3:82266..82296,AC073387.4:1..153300,
            AC073139.6:2001..128635,AC102951.5:2001..22461,
            AC004161.1:2001..112237,AC079893.4:201..20869,
            AC073114.2:201..175645,AC092063.3:2001..11858,
            AC003081.1:2001..149566,AC006339.3:201..21462,
            complement(AC003088.1:1..79004),AC092167.2:2001..103
```

```
                     /db_xref="LocusID:1080"
                     /db_xref="MIM:602421"
```

RefSeq WGS Supercontig: **NW_**

```
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.
```

NCBI

RefSeq WGS Supercontig: **NW_**

```
LOCUS       NW_000272            27755676 bp    DNA     linear   CON 04-JUN-2002
```

```
     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,
                     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
                     /protein_id="XP_145115.1"
                     /db_xref="GI:20915206"
                     /db_xref="InterimID:243722"
```

NCBI

RefSeq WGS Supercontig: **NW_**

```
LOCUS       NW_000272          27755676 bp    DNA     linear   CON 04-JUN-2002

     FEATURES             Location/Qualifiers
     source          1..27755676
                     /organism="Mus musculus"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="6"
     gene            6728..13494
CONTIG      join(CAAA01201740.1:1..4651,gap(100),CAAA01125491.1:1..11772,
            gap(189),CAAA01120088.1:1..4194,gap(122),CAAA01145531.1:1..18939,
            gap(243),CAAA01174720.1:1..20193,gap(100),CAAA01148186.1:1..4432,
            gap(100),CAAA01138802.1:1..26173,gap(100),CAAA01149691.1:1..10320,
            gap(100),CAAA01147818.1:1..1458,gap(1004),CAAA01147815.1:1..6125,
            gap(749),CAAA01147812.1:1..1561,gap(100),CAAA01074766.1:1..2776,
            gap(894),CAAA01074767.1:1..8180,gap(183),CAAA01138806.1:1..2425,
            gap(100),CAAA01144336.1:1..11048,gap(100),CAAA01191698.1:1..2516,
            gap(330),CAAA01048494.1:1..6532,gap(100),CAAA01048497.1:1..1666,
            gap(1506),CAAA01044252.1:1..76338,gap(100),CAAA01171120.1:1..13327,
            gap(100),CAAA01044255.1:1..5744,gap(100),CAAA01222036.1:1..8808,
            gap(100),CAAA01159465.1:1..4183,gap(271),CAAA01044258.1:1..20222,
            gap(583),CAAA01204593.1:1..2875,gap(867),CAAA01114772.1:1..2380,
            AC066688.3:1..176379,AC078790.2:1..173690,
            CAAA01044276.1:58306..74674,gap(100),CAAA01044280.1:1..10478,
            gap(100),CAAA01044283.1:1..1857,gap(1569),CAAA01044288.1:1..126746,
```

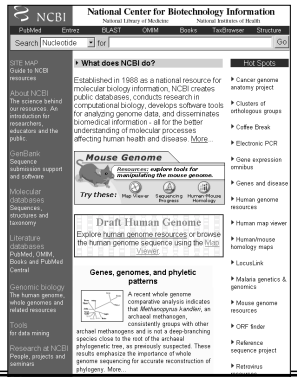
NCBI

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

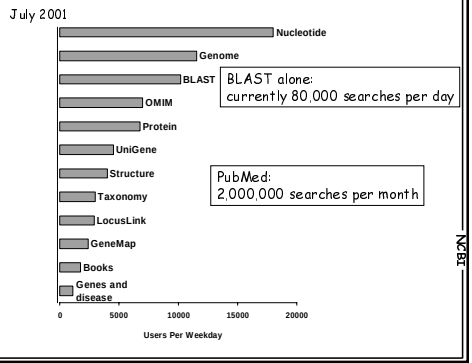
NCBI

Integrated WWW Access: BLAST and Entrez



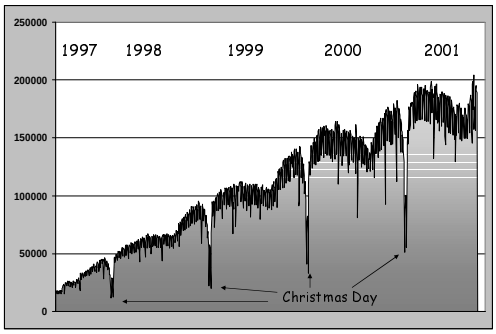
NCBI

Some Web Statistics



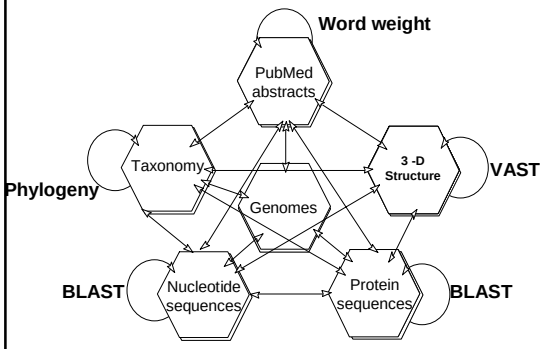
NCBI

Users per day



NCBI

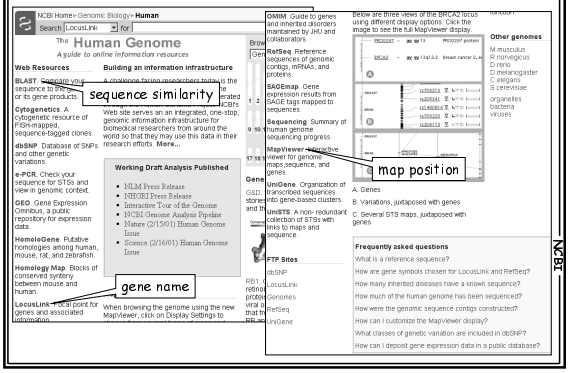
Entrez: Database Integration



The Draft Human Genome

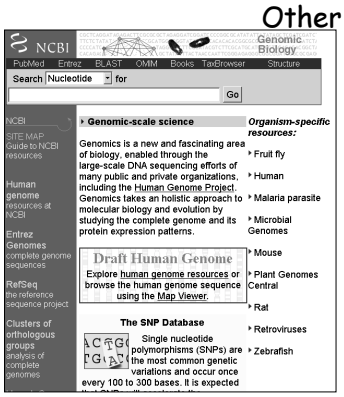
NCBI

Human Genome Resources Find a gene by ...

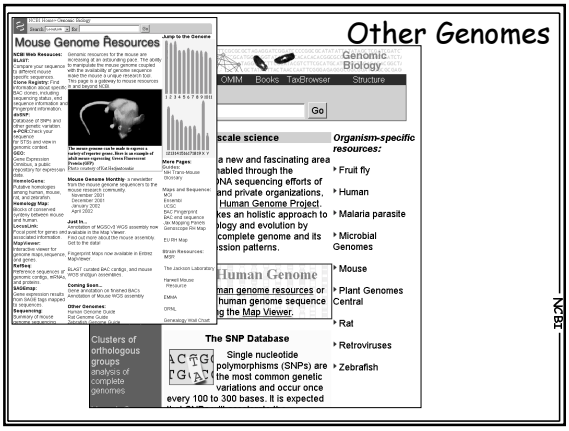


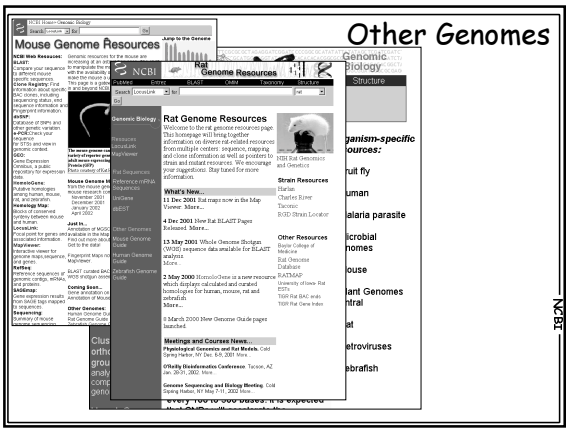
NCBI

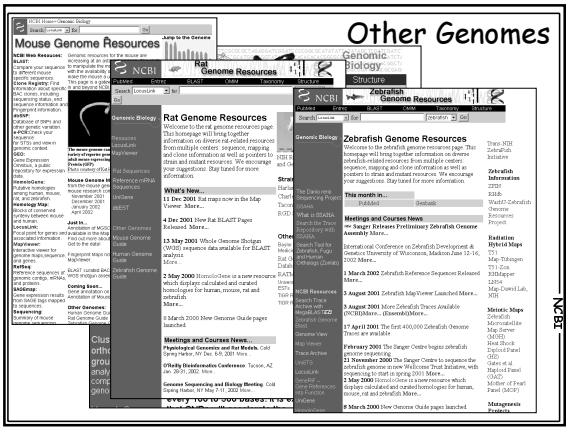
Other Genomes



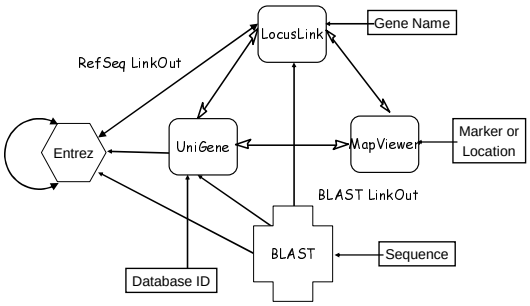
NCBI


Other Genomes
Mouse Genome Resources
Genomic Biology
Organism-specific resources:
Fruit fly
Human
Malaria parasite
Microbial Genomes
Mouse
Plant Genomes Central
Rat
Retroviruses
Zebrafish
Human Genome
The SNP Database
Clusters of orthologous groups analysis of complete genomes
NCBI


Other Genomes
Mouse Genome Resources
Rat Genome Resources
NCBI

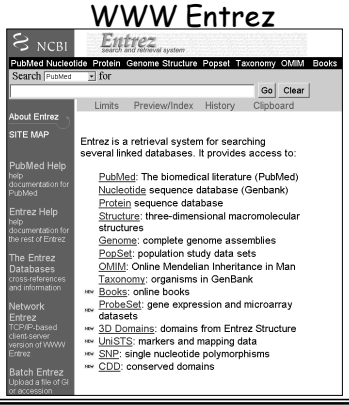

Other Genomes
Mouse Genome Resources
Rat Genome Resources
Zebrafish Genome Resources
NCBI

Genome Resources Integration



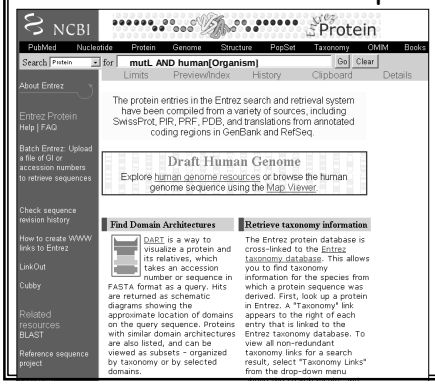
NCBI

WWW Entrez

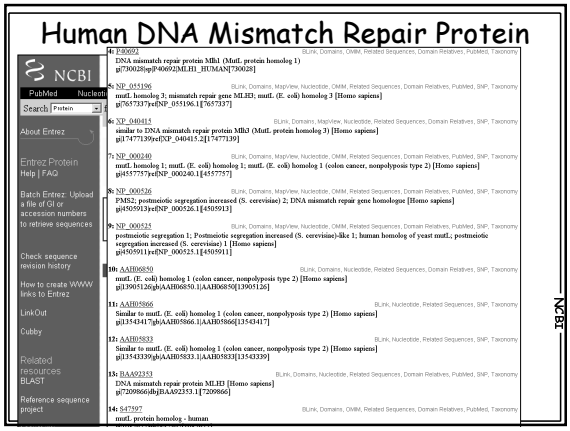


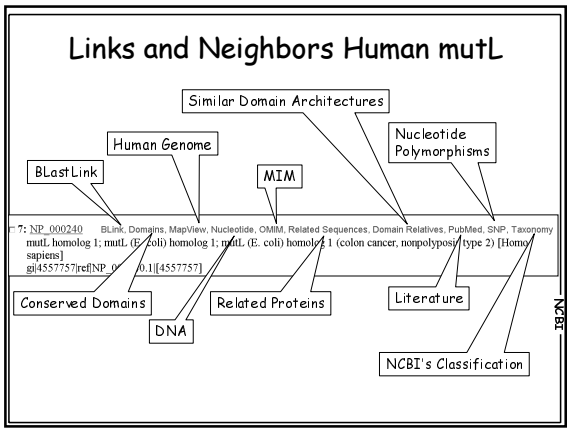
NCBI

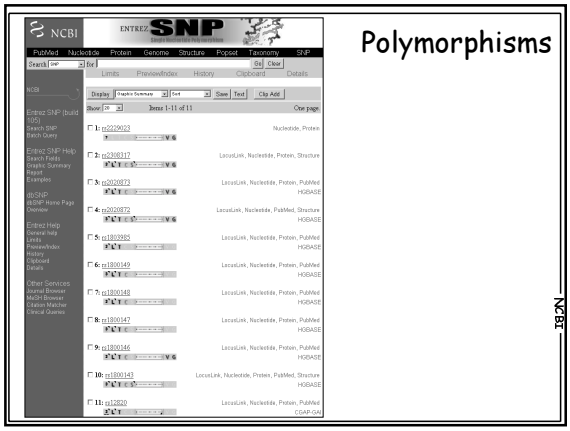
Human DNA Mismatch Repair Protein

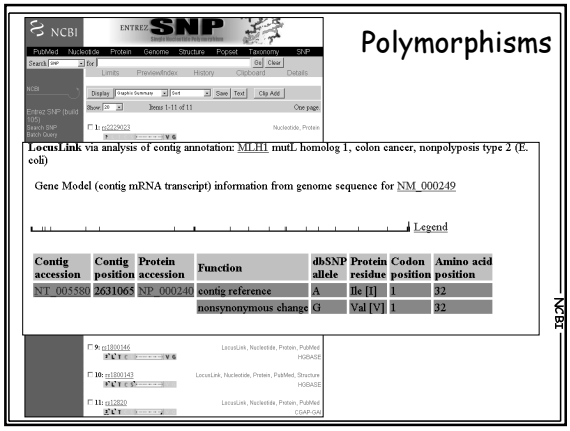


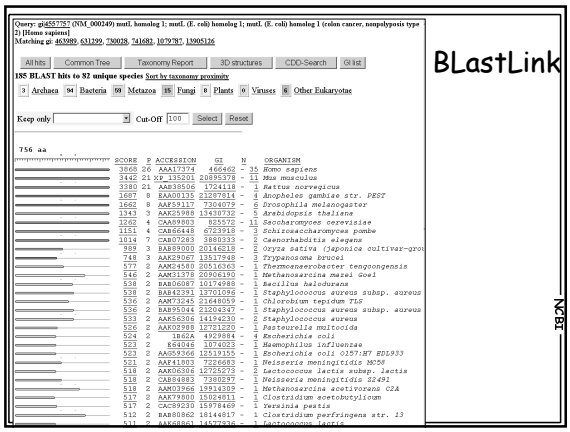
NCBI

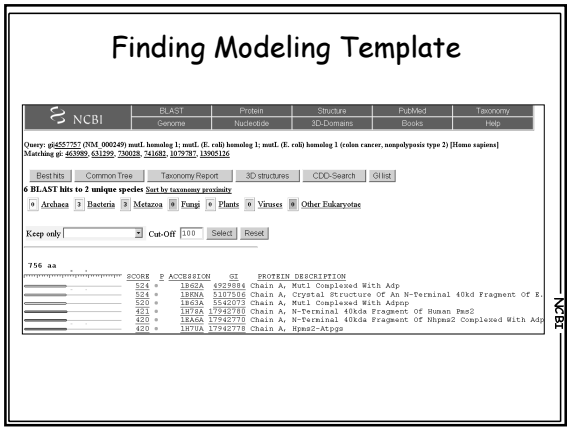

Human DNA Mismatch Repair Protein
NCBI

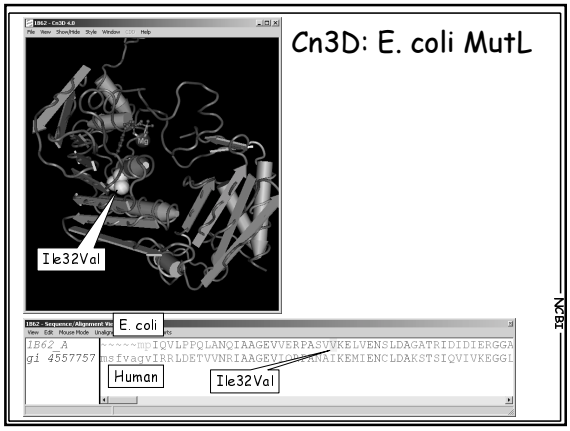

Links and Neighbors Human mutL
Similar Domain Architectures
Human Genome
Nucleotide Polymorphisms
BLastLink
MIM
7: NP_000240 BLink, Domains, MapView, Nucleotide, OMIM, Related Sequences, Domain Relatives, PubMed, SNP, Taxonomy
Conserved Domains
Related Proteins
Literature
DNA
NCBI's Classification
NCBI

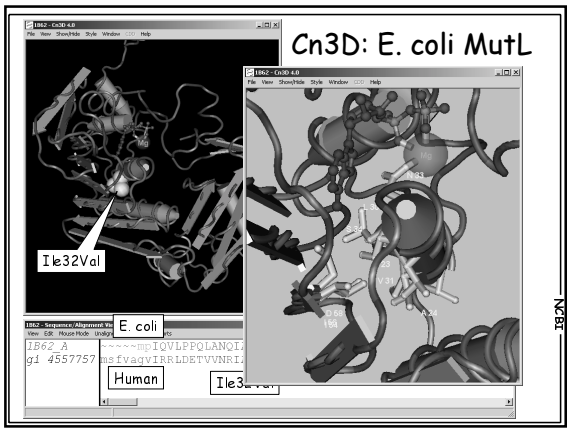

Polymorphisms
NCBI

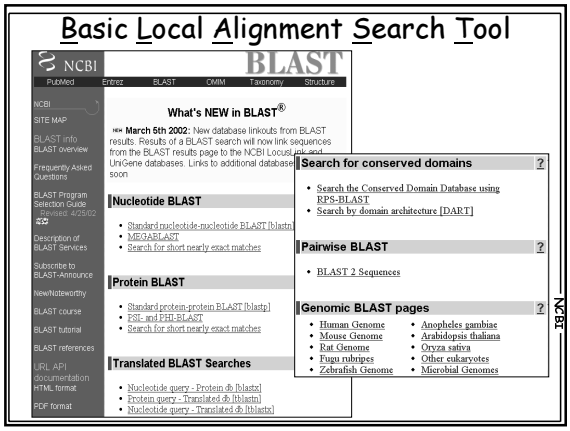

Polymorphisms
NCBI
ENTREZ SNP
LocusLink via analysis of contig annotation: MLH1 mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
Gene Model (contig mRNA transcript) information from genome sequence for NM_000249
Legend
Contig accession | Contig position | Protein accession | Function | dbSNP 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
Query: gi|4557757 (NM_000249) mutL homolog 1; mutL (E. coli) homolog 1; mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2) [Homo sapiens]
Matching gi: 463989, 631299, 730028, 741682, 1079787, 13905126
All hits | Common Tree | Taxonomy Report | 3D structures | CDD-Search | GI list
185 BLAST hits to 82 unique species Sort by taxonomy proximity
3 Archaea 94 Bacteria 59 Metazoa 15 Fungi 8 Plants 0 Viruses 6 Other Eukaryotae
Keep only Cut-Off 100 Select Reset
756 aa
SCORE P ACCESSION GI N ORGANISM
3568 26 AAA17374 466462 - 35 Homo sapiens
3442 21 XP_135201 20895378 - 11 Mus musculus
3380 21 AAB38506 1724118 - 1 Rattus norvegicus
1687 8 EAA00135 21287814 - 4 Anopheles gambiae str. PEST
1662 8 AAF59117 7304079 - 6 Drosophila melanogaster
1343 3 AAK25988 13430732 - 5 Arabidopsis thaliana
1262 4 CAA89803 825572 - 11 Saccharomyces cerevisiae
1151 4 CAB66448 6723918 - 3 Schizosaccharomyces pombe
1014 7 CAB07283 3860333 - 2 Caenorhabditis elegans
989 3 BAB89000 20146218 - 2 Oryza sativa (japonica cultivar-gro
748 3 AAK29067 13517948 - 2 Trypanosoma brucei
577 2 AAM24580 20516363 - 1 Thermoanaerobacter tengcongensis
546 2 AAM31378 20906190 - 1 Methanosarcina mazei Goe1
538 2 BAB06087 10174988 - 1 Bacillus halodurans
538 2 BAB42391 13701096 - 1 Staphylococcus aureus subsp. aureus
536 2 AAM73245 21648059 - 1 Chlorobium tepidum TLS
536 2 BAB95044 21204347 - 2 Staphylococcus aureus subsp. aureus
533 2 AAK56306 14194230 - 2 Staphylococcus aureus
526 2 AAO02988 12721220 - 1 Pasteurella multocida
524 2 1B62A 4929884 - 4 Escherichia coli
523 2 K64046 1074023 - 1 Haemophilus influenzae
523 2 AAG59366 12519155 - 1 Escherichia coli O157:H7 EDL933
521 2 AAF41803 7226683 - 1 Neisseria meningitidis MC58
518 2 AAK06306 12725273 - 2 Lactococcus lactis subsp. lactis
518 2 CAB84883 7380297 - 1 Neisseria meningitidis Z2491
518 2 AAM03966 19914209 - 1 Methanosarcina acetivorans C2A
517 2 AAK79800 15024811 - 1 Clostridium acetobutylicum
517 2 CAC89230 15978469 - 1 Yersinia pestis
512 2 BAB80862 18144817 - 1 Clostridium perfringens str. 13
511 2 AAK68861 14577836 - 1 Lactococcus lactis
NCBI

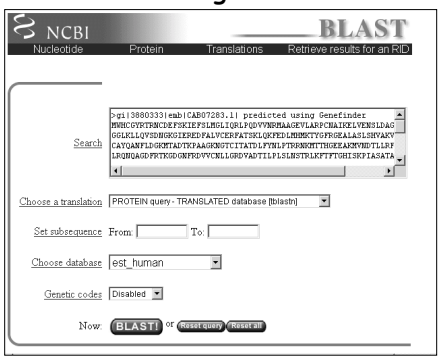

Finding Modeling Template
NCBI
BLAST | Protein | Structure | PubMed | Taxonomy
Genome | Nucleotide | 3D-Domains | Books | Help
Query: gi|4557757 (NM_000249) mutL homolog 1; mutL (E. coli) homolog 1; mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2) [Homo sapiens]
Matching gi: 463989, 631299, 730028, 741682, 1079787, 13905126
Best hits | Common Tree | Taxonomy Report | 3D structures | CDD-Search | GI list
6 BLAST hits to 2 unique species Sort by taxonomy proximity
0 Archaea 3 Bacteria 3 Metazoa 0 Fungi 0 Plants 0 Viruses 0 Other Eukaryotae
Keep only Cut-Off 100 Select Reset
756 aa
SCORE P ACCESSION GI PROTEIN DESCRIPTION
524 * 1B62A 4929884 Chain A, Mutl Complexed With Adp
524 * 1BNNA 5107506 Chain A, Crystal Structure Of An N-Terminal 40kd Fragment Of E.
520 * 1B63A 5542073 Chain A, Mutl Complexed With Adpnp
421 * 1H7XA 17942780 Chain A, N-Terminal 40kda Fragment Of Human Pms2
420 * 1EA6A 17942770 Chain A, N-Terminal 40kda Fragment Of Nhpms2 Complexed With Adp
420 * 1H7UA 17942778 Chain A, Hpms2-Atpgs
NCBI


Cn3D: E. coli MutL
Ile32Val
E. coli
Human
Ile32Val
NCBI

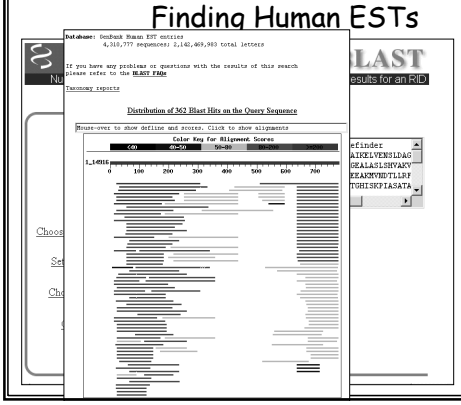

Cn3D: E. coli MutL
Ile32Val
E. coli
Human
NCBI


Basic Local Alignment Search Tool
NCBI
BLAST
What's NEW in BLAST®
Search for conserved domains
Nucleotide BLAST
Pairwise BLAST
Protein BLAST
Genomic BLAST pages
Translated BLAST Searches
NCBI

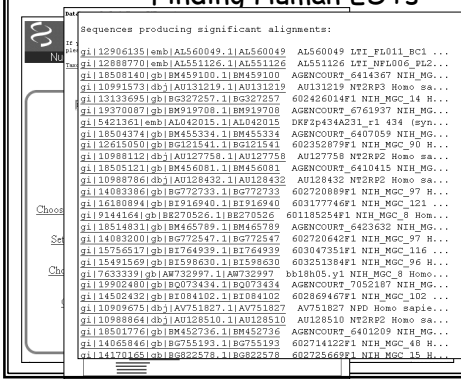
Finding Human ESTs

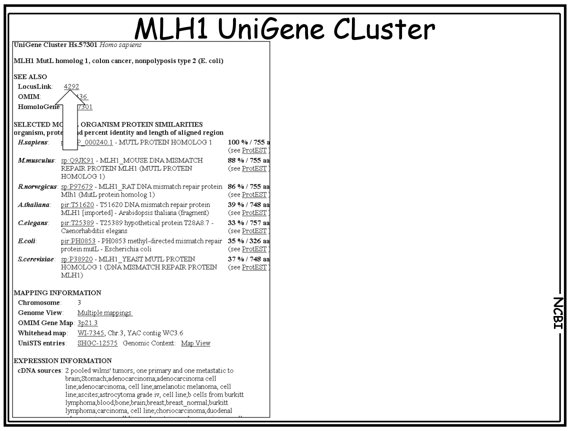


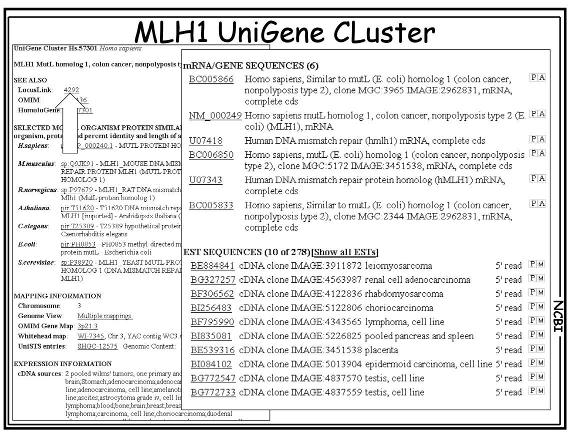
Finding Human ESTs

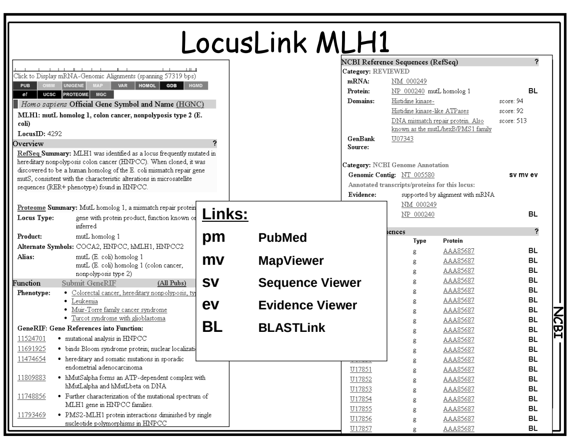


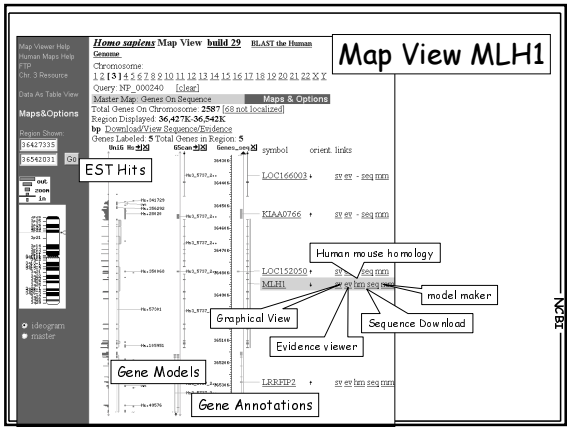
Finding Human ESTs

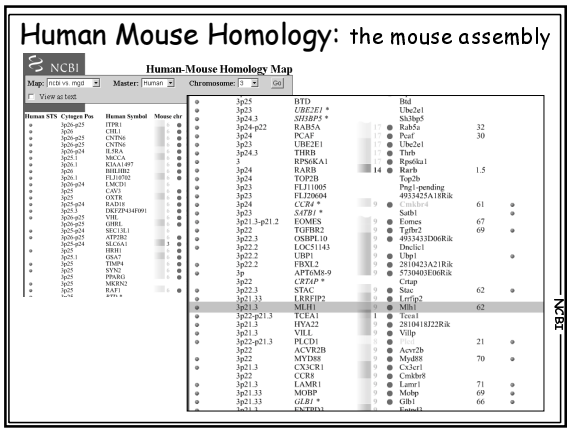


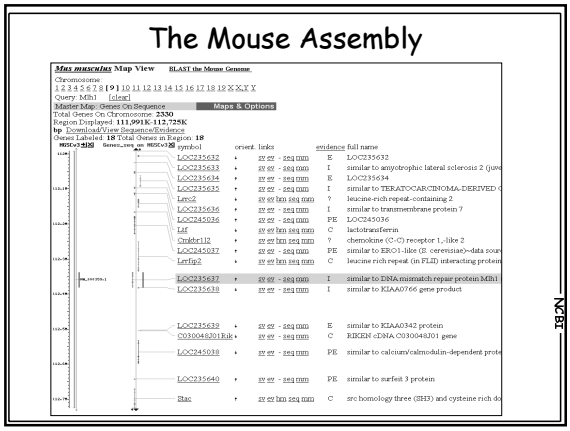

MLH1 UniGene CLuster
UniGene Cluster Hs.57301 Homo sapiens
MLH1 MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
SEE ALSO
MAPPING INFORMATION
EXPRESSION INFORMATION
NCBI


MLH1 UniGene CLuster
mRNA/GENE SEQUENCES (6)
BC005866 Homo sapiens, Similar to mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2), clone MGC:3965 IMAGE:2962831, mRNA, complete cds
NM_000249 Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), mRNA
U07418 Human DNA mismatch repair (hmlh1) mRNA, complete cds
BC006850 Homo sapiens, mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2), clone MGC:5172 IMAGE:3451538, mRNA, complete cds
U07343 Human DNA mismatch repair protein homolog (hMLH1) mRNA, complete cds
BC005833 Homo sapiens, Similar to mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2), clone MGC:2344 IMAGE:2962831, mRNA, complete cds
EST SEQUENCES (10 of 278)[Show all ESTs]
BE884841 cDNA clone IMAGE:3911872 leiomyosarcoma 5' read
BG327257 cDNA clone IMAGE:4563987 renal cell adenocarcinoma 5' read
BF306562 cDNA clone IMAGE:4122836 rhabdomyosarcoma 5' read
BI256483 cDNA clone IMAGE:5122806 choriocarcinoma 5' read
BF795990 cDNA clone IMAGE:4343565 lymphoma, cell line 5' read
BI835081 cDNA clone IMAGE:5226825 pooled pancreas and spleen 5' read
BE539316 cDNA clone IMAGE:3451538 placenta 5' read
BI084102 cDNA clone IMAGE:5013904 epidermoid carcinoma, cell line 5' read
BG772547 cDNA clone IMAGE:4837570 testis, cell line 5' read
BG772733 cDNA clone IMAGE:4837559 testis, cell line 5' read
NCBI


LocusLink MLH1
Links:
pm PubMed
mv MapViewer
sv Sequence Viewer
ev Evidence Viewer
BL BLASTLink
NCBI


Map View MLH1
EST Hits
Human mouse homology
model maker
Graphical View
Sequence Download
Evidence viewer
Gene Models
Gene Annotations
NCBI


Human Mouse Homology: the mouse assembly
NCBI
Human-Mouse Homology Map
NCBI


The Mouse Assembly
NCBI

Live Web Demos

Genetic mapping Fanconi renal syndrome

Finding Prestin Homologues

NCBI