NCBI Molecular Biology 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

GenBank: NCBI's Primary Sequence Database

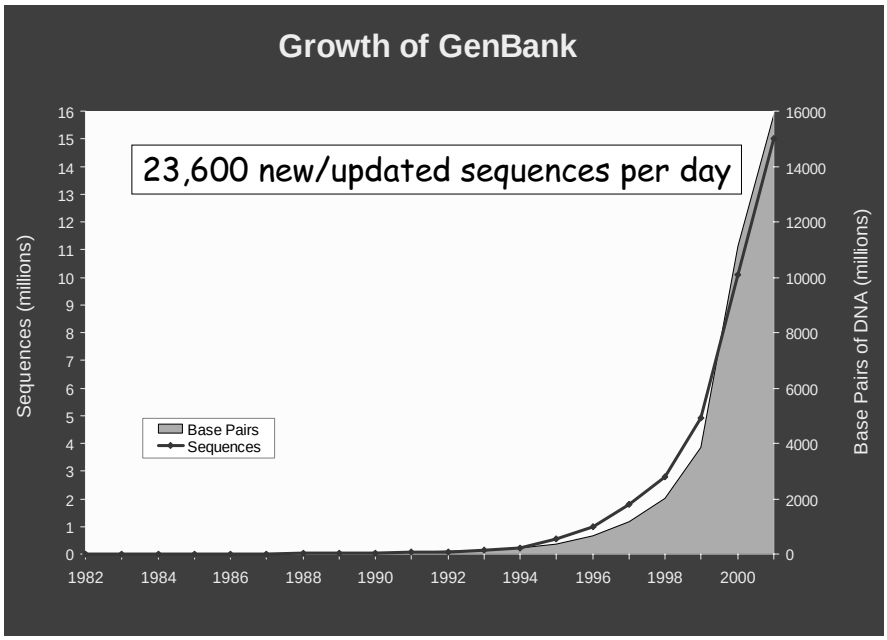
Release 129	April 2002
16,769,983	Records
19,072,679,701	Nucleotides
110,000 +	Species

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

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

66.3 Gigabytes of data

NCBI



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

RefSeq Resource

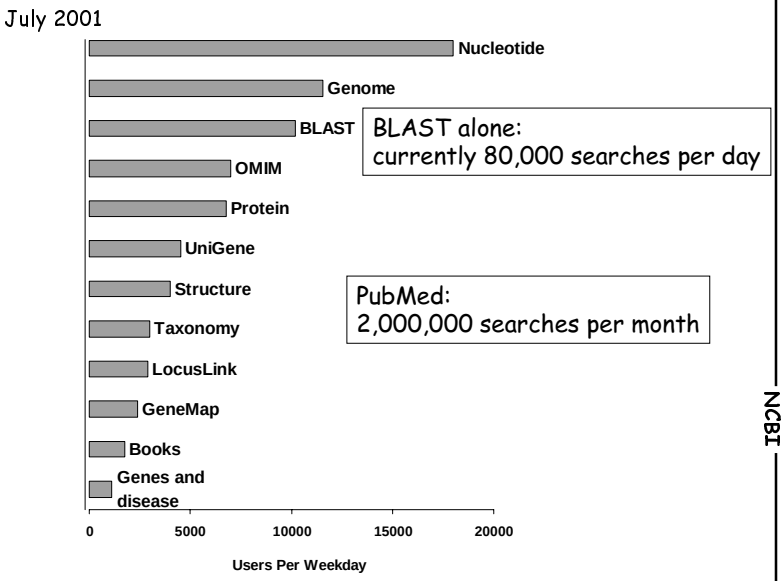
- Genome Oriented Resource
  - A sequence for each macromolecule Central Dogma: Chromosome, mRNA, preprotein, mature protein
  - Linked on a residue by residue basis
  - Objectively non-redundant and comprehensive
- Curated Resource
  - Authoritative source by genome
  - Derived from GenBank but corrected,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

NCBI

Integrated WWW Access: BLAST and Entrez

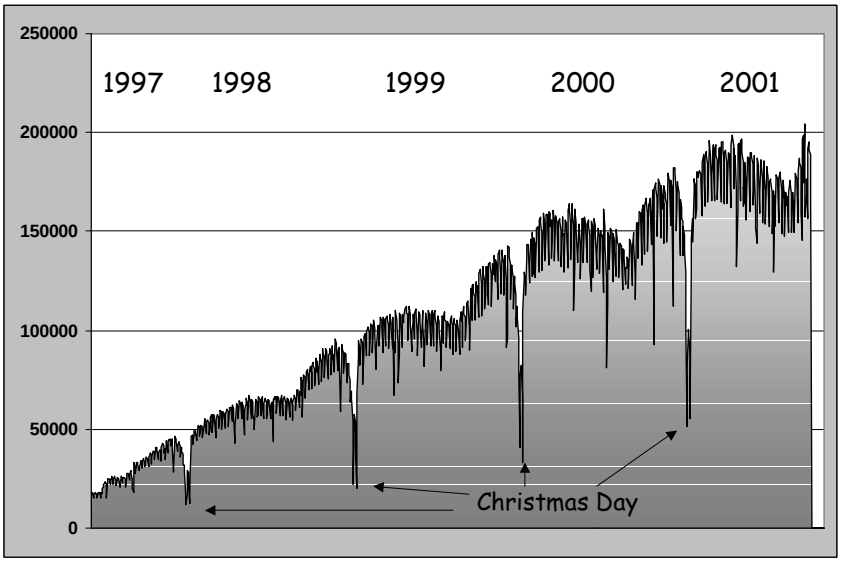
NCBI

Some Web Statistics



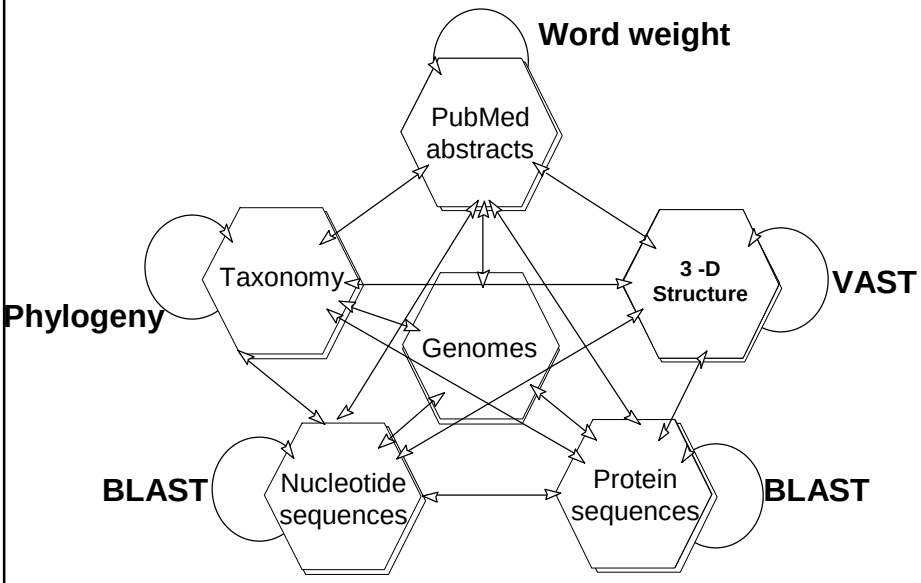
NCBI

Users per day



NCBI

Entrez: Database Integration



Basic Local Alignment Search Tool

NCBI

BLAST

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

NCBI
SITE MAP
BLAST info
BLAST overview
Frequently Asked Questions
New/Noteworthy
Receive e-mail with BLAST announcements
Description of BLAST Services
BLAST course
BLAST tutorial
BLAST references
URL API documentation
HTML format
PDF format
PostScript format

What's NEW in BLAST®

NEW **March 5th 2002:** New database linkouts from BLAST results. Results of a BLAST search will now link sequences from the BLAST results page to the NCBI LocusLink and UniGene databases. Links to additional databases coming soon

Nucleotide BLAST

- Standard nucleotide-nucleotide BLAST [blastn]
- MEGABLAST
- Search for short nearly exact matches

Protein BLAST

- Standard protein-protein BLAST [blastp]
- PSI- and PHI-BLAST
- Search for short nearly exact matches

Translated BLAST Searches

- Nucleotide query - Protein db [blastx]
- Protein query - Translated db [tblastn]
- Nucleotide query - Translated db [tblastx]

Search for conserved domains ?

- Search the Conserved Domain Database using RPS-BLAST
- Search by domain architecture [DART]

Pairwise BLAST ?

- BLAST 2 Sequences

Genomic BLAST pages ?

- Human Genome
- Mouse Genome
- Rat Genome
- Fugu rubripes
- Zebrafish Genome
- Anopheles gambiae
- Arabidopsis thaliana
- Oryza sativa
- Other eukaryotes
- Microbial Genomes

NCBI

Human Genome Resources Find a gene by …

NCBI Home ▸ Genomic Biology ▸ **Human**

Search LocusLink for

The **Human Genome**

A guide to online information resources

Web Resources

BLAST. Compare your sequence to the genome or its gene products.

Cytogenetics. A cytogenetic resource of FISH-mapped, sequence-tagged clones.

dbSNP. Database of SNPs and other genetic variations.

e-PCR. Check your sequence for STSs and view in genomic context.

GEO. Gene Expression Omnibus, a public repository for expression data.

HomoloGene. Putative homologies among human, mouse, rat, and zebrafish.

Homology Map. Blocks of conserved synteny between mouse and human.

LocusLink. Focal point for genes and associated information.

Building an information infrastructure

A challenge facing researchers today is the ability to piece together and analyze the multitudes of data currently being generated through the Human Genome Project. NCBI's Web site serves an an integrated, one-stop, genomic information infrastructure for biomedical researchers from around the world so that they may use this data in their research efforts. **More...**

Working Draft Analysis Published

- NLM Press Release
- NHGRI Press Release
- Interactive Tour of the Genome
- NCBI Genome Analysis Pipeline
- Nature (2/15/01) Human Genome Issue
- Science (2/16/01) Human Genome Issue

MapViewer tips and tricks

When browsing the genome using the new MapViewer, click on Display Settings to

OMIM. Guide to genes and inherited disorders maintained by JHU and collaborators.

RefSeq. Reference sequences of genomic contigs, mRNAs, and proteins.

SAGEmap. Gene expression results from SAGE tags mapped to sequences.

Sequencing. Summary of human genome sequencing progress.

MapViewer. Interactive viewer for genome maps,sequence, and genes.

UniGene. Organization of transcribed sequences into gene-based clusters.

UniSTS. A non- redundant collection of STSs with links to maps and sequence.

FTP Sites

dbSNP
LocusLink
Genomes
RefSeq
UniGene

Below are three views of the BRCA2 locus using different display options. Click the image to see the full MapViewer display.

A. Genes

B. Variations, juxtaposed with genes

C. Several STS maps, juxtaposed with genes

Other genomes

M.musculus
R.norvegicus
D.rerio
D.melanogaster
C.elegans
S.cerevisiae

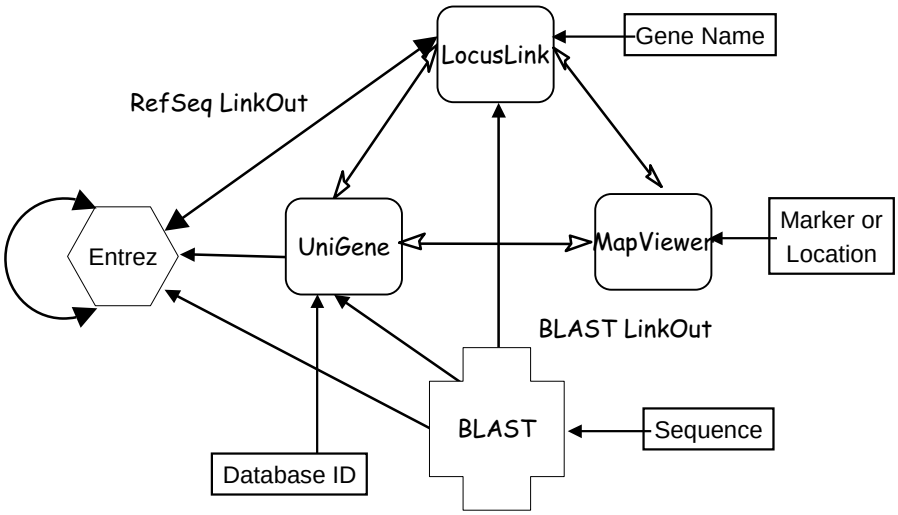
organelles
bacteria
viruses

Frequently asked questions

What is a reference sequence?
How are gene symbols chosen for LocusLink and RefSeq?
How many inherited diseases have a known sequence?
How much of the human genome has been sequenced?
How were the genomic sequence contigs constructed?
How can I customize the MapViewer display?
What classes of genetic variation are included in dbSNP?
How can I deposit gene expression data in a public database?

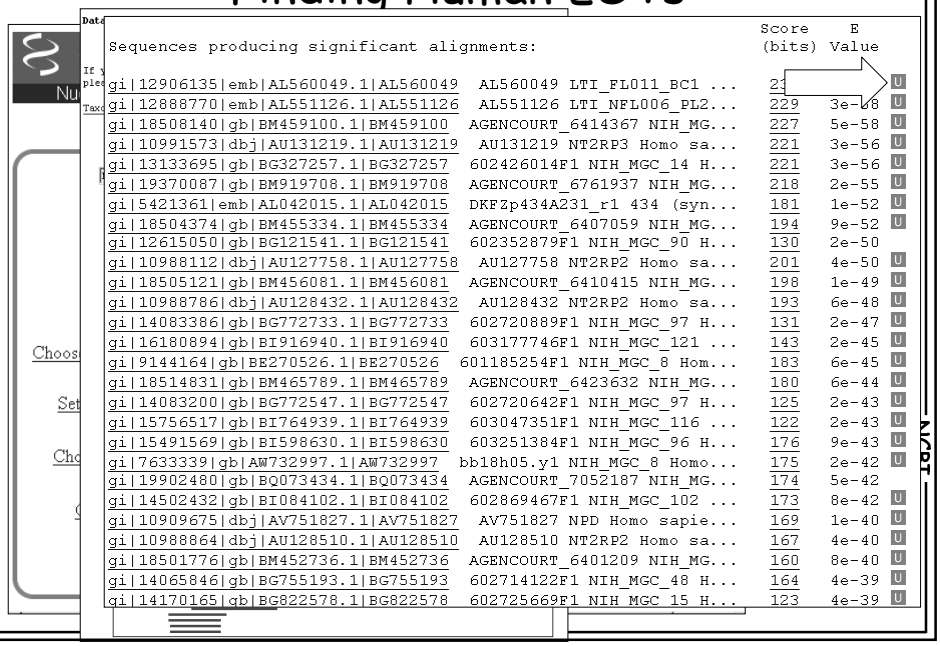
NCBI

Genome Resources Integration



NCBI

Finding Human ESTs



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

gi|12906135|emb|AL560049.1|AL560049  AL560049 LTI_FL011_BC1 ...   23
gi|12888770|emb|AL551126.1|AL551126  AL551126 LTI_NFL006_PL2...   229   3e-58
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|BG822578.1|BG822578  602725669F1 NIH_MGC_15 H...   123   4e-39
```

NCBI

MLH1 UniGene CLuster

UniGene Cluster Hs.57301 *Homo sapiens*

MLH1 MutL homolog 1, colon cancer, nonpolyposis t

SEE ALSO
LocusLink: 4292
OMIM: 36
HomoloGene: 7301

SELECTED MC ORGANISM PROTEIN SIMILAI
organism, prote d percent identity and length of a
H.sapiens: P_000240.1 - MUTL PROTEIN HO
M.musculus: sp:Q9JK91 - MLH1_MOUSE DNA MISN REPAIR PROTEIN MLH1 (MUTL PROT HOMOLOG 1)
R.norvegicus: sp:P97679 - MLH1_RAT DNA mismatch Mlh1 (MutL protein homolog 1)
A.thaliana: pir:T51620 - T51620 DNA mismatch repa MLH1 [imported] - Arabidopsis thaliana (
C.elegans: pir:T25389 - T25389 hypothetical protei Caenorhabditis elegans
E.coli: pir:PH0853 - PH0853 methyl-directed m protein mutL - Escherichia coli
S.cerevisiae: sp:P38920 - MLH1_YEAST MUTL PRO HOMOLOG 1 (DNA MISMATCH REPAI MLH1)

MAPPING INFORMATION
Chromosome: 3
Genome View: Multiple mappings.
OMIM Gene Map: 3p21.3
Whitehead map: WI-7345, Chr.3, YAC contig WC3
UniSTS entries: SHGC-12575 Genomic Context:

EXPRESSION INFORMATION
cDNA sources: 2 pooled wilms' tumors, one primary and brain;Stomach;adenocarcinoma;adenocar line;adenocarcinoma, cell line;amelanot line;ascites;astrocytoma grade iv, cell lin lymphoma;blood;bone;brain;breast;breas lymphoma;carcinoma, cell line;choriocarcinoma;duodenal

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	P A
NM_000249	Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), mRNA	P A
U07418	Human DNA mismatch repair (hmlh1) mRNA, complete cds	P A
BC006850	Homo sapiens, mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2), clone MGC:5172 IMAGE:3451538, mRNA, complete cds	P A
U07343	Human DNA mismatch repair protein homolog (hMLH1) mRNA, complete cds	P A
BC005833	Homo sapiens, Similar to mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2), clone MGC:2344 IMAGE:2962831, mRNA, complete cds	P A

EST SEQUENCES (10 of 278)[Show all ESTs]

BE884841	cDNA clone IMAGE:3911872 leiomyosarcoma	5' read	P M
BG327257	cDNA clone IMAGE:4563987 renal cell adenocarcinoma	5' read	P M
BF306562	cDNA clone IMAGE:4122836 rhabdomyosarcoma	5' read	P M
BI256483	cDNA clone IMAGE:5122806 choriocarcinoma	5' read	P M
BF795990	cDNA clone IMAGE:4343565 lymphoma, cell line	5' read	P M
BI835081	cDNA clone IMAGE:5226825 pooled pancreas and spleen	5' read	P M
BE539316	cDNA clone IMAGE:3451538 placenta	5' read	P M
BI084102	cDNA clone IMAGE:5013904 epidermoid carcinoma, cell line	5' read	P M
BG772547	cDNA clone IMAGE:4837570 testis, cell line	5' read	P M
BG772733	cDNA clone IMAGE:4837559 testis, cell line	5' read	P M

NCBI

LocusLink MLH1

Click to Display mRNA-Genomic Alignments (spanning 57319 bps)

PUB | OMIM | UNIGENE | MAP | VAR | HOMOL | GDB | HGMD
e! | UCSC | PROTEOME | MGC

Homo sapiens **Official Gene Symbol and Name** (HGNC)

MLH1: mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
LocusID: 4292

Overview ?

RefSeq Summary: MLH1 was identified as a locus frequently mutated in hereditary nonpolyposis colon cancer (HNPCC). When cloned, it was discovered to be a human homolog of the E. coli mismatch repair gene mutS, consistent with the characteristic alterations in microsatellite sequences (RER+ phenotype) found in HNPCC.

Proteome Summary: MutL homolog 1, a mismatch repair protein
Locus Type: gene with protein product, function known o inferred
Product: mutL homolog 1
Alternate Symbols: COCA2, HNPCC, hMLH1, HNPCC2
Alias: mutL (E. coli) homolog 1
mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2)

Function Submit GeneRIF (All Pubs)

Phenotype:
- Colorectal cancer, hereditary nonpolyposis, ty
- Leukemia
- Muir-Torre family cancer syndrome
- Turcot syndrome with glioblastoma

GeneRIF: Gene References into Function:

11524701	mutational analysis in HNPCC
11691925	binds Bloom syndrome protein, nuclear localizati
11474654	hereditary and somatic mutations in sporadic endometrial adenocarcinoma
11809883	hMutSalpha forms an ATP-dependent complex with hMutLalpha and hMutLbeta on DNA
11748856	Further characterization of the mutational spectrum of MLH1 gene in HNPCC families.
11793469	PMS2-MLH1 protein interactions diminished by single nucleotide polymorphisms in HNPCC

Links:

pm	PubMed
mv	MapViewer
sv	Sequence Viewer
ev	Evidence Viewer
BL	BLASTLink

NCBI Reference Sequences (RefSeq) ?

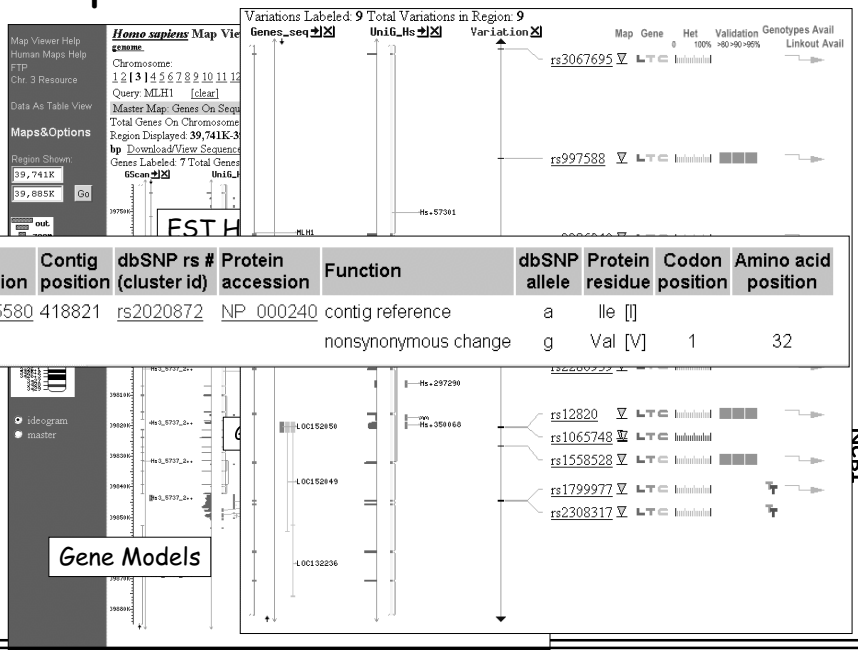
Category: REVIEWED
mRNA: NM_000249
Protein: NP_000240 mutL homolog 1 BL
Domains: Histidine kinase- score: 94
Histidine kinase-like ATPases score: 92
DNA mismatch repair protein. Also known as the mutL/hexB/PMS1 family score: 513
GenBank Source: U07343

Category: NCBI Genome Annotation
Genomic Contig: NT_005580 sv mv ev
Annotated transcripts/proteins for this locus:
Evidence: supported by alignment with mRNA
NM_000249
NP_000240 BL

	Type	Protein	
	g	AAA85687	BL
	g	AAA85687	BL
	g	AAA85687	BL
	g	AAA85687	BL
	g	AAA85687	BL
	g	AAA85687	BL
	g	AAA85687	BL
	g	AAA85687	BL
	g	AAA85687	BL
	g	AAA85687	BL
	g	AAA85687	BL
	g	AAA85687	BL
U17851	g	AAA85687	BL
U17852	g	AAA85687	BL
U17853	g	AAA85687	BL
U17854	g	AAA85687	BL
U17855	g	AAA85687	BL
U17856	g	AAA85687	BL
U17857	g	AAA85687	BL

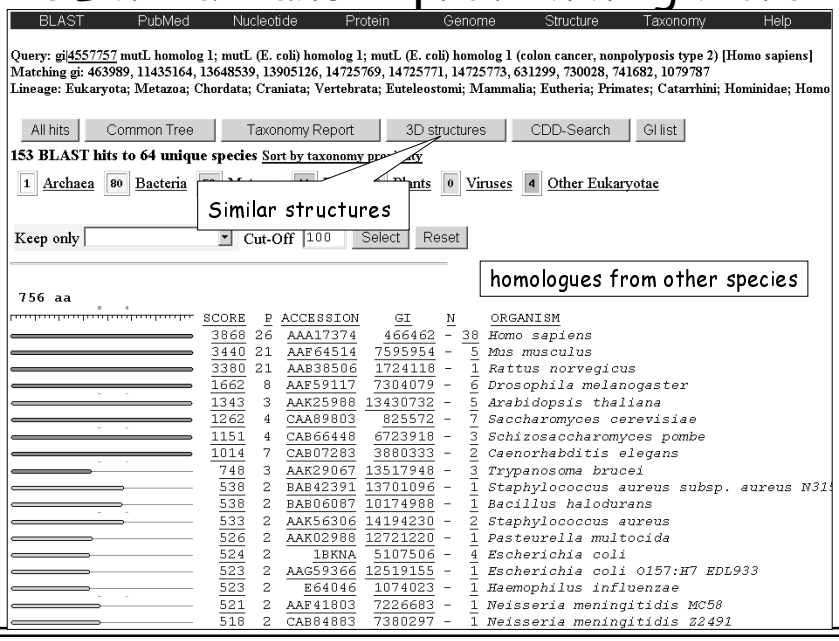
NCBI

Map View MLH1 -> Variation Data

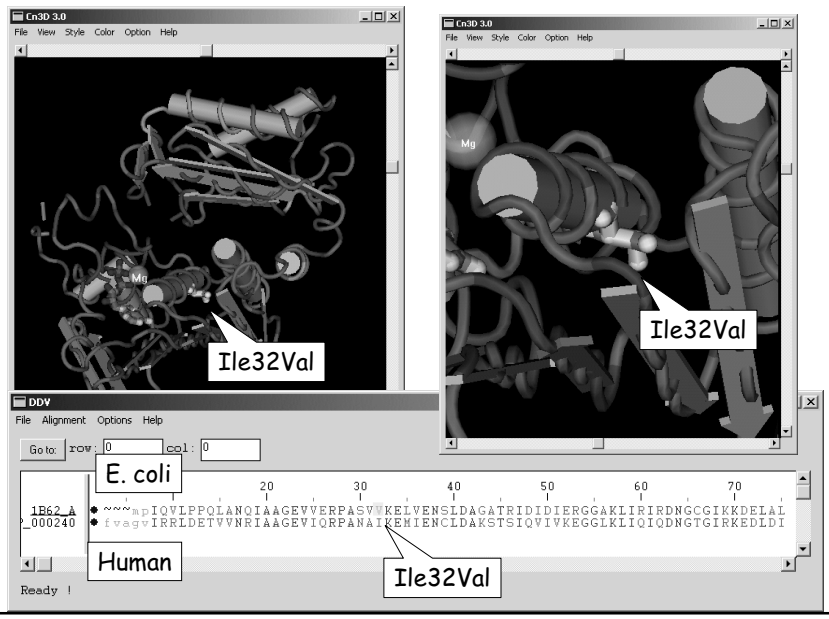


Contig accession	Contig position	dbSNP rs # (cluster id)	Protein accession	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
NT_005580	418821	rs2020872	NP_000240	contig reference	a	Ile [I]		
				nonsynonymous change	g	Val [V]	1	32

BLink: advanced protein neighbors



Cn3D: Finding a Modeling Template



NCBI