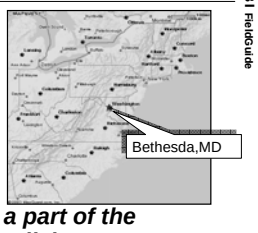
NCBI FieldGuide

NCBI Molecular Biology Resources

April 4, 2006

NCBI FieldGuide

The National Center for Biotechnology Information



Created in 1988 as a part of the National Library of Medicine at NIH

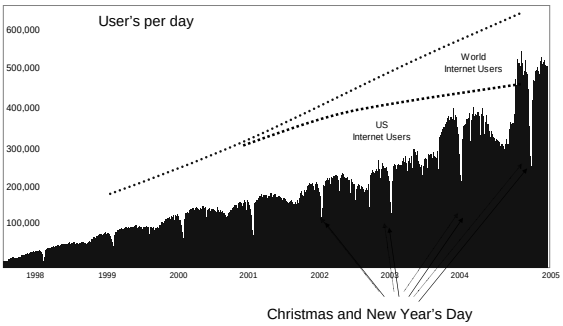
- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

NCBI FieldGuide

Web Access: www.ncbi.nlm.nih.gov

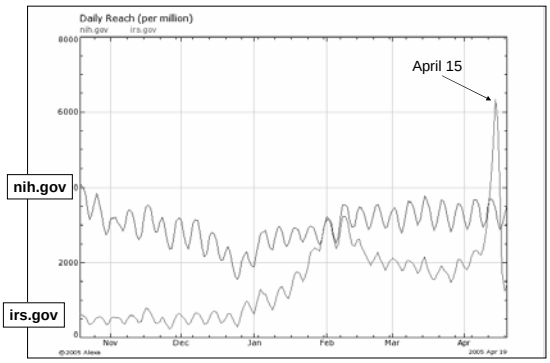
NCBI FieldGuide

Website Traffic



NCBI FieldGuide

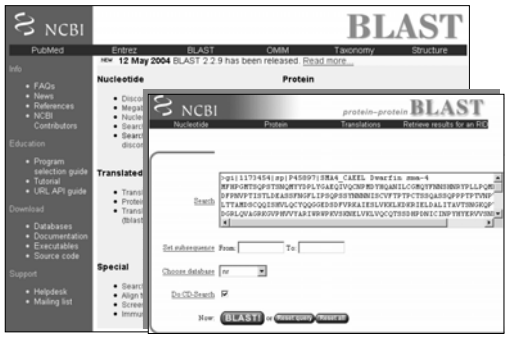
Government Web Traffic



NCBI FieldGuide

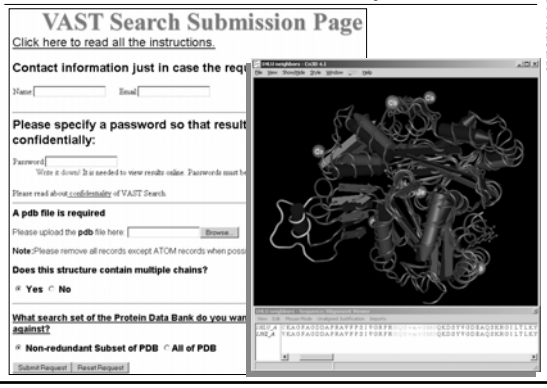
The Entrez System: Text Searches

BLAST: Sequence Similarity Searches



NCBI FieldGuide

VAST: Structure Similarity Searches



NCBI FieldGuide

Types of Databases

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

NCBI FieldGuide

What is GenBank?

NCBI's Primary Sequence Database

- Nucleotide only sequence database
- Archival in nature
  - Historical
  - Reflective of submitter point of view (subjective)
  - Redundant
- GenBank Data
  - Direct submissions (traditional records)
  - Batch submissions (EST, GSS, STS)
  - ftp accounts (genome data)
- Three collaborating databases
  - GenBank
  - DNA Database of Japan (DDBJ)
  - European Molecular Biology Laboratory (EMBL) Database

NCBI FieldGuide

GenBank: NCBI's Primary Sequence Database

Release 152	February 2006
54,584,635	Records (non-WGS)
59,750,386,305	Nucleotides (non-WGS)
63,183,065,091	Nucleotides (WGS)
122,933,451,396	Total

GenBank exceeds 100 GigaBases !

>140,000 Species

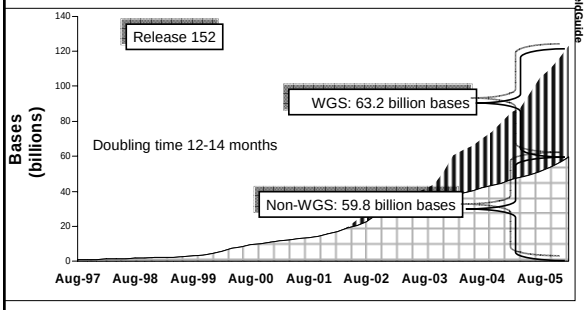
216 Gigabytes *908 files*

- full release every two months
- incremental and cumulative updates daily
- available only via ftp

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

NCBI FieldGuide

The Growth of GenBank



NCBI FieldGuide

Organization of GenBank: Traditional Divisions

NCBI FieldGuide

Records are divided into 17 Divisions.

- 12 Traditional
- 6 Bulk

Traditional Divisions:

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

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

Entrez query: `gbdiv_xxx[Properties]`

Organization of GenBank: Bulk Divisions

NCBI FieldGuide

Records are divided into 17 Divisions.

- 11 Traditional
- 6 Bulk

BULK Divisions:

- Batch Submission (Email and FTP)
- Inaccurate
- Poorly characterized

EST Expressed Sequence Tag
GSS Genome Survey Sequence
HTG High Throughput Genomic
STS Sequence Tagged Site
HTC High Throughput cDNA
PAT Patent

Entrez query: `gbdiv_xxx[Properties]`

EST Division: Expressed Sequence Tags

NCBI FieldGuide

`gbdiv_est[Properties]`

```
>IMAGE:275615 5' mRNA sequence
GACAGCATTCGGGCCGAGATGTCTCGCTCCGTGGCCTTAGCTGTGCTCGCGCTACTCTCTCTTTCTGG
TGGAGGTATCCAGCGTACTCCAAAGATTCAGGTTTACTCACGTCATCCAGCAGAGAATGGAAAGTCAA
TTCCTGAATTGCTATGTGTCTGGGTTTCATCCATCCGACATTGAAGTTGACTTACTGAAGAATGGAGA
GAATTGAAAAAGTGGAGCATTCAGACTTGTCTTTCAGCAAGGACTGGTCTTTCTATCTCTTGTACTAC
TGAATTCACCCCCACTGAAAAAGATGAGTATGCCTGCCGTGTTGAACCATGTNGACTTTGTCACAGNC
AAGTTNAGTTTAAGTGGGNATCGAGACATGTAAGGCAGGCATCATGGGAGGTTTTGAAGNATGCCGCN
TTGGATTGGGATGAATTCCAAATTTCTGGTTTGCTTGNTTTTTTAATATTGGATATGCTTTTG
```

```
>IMAGE:275615 3', mRNA sequence
NNTCAAGTTTTATGATTTATTTAACTTGTGGAACAAAAATAAACCAGATTAACCACAACCATGCCTTA
TTATCAAATGTATAAGANGTAAATATGAATCTTATATGACAAAATGTTTCATTCATTATAACAAATTT
AATAATCCTGTCAATNATATTTCTAAATTTTCCCCCAAATTCTAAGCAGAGTATGTAAATTGGAAGTT
CTTATGCACGCTTAACTATCTTAACAAGCTTTGAGTGCAAGAGATTGANGAGTTCAAATCTGACCAAG
GTTGATGTTGGATAAGAGAATTCTCTGCTCCCCACCTCTANGTTGCCAGCCCTC
```

make cDNA library ⇨ **80-100,000 unique cDNA clones in library**

What is UniGene?

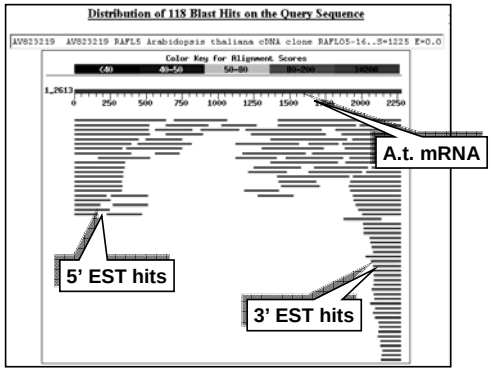
A gene-oriented view of sequence entries

NCBI FieldGuide

- MegaBlast based automated sequence clustering
- Now informed by genome hits *New!*
- 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

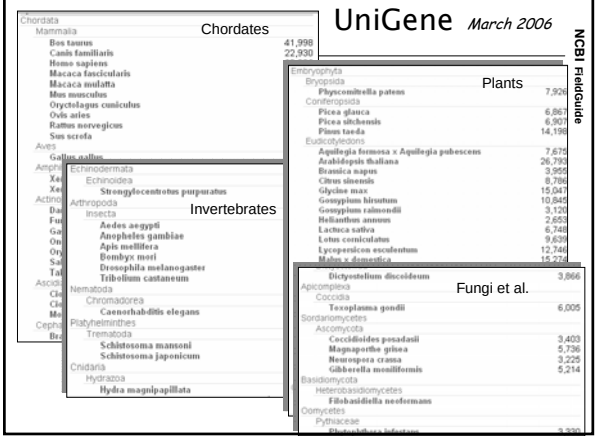
EST hits: *Arabidopsis* serine protease mRNA

NCBI FieldGuide



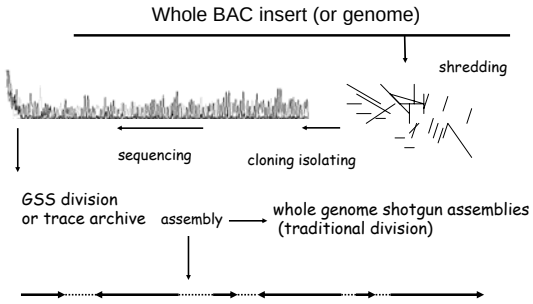
UniGene *March 2006*

NCBI FieldGuide



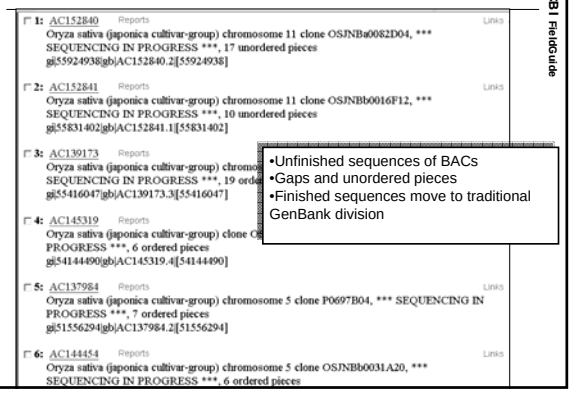
Genome Sequencing – HTG, GSS,(WGS)

NCBI FieldGuide



HTG Division: Rice Draft Sequences

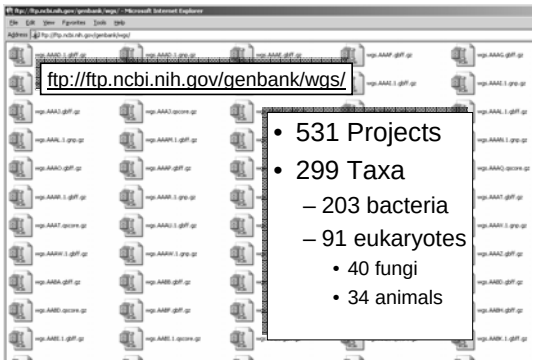
NCBI FieldGuide



- Unfinished sequences of BACs
- Gaps and unordered pieces
- Finished sequences move to traditional GenBank division

Whole Genome Shotgun Projects

NCBI FieldGuide



- 531 Projects
- 299 Taxa
  - 203 bacteria
  - 91 eukaryotes
    - 40 fungi
    - 34 animals

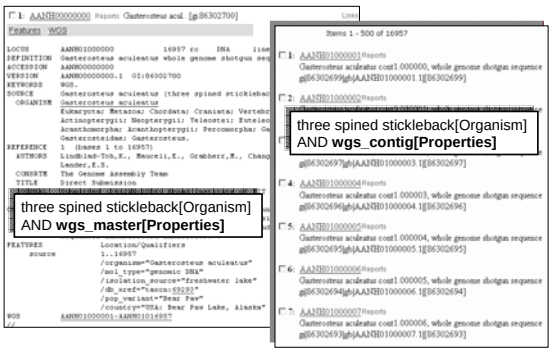
Mammalian WGS

NCBI FieldGuide

- Nine banded armadillo
- Hedgehog
- African elephant
- Dog
- Eurasian shrew
- Cow
- Domestic rabbit
- Guinea pig
- Mouse
- Rat
- Human
- Chimpanzee
- Gray short-tailed opossum

Accessing WGS Records: Entrez

NCBI FieldGuide



NCBI FieldGuide

Derivative Sequence Databases

GenPept
RefSeq

NCBI Protein: Derivative Database

NCBI FieldGuide

GenPept	**5,223,665**
RefSeq	**2,429,517**
Third Party Annotation	**4,495**
Swiss Prot	**207,038**
PIR	**39,615**
PRF	**12,079**
PDB	**76,628**
Total	**7,993,037**
BLAST nr total	**3,310,354**

GenPept: GenBank CDS translations

NCBI FieldGuide

```
FEATURES             Location/Qualifiers
     source          1..2484
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="3"
                     /map="3p22-p23"
     gene            1..2484
                     /gene="M
     CDS             22..2292
                     /gene="M
                     /note="homolog of S. cerevisiae PMS1 (Swiss-Prot Accession
                     Number P14242), S. cerevisiae MLH1 (GenBank Accession
                     Number U07187), E. coli MUTL (Swiss-Prot Accession Number
                     P23367), Salmonella typhimurium MUTL (Swiss-Prot Accession
                     Number P14161) and Streptococcus pneumoniae (Swiss-Prot
                     Accession Number P14160)"
                     /codon_start=1
                     /product="DNA mismatch repair protein homolog"
                     /protein_id="AAC50285.1"
                     /db_xref="GI:463989"
                     /translation="MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKS
                     TSIQVIVKEGGLKLIQIQDNGTGIRKEDLDIVCERFTTSKLQSFEDLASISTYGFRGE
                     ALASISHVAHVTITTKTADGKCAYRASYSDGKLKAPPKPCAGNQGTQITVEDLFYNIA
                     TRRKALKNPSEEYGKILEVVGRYSVHNAGISFSVKKQGETVADVRTLPNASTVDNIRS
```

```
>gi|463989|gb|AAC50285.1| DNA mismatch repair prote...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

Redundant Proteins

NCBI FieldGuide

```
>gi|463989|gb|AAC50285.1| DNA mismatch repair prote...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

```
>gi|13905126|gb|AAH06850.1| MutL protein homolog 1 ...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

```
>gi|1079787|gb|AAA82079.1| DNA mismatch repair prot...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

GenPept

```
>gi|4557757|ref|NP_000240.1| MutL protein homolog 1...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

NCBI RefSeq

```
>gi|730028|sp|P40692|MLH1_HUMAN DNA mismatch repair...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

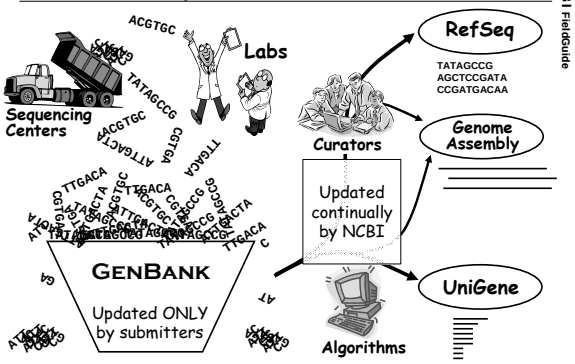
Swiss-Prot

```
>gi|741682|prf||2007430A DNA mismatch repair protei...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

PRF

Primary vs. Derivative Sequence Databases

NCBI FieldGuide



RefSeq: NCBI's Derivative Sequence Database

NCBI FieldGuide

- **Curated transcripts and proteins [NM_, NP_]**
  - reviewed
  - human, mouse, rat, cow, fruit fly, zebrafish, arabidopsis *C.elegans*
- **Model transcripts and proteins [XM_, XP_]**
- **Assembled Genomic Regions [NT_, NW_]**
  - draft human genome
  - mouse genome
- **Chromosome records [NC_]**
  - microbial
  - organelle

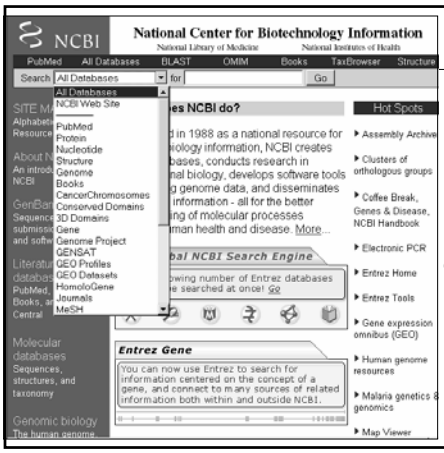
`srcdb_refseq[Properties]`

ftp://ftp.ncbi.nih.gov/refseq/release/

Other NCBI Databases

NCBI FieldGuide

•dbSNP:	nucleotide polymorphism
•Geo:	Gene Expression Omnibus microarray and other expression data
•Gene:	gene records Unifies LocusLink and Microbial Genomes
•Structure:	imported structures (PDB) Cn3D viewer, NCBI curation
•CDD:	conserved domain database Protein families (COGs and KOGs) Single domains (PFAM, SMART, CD)

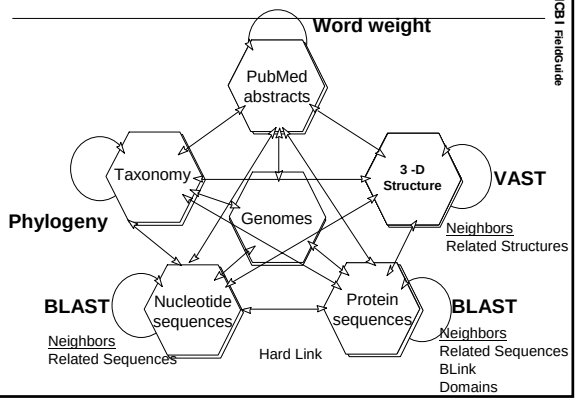


WWW Access

Entrez & BLAST

NCBI FieldGuide

Entrez: Database Integration



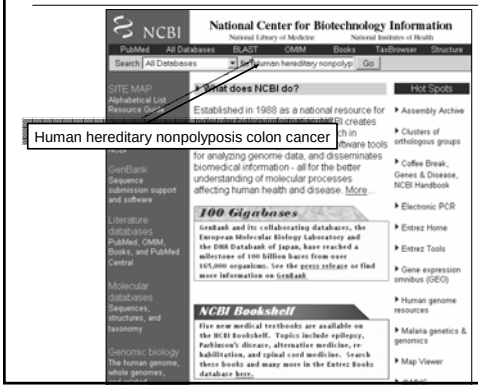
NCBI FieldGuide

Database Searching with Entrez

- ◆Using limits and field restriction to find human MutL homolog
- ◆Linking and neighboring with MutL
- ◆Mapping SNPs onto structure and the genome

NCBI FieldGuide

Global NCBI (Entrez) Search

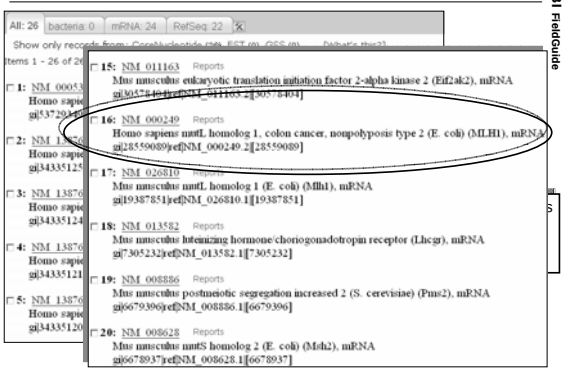


NCBI FieldGuide

Global Entrez Search Results

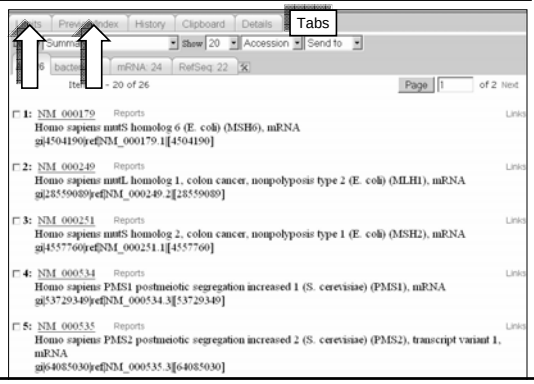
NCBI FieldGuide

Nucleotide Sequences



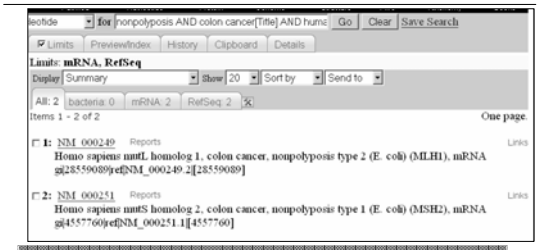
NCBI FieldGuide

Advanced Search Options



NCBI FieldGuide

More Precise Nucleotides Search



nonpolyposis[All Fields] AND colon cancer[Title] AND human[Organism] AND biomol_mrna[Properties] AND srcdb_refseq[Properites]

NCBI FieldGuide

Useful Field Restrictions

[Title]: Definition line in GenBank / GenPept format shown in Summary format

glyceraldehyde 3 phosphate dehydrogenase[Title]

[Organism]: NCBI's taxonomy. Organizing system for molecular databases

mouse[organism]; green plants[organism]; Streptomyces coelicolor[organism]

[Properties]: molecule type, location, database source

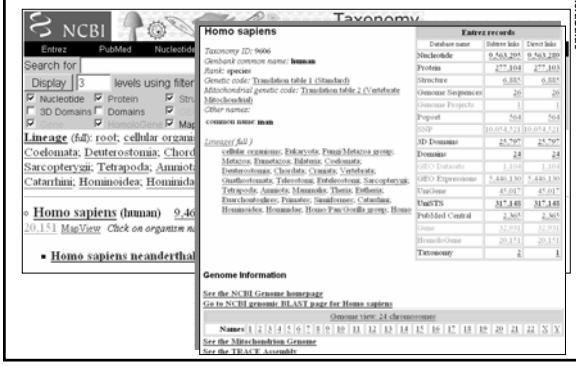
biomol_mrna[properties]; biomol_genomic[properties]; gene_in_mitochondrion[properties]; srcdb pdb[properties]

[Filter]: subsets of data, Entrez links

all[filter]; nucleotide mapview[filter]; nucleotide omim[filter]

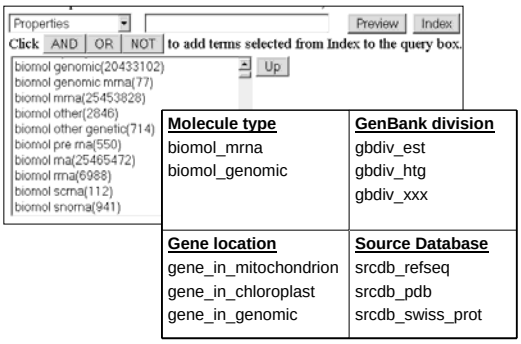
NCBI FieldGuide

Organism Field: NCBI's Taxonomy



NCBI FieldGuide

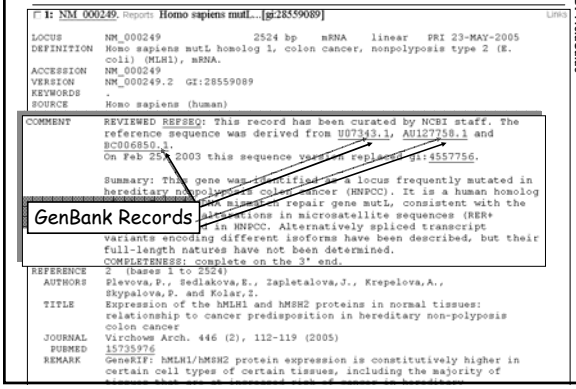
Useful Properties Field Terms



Molecule type	GenBank division
biomol_mrna	gbdiv_est
biomol_genomic	gbdiv_htg
	gbdiv_xxx
Gene location	**Source Database**
gene_in_mitochondrion	srcdb_refseq
gene_in_chloroplast	srcdb_pdb
gene_in_genomic	srcdb_swiss_prot

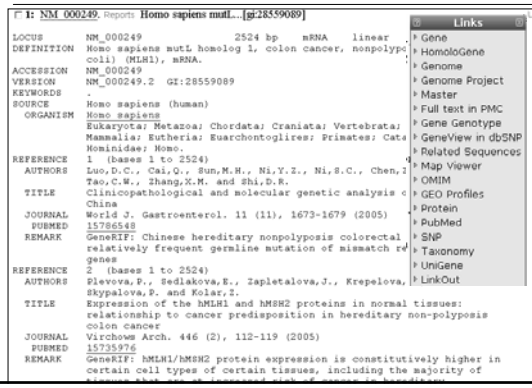
NCBI FieldGuide

Human MutL RefSeq



NCBI FieldGuide

NM_000249: Links



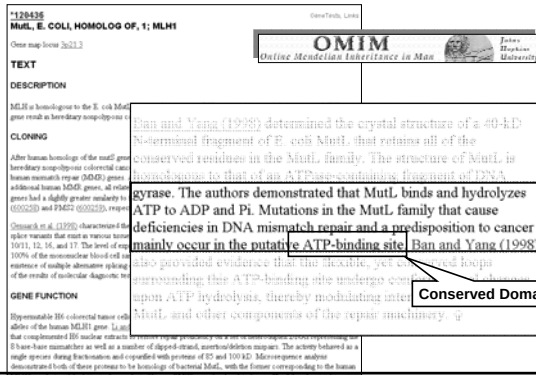
NCBI FieldGuide

Literature Links

OMIM

NCBI FieldGuide

OMIM: Human Disease Genes



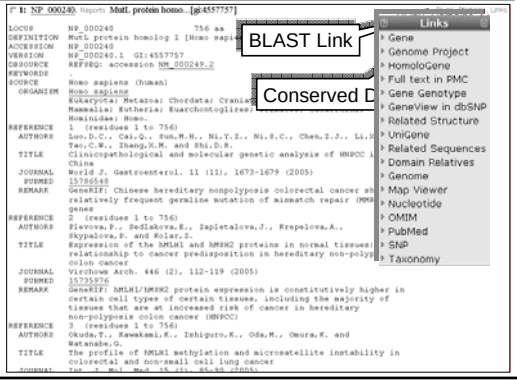
NCBI FieldGuide

Sequence Links

Finding Homologs and Structures

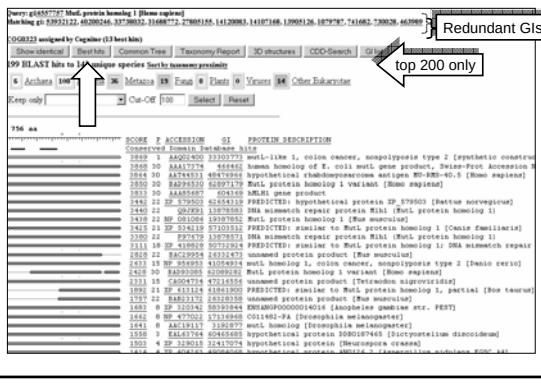
NCBI FieldGuide

Protein Link



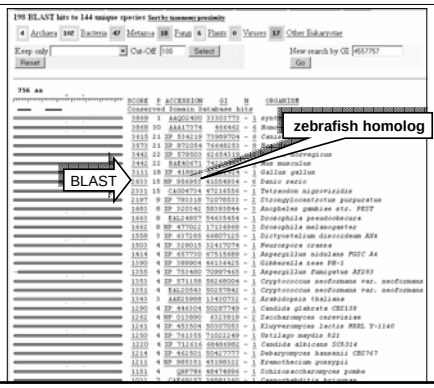
NCBI FieldGuide

BLink: BLAST Link



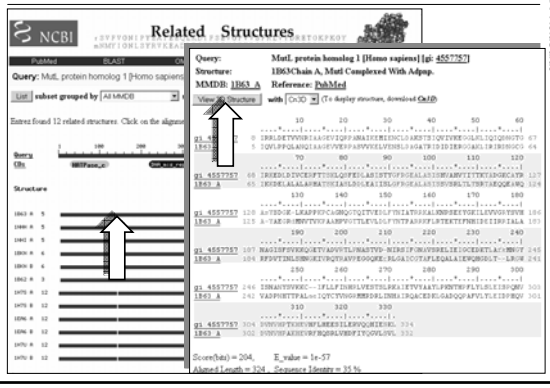
NCBI FieldGuide

BLink: non-redundant relatives



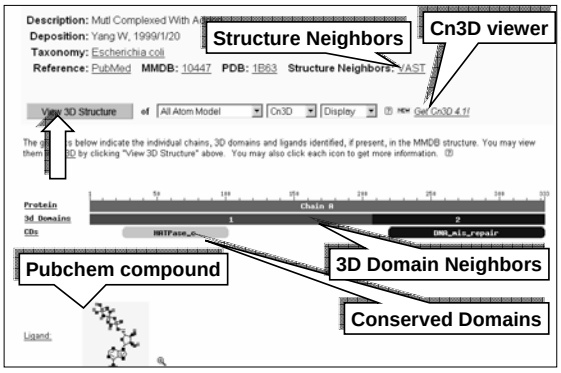
NCBI FieldGuide

Related Structures



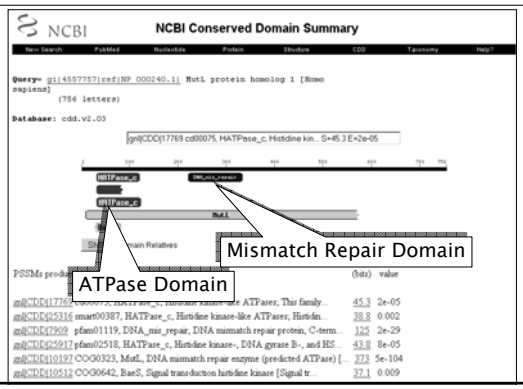
NCBI FieldGuide

E. coli MutL Structure



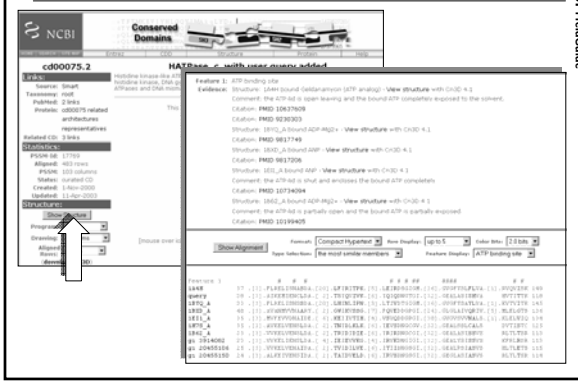
NCBI FieldGuide

MLH1 Domain Structure: CDD



NCBI FieldGuide

MLH1: ATPase Domain

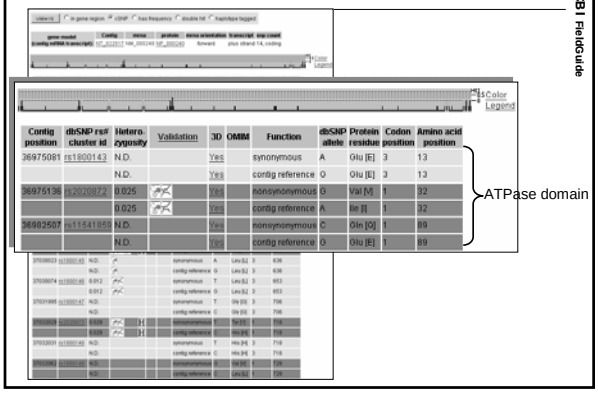


NCBI FieldGuide

Mapping Polymorphisms onto Structure

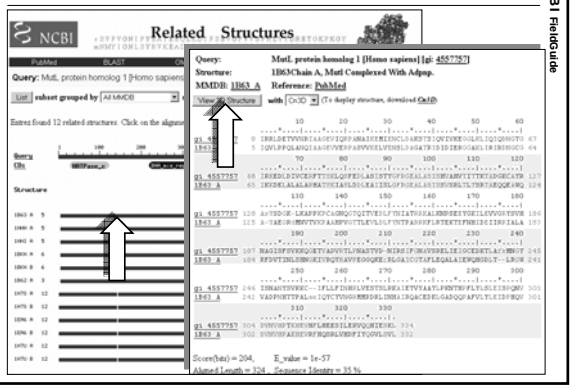
NCBI FieldGuide

GeneView: Variations Human MLH1



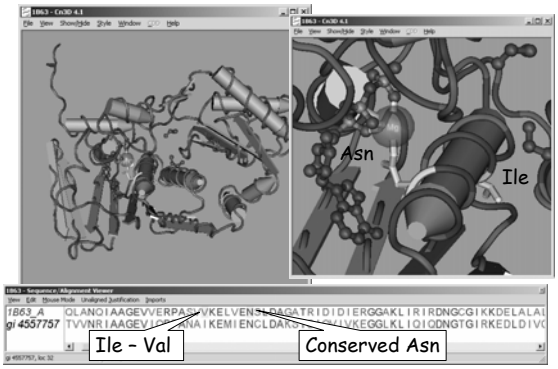
NCBI FieldGuide

Related Structures



NCBI FieldGuide

Mapping Variation Onto Structure

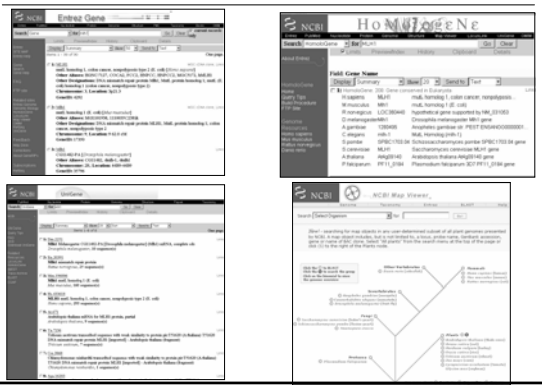


NCBI FieldGuide

Genome Resources

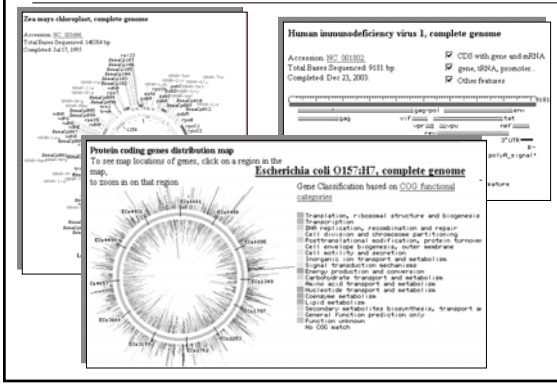
NCBI FieldGuide

Higher Genome Resources



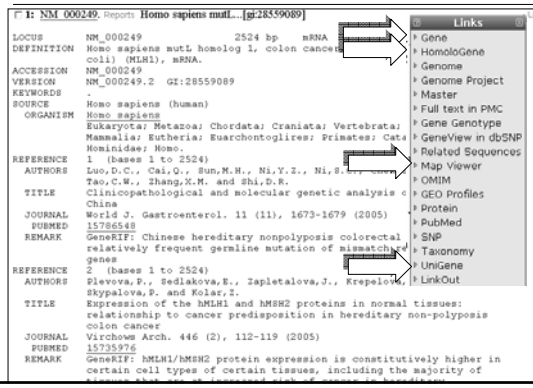
NCBI FieldGuide

Microbial Genomes



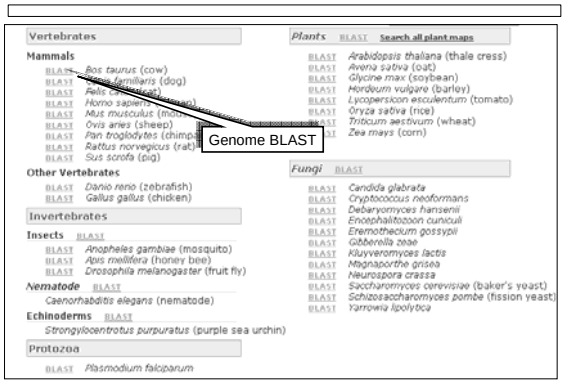
NCBI FieldGuide

NM_000249: Genome Links



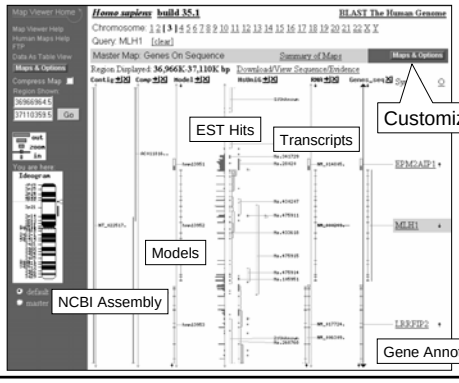
NCBI FieldGuide

The Map Viewer



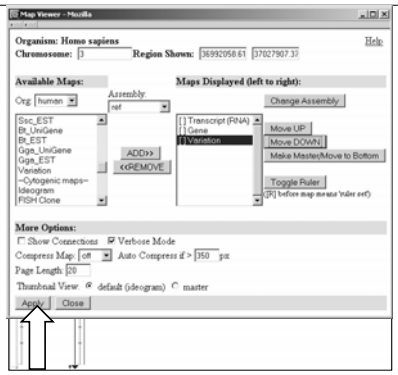
NCBI FieldGuide

Map Viewer: Human MLH1



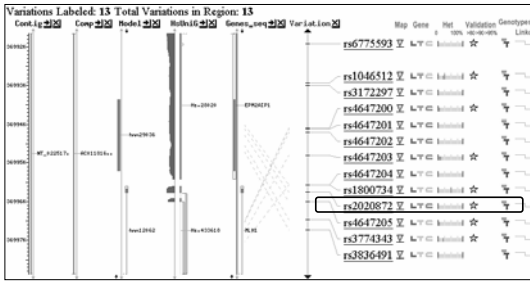
NCBI FieldGuide

Maps and Options



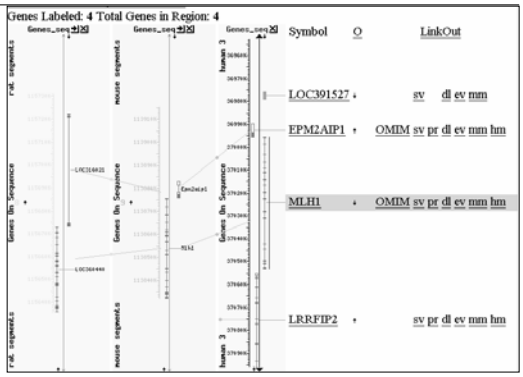
NCBI FieldGuide

Mapped Variations



NCBI FieldGuide

Synteny: Mammalian Genomes

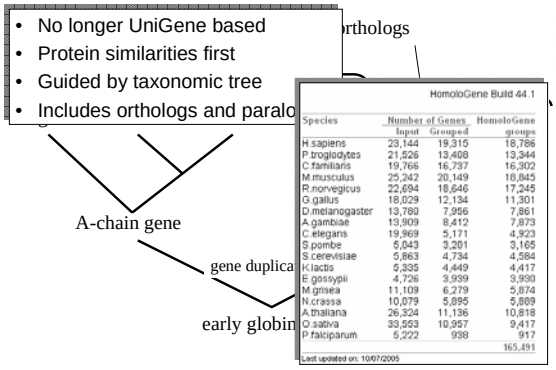


NCBI FieldGuide

The New Homologene

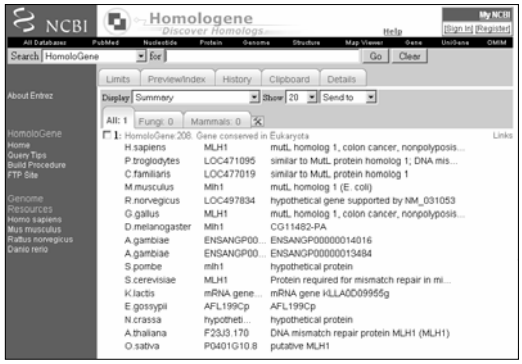
NCBI FieldGuide

- No longer UniGene based
- Protein similarities first
- Guided by taxonomic tree
- Includes orthologs and paralo



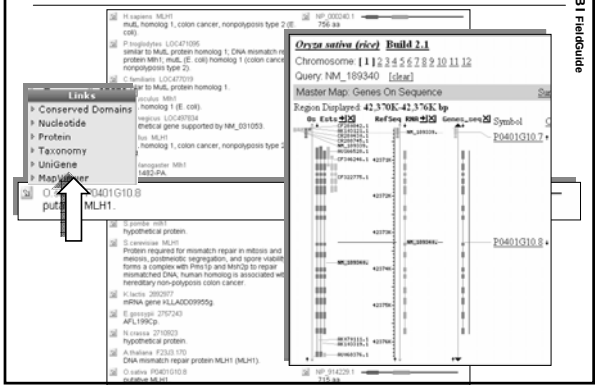
The New Homologene

NCBI FieldGuide



Rice Homolog

NCBI FieldGuide



The Gene Database

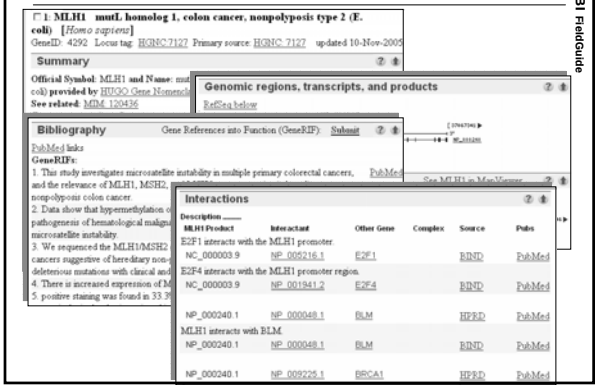
NCBI FieldGuide

- Gene Centered Information
- Unifies LocusLink and microbial Genomes
- 1.5 million records for 3,038 taxa

Human	32,924	Chicken	18, 444
Chimpanzee	24,158	Zebrafish	37, 715
Mouse	47,988	Sea Urchin	20,636
Rat	28,584	Mosquito	13,897
Dog	20,153	Fruit Fly	20,763
Cow	23, 647	*C. elegans*	20,846

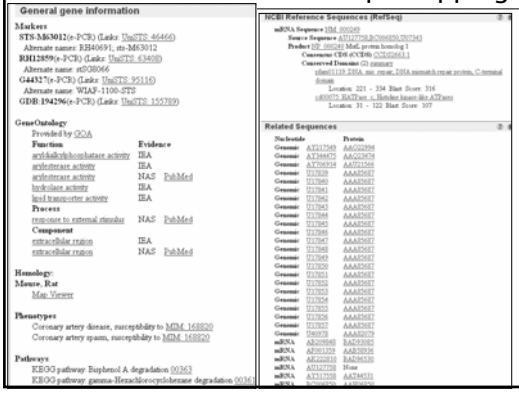
Genes MLH1: One Stop Shopping

NCBI FieldGuide

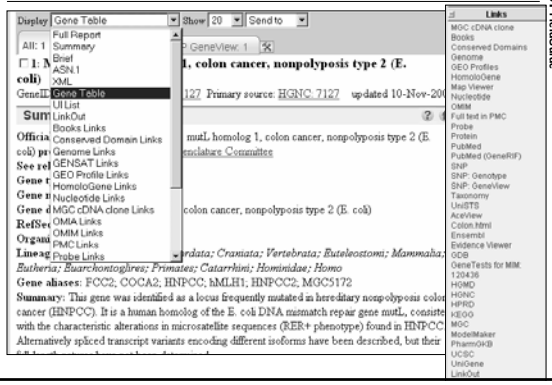


Genes MLH1: One Stop Shopping (cont.)

NCBI FieldGuide



Genes: Display Options and Links



NCBI FieldGuide

Taxonomy Link

The Tax Browser

NCBI's Taxonomy

NCBI FieldGuide

Sequence Similarity Searching

NCBI FieldGuide

What BLAST tells you

- BLAST reports surprising alignments
  - Different than chance
- Assumptions
  - Random sequences
  - Constant composition
- Conclusions
  - Surprising similarities imply **evolutionary homology**

Evolutionary Homology: descent from a common ancestor
Does not always imply similar function

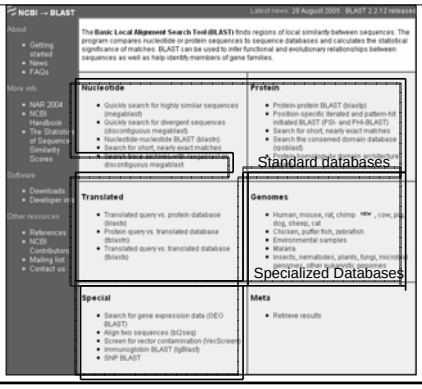
NCBI FieldGuide

Basic Local Alignment Search Tool

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best **local** alignments
- Provides statistical **significance**
- All combinations (DNA/Protein) query and database.
  - DNA vs DNA
  - DNA translation vs Protein
  - Protein vs Protein
  - Protein vs DNA translation
  - DNA translation vs DNA translation
- www, standalone, and network clients

NCBI FieldGuide

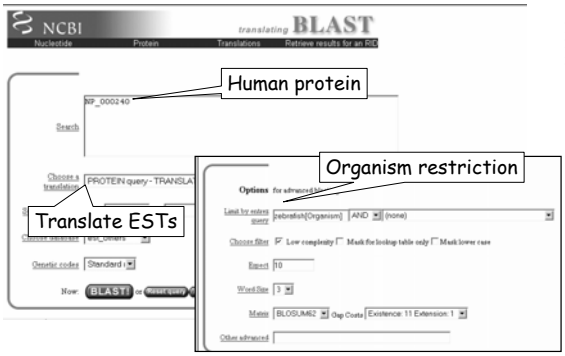
The BLAST homepage



NCBI FieldGuide

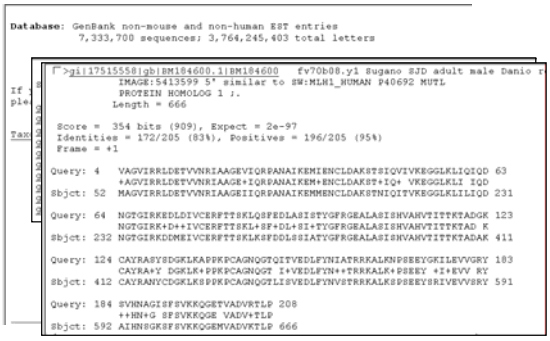
Finding Zebrafish mlh1

NCBI FieldGuide



BLAST Results Zebrafish ESTs

NCBI FieldGuide



Service Addresses

NCBI FieldGuide

•***General Help***	`info@ncbi.nlm.nih.gov`
•***BLAST***	`blast-help@ncbi.nlm.nih.gov`