NCBI FieldGuide

NCBI Molecular Biology Resources

A Field Guide

May 26, 2004

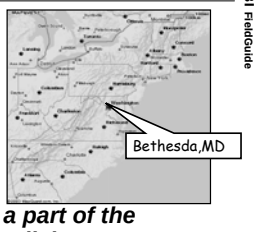
NCBI FieldGuide

NCBI Resources

- About NCBI
- NCBI Sequence Databases
- Other NCBI Databases
- Entrez Databases and Text Searching
- Genomic Resources
- BLAST Services

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

Types of Databases

NCBI FieldGuide

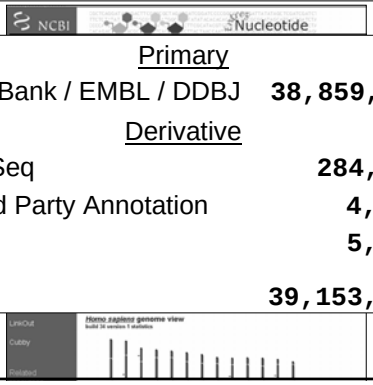
- Primary Databases
  - Original submissions by experimentalists
  - Content controlled by the submitter
    - Examples: GenBank, SNP, GEO
- Derivative Databases
  - Built from primary data
  - Content controlled by third party (NCBI)
    - Examples: Refseq, TPA, RefSNP, UniGene, NCBI Protein, Structure,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

The Entrez System

NCBI FieldGuide

Entrez Nucleotides

NCBI FieldGuide



<u>Primary</u>	
• GenBank / EMBL / DDBJ	**38,859,434**
<u>Derivative</u>	
• RefSeq	**284,238**
• Third Party Annotation	**4,255**
• PDB	**5,000**
Total	**39,153,111**

Entrez Protein

NCBI FieldGuide

• GenPept (GB,EMBL, DDBJ)	3,240,758
• RefSeq	933,905
• Third Party Annotation	4,676
• Swiss Prot	151,524
• PIR	282,819
• PRF	12,079
Total	4,728,703
BLAST nr	1,829,482

select "Taxonomy Links" from the drop-down menu above

What is GenBank?

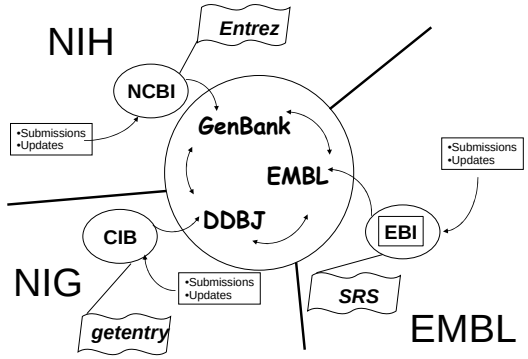
NCBI's Primary Sequence Database

NCBI FieldGuide

- Nucleotide only sequence database
- Archival in nature
- 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

International Sequence Database Collaboration

NCBI FieldGuide



GenBank: NCBI's Primary Sequence Database

NCBI FieldGuide

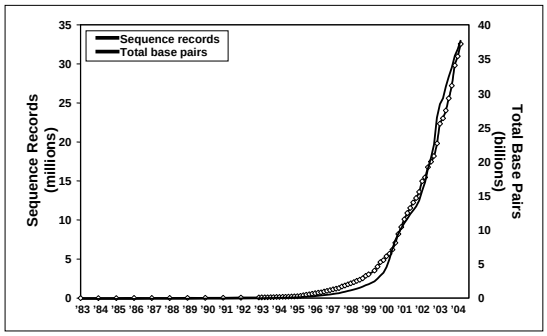
Release 141	April 2004
33,676,218	Records
38,989,342,565	Nucleotides
>140,000	Species
146 Gigabytes	*606 files*

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

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

The Growth of GenBank

NCBI FieldGuide



Organization of GenBank: GenBank Divisions

NCBI FieldGuide

Records are divided into 17 Divisions.

- 11 Traditional
- 6 Bulk

Traditional Divisions:

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

PRI (27) Primate
PLN (11) Plant and Fungal
BCT (8) Bacterial and Archeal
INV (6) Invertebrate
ROD (12) Rodent
VRL (4) Viral
VRT (6) Other Vertebrate
MAM (1) Mammalian (ex. ROD and PRI)
PHG (1) Phage
SYN (1) Synthetic (cloning vectors)
UNA (1) Unannotated

Entrez query: `gbdiv_xxx[Properties]`

Organization of GenBank: GenBank Divisions

NCBI FieldGuide

Records are divided into 17 Divisions.

- 11 Traditional
- 6 Bulk

EST (305)	Expressed Sequence Tag
GSS (104)	Genome Survey Sequence
HTG (62)	High Throughput Genomic
STS (4)	Sequence Tagged Site
HTC (4)	High Throughput cDNA
PAT (15)	Patent

BULK Divisions:

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

Entrez query: `gbdiv_xxx[Properties]`

A Traditional GenBank Record

NCBI FieldGuide

```
LOCUS       AF062069                3808 bp    mRNA    linear   INV 23-OCT-2002
DEFINITION  Limulus polyphemus myosin III mRNA, complete cds.
ACCESSION   AF062069
VERSION     AF062069.2  GI:7144484
KEYWORDS    .
SOURCE      Limulus polyphemus (Atlantic horseshoe crab)
  ORGANISM  Limulus polyphemus
            Eukaryota; Metazoa; Arthropoda; Chelicerata; Merostomata;
            Xiphosura; Limulidae; Limulus.
REFERENCE   1  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     A myosin III from Limulus eyes is a clock-regulated phosphoprotein
  JOURNAL   J. Neurosci. 18 (12), 4548-4559 (1998)
  MEDLINE   98279067
   PUBMED   9614231
REFERENCE   2  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-APR-1998) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
REFERENCE   3  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-MAR-2000) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
  REMARK    Sequence update by submitter
COMMENT     On Mar 2, 2000 this sequence version replaced gi:3132700.
```

GenBank Record: Locus

```
LOCUS       AF062069        3808 bp    mRNA    linear   INV 23-OCT-2002
  VERSION     AF062069.2  GI:7144484
  KEYWORDS
  SOURCE      Limulus polyphemus (Atlantic horseshoe crab)
    ORGANISM  Limulus polyphemus
              Eukaryota; Metazoa; Arthropoda; Chelicerata; Merostomata;
              Xiphosura; Limulidae; Limulus.
  REFERENCE   1  (bases 1 to 3808)
    AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
              Greenberg,R.M. and Smith,W.C.
    TITLE     A myosin III from Limulus eyes is a clock-regulated phosphoprotein
    JOURNAL   J. Neurosci. 18 (12), 4548-4559 (1998)
    MEDLINE   98279067
     PUBMED   9614231
  REFERENCE   2  (bases 1 to 3808)
    AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
              Greenberg,R.M. and Smith,W.C.
    TITLE     Direct Submission
    JOURNAL   Submitted (29-APR-1998) Whitney Laboratory, University of Florida,
              9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
  REFERENCE   3  (bases 1 to 3808)
    AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
              Greenberg,R.M. and Smith,W.C.
    TITLE     Direct Submission
    JOURNAL   Submitted (02-MAR-2000) Whitney Laboratory, University of Florida,
              9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
    REMARK    Sequence update by submitter
  COMMENT     On Mar 2, 2000 this sequence version replaced gi:3132700.
```

Length

Division

Molecule type

Locus name

Modification Date

GenBank Record: Identifiers

```
LOCUS       AF062069                 3808 bp    mRNA    linear   INV 23-OCT-2002
DEFINITION  Limulus polyphemus myosin III mRNA, complete cds.
```

```
ACCESSION   AF062069
VERSION     AF062069.2  GI:7144484
```

```
  TITLE     A myosin III from Limulus eyes is a clock-regulated phosphoprotein
  JOURNAL   J. Neurosci. 18 (12), 4548-4559 (1998)
  MEDLINE   98279067
   PUBMED   9614231
REFERENCE   2  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-APR-1998) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
REFERENCE   3  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-MAR-2000) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
  REMARK    Sequence update by submitter
COMMENT     On Mar 2, 2000 this sequence version replaced gi:3132700.
```

GenBank Record: Organism

```
LOCUS       AF062069                 3808 bp    mRNA    linear   INV 23-OCT-2002
DEFINITION  Limulus polyphemus myosin III mRNA, complete cds.
ACCESSION   AF062069
VERSION     AF062069.2  GI:7144484
KEYWORDS    .
```

```
SOURCE    Limulus polyphemus (Atlantic horseshoe crab)
 ORGANISM Limulus polyphemus
          Eukaryota; Metazoa; Arthropoda; Chelicerata; Merostomata;
          Xiphosura; Limulidae; Limulus.
```

NCBI's Taxonomy

```
   PUBMED   9614231
REFERENCE   2  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Ca
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-APR-1998) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
REFERENCE   3  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-MAR-2000) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
  REMARK    Sequence update by submitter
COMMENT     On Mar 2, 2000 this sequence version replaced gi:3132700.
```

GenBank Record: Feature Table

```
FEATURES             Location/Qualifiers
     source          1..3808
                     /organism="Limulus polyphemus"
                     /db_xref="taxon:6850"
                     /tissue_type="lateral eye"
     CDS             258..3302
                     /note="N-terminal protein kinase domain; C-terminal myosin
                     heavy chain head; substrate for PKA"
                     /codon_start=1
                     /product="myosin III"
                     /protein_id="AAC16332.2"
                     /db_xref="GI:7144485"
                     TGTYATVYSAIDKQA
                     KRGERESDNEVWLGI
                     NSIIHRDIRAANIMF
                     SKEGYVKLIDFGLSASVKNTNGKAQSSVGSPYWMAP  SCDCLQEPYNYTCDVWSIG
                     ITAIELADTVPSLSDIHALRAMFRINRNPPPSVKRET  TLKDFISECLVKNPEYR
                     PCIQEIPQHPFLAQVEGKEDQLRSELVDILKKNPG  TISGQ
BASE COUNT     1201 a    689 c    782 g   1136 t
ORIGIN
        1 tcgacatctg tggtcgcttt ttttagtaat aaaaaattgt attatgacgt cctatctgtt
     3781 aagatacagt aactagggaa aaaaaaaa
//
```

GenPept IDs

GenPept: FASTA format

NCBI FieldGuide

```
>gi|7144485|gb|AAC16332.2| myosin III [Limulus polyphemus]
MEYKCISEHLPFETLPDPGDRFEVQELVGTGTYATVYSAIDKQANKKVALKIIGHIAENLLDIETEYRIY
KAVNGIQFFPEFRGAFFKRGERESDNEVWLGIEFLEEGTAADLLATHRRFGIHLKEDLIALIIKEVVRAV
QYLHENSIIHRDIRAANIMFSKEGYVKLIDFGLSASVKNTNGKAQSSVGSPYWMAPEVISCDCLQEPYNY
TCDVWSIGITAIELADTVPSLSDIHALRAMFRINRNPPPSVKRETRWSETLKDFISECLVKNPEYRPCIQ
EIPQHPFLAQVEGKEDQLRSELVDILKKNPGEKLRNKPYNVTFKNGHLKTISGQPHEKIYVDDLAFLDSP
TEEVVLENLEQRYRKGEIYTFAGDVLLTLNPGKVLPLYGDQTAVKYCERGRSDNPPHVFAVADRAYQQML
HHKSPQAVILSGVSGSGKSFCTHQVIRHLAFLGAQNKEGMREKLEYLCPLLDTLGNAYTSTNPNSSHFVK
ILEVTFTKTGKITGAILFTFLLEARRLTDIPKGERNFHVFYYFYEGLRSEGRLKEFGLEEKNYRYLPELK
SSNSPEYVKGYQQFLRALTSLAFTEEEIFAIQKVLAAILLLGETEIQNSAAFKLLGAESSELENTLTQDV
NARDVYARAMYLRLFSWIVAVVNRQLSFSRLVFGDVYSVTVIDSPGFENGLHNSLHQLCANVISDNLQNY
IQQIIFFKELEEYGEEGVNVPFNLEGGVDHRTLVNKLMDSGQGLLTAISKATQYQRKGESGWMESLQEAD
SEELVEFSNVNGKPIVSVKHIFRKVSYDATDLVKKNVEDKTRALTSTMQRSCDPRIRAIFSSENPSPFLS
SPRRSSIQENMLLPERTVTDSLHSALSSVLNLASTEDPPHLILCMRPQKKELINDYDSKSVQIQLHALNV
LETILIRQFGFARRISFVDFLNRYQYLAFDFNENVELTKENCRLLLLRLKMDGWTLGKNKVFLKYYSEEY
LSRIYETHIKKIVKVQAIARKYFVKVRQSKTKPH
>gi|7144486|gb|AAA23731.2| metC peptide [Escherichia coli
MADKKLDTQLVNAGRSKKYSLGAVNSVIQRASSLVFDSVEAKKHATRNRANGELFYGRRGTLTHFSLQQA
MCELEGGAGCVLFPCGAAAVANSILAFIEQGDPRVPSSNS
```

Abstract Syntax Notation: ASN.1

NCBI FieldGuide

```
Seq-entry ::= set {
  class nuc-prot ,
  descr {
    title                    yphemus myosin III mRN                    ds." ,
    sourc
      org {
        taxname "Limulus polyphemus" ,
        common "Atlantic horseshoe crab" ,
        db {
          {
            db "taxon" ,
            tag
              id 6850 } } ,
        orgname {
          name
                                  imulus" ,
                                  "polyphemus" } ,
                                  karyota; Metazoa; Arthro                    ;
Merostomata; Xiphosura; Limulidae; Limulus" ,
          gcode 1 ,
          mgcode 5 ,
          div "INV" } } ,
```

GenPept

GenBank

ASN.1

FASTA Protein

FASTA Nucleotide

NCBI Toolbox

NCBI FieldGuide

```
/*****************************************************************************
*
*   asn2ff.c
*         convert an ASN.1 entry to flat file format, using the FFPrintArray.
*
*****************
#include <accent
#include "asn2ff
#include "asn2ff
#include "ffprin
#include <subuti
#include <objall
#include <objcod
#include <lsqfet
#include <explor

#ifdef ENABLE_ID
#include <accid1
#endif

FILE *fpl;

Args myar
```

Toolbox Sources

```
ftp> open ftp.ncbi.nih.gov
.
.
ftp> cd toolbox
ftp> cd ncbi_tools
```

ftp://ftp.ncbi.nlm.gov/toolbox/ncbi_tools

```
{"Input asnfile in binary mode","F",NULL,NULL,TRUE,'b',ARG_BOOLEAN,0.0,0,NULL},
{"Output Filename","stdout", NULL,NULL,TRUE,'o',ARG_FILE_OUT,0.0,0,NULL},
{"Show Sequence?","T", NULL ,NULL ,TRUE,'h',ARG_BOOLEAN,0.0,0,NULL},
```

Bulk Divisions

NCBI FieldGuide

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

- Expressed Sequence Tag
  - 1st pass single read cDNA
- Genome Survey Sequence
  - 1st pass single read gDNA
- High Throughput Genomic
  - incomplete sequences of genomic clones
- Sequence Tagged Site
  - PCR-based mapping reagents

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**

ESTs in Entrez

NCBI

What is UniGene?

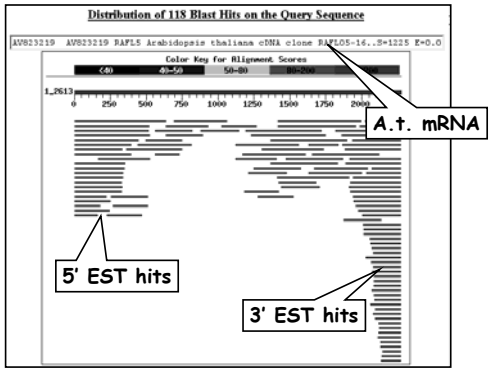
A gene-oriented view of sequence entries

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 A.t. serine protease mRNA

NCBI FieldGuide



UniGene

March 2004

NCBI FieldGuide

Species	UniGene Entries
Chordata	
Mammalia	
Bos taurus	23,572
Canis familiaris	3,905
Homo sapiens	105,651
Mus musculus	79,846
Rattus norvegicus	39,874
Sus scrofa	20,426
Aves	
Gallus gallus	12,220
Amphibia	
Xenopus laevis	22,129
Xenopus tropicalis	17,102
Actinopterygii	
Danio rerio	17,951
Oncorhynchus mykiss	13,458
Oryzias latipes	8,127
Salmo salar	1,033
Ascidiacea	
Ciona intestinalis	13,600
Echinodermata	
Echinoidea	
Strongylocentrotus purpuratus	2,592
Arthropoda	
Insecta	
Anopheles gambiae	14,085
Apis mellifera	4,756
Bombyx mori	2,008
Drosophila melanogaster	12,322
Nematoda	
Chromadorea	
Caenorhabditis elegans	15,655
Platyhelminthes	
Trematoda	
Schistosoma mansoni	1,185
Embryophyta	
Bryopsida	
Physcomitrella patens	6,946
Coniferopsida	
Pinus taeda	8,669
Eudicotyledons	
Arabidopsis thaliana	19,666
Glycine max	12,567
Helianthus annuus	1,897
Lactuca sativa	9,629
Lycopersicon esculentum	3,658
Medicago truncatula	5,398
Populus tremula x Populus tremuloides	2,711
Solanum tuberosum	5,476
Vitis vinifera	11,832
Liliopsida	
Hordeum vulgare	11,980
Oryza sativa	25,089
Saccharum officinarum	4,771
Sorghum bicolor	7,075
Triticum aestivum	24,543
Zea mays	12,941
Chlorophyta	
Chlorophyceae	
Chlamydomonas reinhardtii	5,874
Mycetozoa	
Dictyostelida	
Dictyostelium discoideum	3,853
Apicomplexa	
Coccidia	
Toxoplasma gondii	6,001

Human UniGene

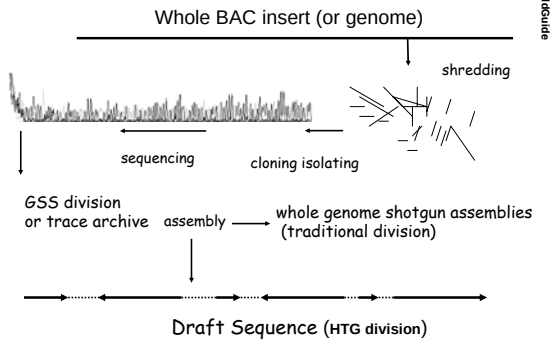
UniGene Build 170
Apr. 24, 2004

```
    136,416 mRNAs
      5,187 models
      7,415 HTC
  1,416,936 EST, 3'reads
  2,092,475 EST, 5'reads
+   775,747 EST, other/unknown
----------
  4,434,176 total sequences in clusters

Final Number of Clusters (sets)
==============================
            total
106,219
 28,126   c                ne mRNA
  5,750   c                ne HTC
104,890   c                ne EST
 26,839   contain both mRNAs and ESTs
```

3,000,000,000 bp
30 K expected genes
75% excess

Genome Sequencing – HTG, GSS,(WGS)



HTG Division: Honeybee Draft Sequences

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

NCBI FieldGuide

Other Genome Sequencing Products

Trace Archive

Whole Genome Shotgun

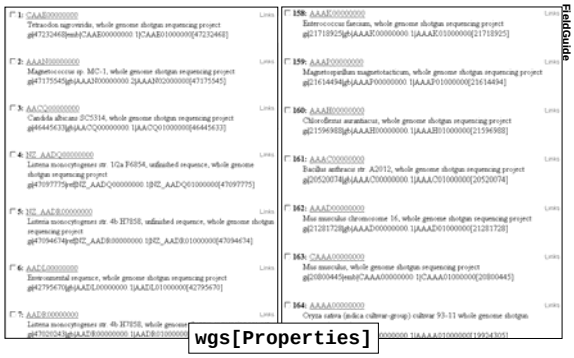
NCBI FieldGuide

Whole Genome Shotgun Projects

- Traditional GenBank Divisions
- 164 + projects
  - Virus
  - Bacteria
  - Environmental sequences
  - Archaea
  - 44 Eukaryotes featuring:
    - Chicken, Rat, Mouse, Dog, Chimpanzee, Human
    - Pufferfish (2)
    - Honeybee, Anopheles, Fruit Flies (2), Silkworm
    - Nematode (*C. briggsae*)
    - Yeasts (8), Aspergillus (2)
    - Rice

NCBI FieldGuide

Whole Genome Shotgun Projects



NCBI FieldGuide

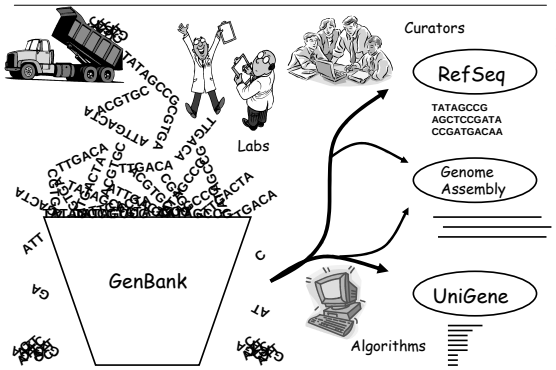
Derivative Sequence Databases

RefSeq

TPA

NCBI FieldGuide

NCBI Derivative Sequence Data



NCBI FieldGuide

RefSeq: NCBI's Derivative Sequence Database

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

`srcdb_refseq[Properties]`

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

RefSeq Benefits

NCBI FieldGuide

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current sequence data and biology
- data validation
- format consistency
- distinct accession series
- stewardship by NCBI staff and collaborators

RefSeq Accession Numbers

NCBI FieldGuide

mRNAs and Proteins	
NM_123456	Curated mRNA
NP_123456	Curated Protein
NR_123456	Curated non-coding RNA
XM_123456	Predicted mRNA
XP_123456	Predicted Protein
XR_123456	Predicted non-coding RNA
Gene Records	
NG_123456	Reference Genomic Sequence
Chromosome	
NC_123455	Microbial replicons, organelle
Assemblies	
NT_123456	Contig
NW_123456	WGS Supercontig

Third Party Annotation (TPA) Database

NCBI FieldGuide

- Annotations of existing GenBank sequences
- Allows for community annotation of genomes
- Direct submissions
  - BankIt
  - Sequin

`tpa[Properties]`

TPA record: WGS Assembly

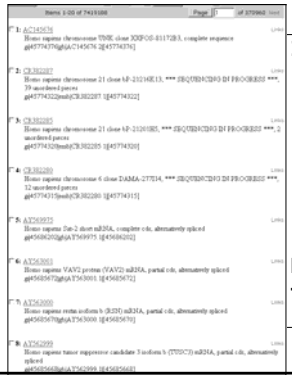
NCBI FieldGuide

```
LOCUS       BK001609                3412 bp    DNA     linear   INV 03-DEC-2003
DEFINITION  TPA: Apis mellifera troponin C type I (TpnCI) gene, complete cds.
ACCESSION   BK001609
VERSION     BK001609.1  GI:38639836
KEYWORDS    Third Party Annotation; TPA.
SOURCE      Apis mellifera (honey bee)
  ORGANISM  Apis mellifera
            Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
FEATURES             Location/Qualifiers
     source          1..3412
                     /organism="Apis mellifera"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:7460"
     gene            <1..>3361
                     /gene="TpnCI"
     mRNA            join(<1..13,928..1071,1231..1350,2063..2194,
                     3333..>3361)
                     /gene="TpnCI"
                     /product="troponin C type I"
     CDS             join(1..13,928..1071,1231..1350,2063..2194,3333..3361)
                     /gene="TpnCI"
                     /note="AmTpnCI; expressed mai          tage and
                     abdomen muscles; putative fun          s: second and
                     fourth; similar to DmTpnC73F a          CG5381"
                     /codon_start=1
                     /product="troponin C type I"
                     /protein_id="DAA01875.1"
                     /db_xref="GI:38639837"
                     /translation="MFHQFLRKAFDSFDREKSGSIPTDMVADILRLMGQPFNKKILDE
                     LIEEVDADKSGRLEFEEFVTLAAKFIVEEDAEALEKELREAFRLYDKEGNGYIPTTCL
                     REILRELDDQLTDEELDIMIEEIDSDGSGTVDFDEFMEMMTGE"
```

CDS Feature
TPA protein

Human Nucleotide Sequences

NCBI FieldGuide



ISDC	**8,480,614**
(GenBank/EMBL/DDBJ)	
PRI	**901,843**
(WGS	**601,710)**
EST	**5,639,828**
GSS	**905,594**
HTG	**18,367**
STS	**76,333**
PAT	**930,108**
RefSeq	**29,935**
TPA	**864**

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)

PDB → Structure

NCBI Structures and Domains

Pfam → CDD

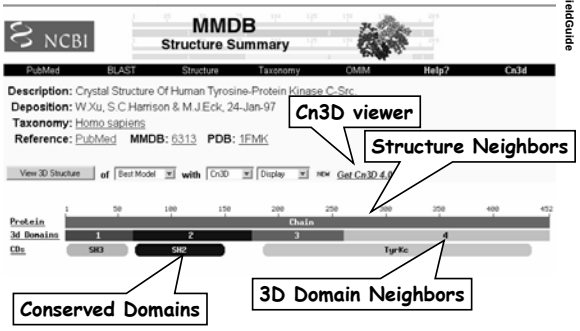
SMART

Clusters of Orthologous Groups

MMDB: Molecular Modeling Data Base

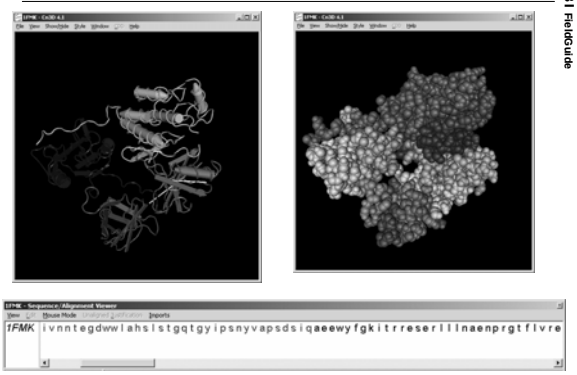
- Derived from experimentally determined PDB records
- Value added to PDB records including:
  - Addition of explicit chemical graph information
  - Validation
  - Inclusion of Taxonomy, Citation,
  - Conversion to ASN.1 data description language
- Structure neighbors determined by
  Vector Alignment Search Tool (VAST)

Structure Summary



Cn3D 4.1: C-Src

NCBI FieldGuide



VAST: Structure Neighbors

NCBI FieldGuide

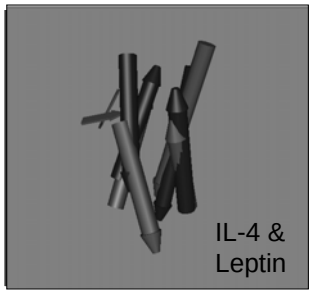
Vector Alignment Search Tool

For each protein chain,

locate SSEs (secondary structure elements),

and represent them as individual vectors.

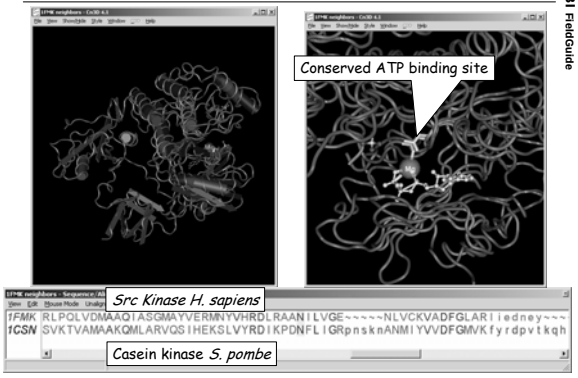
align the vectors



Human IL-4

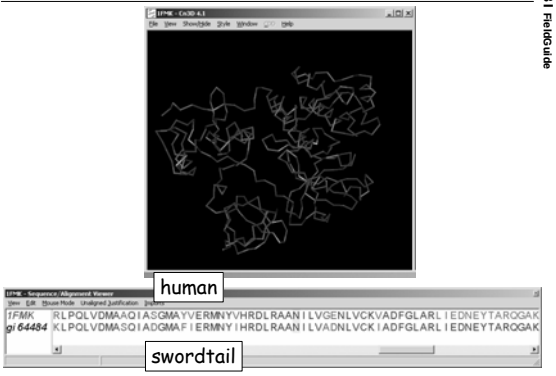
Cn3D 4.1: Structural Alignment

NCBI FieldGuide



Cn3D: Simple Homology Modeling

NCBI FieldGuide



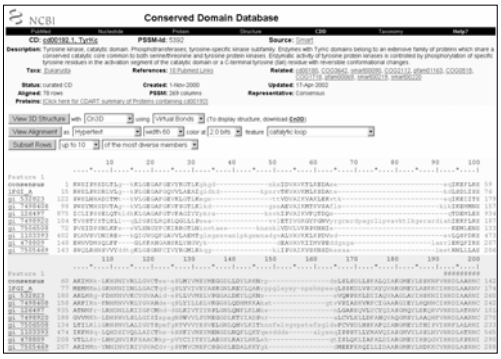
NCBI's Conserved Domain Database

NCBI FieldGuide

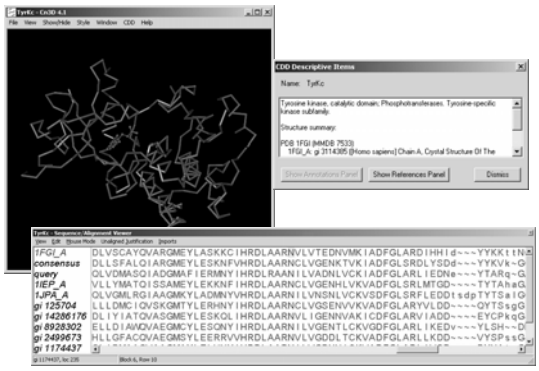
- Multiple sequence alignments
- PSI-BLAST –based score matrices
- Sources SMART, PFAM, COGs, New NCBI curated domains
  - structure informed alignments
- Stats:
  - COGS 4,873
  - Pfam 5,193
  - Smart 653
  - NCBI CDD 316

NCBI CD: Tyrosine Kinase

NCBI FieldGuide



Using Cn3D to model domains



NCBI FieldGuide

WWW Access

Entrez & BLAST

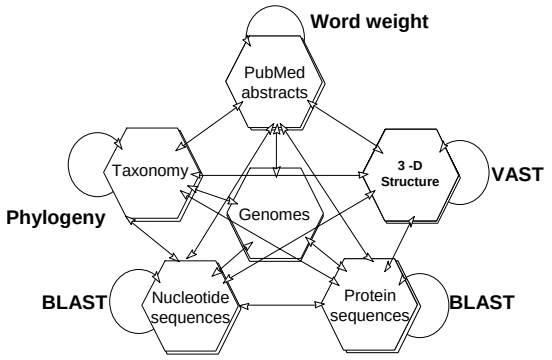
NCBI FieldGuide

Using Entrez

An integrated database search and retrieval system

NCBI FieldGuide

Entrez: Database Integration



NCBI FieldGuide

Database Searching with Entrez

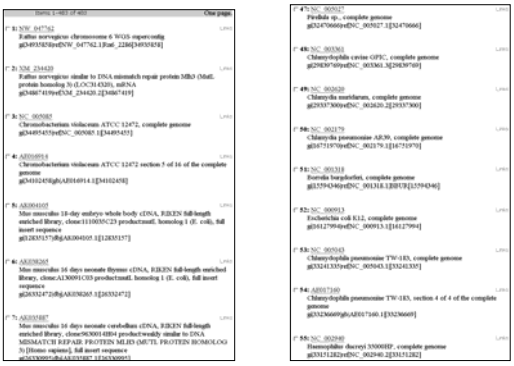
- Using limits and field restriction to find human MutL homolog
- Linking and neighboring with MutL
- Mapping SNPs onto structure and the genome

NCBI FieldGuide

Global Entrez Search

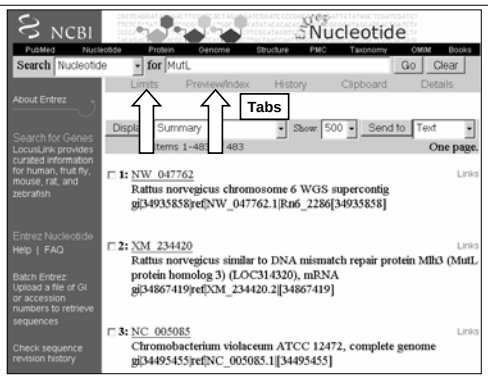
Document Summaries:

MutL[All Fields]



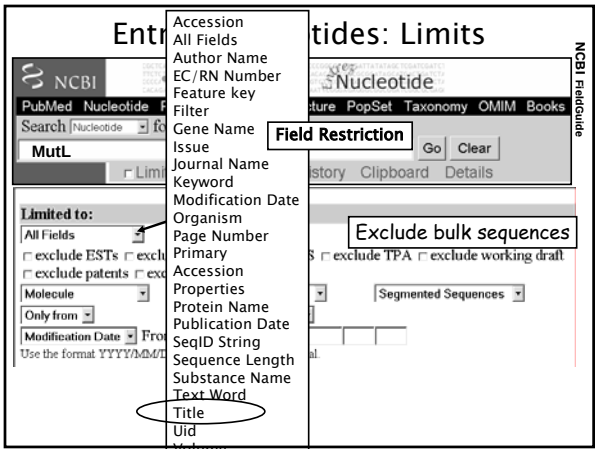
NCBI FieldGuide

Entrez Nucleotides: Limits & Preview/Index



NCBI FieldGuide

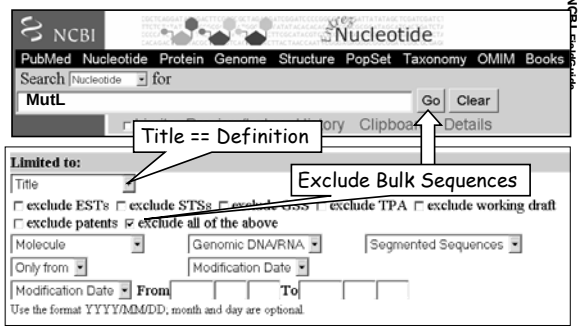
Entr tides: Limits



NCBI FieldGuide

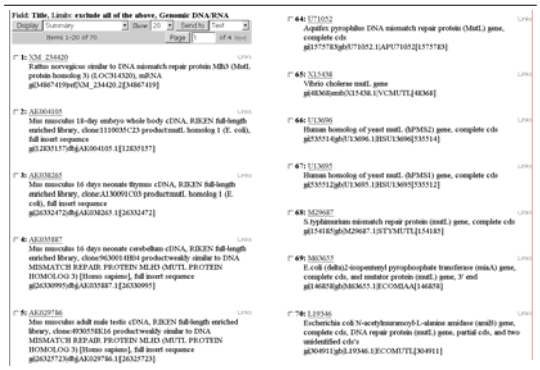
Entrez Nucleotides: Limits

NCBI FieldGuide



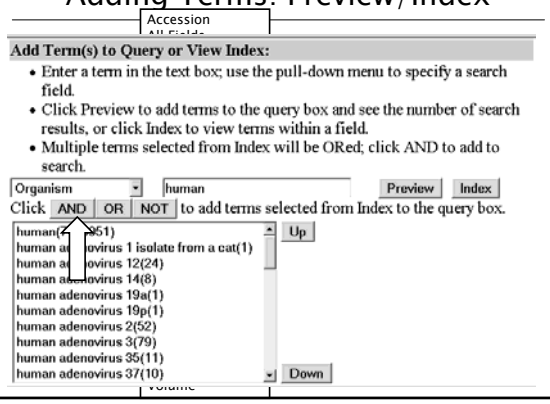
Document Summaries: Limits

NCBI FieldGuide



Adding Terms: Preview/Index

NCBI FieldGuide



Human MutL Search Results

NCBI FieldGuide

Display Summary Show 20 Send to Text

Items 1-7 of 7 One page.

1: NM_000249 Links
Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), mRNA
gi|28559089|ref|NM_000249.2|[28559089]

2: NM_014381 Links
Homo sapiens mutL homolog 3 (E. coli) (MLH3), mRNA
gi|76573306|ref|NM_014381.1|[76573306]

3: BC006850 Links
Homo sapiens, mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2), clone MGC:5172 IMAGE:3451538, mRNA, complete cds
gi|13905125|gb|BC006850.1|BC006850[13905125]

4: AY217549 Links
Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1) gene, complete cds
gi|27805154|gb|AY217549.1|[27805154]

5: S77856 Links
hMLH1=DNA mismatch repair protein MutL homolog (alternatively spliced) [human, peripheral blood lymphocytes, mRNA Partial, 60 nt]
gi|994493|bbm|370506|bbs|166341|gb|S77856.1|S77856[994493]

6: U13696 Links
Human homolog of yeast mutL (hPMS2) gene, complete cds
gi|535514|gb|U13696.1|HSU13696[535514]

7: U13695 Links
Human homolog of yeast mutL (hPMS1) gene, complete cds
gi|535512|gb|U13695.1|HSU13695[535512]

Human MutL RefSeq

NCBI FieldGuide

1: NM_000249. Homo sapiens mutL... [gi:28559089] Links

```
LOCUS       NM_000249                2524 bp    mRNA    linear   PRI 06-SEP-2003
DEFINITION  Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E.
            coli) (MLH1), mRNA.
ACCESSION   NM_000249
VERSION     NM_000249.2  GI:28559089
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from U07343.1, AU127758.1 and
            BC006850.1.
            On Feb 25, 2003 this sequence version replaces gi:4557756.

            Summary: This gene was identified as a locus frequently mutated in
            hereditary nonpolyposis colon cancer (HNPCC). It is a human homolog
            of DNA mismatch repair gene mutL, consistent with the
            alterations in microsatellite sequences (RER+
            in HNPCC. Alternatively spliced transcript
            variants encoding different isoforms have been described, but their
            full-length natures have not been determined.
            COMPLETENESS: complete on the 3' end.
  TITLE     Differential expression of hMLH1 and hMSH2 is related to bladder
            cancer grade, stage and prognosis but not microsatellite
            instability
  JOURNAL   Int. J. Cancer 105 (4), 484-490 (2003)
  MEDLINE   22597635
   PUBMED   12712438
  REMARK    GeneRIF: Our findings indicate that reduced expression of the MMR
            proteins may have an important contribution in the development of a
            subset of TCCs and suggest a potential role for MMR expression as
            prognostic indicators.
```

GenBank Records

NM_000249: Links

NCBI FieldGuide

1: NM_000249. Homo sapiens mutL... [gi:28559089]

```
LOCUS       NM_000249                2524 bp    mRNA    linear
DEFINITION  Homo sapiens mutL homolog 1, colon cancer, nonpoly
            coli) (MLH1), mRNA.
ACCESSION   NM_000249
VERSION     NM_000249.2  GI:28559089
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Mammalia; Eutheria; Primates; Catarrhini; Hominida
REFERENCE   1  (bases 1 to 2524)
  AUTHORS   Strazzullo,M., Cossu,A., Baldinu,P., Colombino,M.,
            Tanda,F., De Bonis,M.L., Cerase,A., D'Urso,M., D'
            Palmieri,G.
  TITLE     High-resolution methylation analysis of the hMLH1 promoter in
            sporadic endometrial and colorectal carcinomas
  JOURNAL   Cancer 98 (7), 1540-1546 (2003)
   PUBMED   14508843
  REMARK    GeneRIF: hMLH1 is one of the most commonly mutated MMR genes in
            endometrial carcinomas and colorectal carcinomas.
REFERENCE   2  (bases 1 to 2524)
  AUTHORS   Scarisbrick,J.J., Mitchell,T.J., Calonje,E., Orchard,G.,
            Russell-Jones,R. and Whittaker,S.J.
  TITLE     Microsatellite instability is associated with hypermethylation of
            the hMLH1 gene and reduced gene expression in mycosis fungoides
  JOURNAL   J. Invest. Dermatol. 121 (4), 894-901 (2003)
   PUBMED   14632210
  REMARK    GeneRIF: Microsatellite instability may be a consequence of hMLH1
            promoter hypermethylation in mycosis fungoides patients and may
            prevent transcription in a subset of patients.
```

Links
- Gene
- HomoloGene
- Master
- Full text in PMC
- Related Sequences
- Map Viewer
- OMIM
- GEO
- Protein
- PubMed
- SNP
- Taxonomy
- UniGene
- UniSTS
- LinkOut

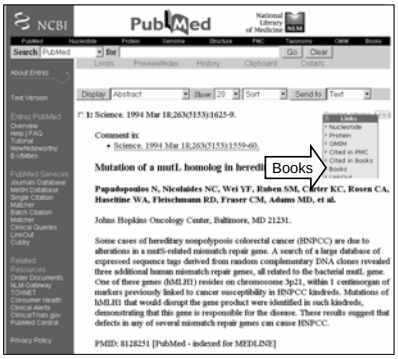
NCBI FieldGuide

Literature Links

PubMed
OMIM

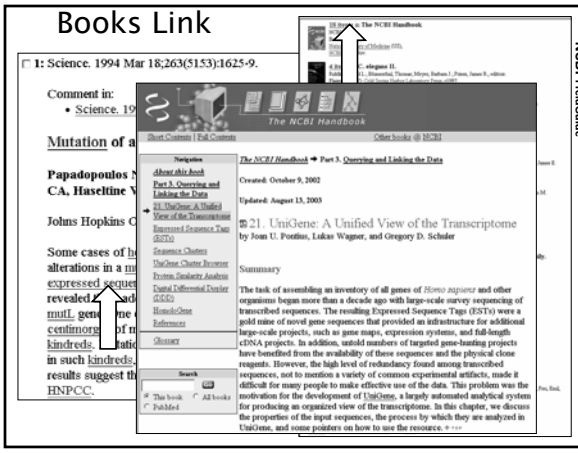
NCBI FieldGuide

NM_000249: PubMed



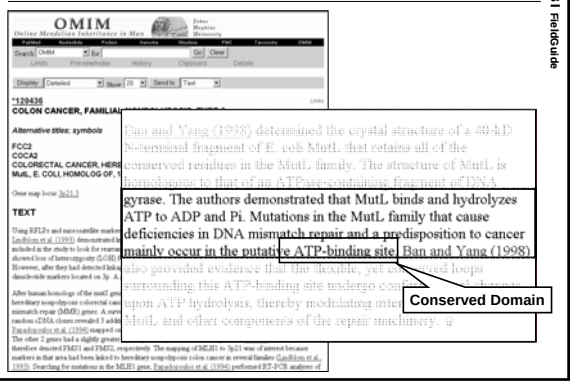
NCBI FieldGuide

Books Link



OMIM: Human Disease Genes

NCBI FieldGuide



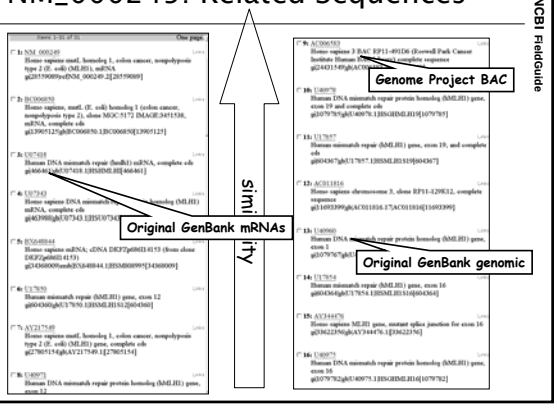
NCBI FieldGuide

Sequence Links

Nucleotide Protein

NCBI FieldGuide

NM_000249: Related Sequences



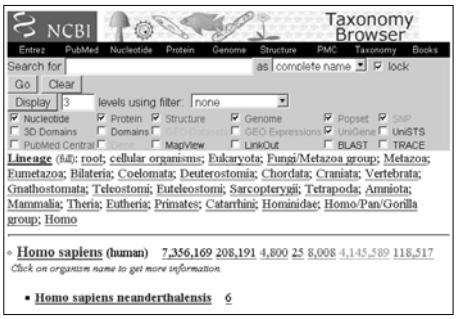
NCBI FieldGuide

Taxonomy Link

The Tax Browser
NCBI's Taxonomy

NCBI FieldGuide

Taxonomy Link



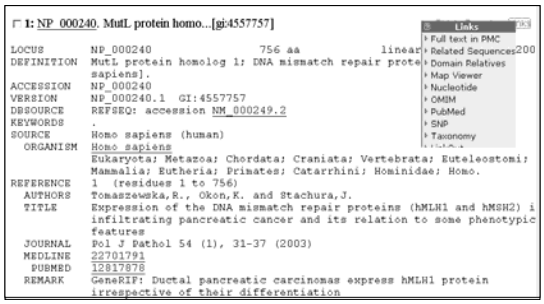
NCBI FieldGuide

NCBI Protein Databases

- GenPept GenBank, EMBL, DDBJ CDS translations
- RefSeq mRNA based (NP_) and genome based (XP_)
- Swiss-Prot curated high quality protein reviews
- PIR protein information resource Georgetown University
- PRF protein resource foundation
- PDB Protein Databank sequences from structures

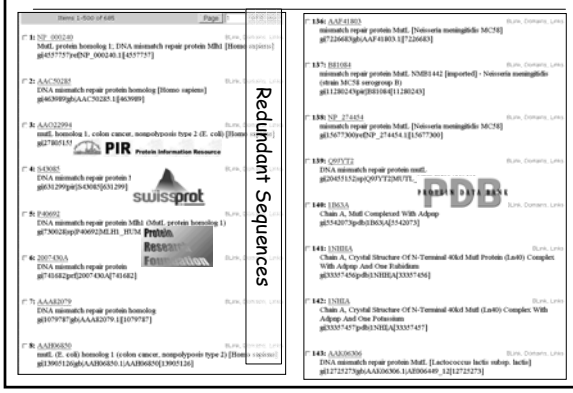
Protein Link

NCBI FieldGuide



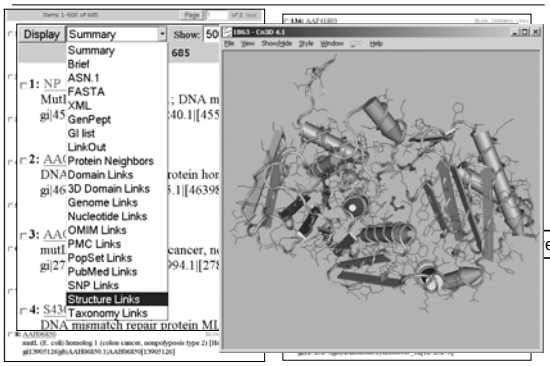
Related Proteins: Redundancy

NCBI FieldGuide



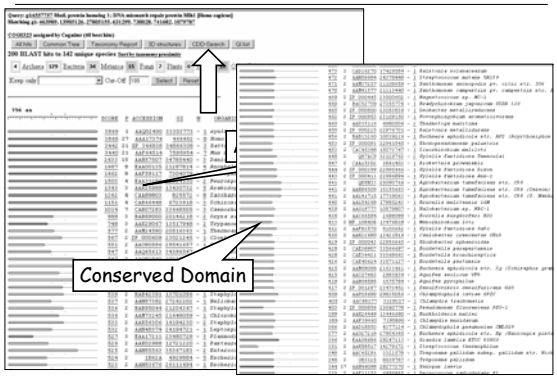
Related Proteins: Links

NCBI FieldGuide



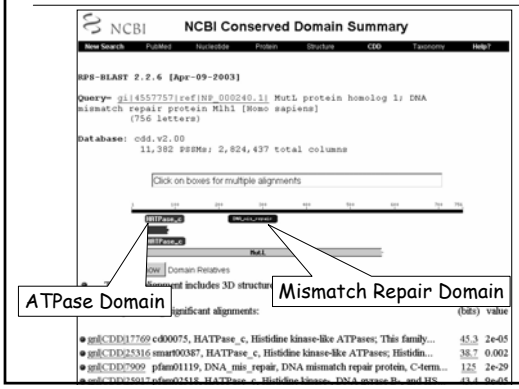
BLink: non-redundant relatives

NCBI FieldGuide



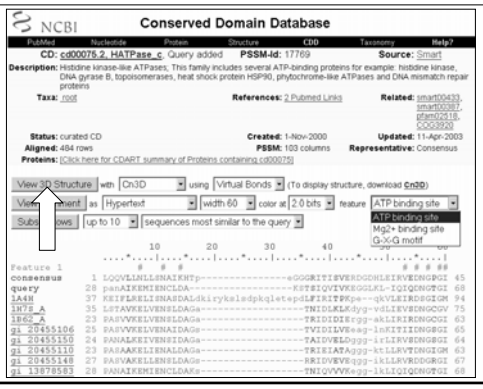
MLH1 Domain Structure: CDD

NCBI FieldGuide



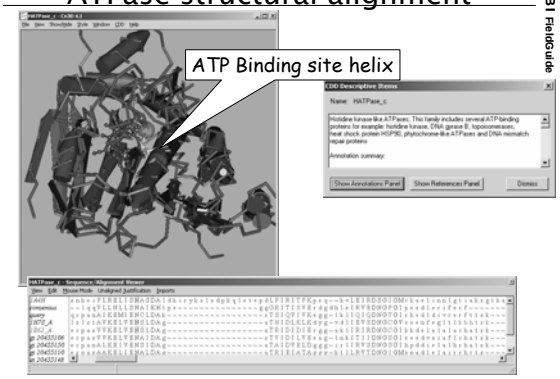
MLH1: ATPase Domain

NCBI FieldGuide



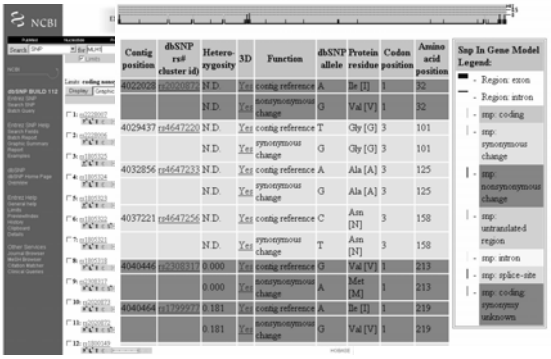
ATPase structural alignment

NCBI FieldGuide



Variations Human MLH1

NCBI FieldGuide

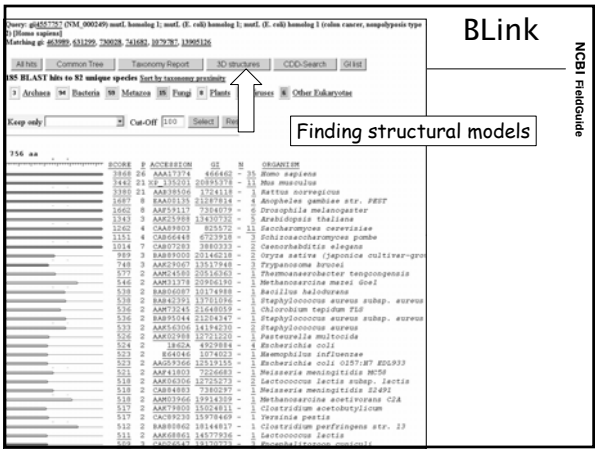


Contig position	dbSNP rs# (cluster id)	Heterozygosity	3D	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
4029028	rs2308819	N.D.	Yes	contig reference	A	Ile [I]	1	32
		N.D.	Yes	nonsynonymous change	G	Val [V]	1	32
4029437	rs4647220	N.D.	Yes	contig reference	T	Gly [G]	3	101
		N.D.	Yes	synonymous change	G	Gly [G]	3	101
4032856	rs4647233	N.D.	Yes	contig reference	A	Ala [A]	3	125
		N.D.	Yes	synonymous change	G	Ala [A]	3	125
4037221	rs4647256	N.D.	Yes	contig reference	C	Asn [N]	3	158
		N.D.	Yes	synonymous change	T	Asn [N]	3	158
4040446	rs2308817	0.000	Yes	contig reference	G	Val [V]	1	213
		0.000	Yes	nonsynonymous change	A	Met [M]	1	213
4040464	rs1799977	0.181	Yes	contig reference	A	Ile [I]	1	219
		0.181	Yes	nonsynonymous change	G	Val [V]	1	219

Snp In Gene Model Legend:
- Region: exon
- Region: intron
- snp: coding
- snp: synonymous change
- snp: nonsynonymous change
- snp: untranslated region
- snp: intron
- snp: splice-site
- snp: coding: synonymy unknown

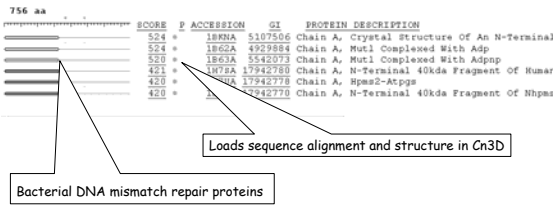
BLink

NCBI FieldGuide



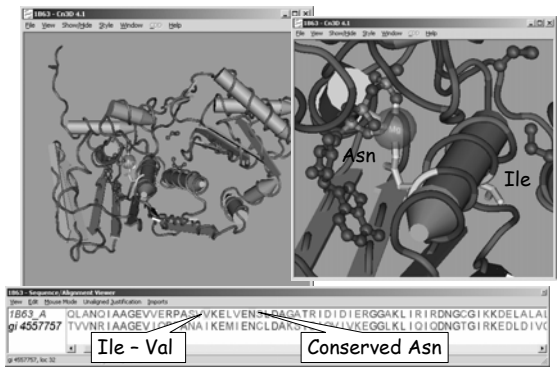
Mapping Variation Onto Structure

NCBI FieldGuide



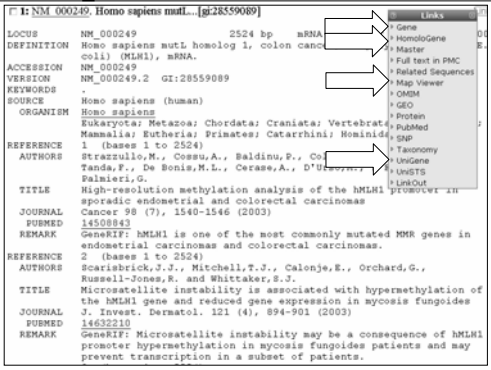
Mapping Variation Onto Structure

NCBI FieldGuide



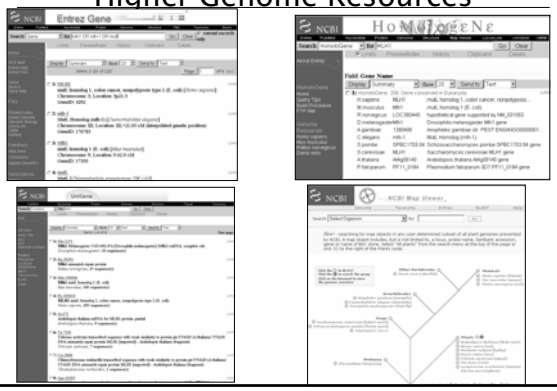
NM_000249: Genome Links

NCBI FieldGuide



Higher Genome Resources

NCBI FieldGuide



MLH1: UniGene Cluster

NCBI FieldGuide

MLH: MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)

LINKS: Gene | LocusLink | HomoloGene

SELECT MODEL ORGANISM protein and percent identity and

H.sapiens ref|NP_000249 mutL

M.musculus sp|Q97K91 - MO

R.norvegicus ref|NP_112315.1 - mismatch

A.thaliana: pir|T51620 - T51620 DNA Arabidopsis thaliana

C.elegans: ref|NP_499796.1 - DNA

D.melanogaster: ref|NP_477022.1 - MMR

E. coli: sp|P23367 - MUTL_ECO

S. cerevisiae: pir|S54525 - S54525 mismatch

MAPPING INFORMATION

Genome View:	3
Genetic map:	3p21.3 cM
UniSTS entry:	Chr 3
UniSTS entry:	

EXPRESSION INFORMATION

cDNA sources: prostate ; blood ; testis ; p

description) ; Avenlar Mar

Pooled-Glandular ; Human Placenta ; Pituitary ; NEUROBLASTOMA COT

25-NORMALIZED ; PLACENTA COT 25-NORMALIZED ; Pooled human melanocyte

; Liver and Spleen ; carcinoma, cell line ; parathyroid tumor ; kidney tumor

SAGE : Gene to Tag mapping

SEQUENCE INFORMATION

mRNA SEQUENCES (5)

BX648844.1	Homo sapiens mRNA; cDNA DKFZp686I14153 (from clone DKFZp686I14153)	PA
NM_000249.2	Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), mRNA	P
BC006850.1	Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli), mRNA (cDNA clone MGC:5172 IMAGE:3451538), complete cds	PA
U07418.1	Human DNA mismatch repair (hmlh1) mRNA, complete cds.	P
U07343.1	Homo sapiens DNA mismatch repair protein homolog (MLH1) mRNA, complete cds	P

EST SEQUENCES (10 of 296)[Show all ESTs]

BI835081.1	cDNA clone IMAGE:5226825	pooled pancreas and spleen	5' read	PM
BF795990.1	cDNA clone IMAGE:4343565	lymphoma, cell line	5' read	PM
BG335580.1	cDNA clone IMAGE:4541662	choriocarcinoma	5' read	PM
BG772547.1	cDNA clone IMAGE:4837570	testis	5' read	PM
BG772733.1	cDNA clone IMAGE:4837559	testis	5' read	PM
BI764939.1	cDNA clone IMAGE:5187469	pooled colon, kidney, stomach	5' read	PM
BE884841.1	cDNA clone IMAGE:3911872	leiomyosarcoma	5' read	PM
BI084102.1	cDNA clone IMAGE:5013904	epidermoid carcinoma, cell line	5' read	PM
BI916940.1	cDNA clone IMAGE:5242115	brain	5' read	PM
BF306562.1	cDNA clone IMAGE:4122836	rhabdomyosarcoma	5' read	PM

ESTs in UniGene

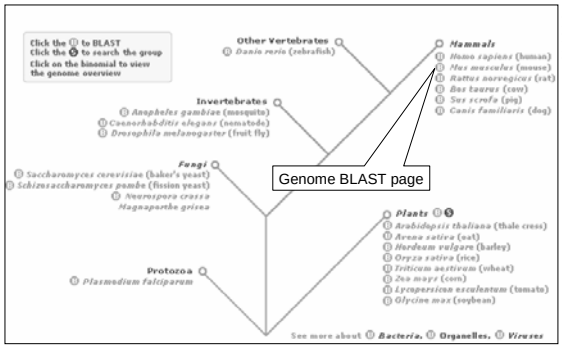
NCBI FieldGuide

1: BI835081, 603087730F1 NIH_M...[gi:15946631]

Links
- Map Viewer
- Taxonomy
- UniGene
- Blast

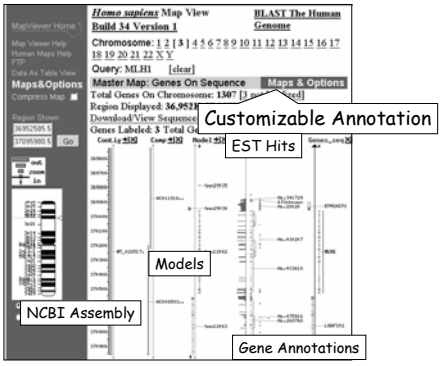
```
LOCUS       BI835081                 667 bp    mRNA    linear   EST
DEFINITION  603087730F1 NIH_MGC_120 Homo sapiens cDNA clone IMAGE:52
            mRNA sequence.
ACCESSION   BI835081
VERSION     BI835081.1  GI:15946631
KEYWORDS    EST.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo;
REFERENCE   1  (bases 1 to 667)
  AUTHORS   NIH-MGC http://mgc.nci.nih.gov/.
  TITLE     National Institutes of Health, Mammalian Gene Collection (MGC)
  JOURNAL   Unpublished (1999)
COMMENT     Contact: Robert Strausberg, Ph.D.
            Email: cgap@c-r@mail.nih.gov
            Tissue Procurement: Life Technologies, Inc.
             cDNA Library Preparation: Life Technologies, Inc.
             cDNA Library Arrayed by: The I.M.A.G.E. Consortium (LLNL)
             DNA Sequencing by: Incyte Genomics, Inc.
             Clone distribution: MGC clone distribution information can be
            found through the I.M.A.G.E. Consortium/LLNL at:
            http://image.llnl.gov
            Plate: LLAM13570  row: a  column: 10
            High quality sequence stop: 667.
FEATURES             Location/Qualifiers
     source          1..667
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /clone="IMAGE:5226825"
                     /lab_host="DH10B"
                     /clone_lib="NIH_MGC_120"
                     /note="Organ: pooled pancreas and spleen; Vector:
                     pCMV-SPORT6; Site_1: NotI; Site_2: EcoRV (destroyed); RNA
                     source anonymous pool of spleen and pancreas from 28 yo
                     male.  Library is oligo-dT primed and directionally cloned
                     (EcoRV  site is destroyed upon cloning).  Average insert
                     size 1.5 kb, insert size range 1-2.5 kb.  Library is
                     normalized and enriched for  full-length clones and was
                     constructed by C. Gruber (Invitrogen). Research Genetics
                     tracking code 025. Note: this is a NIH_MGC Library."
```

The Map Viewer



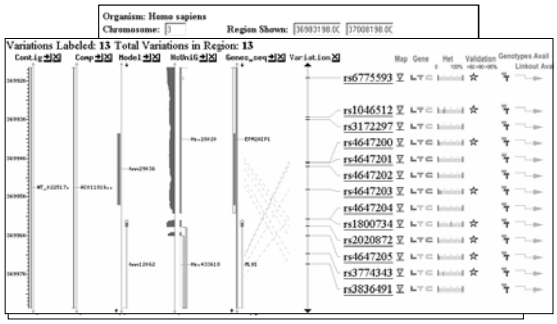
NCBI FieldGuide

Map Viewer: Human MLH1



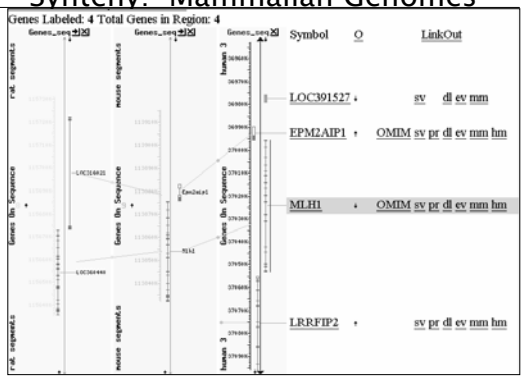
NCBI FieldGuide

Maps and Options



NCBI FieldGuide

Synteny: Mammalian Genomes



NCBI FieldGuide

The New Homologene

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

ologs

chick B frog B

A-chain gene

gen

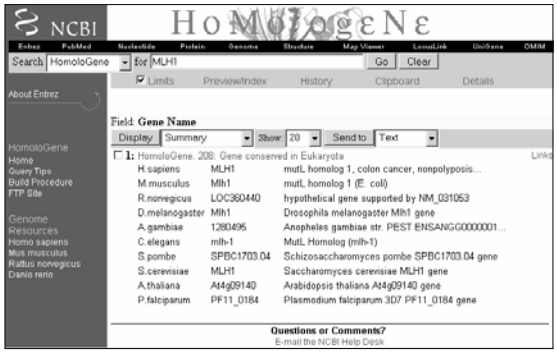
earl

Species	Number of Genes input	grouped	HomoloGene Groups
H.sapiens	21,450	16,793	14,431
M.musculus	23,650	18,287	15,003
R.norvegicus	20,885	15,880	13,424
D.melanogaster	12,386	7,846	6,756
A.gambiae	11,784	7,562	6,657
C.elegans	18,710	5,910	4,787
S.pombe	4,942	3,160	2,929
S.cerevisiae	5,803	3,289	2,832
A.thaliana	27,133	7,099	4,114
P.falciparum	5,206	1,492	1,408

Last updated on: 02/25/2004

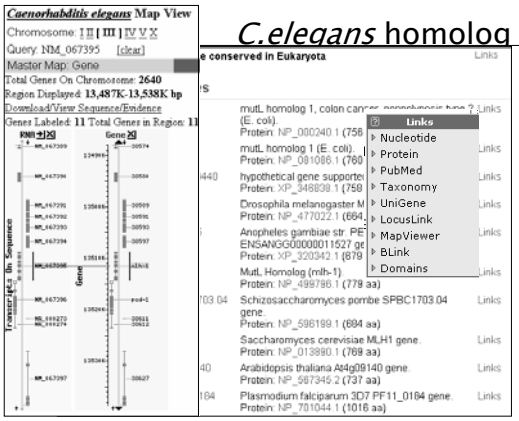
NCBI FieldGuide

The New Homologene



NCBI FieldGuide

C.elegans homolog

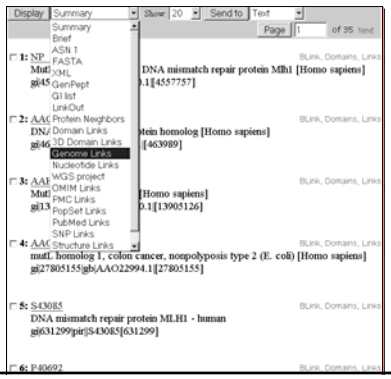


NCBI FieldGuide

NCBI FieldGuide

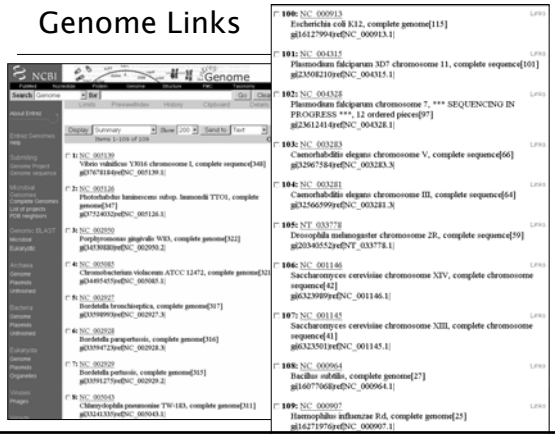
Microbial Genomes

Related Proteins->Genome Links



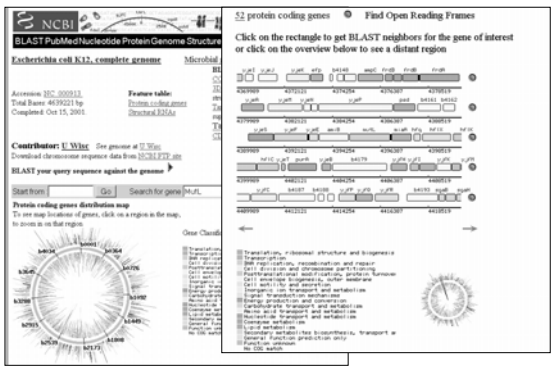
NCBI FieldGuide

Genome Links



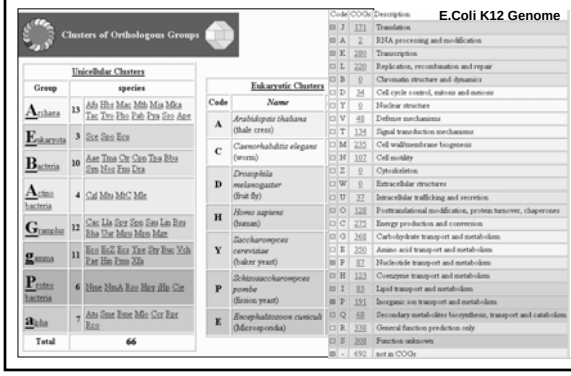
NCBI FieldGuide

Microbial Genomes



NCBI FieldGuide

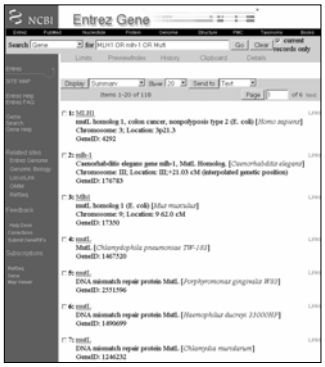
COGs Analysis



NCBI FieldGuide

Entrez Genes: integrated gene-based access

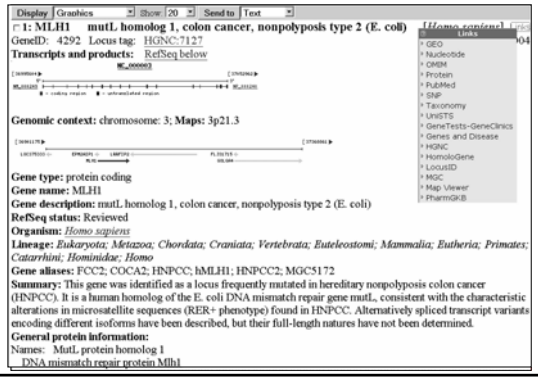
NCBI FieldGuide



LocusLink
Complete Genomes
- •eukaryotic
- •microbial
- •organelle

Genes MLH1: Central Resource

NCBI FieldGuide



NCBI FieldGuide

Intermission