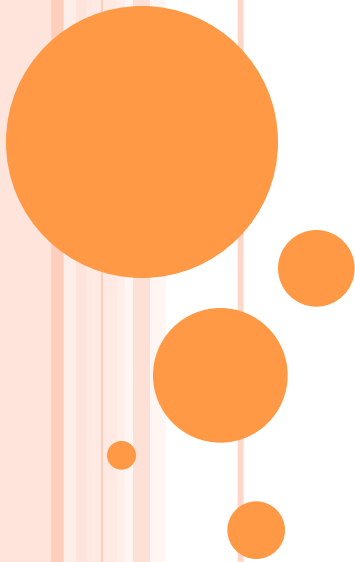


# INTRODUCTION TO BIOLOGICAL DATABASES

Etienne de Villiers, PhD

Kemri Wellcome Trust Research Programme  
University of Oxford

*NOTE: Lecture adapted from SANBI course*



# WHAT YOU NEED TO LEARN:

- What is a database and what are the features of an ideal db?
- What are the relationships/differences between primary and derived sequence databases?
- What are the benefits of RefSeq?
- Why is data integration useful?



# TOUR OF MAJOR BIOLOGICAL DATABASES

- There is a tremendous amount of information about biomolecules in publicly available databases.
- Today, we will look at a few of the main databases and what kind of information they contain.



# WHAT CAN BE DISCOVERED ABOUT A GENE BY A DATABASE SEARCH?

- A little or a lot, depending on the gene
  - Evolutionary information: homologous genes, taxonomic distributions, allele frequencies, synteny, etc.
  - Genomic information: chromosomal location, introns, UTRs, regulatory regions, shared domains, etc.
  - Structural information: associated protein structures, fold types, structural domains
  - Expression information: expression specific to particular tissues, developmental stages, phenotypes, diseases, etc.
  - Functional information: enzymatic/molecular function, pathway/cellular role, localization, role in diseases



# USING A DATABASE

- How to get information out of a database:
  - Browsing: no targeted information to retrieve
  - Search: looking for particular information
- Searching a database:
  - Must have a key that identifies the element(s) of the database that are of interest.
    - **Name of gene**
    - **Sequence of gene**
    - **Other information**
  - Helps to have particular *informational goals*



# SEARCHING FOR INFORMATION ABOUT GENES AND THEIR PRODUCTS

- Gene and gene product databases are often organized by sequence
  - Genomic sequence encodes all traits of an organism.
  - Gene products are uniquely described by their sequences.
  - Similar sequences among biomolecules indicates both similar function and an evolutionary relationship
- Macromolecular sequences provide biologically meaningful keys for searching databases

# SEARCHING SEQUENCE DATABASES

- Start from sequence, find information about it
- Many kinds of input sequences
  - Could be amino acid or nucleotide sequence
  - Genomic or mRNA/cDNA or protein sequence
  - Complete or fragmentary sequences
- Exact matches are rare (even uninteresting in many cases), so often goal is to retrieve a set of similar sequences.
  - Both small (mutations) and large (required for function) differences within “similar” can be interesting.




# WHAT MIGHT WE WANT TO KNOW ABOUT A SEQUENCE?

- Is this sequence similar to any known genes?  
How close is the best match? Significance?
- What do we know about that gene?
  - Genomic (chromosomal location, allelic information, regulatory regions, etc.)
  - Structural (known structure? structural domains? etc.)
  - Functional (molecular, cellular & disease)
- Evolutionary information:
  - Is this gene found in other organisms?
  - What is its taxonomic tree?






## NCBI AND ENTREZ

- One of the most useful and comprehensive sources of databases is the NCBI, part of the National Library of Medicine.
- NCBI provides interesting summaries, browsers for genome data, and search tools
- Entrez is their database search interface  
<http://www.ncbi.nlm.nih.gov/Entrez>
- Can search on gene names, sequences, chromosomal location, diseases, keywords, ... 

# WEB ACCESS: WWW.NCBI.NLM.NIH.GOV



Resources ▾ How To ▾

My NCBI | Sign In

National Center for Biotechnology Information

Search  for

Resources

NCBI Home

All Resources (A-Z)

Literature

DNA & RNA

Proteins

Sequence Analysis

Genes & Expression

Genomes

Maps & Markers

Domains & Structures

Genetics & Medicine

Taxonomy

Data & Software

Training & Tutorials

Homology

Small Molecules

Variation


Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[More about the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS](#)

Genome

1000 prokaryotic genomes are now completed and available in the Genome database



Popular Resources

- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene
- Nucleotide
- Protein
- GEO
- Conserved Domains
- Structure
- PubChem

You are here: NCBI

GETTING STARTED

[Site Map](#)  
[NCBI Help Manual](#)  
[NCBI Handbook](#)  
[Training & Tutorials](#)

RESOURCES

[Literature](#)  
[DNA & RNA](#)  
[Proteins](#)  
[Sequence Analysis](#)  
[Genes & Expression](#)  
[Genomes](#)  
[Maps & Markers](#)  
[Domains & Structures](#)  
[Genetics & Medicine](#)  
[Taxonomy](#)  
[Data & Software](#)  
[Training & Tutorials](#)  
[Homology](#)  
[Small Molecules](#)  
[Variation](#)

POPULAR

[PubMed](#)  
[PubMed Central](#)  
[Bookshelf](#)  
[BLAST](#)  
[Gene](#)  
[Nucleotide](#)  
[Protein](#)  
[GEO](#)  
[Conserved Domains](#)  
[Structure](#)  
[PubChem](#)

FEATURED

[GenBank](#)  
[Reference Sequences](#)  
[Map Viewer](#)  
[Genome Projects](#)  
[Human Genome](#)  
[Mouse Genome](#)  
[Influenza Virus](#)  
[Primer-BLAST](#)  
[Short Read Archive](#)

NCBI INFORMATION

[About NCBI](#)  
[Research at NCBI](#)  
[NCBI Newsletter](#)  
[NCBI FTP Site](#)  
[Contact Us](#)

Help Desk

New pages!

Common footer

# WHAT ARE DATABASES?

- **Structured** collection of information.
- Consists of basic units called records or entries.
- Each record consists of fields, which hold **pre-defined** data related to the record.
- For example, a protein database would have protein entries as records and protein properties as fields (e.g., name of protein, length, amino-acid sequence)

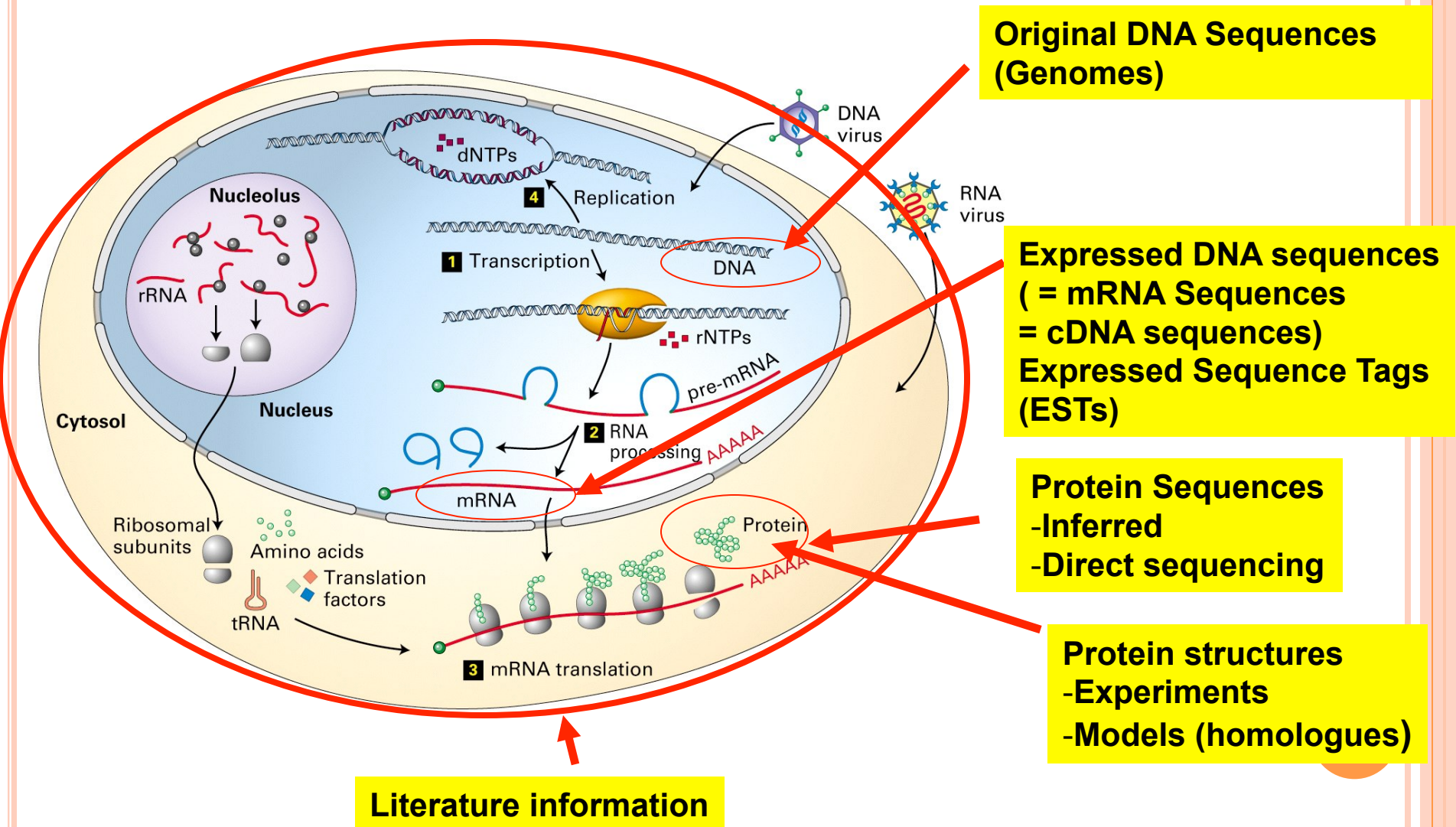


# THE ‘PERFECT’ DATABASE

- Comprehensive, but **easy to search**.
- Annotated, but not “too annotated”.
- A simple, easy to understand structure.
- **Cross-referenced.**
- Minimum redundancy.
- **Easy retrieval** of data.



# THE CENTRAL DOGMA & BIOLOGICAL DATA



# NCBI DATABASES AND SERVICES

- GenBank primary sequence database
- Free public access to biomedical literature
  - PubMed free Medline (3 million searches per day)
  - PubMed Central full text online access
- Entrez integrated molecular and literature databases



# TYPES OF MOLECULAR DATABASES

## ○ Primary Databases

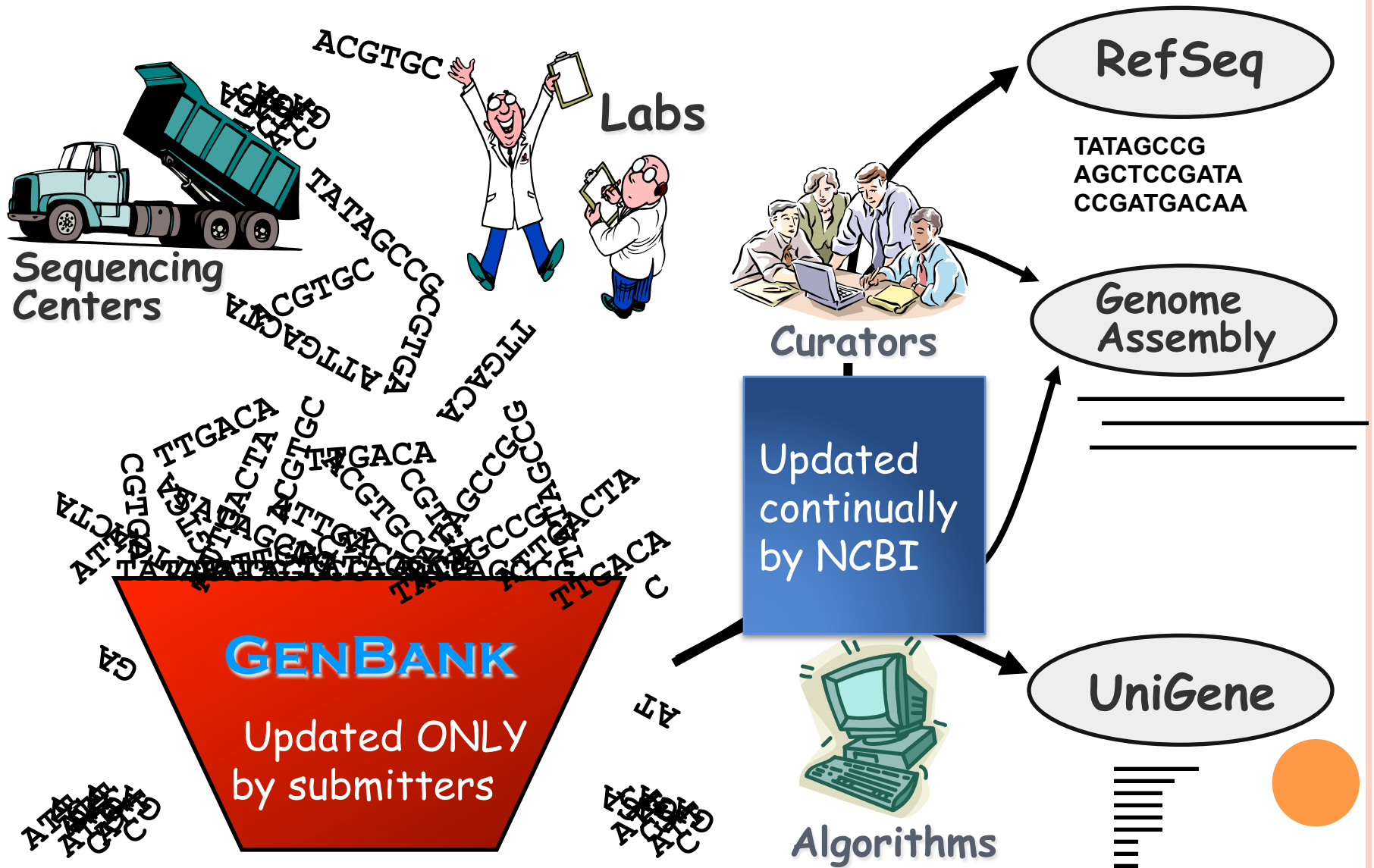
- Original submissions by experimentalists
- Content controlled by the **submitter**
  - Examples: GenBank, Trace, SRA, SNP, GEO

## ○ Derivative Databases

- *Derived* from primary data
- Content controlled by **third party** (NCBI)
  - Examples: NCBI Protein, Refseq, Ensembl, RefSNP, GEO datasets, UniGene, Homologene, Structure, Conserved Domain



# PRIMARY VS. DERIVATIVE SEQUENCE DATABASES





# SEQUENCE DATABASES AT NCBI

## ○ **Primary**

- GenBank: NCBI's primary sequence database
- Trace Archive: reads from capillary sequencers
- Sequence Read Archive: next generation data

## ○ **Derivative**

- GenPept (GenBank translations)
- Outside Protein (UniProt—Swiss-Prot, PDB)
- NCBI Reference Sequences (**RefSeq**)



# GENBANK - PRIMARY SEQUENCE DB

- **Nucleotide only** sequence database
- **Archival** in nature
  - Historical
  - Reflective of submitter point of view (subjective)
  - **Redundant**
- **Data**
  - Direct submissions (traditional records)
  - Batch submissions
  - FTP accounts (genome data)



# GENBANK - PRIMARY SEQUENCE DB (2)

- Three collaborating databases
  1. GenBank
  2. European Molecular Biology Laboratory (EMBL) Database
  3. DNA Database of Japan (DDBJ)



# TRADITIONAL GENBANK RECORD

LOCUS HSHMLHI 2503 bp mRNA linear PRI 31-MAR-1994

DEFINITION Human DNA mismatch  
ACCESSION U07418  
VERSION U07418.1 GI:461185  
KEYWORDS .  
SOURCE Homo sapiens (Homo sapiens)  
ORGANISM [Homo sapiens](#)  
Eukaryota; Metazoa; Chordata; Mammalia; Eutheria  
REFERENCE 1 (bases 1 to 2503)

## FEATURES

source

[gene](#)

[CDS](#)

## Location/Qualifiers

1..2503

/organism="Homo sapiens"

/db\_xref="taxon:9606"

/chromosome

/map

/tissue

/development

/gene

42..2503

/gene

/function

/note

Swiss

/codon

/protein

/db\_xref

/translation

TSIQV

ALASIS

TRRKAL

VFGNAV

ETVYAA

LGSNS

LDAPFL

TTKGT

LQEEIN

DFANFC

YFSLEI

FYSIRP

LQLANT

## BASE COUNT

723 a 539 c 599 g 642 t

## ORIGIN

```
1 gttgaacatc tagacgtttc cttggctctt ctggcgccaa aatgtcgttc gtggcagggg
61 ttattcggcg gctggacgag acagtgttga accgcatcgc ggcgggggaa gttatccagc
121 ggccagctaa tgctatcaaa gagatgattg agaactgttt agatgcaaaa tccacaagta
181 ttcaagtgat tgtaaagag ggaggcctga agttgattca gatccaagac aatggcaccg
241 ggatcaggaa agaagatctg gatattgtat gtgaaagggt cactactagt aaactgcagt
301 cctttgagga tttagccagt atttctacct atggccttcg aggtgaggct ttggccagca
361 taagccatgt ggctcatgtt actattacaa cgaaaacagc tgatggaaag tgtgcataca
421 gagcaagtta ctcatatgga aaactgaaag cccctcctaa accatgtgct ggcaatcaag
481 ggaccagat caccgtggag gacctttttt acaacatagc caccaggaga aaagctttaa
541 aaaatccaag tgaagaatat gggaaaaatt tgggaagttg tggcagggtat tcagtacaca
601 atgcaggcat tagtttctca gttaaaaaac aaggagagac agtagctgat gttaggacac
661 tacccaatgc ctcaaccgtg gacaatattc gctccgtctt tggaaatgct gttagtcgag
721 aactgataga aattggatgt gaggataaaa ccctagcctt caaaatgaat gggtacatat
781 ccaatgcaaa ctactcagt aagaagtgca tcttcttact cttcatcaac caatgctctg
841 tagaatcaac ttcttgaga aaagccatag aaacagtgtg tgcagcctat ttgccaaaaa
901 acacacaccc attcctgtac ctcagtttag aaactcagtc ccagaatgtg gatgttaagt
961 tgcacccccc aaagcatgaa gttcacttcc tgcacgagga gagcatcctg gagcgggtgc
1021 agcagcacat cgagagcaag ctctggggtt ccaattctct caggatgtac ttcaccacag
1081 ctttgctacc aggacttgct ggcccctctg gggagatggt taaatccaca acaagtctga
1141 cctcgtcttc tacttctgga agtagtgata aggtctatgc caccagatg gttcgtacag
1201 attcccgga acagaagctt gatgcatttc tgcagcctct gagcaaaccc ctgtccagtc
1261 agccccaggc cattgtcaca gaggataaga cagatatttc tagtggcagg gctaggcagg
1321 aagatgagga gatgcttgaa ctcccagccc ctgctgaagt ggctgccaaa aatcagagct
1381 tggaggggga tacaacaaag gggacttcag aaatgtcaga gaagagagga cctacttcca
1441 gcaaccccag aaagagacat cgggaagatt ctgatgtgga aatggtggaa gatgattccc
1501 gaaaggaaat gactgcagct tgtaccccc ggagaaggat cattaacctc actagtgttt
1561 tgagtctcca ggaagaaatt aatgagcagg gacatgaggt tctccgggag atgttgcata
1621 accactcctt cgtgggctgt gtaatcctc agtgggcctt ggcacagcat caaaccaggt
1681 tataccttct caacaccacc aagcttagtg aagaactgtt ctaccagata ctctttatg
1741 attttgccaa ttttggtgtt ctcaggttat cggagccagc accgctcttt gaccttgcca
1801 tgcttgccct agatagtcga gagagtggct ggacagagga agatggtccc aaagaaggac
1861 ttgctgaata cattgttgag ttctgaaga agaaggctga gatgcttgca gactatttct
1921 ctttggaat tgatgaggaa gggaaacctga ttggattacc ccttctgatt gacaactatg
1981 tgcccccttt ggagggactg cctatcttca ttcttcgact agcccactgag gtgaattggg
2041 acgaagaaaa ggaatgtttt gaaagcctca gtaagaatg cgctatgttc tattccatcc
2101 ggaagcagta catatctgag gactcgaccc tctcaggcca gcagagtga gtcctggctc
2161 ccattccaaa ctctggaag tggactgtgg aacacattgt ctataaagcc ttgcgctcac
2221 acattctgcc tcctaaacct ttcacagaag atggaaatat cctgcagctt gctaacctgc
2281 ctgactcata caaagtcttt gagaggtgtt aaataggtt ccttctgac tgtggaggtg
2341 gttcttcttt ctctgtattc cgatacaaa gtgtgtatca aagtgtgata taaaagtgtg
2401 accaacataa gtgtgtgtag cacttaagac ttatacttgc cttctgatag tattccttta
2461 tacacagtgg attgattata aataaataga tgtgtcttaa cat
```

ACCESSION

VERSION

Version

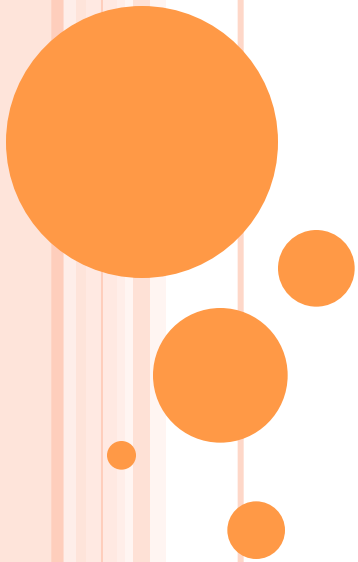
Tracks changes in

20850, USA

well annotated

the sequence is the data


# DERIVATIVE SEQUENCE DATABASES



# GENPEPT: GENBANK CDS TRANSLATIONS

```
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="MSFVAGVIRRLDET VVNRIAAGEVIQRPANA IKEMIENCLDAKS
                        TSIQVIVKEGGLKLIQIQDNGTGIRKEDLDIVCERFTTSK LQSFEDLASISTYGF
                        RGEALASISHVAHVTTITTKTADGKCA YRASYS DGKLKAPPKPCAGNQGTQIT
                        VEDL FYNIA TRRKALKNPSEEY GKILEVVGRYSVHNAGISFSVKKQGETVAD
                        VRTL PNASTVDNIRS
```

**>gi|463989|gb|AAC50285.1| DNA mismatch repair prote...**  
MSFVAGVIRRLDET VVNRIAAGEVIQRPANA IKEMIENCLDAKSTSIQVIV...  
EDLDIVCERFTTSK LQSFEDLASISTYGF RGEALASISHVAHVTTITTKTAD...



# REFSEQ: *DERIVATIVE* SEQUENCE DATABASE

- **Curated** transcripts and proteins
- **Model** transcripts and proteins
- **Assembled Genomic Regions**
- **Chromosome records**
  - Human genome
  - microbial
  - organelle

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



# SELECTED REFSEQ ACCESSION NUMBERS

## 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  
genomes, human chromosomes

AC\_123455

Alternate assemblies

## Assemblies

NT\_123456

Contig

NW\_123456

WGS Supercontig



# GENBANK TO REFSEQ

☐ [Human apolipoprotein E \(epsilon-4 allele\) gene, complete cds](#)  
1. 5,515 bp linear DNA  
M10065.1 GI:178852

☐ [Human mRNA fragment for apolipoprotein E \(apo E\)](#)  
2. 528 bp linear mRNA  
X00199.1 GI:28808

☐ [H.sapiens mRNA](#)  
3. 275 bp linear mRNA  
Z70760.1 GI:12631

☐ [Homo sapiens c](#)  
4. 1,023 bp linear  
AK314898.1 GI:16

[Homo sapiens apolipoprotein E \(APOE\), mRNA](#)  
1,223 bp linear mRNA  
NM\_000041.2 GI:48762938

☐ [Human apolipoprotein E mRNA, complete cds](#)  
5. 1,157 bp linear mRNA  
M12529.1 GI:178848

☐ [Homo sapiens preapolipoprotein E \(APOE\) mRNA, complete cds](#)  
6. 1,156 bp linear mRNA  
K00396.1 GI:178850

☐ [Homo sapiens apolipoprotein E, mRNA \(cDNA clone MGC:1571 IMAGE:3355712\), complete cds](#)  
7. 1,186 bp linear mRNA  
BC003557.1 GI:13097698



# REFSEQS: ANNOTATION REAGENTS



Scanning....

Genomic DNA  
(**NC**, **NT**, **NW**)

Model mRNA (**XM**)  
(**XR**)



Model protein (**XP**)



= ?

Curated mRNA (**NM**)  
(**NR**)



Curated Protein (**NP**)

RefSeq

GenBank  
Sequences

**RNM\_002467**

[GBC000141](#)

[GBC000917](#)

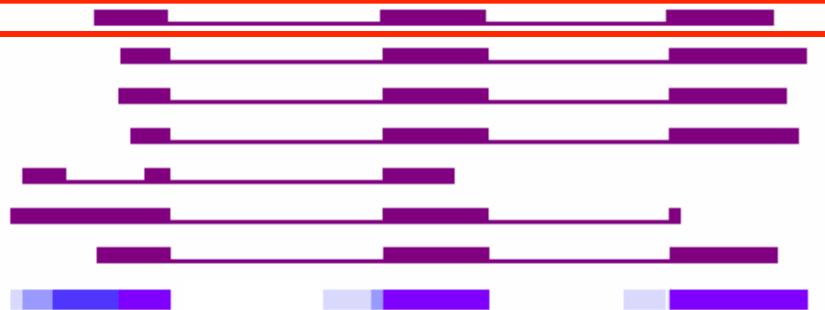
[GK02276](#)

[GM13929](#)

[GM13930](#)

[GV00568](#)

EEESTs



# REFSEQ BENEFITS

- Non-redundancy
- Updates to reflect current sequence data and *biology*
- Data *validation*
- Format *consistency*
- Distinct accession series
- ***Stewardship* by NCBI staff and collaborators**



# OTHER DERIVATIVE DATABASES

- Expressed Sequences
- dbSNP
- Structure
- Gene
- and more...

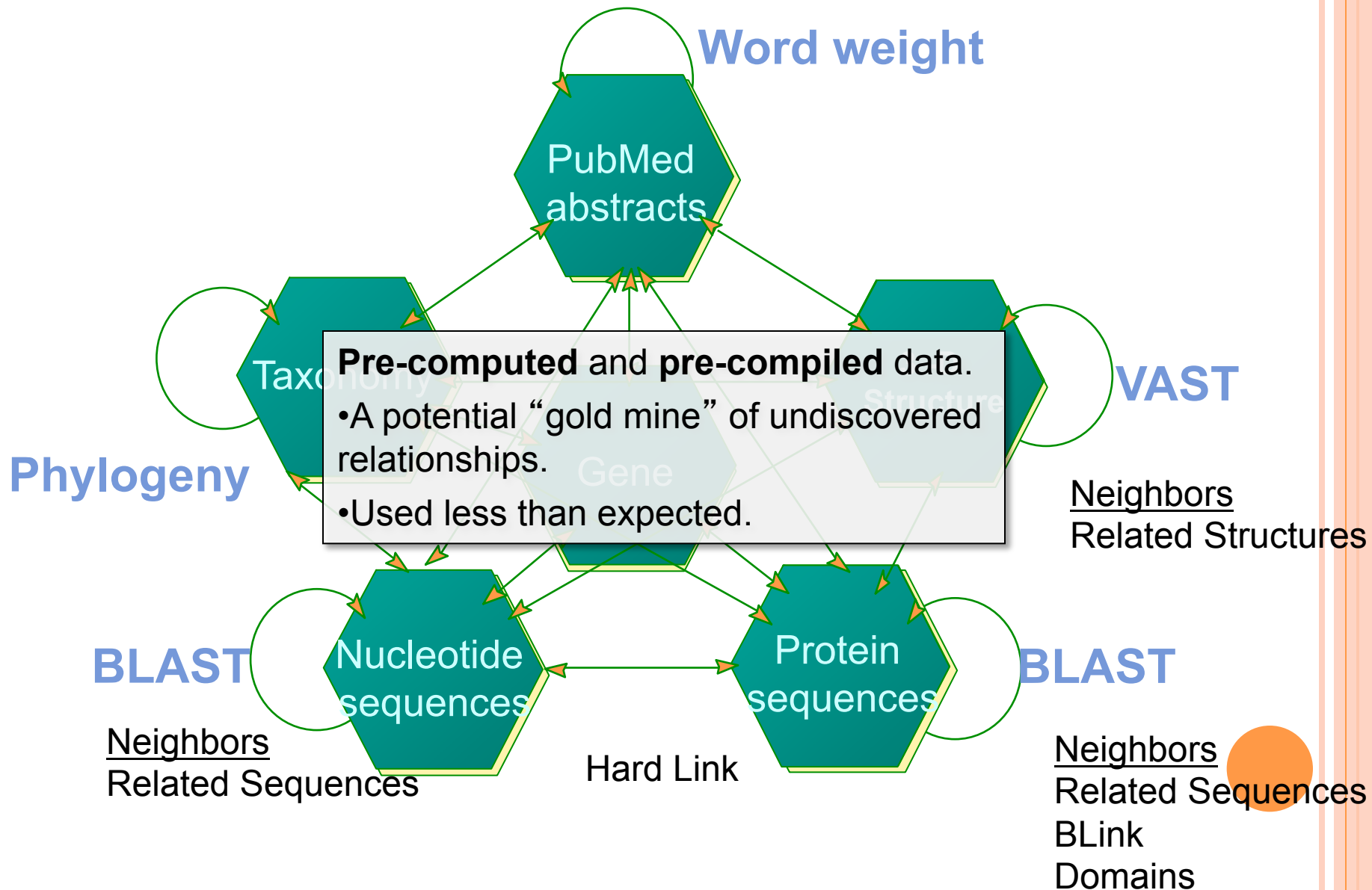




*ENTREZ*

**FINDING RELEVANT  
INFORMATION IN NCBI  
DATABASES**

# ENTREZ: A DISCOVERY SYSTEM



# GLOBAL QUERY: ALL NCBI DATABASES

Search across databases    [Help](#)

- Result counts displayed in gray indicate one or more terms not found

<b>66868</b> <b>PubMed:</b> biomedical literature citations and abstracts	<b>892</b> <b>Books:</b> online books
<b>12588</b> <b>PubMed Central:</b> free, full text journal articles	<b>145</b> <b>OMIM:</b> online Mendelian Inheritance in Man
<b>none</b> <b>Site Search:</b> NCBI web and FTP sites	<b>165</b> <b>OMIA:</b> online Mendelian Inheritance in Animals
<b>10785</b> <b>Nucleotide:</b> Core subset of nucleotide sequence records	<b>180</b> <b>dbGaP:</b> genotype and phenotype
<b>none</b> <b>EST:</b> Expressed Sequence Tag records	<b>5</b> <b>UniGene:</b> gene-oriented clusters of transcript sequences
<b>none</b> <b>GSS:</b> Genome Survey Sequence records	<b>31</b> <b>CDD:</b> conserved protein domain database
<b>4445</b> <b>Protein:</b> sequence database	<b>435</b> <b>3D Domains:</b> domains from Entrez Structure
<b>1</b> <b>Structure:</b> three-dimensional macromolecular structures	<b>1</b> <b>PopSet:</b> population study data sets
<b>none</b> <b>Taxonomy:</b> organisms in GenBank	<b>48581</b> <b>GEO Profiles:</b> expression and molecular abundance profiles
<b>none</b> <b>SNP:</b> single nucleotide polymorphism	<b>58</b> <b>GEO DataSets:</b> experimental sets of GEO data
<b>525</b> <b>Gene:</b> gene-centered information	<b>none</b> <b>Cancer Chromosomes:</b> cytogenetic databases
<b>none</b> <b>SRA:</b> Short Read Archive	<b>82</b> <b>PubChem BioAssay:</b> bioactivity screens of chemical substances
<b>15</b> <b>BioSystems:</b> Pathways and systems of interacting molecules	<b>5</b> <b>PubChem Compound:</b> unique small molecule chemical structures
<b>none</b> <b>HomoloGene:</b> eukaryotic homology groups	<b>28</b> <b>PubChem Substance:</b> deposited chemical substance records
<b>114</b> <b>GENSAT:</b> gene expression atlas of mouse central nervous system	<b>none</b> <b>Protein Clusters:</b> a collection of related protein sequences
<b>61</b> <b>Probe:</b> sequence-specific reagents	<b>none</b> <b>Peptidome:</b> MS/MS proteomic experiments
<b>1</b> <b>Genome Project:</b> genome project information	
<b>3</b> <b>Journals:</b> detailed information about the journals indexed in PubMed and other Entrez databases	<b>3</b> <b>MeSH:</b> detailed information about NLM's controlled vocabulary
<b>1476</b> <b>NLM Catalog:</b> catalog of books, journals, and audiovisuals in the NLM collections	

**The Entrez system: 38 (and counting) integrated databases**

# TRADITIONAL METHOD: THE LINKS MENU

## DNA Sequence

[U07343](#) Reports

Homo sapiens DNA mismatch repair protein homolog (MLH1) mRNA, complete cds  
gil463988|gb|U07343.1|HSU07343[463988]

Links

Nucleotide – Protein Link

Related Proteins

[AAC50285](#) Reports

Conserved Domains, BLink, Links

DNA mismatch repair protein homolog [Homo sapiens]  
gil463989|gb|AAC50285.1|[463989]

[1B63 A](#) Reports

Conserved Domains, BLink, Links

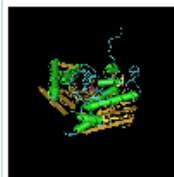
Chain A, Mutl Complexed With Adpnp  
gil5542073|pdb|1B63|A[5542073]

Protein – Structure Link

3-D Structure

1B63

Related Structures, Literature, Domains, Chemicals



Mutl Complexed With Adpnp [Dna Mismatch Repair]

Taxonomy: [Escherichia coli](#)

Proteins: 1; Chemicals: 3

modified: 2007/10/06; MMDB ID: 10447



# THE PROBLEM




- Rapidly growing databases with complex and **changing relationships**
- Rapidly changing interfaces to match the above

## Result

- Many people don't know:
  - **Where to begin**
  - Where to click on a Web page
  - Why it might be useful to click there



# GLOBAL NCBI (ENTREZ) SEARCH

 NCBI [Resources](#)  [How To](#) 

National Center for Biotechnology Information

Search  for

## Resources

- NCBI Home
- All Resources (A-Z)
- Literature
- DNA & RNA
- Proteins
- Sequence Analysis
- Genes & Expression
- Genomes
- Maps & Markers
- Domains & Structures
- Genetics & Medicine
- Taxonomy
- Data & Software
- Training & Tutorials
- Homology
- Small Molecules
- Variation

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing free access to biomedical and genomic information.

[More about the NCBI mission](#) | [Organization](#) | [Research](#) | [RSS](#)

### Genome Reference Consortium

Formed to improve human and mouse reference assemblies, GRC will fix loci misrepresented in reference assembly, fill remaining gaps, and make alternate representations of complex loci.

II 1 2 3 4

## Popular Resources

- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene
- Nucleotide
- Protein
- GEO
- Conserved Domains
- Structure
- PubChem

## NCBI News

[November and October News](#) 02 Dec 2009  
Featured: New Discovery-oriented PubMed and NCBI Homepage. T...



[NCBI News - September 2009](#) 05 Oct 2009  
The September 2009 issue of the NCBI News is available ...

[NCBI News - August](#) 19 Aug 2009

## How To...

- Obtain the full text of an article
- Retrieve all sequences for an organism or taxon
- Find a homolog for a gene in another organism
- Find genes associated with a phenotype or disease
- Design PCR primers and check them for specificity
- Find the function of a gene or gene product
- Find syntenic regions between the genomes of two organisms

# GLOBAL ENTREZ SEARCH RESULTS



Entrez, The Life Sciences Search Engine

[HOME](#)
[SEARCH](#)
[SITE MAP](#)

[PubMed](#)
[All Databases](#)
[Human Genome](#)
[GenBank](#)
[Map Viewer](#)
[BLAST](#)

Search across databases


[Help](#)

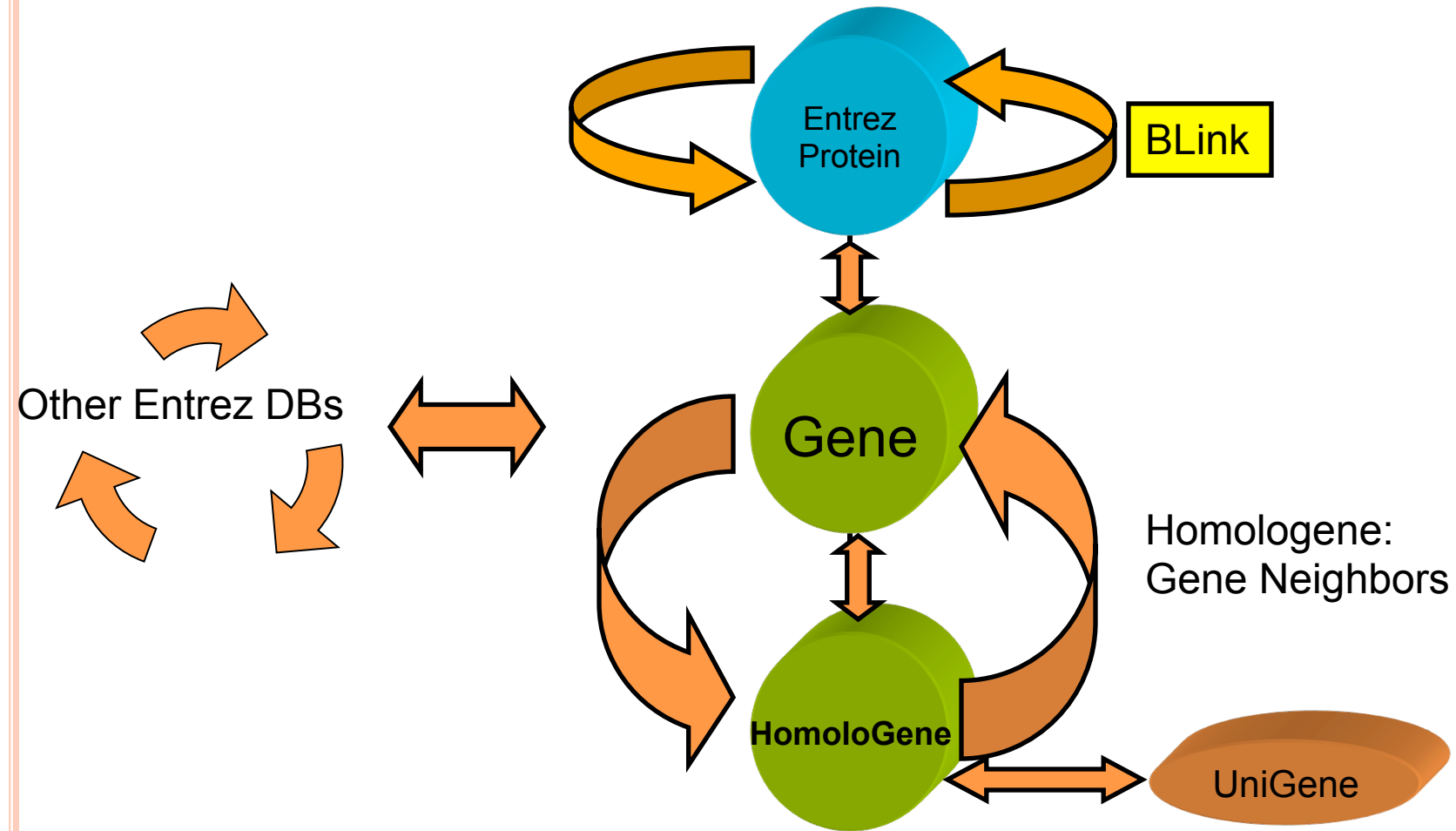
- Result counts displayed in gray indicate one or more terms not found

<b>83359</b> <b>PubMed:</b> biomedical literature citations and abstracts	<b>802</b> <b>Books:</b> online books
<b>13740</b> <b>PubMed Central:</b> free, full text journal articles	<b>445</b> <b>OMIM:</b> online Mendelian Inheritance in Man
<b>11</b> <b>Site Search:</b> NCBI web and FTP sites	<b>1</b> <b>OMIA:</b> online Mendelian Inheritance in Animals


<b>941100</b> <b>Nucleotide:</b> Core subset of nucleotide sequence records	<b>87</b> <b>dbGaP:</b> genotype and phenotype
<b>3852867</b> <b>EST:</b> Expressed Sequence Tag records	<b>199</b> <b>UniGene:</b> gene-oriented clusters of transcript sequences
<b>310846</b> <b>GSS:</b> Genome Survey Sequence records	<b>13</b> <b>CDD:</b> conserved protein domain database
<b>398899</b> <b>Protein:</b> sequence database	<b>none</b> <b>3D Domains:</b> domains from Entrez Structure
<b>8</b> <b>Genome:</b> whole genome sequences	<b>39</b> <b>UniSTS:</b> markers and mapping data
<b>3</b> <b>Structure:</b> three-dimensional macromolecular structures	<b>1</b> <b>PopSet:</b> population study data sets
<b>none</b> <b>Taxonomy:</b> organisms in GenBank	<b>86766</b> <b>GEO Profiles:</b> expression and molecular abundance profiles
<b>none</b> <b>SNP:</b> single nucleotide polymorphism	<b>136</b> <b>GEO DataSets:</b> experimental sets of GEO data
<b>780</b> <b>Gene:</b> gene-centered information	<b>162</b> <b>Cancer Chromosomes:</b> cytogenetic databases
<b>none</b> <b>SRA:</b> Sequence Read Archive	<b>16</b> <b>PubChem BioAssay:</b> bioactivity screens of chemical substances
<b>31</b> <b>BioSystems:</b> Pathways and systems of interacting molecules	<b>none</b> <b>PubChem Compound:</b> unique small molecule chemical structures
<b>17</b> <b>HomoloGene:</b> eukaryotic homology groups	<b>none</b> <b>PubChem Substance:</b> deposited chemical substance records
<b>none</b> <b>GENSAT:</b> gene expression atlas of mouse central nervous system	<b>none</b> <b>Protein Clusters:</b> a collection of related protein sequences
<b>1995</b> <b>Probe:</b> sequence-specific reagents	<b>none</b> <b>Peptidome:</b> MS/MS proteomic experiments
<b>1</b> <b>Genome Project:</b> genome project information	

<b>none</b> <b>Journals:</b> detailed information about the journals indexed in PubMed and other Entrez databases	<b>18</b> <b>MeSH:</b> detailed information about NLM's controlled vocabulary
<b>423</b> <b>NLM Catalog:</b> catalog of books, journals, and audiovisuals in the NLM collections	

# ENTREZ TIP: START SEARCHES IN GENE



# PRECISE RESULTS



## Entrez Gene

My NCBI  
[\[Sign In\]](#) [\[Register\]](#)

All DatabasesPubMedNucleotideProteinGenomeStructureOMIMPMCIJournalsBooks

Search  for    [Save Search](#)

LimitsPreview/IndexHistoryClipboardDetails

Display

☐ 1: **MLH1 mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)** [*Homo sapiens*]  
GeneID: 4292 updated 8-Dec-2009

**Summary**

<b>Official Symbol</b>	MLH1	provided by <a href="#">HGNC</a>
<b>Official Full Name</b>	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	provided by <a href="#">HGNC</a>
<b>Primary Source</b>	<a href="#">HGNC:7127</a>	
<b>See related</b>	<a href="#">Ensembl:ENSG00000076242</a> ; <a href="#">HPRD:00390</a> ; <a href="#">MIM:120436</a>	
<b>Gene type</b>	protein coding	
<b>RefSeq status</b>	REVIEWED	
<b>Organism</b>	<a href="#">Homo sapiens</a>	
<b>Lineage</b>	<a href="#">Eukaryota</a> <a href="#">Mammalia</a> <a href="#">Hominidae</a> ; <a href="#">Homo</a>	
<b>Also known as</b>	FCC2; COCA2; HNPCC; hMLH1; HNPCC2; MGC5172; MLH1	
<b>Summary</b>	This gene was identified as a locus frequently mutated in hereditary nonpolyposis colon cancer (HNPCC). It is a human homolog of the E. coli DNA mismatch repair gene mutL, consistent with the characteristic alterations in microsatellite sequences (RER+phenotype) found in HNPCC. Alternative splicing results in multiple transcript variants encoding distinct isoforms. Additional transcript variants have been described, but their full-length natures have not been determined.	

[Entrez Gene Home](#)  
[Table Of Contents](#)  
**Links**  
[Order cDNA clone](#)  
[BioAssay, by Gene target](#)  
[BioSystems](#)  
[Books](#)  
[CCDS](#)  
[Conserved Domains](#)  
[EST](#)  
[Full text in PMC](#)  
[GEO Profiles](#)  
[Gene Genotype](#)  
[GeneView in dbSNP](#)  
[Genome](#)  
[BioGene](#)  
[Viewer](#)  
[Nucleotide](#)  
[OMIM](#)  
[Probe](#)  
[Protein](#)  
[PubChem Compound](#)  
[PubChem Substance](#)  
[PubMed](#)  
[PubMed \(GeneRIF\)](#)  
[PubMed \(OMIM\)](#)  
[SNP](#)  
[SNP VarView](#)

# MLH1 GENE RECORD

1: MLH1 mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [ *Homo sapiens* ]

GeneID: 4292

updated 8-Dec-2009

## Summary

**Official Symbol** MLH1

provided by [HGNC](#)

**Official Full Name** mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)

provided by [HGNC](#)

**Primary Source**

**See related**

**Gene type**

**RefSeq status**

**Organism**

**Lineage**

**Also known as**

**Summary**

## Genomic regions, transcripts, and products

(plus) Go to [reference sequence details](#)

[Try our new Sequence Viewer](#)

## Genomic context

**chromosome: 3;**

[ 37017787 ▶

LOC645571 ▶

EPM2A1

## GeneRIFs: Gene References Into Function

[What's a GeneRIF?](#)

1. BRCA1 methylation correlated with age at diagnosis (P = .015) and 5-years disease free survival (P = .016) while hMLH1 methylation was more frequent in larger tumors (P = .002) and in presence of distant metastasis (P = .004).
2. Co-immunoprecipitation analyses revealed that MutLalpha, and also MSH2 and MSH6, components of the MutSalpa heterodimer, form complexes with Poleta in human cells.
3. MutLalpha (MLH1-PMS2), replication protein A (RPA), and HMGB1 have roles in 5'-directed mismatch repair
4. the present study reveals a possible function of MLH1 protein in protecting colon tumor cells from resistance acquisition by trichostatin A
5. Assessing high rate microsatellite instability and expressions of MLH1 could be used to distinguish benign and malignant insulinomas and to predict the outcome of patients.
6. New hMLH1 missense mutation in Muir-Torre syndrome associated with familial transmission of different gastrointestinal adenocarcinomas.
7. microsatellite instability was associated with hMLH1 in gastric carcinomas
8. microsatellite instability in BAT26 and BAT25 were predictive of MSH1 mutated Lynch syndrome
9. the expression of hMLH1 variant types 1 and 3 did not completely follow the same transcription

**Submit:** [New GeneRIF](#) [Correction](#)

# MLH1:LINKS TO SEQUENCE

**Genomic regions, transcripts, and products**

Go to [reference sequence details](#)

**NC\_000003.10**

5' [369983] 3' [37383246]

[NM\\_000249.2](#) [NP\\_000240.1](#) [CCDS2663.1](#)

■ - coding region ■ - untranslated region

**Links**

**mRNA LINKS**

- ▶ FASTA
- ▶ GENBANK

**Links**

**PROTEIN LINKS**

- ▶ FASTA
- ▶ GENPEPT
- ▶ Blink
- ▶ Conserved Domains

**chromosome: 3; Location: 3p21.3**

[36992791] [37383246]

LOC645571 LRRFIP2 GOLGA4 TCEA1P2

EPM2AIP1 MLH1

**Links**


**Links**

Order cDNA clone  
Books  
Conserved Domains  
Genome  
GEO Profiles  
HomoloGene  
Map Viewer  
Nucleotide  
OMIM  
Full text in PMC  
Probe  
Protein  
PubMed  
PubMed (GeneRIF)  
SNP  
SNP: Genotype  
SNP: GeneView  
Taxonomy  
UniSTS  
AceView  
CCDS  
Colon.html  
Evidence Viewer  
GDB  
GeneTests for MIM:  
120436  
HGMD  
HGNC  
HPRD  
KEGG  
MGC  
ModelMaker  
PharmGKB  
UniGene  
LinkOut

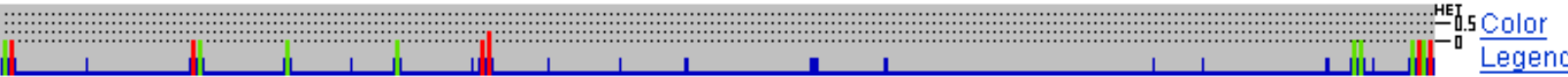
# GENEVIEW: HUMAN MLH1 VARIATIONS



view rs ☐ in gene region ☒ cSNP ☐ has frequency ☐ double hit ☐ haplotype tagged

gene model Contig mRNA protein mRNA orientation transcript snp count  
(contig mRNA transcript): [NT\\_022517](#) [NM\\_000249](#) [NP\\_000240](#) forward plus strand 14, coding







Contig position	dbSNP rs# cluster id	Heterozygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
36975081	<a href="#">rs1800143</a>	N.D.		Yes		synonymous	A	Glu [E]	3	13



Contig position	dbSNP rs# cluster id	Heterozygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
36975081	<a href="#">rs1800143</a>	N.D.		Yes		synonymous	A	Glu [E]	3	13
		N.D.		Yes		contig reference	G	Glu [E]	3	13
36975136	<a href="#">rs2020872</a>	0.025		Yes		nonsynonymous	G	Val [M]	1	32
		0.025		Yes		contig reference	A	Ile [I]	1	32
36982507	<a href="#">rs11541859</a>	N.D.		Yes		nonsynonymous	C	Gln [Q]	1	89
		N.D.		Yes		contig reference	G	Glu [E]	1	89

ATPase domain

37030074	<a href="#">rs1800146</a>	0.012				synonymous	T	Leu [L]	3	653
		0.012				contig reference	G	Leu [L]	3	653
37031995	<a href="#">rs1800147</a>	N.D.				synonymous	T	Gly [G]	3	706
		N.D.				contig reference	C	Gly [G]	3	706
37032029	<a href="#">rs2020873</a>	0.029		H		nonsynonymous	T	Tyr [Y]	1	718
		0.029		H		contig reference	C	His [H]	1	718
37032031	<a href="#">rs1800148</a>	N.D.				synonymous	T	His [H]	3	718
		N.D.				contig reference	C	His [H]	3	718
37032062	<a href="#">rs1800149</a>	N.D.				nonsynonymous	G	Val [M]	1	729
		N.D.				contig reference	C	Leu [L]	1	729



# 'TAKE HOME MESSAGE' ADVANTAGES OF DATA INTEGRATION

- More relevant *inter-related* information in one place
- Makes it easier to find additional relevant information related to your initial query
- Potentially find information *indirectly* linked, but *relevant* to your subject of interest
  - uncover *non-obvious* genetic features that explain phenotype or disease
- Easier to build a 'story' based on *multiple* pieces of biological evidence



# ENSEMBL - INTRODUCTION

- Ensembl is a joint scientific project between the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute.
- Ensembl's aim is to provide a centralized resource for geneticists, molecular biologists and other researchers studying the genomes the human and other vertebrates and model organisms.
- Ensembl now also contain genome data of several plant species.
- The Ensembl gene set is based on protein and mRNA evidence in UniProtKB and NCBI RefSeq databases.



# PAN-TAXONOMIC COMPARA

## Ensembl

*Anolis carolinensis*  
*Ciona savignyi*  
*Danio rerio*  
*Equus caballus*  
*Gallus gallus*  
*Homo sapiens*  
*Macaca mulatta*  
*Monodelphis domestica*  
*Mus musculus*  
*Ornithorhynchus anatinus*  
*Pan troglodytes*  
*Pongo pygmaeus*  
*Xenopus tropicalis*

## EnsemblMetazoa

*Anopheles gambiae*  
*Caenorhabditis elegans*  
*Drosophila melanogaster*

## EnsemblProtists

*Dictyostelium discoideum*  
*Plasmodium falciparum*  
*Plasmodium vivax*

## EnsemblBacteria

*B\_aphidicola\_Tokyo\_1998*  
*B\_burgdorferi\_DSM\_4680*  
*B\_subtilis*  
*E\_coli\_K12*  
*M\_tuberculosis\_H37Rv*  
*N\_meningitidis\_A*  
*P\_horikoshii*  
*S\_aureus\_N315*  
*S\_pneumoniae\_TIGR4*  
*S\_pyogenes\_SF370*  
*W\_pipientis\_wMel*

## EnsemblPlants

*Arabidopsis thaliana*  
*Oryza sativa*  
*Vitis vinifera*

## EnsemblFungi

*Aspergillus nidulans*  
*Neurospora crassa*  
*Saccharomyces cerevisiae*  
*Schizosaccharomyces pombe*

Browsing Genes & Genomes with

 Ensembl



# ENSEMBL PLANTS

- See talk “Browsing Genomic Information with Ensembl Plants”

