joanne@msl.ubc.ca

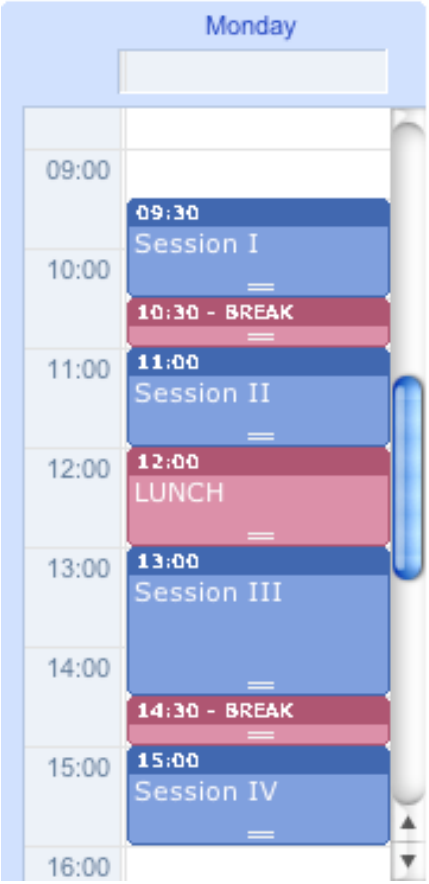
Laboratory Bioinformatics

Common tools, useful databases, and tricks of the trade for practical use in the laboratory.

bioteach.ubc.ca/bioinfo2009

Workshop Schedule

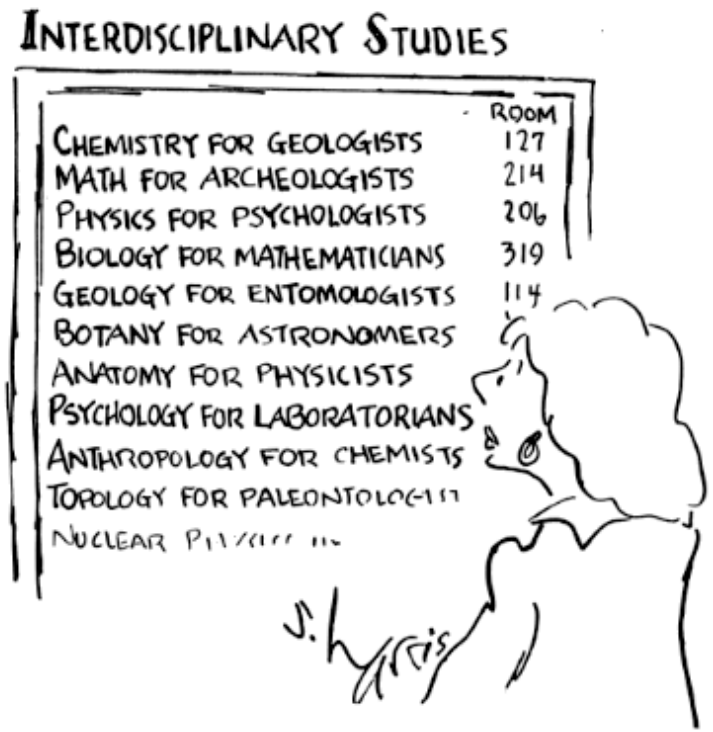
- Laptops, available here for your use 9am - 4:30pm
- wireless login



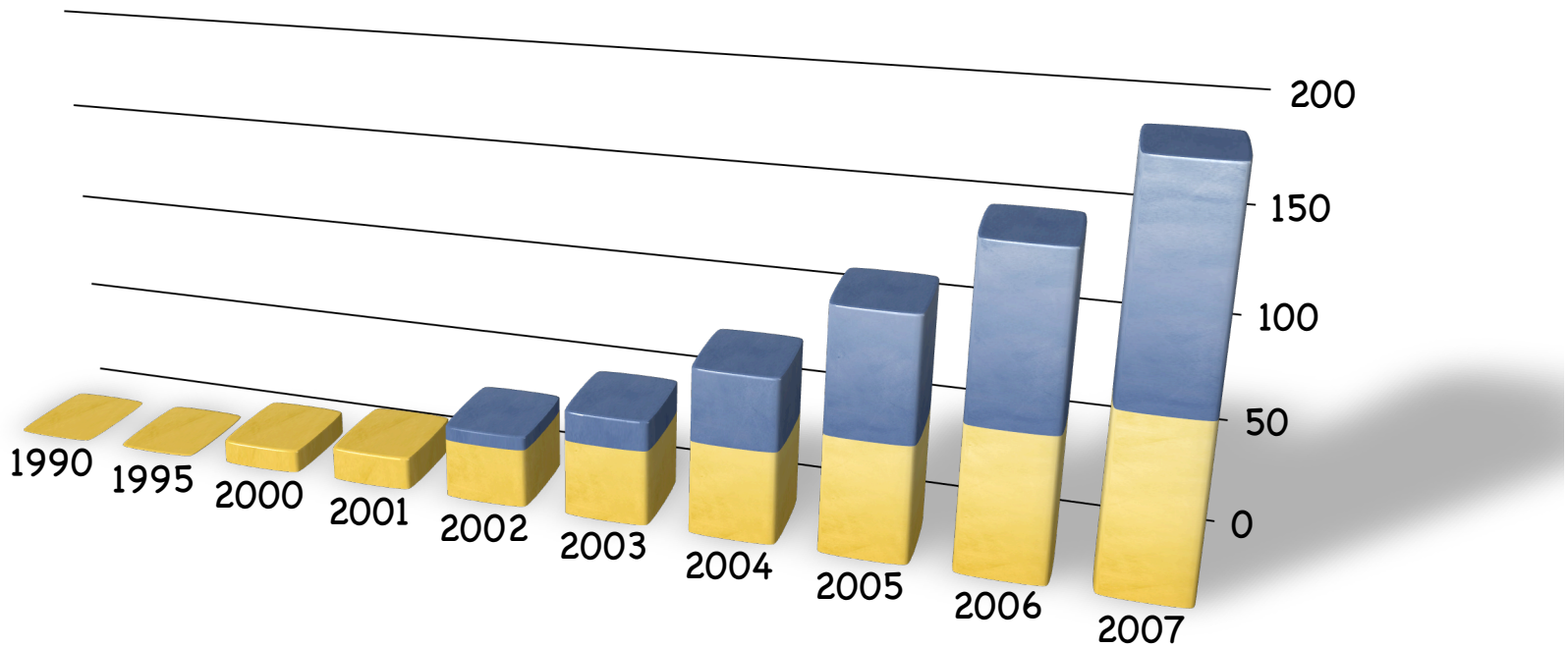
Today's Plan

- **Intro Activity**
- **Subject** - Public Resources at the NCBI
- **GUIDED TOUR** - Database Searching with Entrez
- **PRACTICAL EXERCISES** - Data Retrieval
- **TIPS & TRICKS** - PubMed, MyNCBI, Bookshelf...

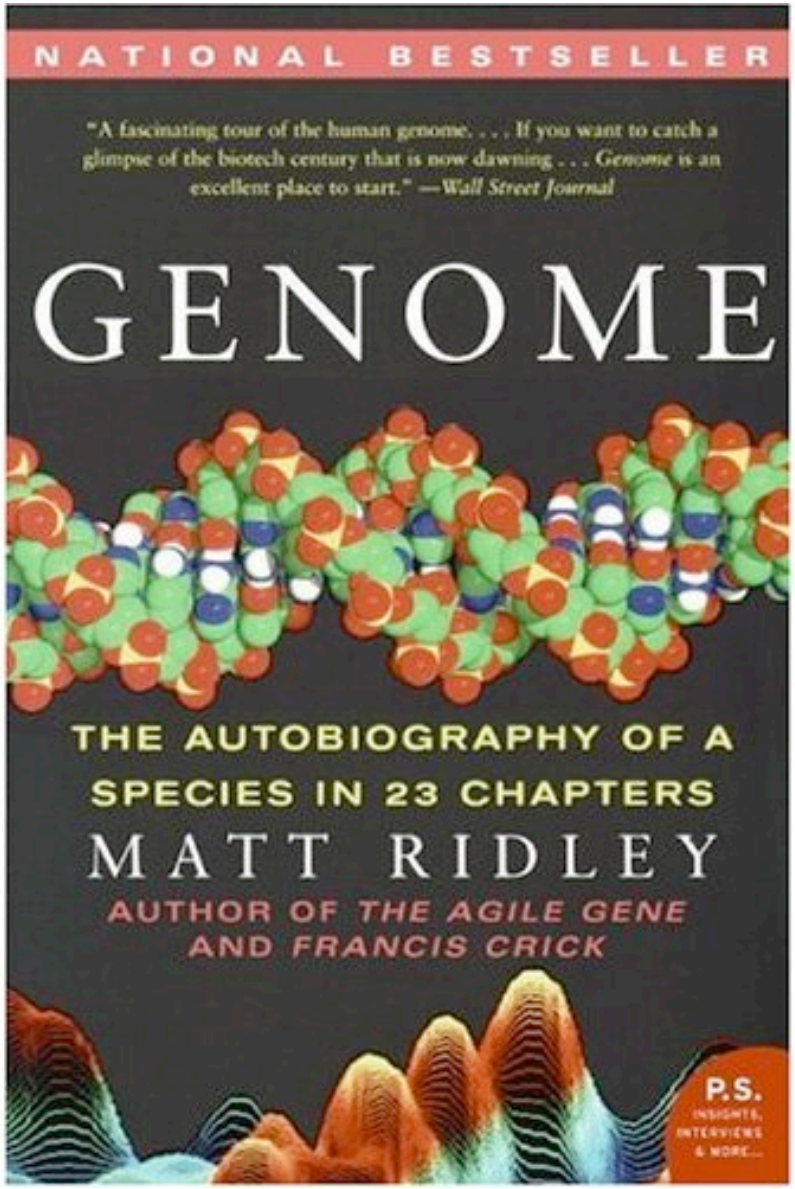
Bioinformatics for Biologists



Growth of GenBank



In 2005, International sequence databases exceed 100 gigabases


NATIONAL BESTSELLER
"A fascinating tour of the human genome. . . . If you want to catch a glimpse of the biotech century that is now dawning . . . Genome is an excellent place to start." —Wall Street Journal
GENOME
THE AUTOBIOGRAPHY OF A SPECIES IN 23 CHAPTERS
MATT RIDLEY
AUTHOR OF THE AGILE GENE AND FRANCIS CRICK
P.S.
INSIGHTS, INTERVIEWS & MORE...

Personalized Medicine?

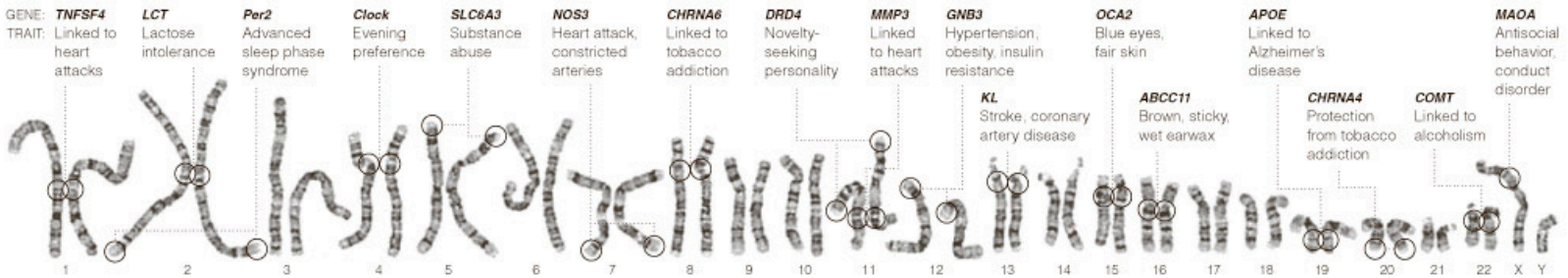
In the Genome Race, the Sequel Is Personal

Thor Swift for The New York Times

A team led by J. Craig Venter, above, has finished the first mapping of a full, or diploid, genome, made up of DNA inherited from both parents. The genome is Dr. Venter's own.

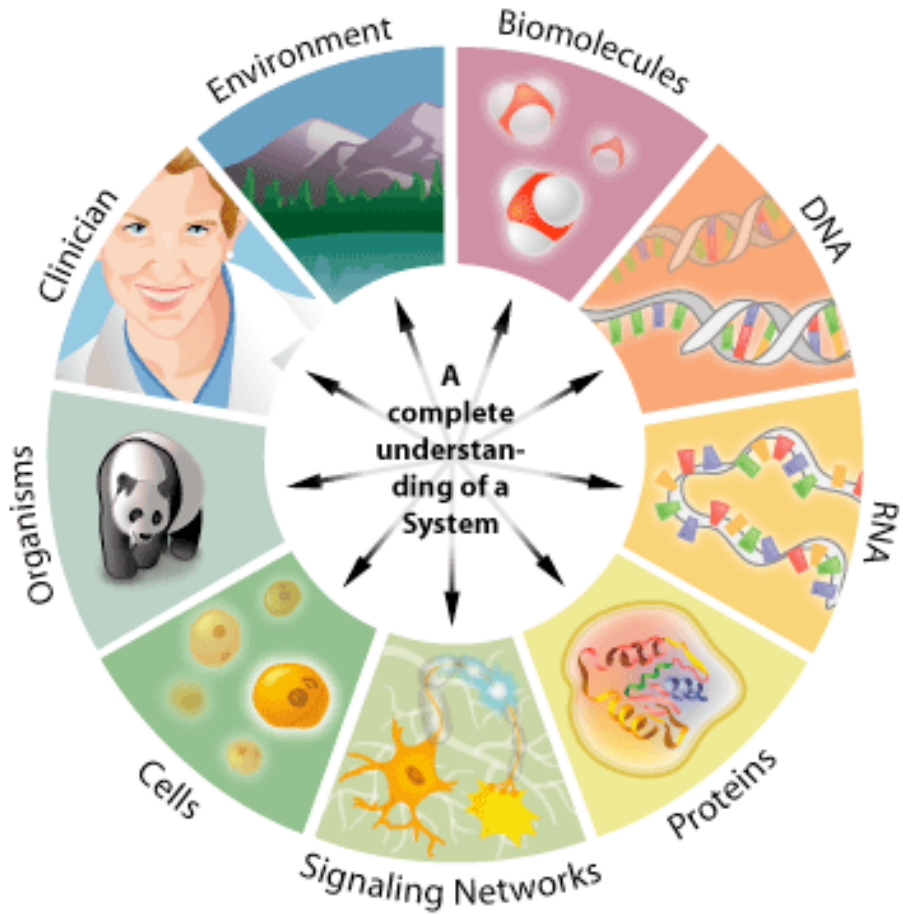
The New York Times — September 3, 2007

DECODING HIMSELF A team led by J. Craig Venter, above, has finished the first mapping of a full, or diploid, genome, made up of DNA inherited from both parents. The genome is Dr. Venter's own.

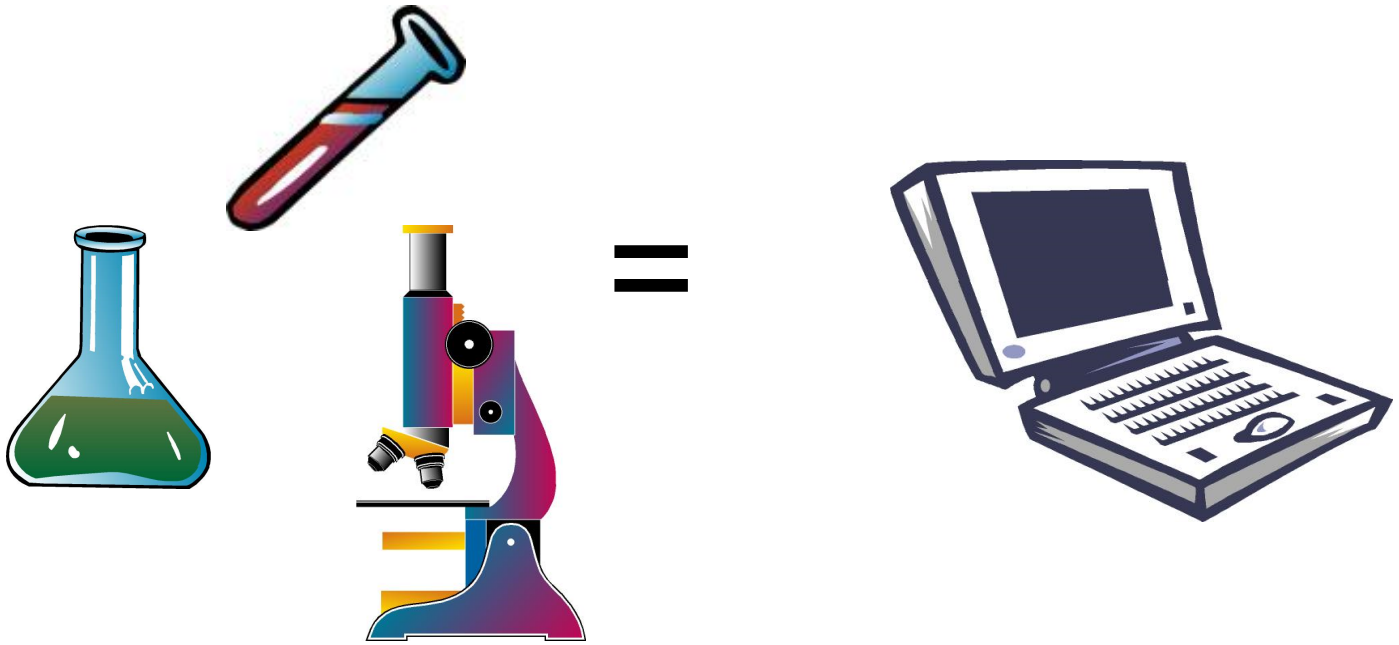


Sources: PLoS Biology; GeneCards — CHROMOSOMES — *Genes highlighted above are linked to specific traits but do not mean the conditions are inevitable.* — THE NEW YORK TIMES

What is Bioinformatics?

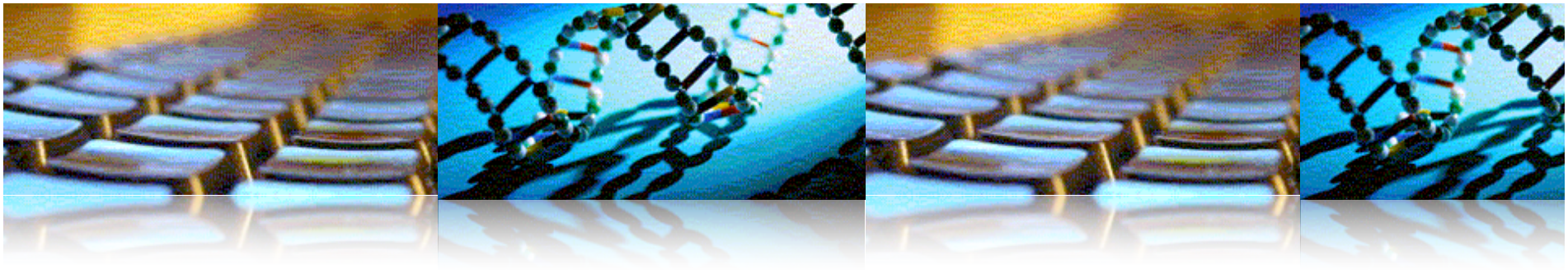


Laboratory Bioinformatics



What is Bioinformatics?

Goals & Priorities



Bioinformatics is an interdisciplinary research field that involves the integration of computers, software tools, and databases in an effort to address biological questions.

Genomics refers to the analysis of all of the genes and transcripts included within the genome. Proteomics, on the other hand, refers to the analysis of the complete set of proteins or proteome.

Bioinformatics Questions

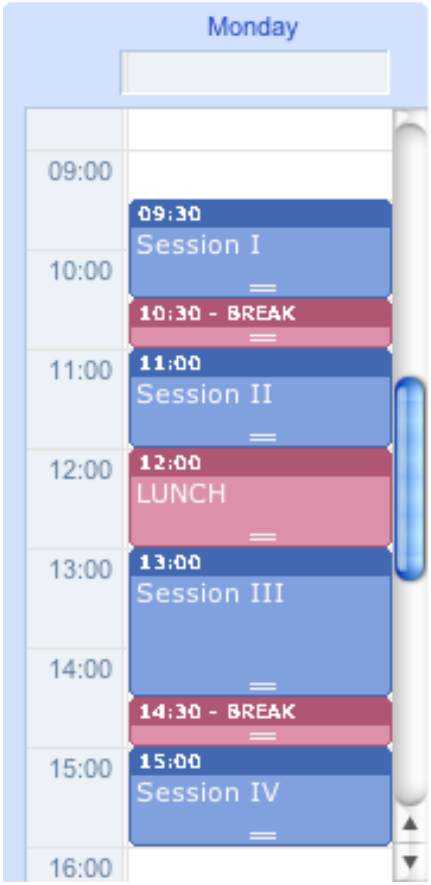
- What is encoded by the genome?
  - Links between genes, regulatory, and functional regions
- How is genome information expressed?
  - Function of genes and gene products (proteins)
  - Structure of proteins
- How can we interpret the information encoded in the genome?
  - Linking knowledge to the biological entities.
  - Systems biology approach
  - drugs, metabolites, ...
- How does the genome interact with its environment?

How do we best educate ourselves/others to take advantage of the latest 'omics research?

Overview of Topics*

- ✓ Day 1 - Public Database Resources NCBI
- ✓ Day 2 - BLAST, BLAST, more BLAST
- ✓ Day 3 - MSA, Genome Browsers, GEO

*additional topics can be scheduled as necessary



Summary

An article called, "What is Bioinformatics?" is available from the Science Creative Quarterly.
http://www.scq.ubc.ca/what-is-bioinformatics/

Sequence Databases

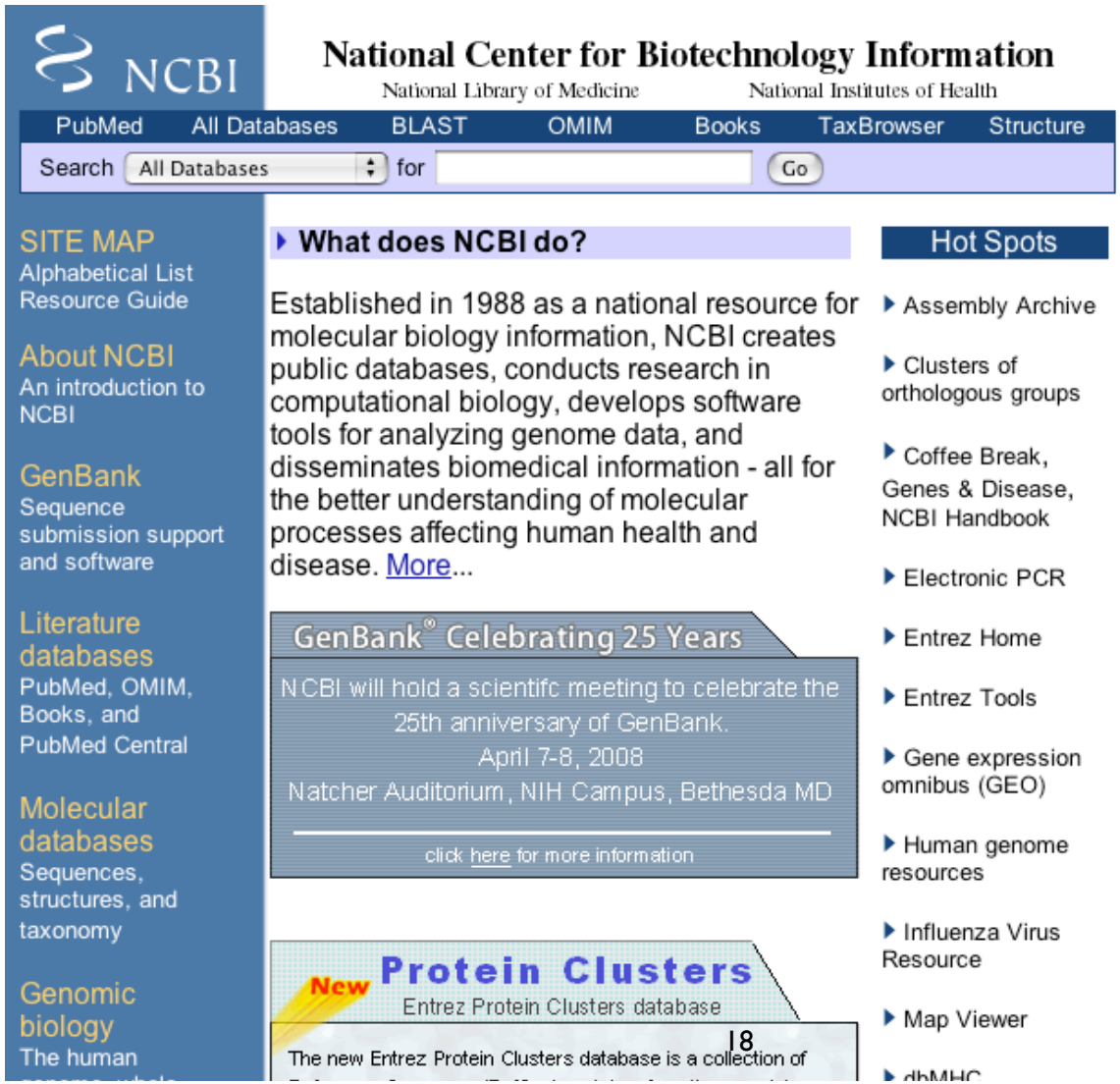
Public Resources at the NCBI

The National Center for Biotechnology Information

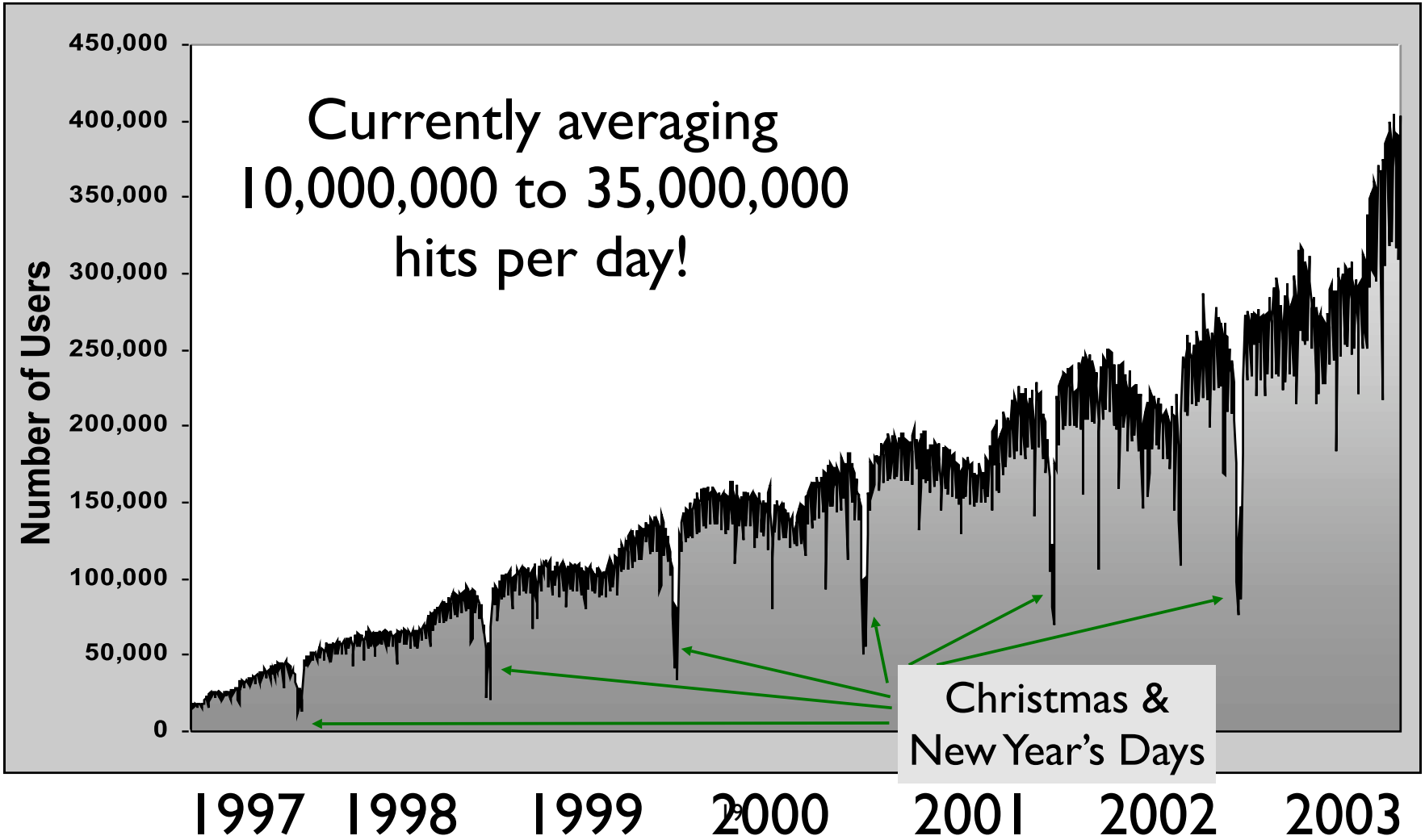
NCBI

- **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

www.ncbi.nlm.nih.gov



Number of Users and Hits Per Day



NCBI Site Map

Search NCBI

Resource Guide Complete resource listing and descriptions
Alphabetical List of major or commonly used resources
Entrez Database · Entrez Database subset (filtered query)

Databases

- Nucleotides
  - GenBank
  - RefSeq (Reference Sequences)
  - dbEST (Expressed Sequence Tags)
  - dbGSS (Genome Survey Sequences)
  - dbMHC (Major Histocompatibility Complex)
  - dbSNP (Single Nucleotide Polymorphisms)
  - dbSTS (Sequence Tagged Sites)
  - Probe
  - TPA (Third Party Annotation Database)
  - Trace Archive
  - UniSTS (Sequence Tagged Sites)
  - PopSet (Evolutionary Relatedness)
  - UniVec (Vector Sequences)
  - WGS (Whole Genome Shotgun Sequences)
- Proteins
  - RefSeq (Reference Sequences)
  - CDD (Conserved Domain Database)
  - Protein Clusters
- MMDB (Molecular Modeling DataBase)
  - 3D Domains
  - PubChem BioAssay
  - PubChem Compound
  - PubChem Substance
- Gene
  - UniGene
  - HomoloGene
  - CCDS (Consensus CoDing Sequence)
- GEO (Gene Expression Omnibus)
  - Entrez GEO Profiles
  - Entrez GEO DataSets
  - GENSAT
- TaxBrowser
  - Entrez Taxonomy

Literature Databases
- PubMed
- PubMed Central
- OMIM
- Books

Molecular Databases
- Nucleotide Sequences
- Protein Sequences
- Structures
- Genes
- Gene Expression
- Taxonomy

Genomes
- Entrez Genome
- Entrez Genome Project
- Map Viewer
- Cancer Chromosomes
- SKY/M-FISH & CGH Database
- dbGAP (Genotypes and Phenotypes)

Query all Entrez Databases

NCBI
- About
- Contact
- Research
- Software Engineering
- Site Search

Tools

Entrez Tools
- Query all Entrez Databases
- Entrez Data Model
- My NCBI (Cubby)
- LinkOut
- Batch Entrez Nucleotides
- Batch Entrez Proteins
- Citation Matcher
- Batch Citation Matcher
- Entrez Utilities

Data Analysis Tools
- Similarity Searching (BLAST)
- Nucleotide Sequence Analysis
- Protein Sequence Analysis and Proteomics
- Molecular Structure Analysis
- Genome Analysis
- Gene Expression

FTP
- Databases and Software

Programming Tools
- Entrez Utilities
- NCBI Toolbox
- Book: NCBI C++ Toolkit
- XML at NCBI
- Information Engineering Branch

Data Submissions

- BankIt (submissions and updates via WWW)
- Sequin (software for submissions and updates)
- dbEST (Expressed Sequence Tags)
- dbGSS (Genome Survey Sequences)
- dbSTS (Sequence Tagged Sites)
- Genomes
- Alignments
- WGS (Whole Genome Shotgun Sequences)
- TPA (Third Party Annotations)

Nucleotide Sequences
- GenBank (general submission tools)
- GenBank (specialized submissions)

Polymorphisms
- dbSNP (Single Nucleotide Polymorphisms)

Expression
- GEO (Gene Expression Omnibus)

Cytogenetic Data
- SKY/M-FISH & CGH Database

Trace Data
- Trace Archive

Education

News
- NCBI News
- What's New
- Announcement Mailing Lists
- RSS Feeds

Bookshelf
- Coffee Break
- Genes and Disease
- NCBI Handbook

Resources
- Tutorials, Courses, and more

The NCBI ftp site

NCBI **FTP site**

PubMed Entrez BLAST OMIM Books TaxBrowser Structure

Search All Databases for Go

NCBI

SITE MAP
Guide to NCBI resources

About NCBI
The science behind our resources. An introduction for researchers, educators and the public.

GenBank
sequence submission support and software

Molecular databases
sequences, structures and taxonomy

Literature databases
PubMed and OMIM

Genomic Biology

Major resources available by ftp (ftp.ncbi.nih.gov):

- BLAST Basic Local Alignment Search Tool

Download the BLAST database and stand-alone sequence comparison software.

- CDD Data

Download data from the Conserved Domain Database.

- CD-Tree

Download the protein domain hierarchy viewer and editor.

- Cn3D

Download the stand-alone software for viewing 3-dimensional structures.

- Data Repository

Download collections of contributed molecular biology

- dbGaP

Download open access Genotype and Phenotype data

- GenBank

Download the full release database, daily updates, or WGS files.

Note: there is a mirror site for GenBank files at Indiana University (bio-mirror.net/biomirror/genbank).

- 30,000 files per day
- 620 Gigabytes per day

NCBI Databases & Services

- GenBank largest sequence database
- Free public access to biomedical literature
  - PubMed free Medline
  - PubMed Central full text online access
- Entrez integrated molecular & literature databases
- BLAST highest volume sequence search service
- VAST structure similarity searches
- Software and Databases

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

What is GenBank?
NCBI's Primary Sequence Database

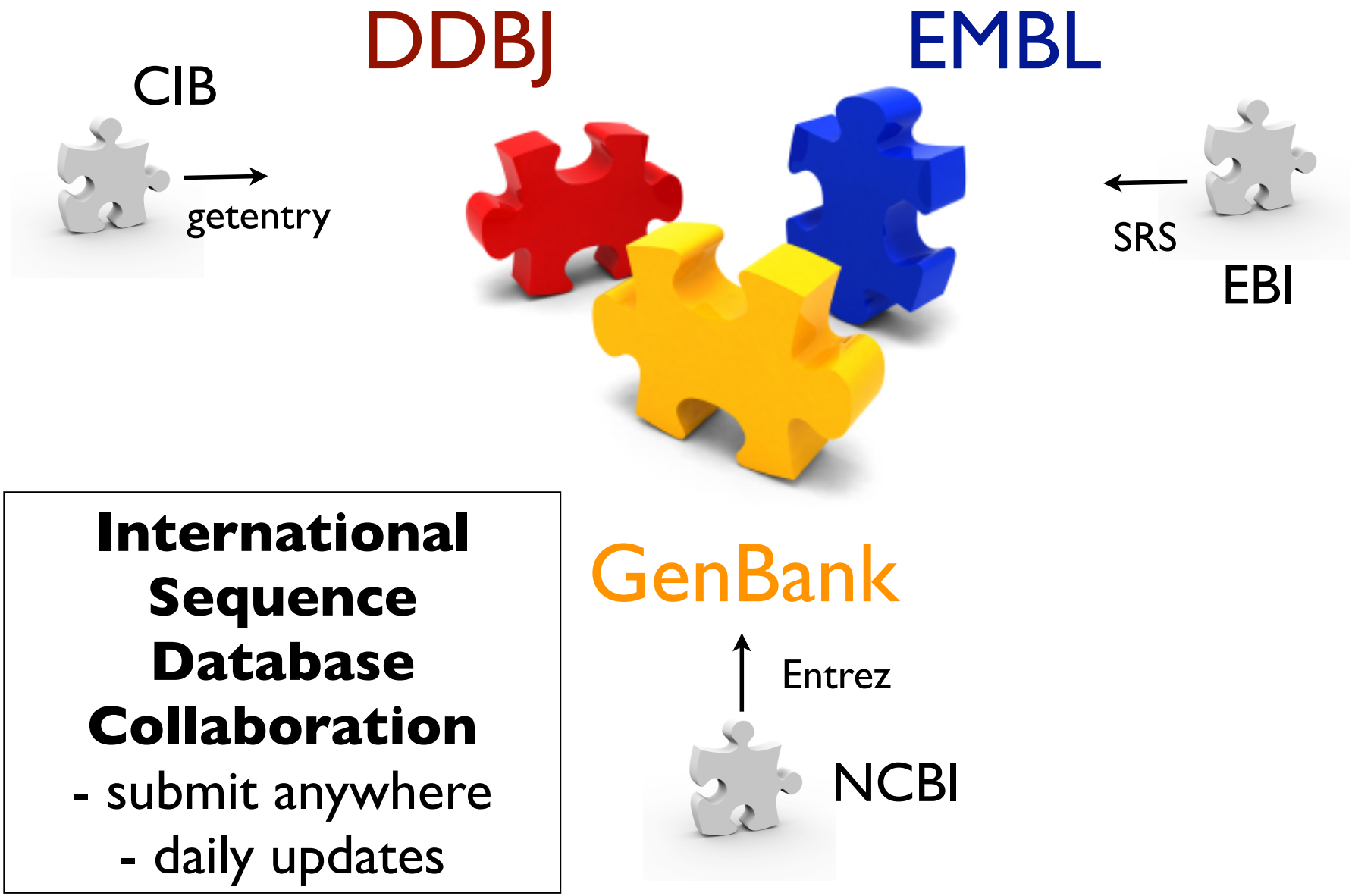
- Nucleotide only sequence database
- Archival in nature
- Historical
- Reflective of submitter point of view (subjective)
- Redundant

GenBank Data

✓ Direct submissions (traditional records)

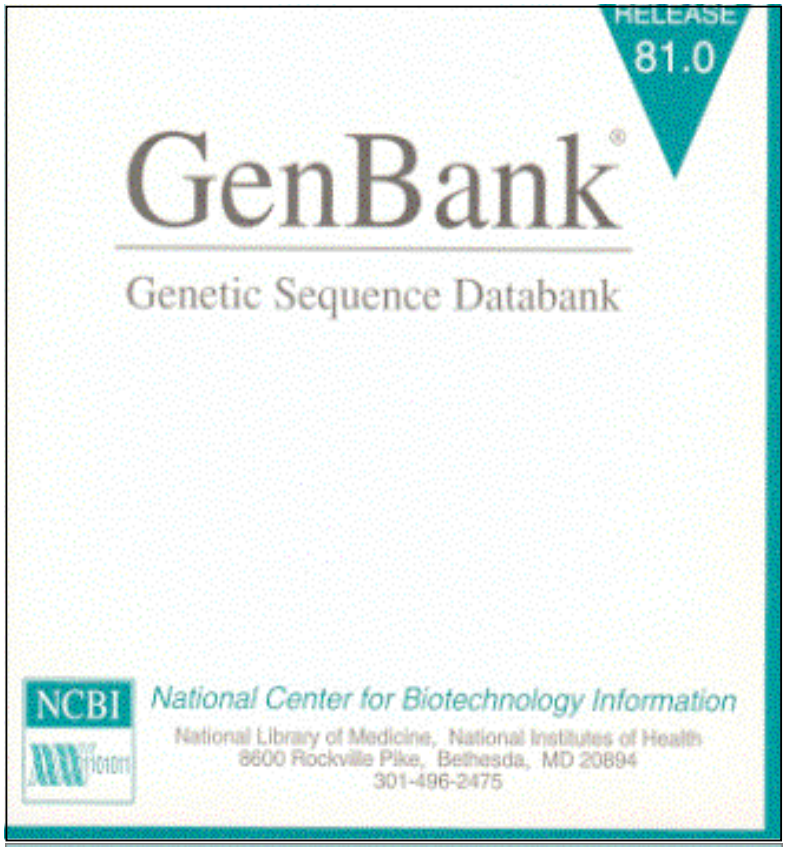
✓ Batch submissions (EST, GSS, STS)

✓ ftp accounts (genome data)



International Sequence Database Collaboration

- submit anywhere
- daily updates



GenBank: NCBI's Primary Sequence Database

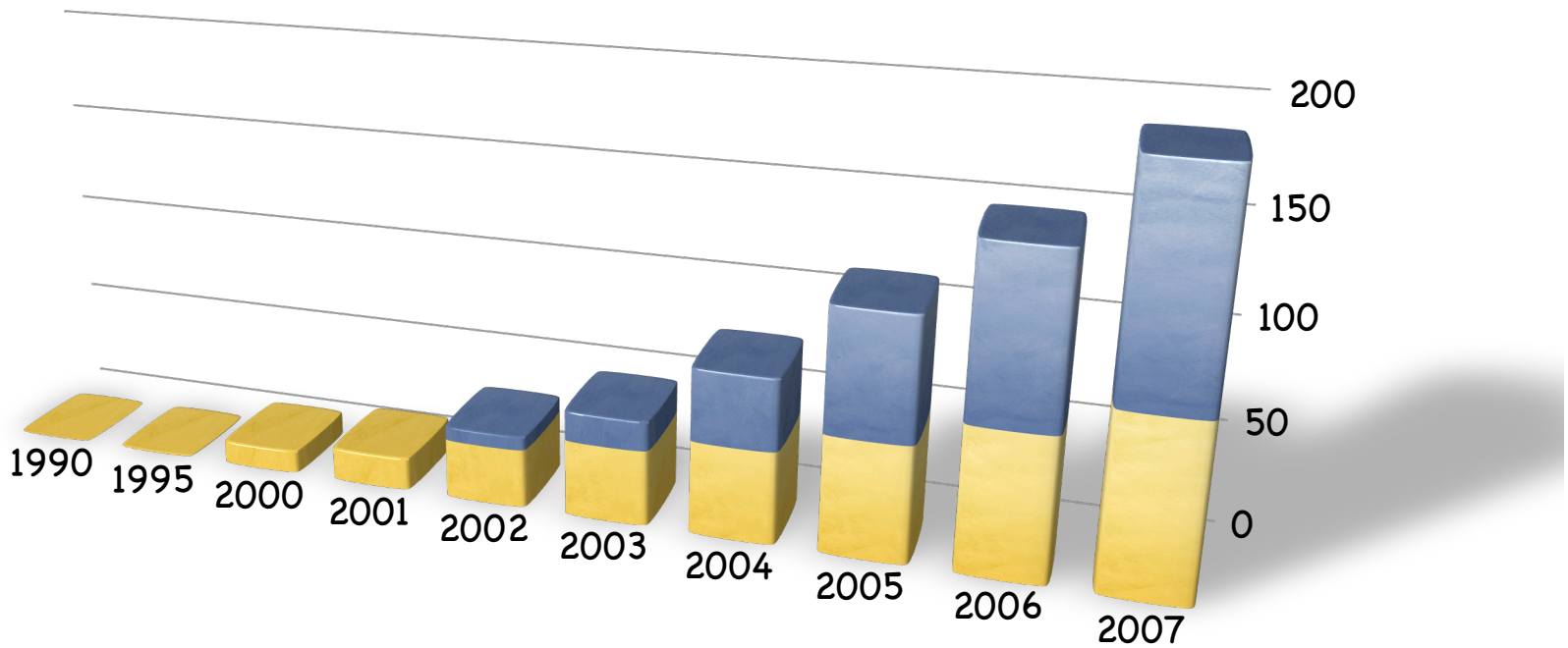
ftp://ftp.ncbi.nih.gov/genbank/

Release 169	Dec 2008
147,263,303	Records
240,491,402,946*	Total Bases

*includes WGS

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

Growth of GenBank



Current Release 169

Doubling time 12-14 months

GenBank WGS

Organization of GenBank

Records are divided into 18 Divisions.

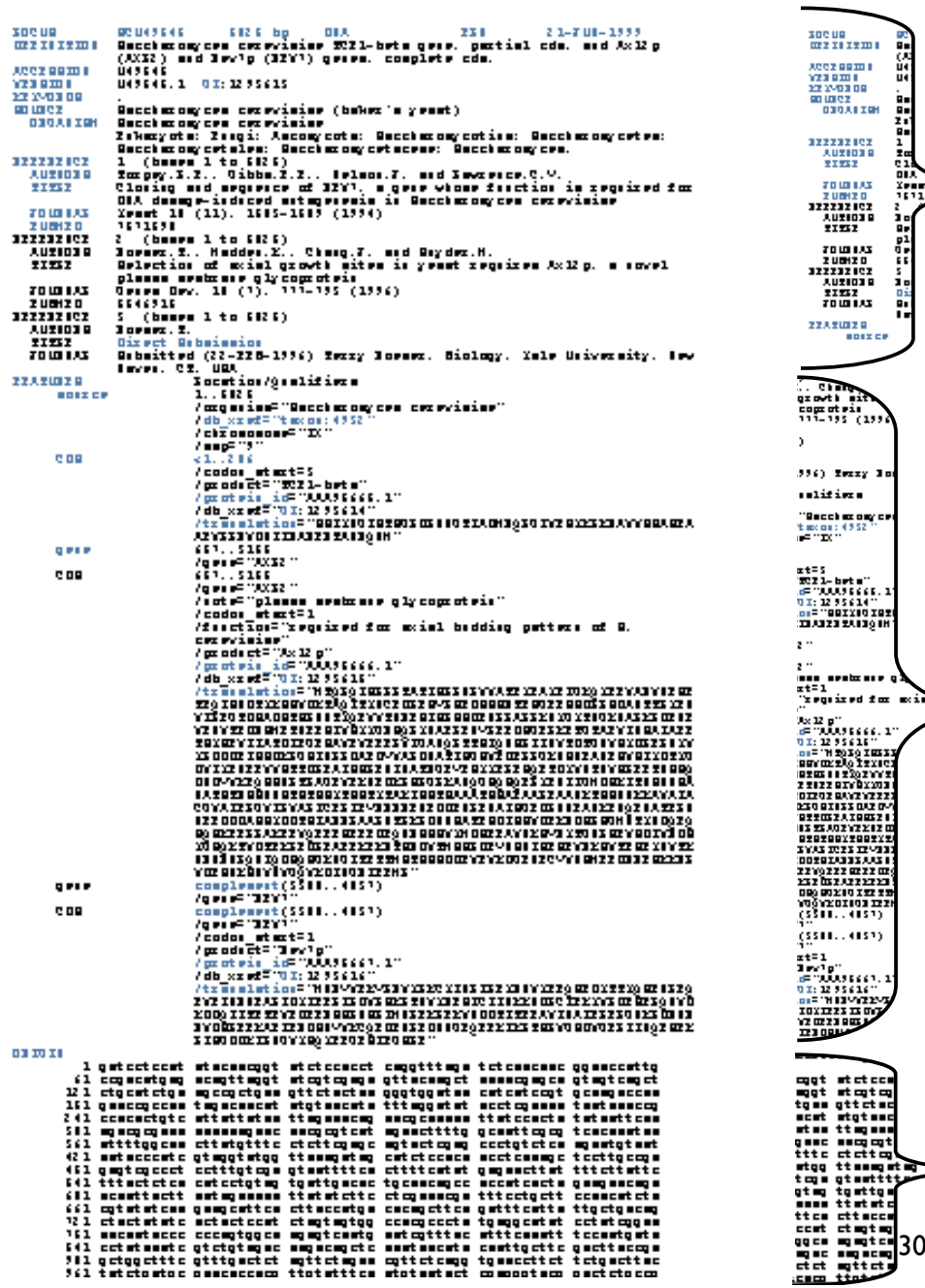
☑ Traditional:

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

☑ BULK Divisions:

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]



A Traditional GenBank Record

Header

Feature Table

Sequence

Traditional GenBank Record

```
LOCUS       HSHMLHI                 2503 bp    mRNA    linear   PRI 31-MAR-1994
DEFINITION  Human DNA mismatch repair (hmlh1) mRNA, complete cds.
ACCESSION   U07418
VERSION     U07418.1  GI:466461
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vert          teleostomi;
            Mammalia; Eutheria; Primates; Catarrhini
REFERENCE   1  (bases 1 to 2503)
  AUTHORS   Papadopoulos,N., Nicolaides,N.C., Wei
            Carter,K.C., Rosen,C.A., Haseltine,W                  .M.,
            Fraser,C.M., Adams,M.D., Venter,J.C                   .,
            Watson,P., Lynch,H.T., Peltomaki,P.
            Kinzler,K.W. and Vogelstein,B.
  TITLE     Mutation of a mutL homolog in heredi
  JOURNAL   Science 263 (5153), 1625-1629 (1994)
```

Accession
- Stable
- Reportable
- Universal

```
ACCESSION    U07418

VERSION      U07418.1   GI:466461
```

Version
- Tracks changes in sequence

GI number
- NCBI internal use

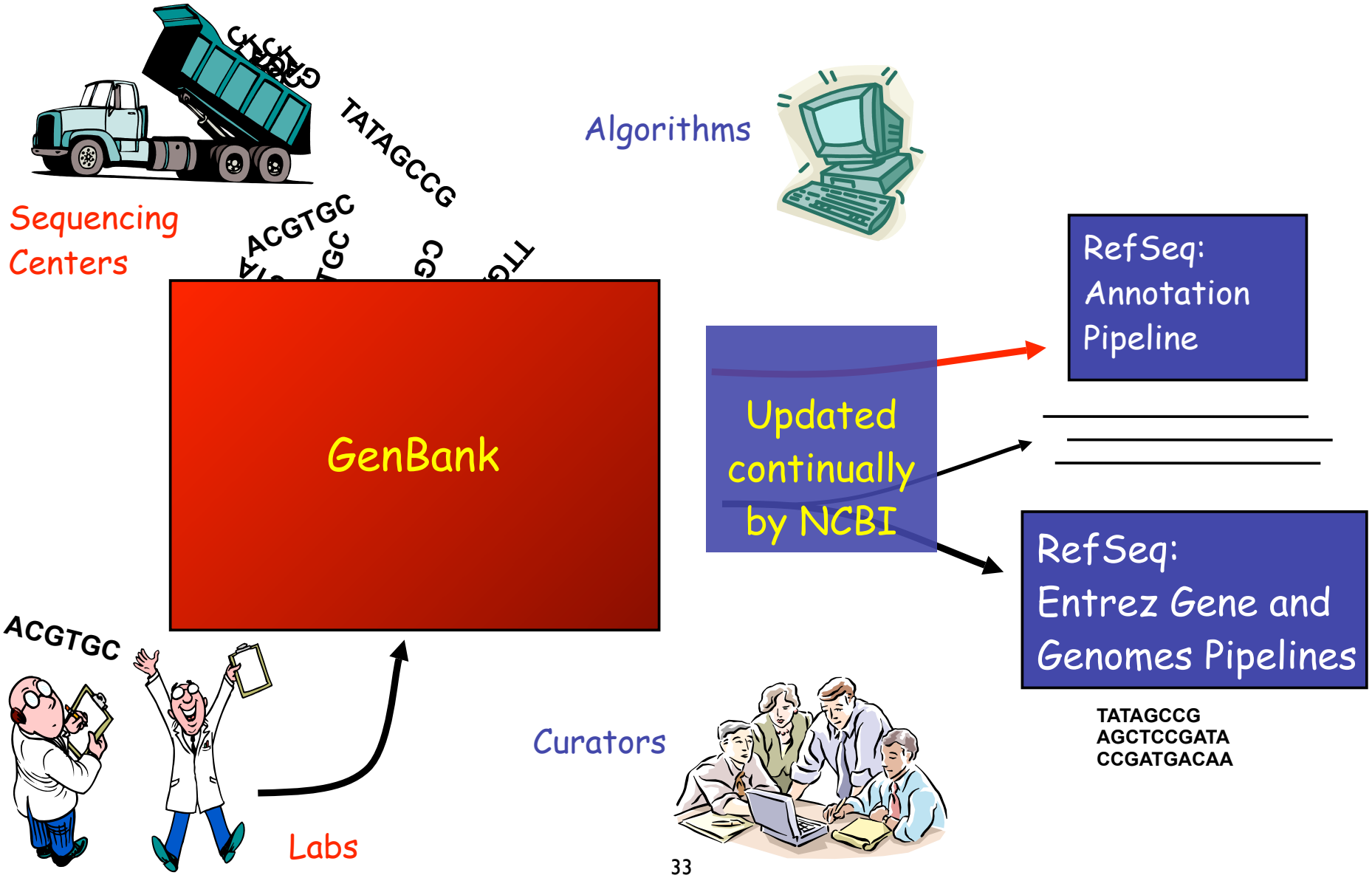
```
FEATURES             Location/Qualifiers
     source          1..2503
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="3"
                     /map="p21"
                     /tissue_type="gall bladder"
                     /dev_stage="adult"
     gene            1..2503
                     /gene="hmlh1"
     CDS             42..2312
                     /gene="hmlh1"
                     /function="DNA mismatch repair"
                     /note="human homolog of E. coli mutL gene product,
                     Swiss-Prot Accession Number P23367"
                     /codon_start=1
                     /protein_id="AAA17374.1"
                     /db_xref="GI:466462"
                     /translation="MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKS
                     TSIQVIVKEGGLKLIQIQDNGTGIRKEDLDIVCERFTTSKLQSFEDLASISTYGFRGE
                     ALASISHVAHVTITTKTADGKCAYRASYSDGKLKAPPKPCAGNQGTQITVEDLFYNIA
                     TRRKALKNPSEEYGKILEVVGRYSVHNAGISFSVKKQGETVADVRTLPNASTVDNIRS
                     VFGNAVSRELIEIGCEDKTLAFKMNGYISNANYSVKKCIFLLFINHRLVESTSLRKAI
                     ETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKL
                     LGSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQK
                     LDAFLQPLSKPLSSQPQAIVTEDKTDISSGRARQQDEEMLELPAPAEVAAKNQSLEGD
                     TTKGTSEMSEKRGPTSSNPRKRHREDSDVEMVEDDSRKEMTAACTPRRRIINLTSVLS
                     LQEEINEQGHEVLREMLHNHSFVGCVNPQWALAQHQTKLYLLNTTKLSEELFYQILIY
                     DFANFGVLRLSEPAPLFDLAMLALDSPESGWTEEDGPKEGLAEYIVEFLKKKAEMLAD
                     YFSLEIDEEGNLIGLPLLIDNYVPPLEGLPIFILRLATEVNWDEEKECFESLSKECAM
                     FYSIRKQYISEESTLSGQQSEVPGSIPNSWKWTVEHIVYKALRSHILPPKHFTEDGNI
                     LQLANLPDLYKVFERC"
```

well annotated

```
BASE COUNT      723 a    539 c    599 g    642 t
ORIGIN
        1 gttgaacatc tagacgtttc cttggctctt ctggcgccaa aatgtcgttc gtggcagggg
       61 ttattcggcg gctggacgag acagtggtga accgcatcgc ggcgggggaa gttatccagc
      121 ggccagctaa tgctatcaaa gagatgattg agaactgttt agatgcaaaa tccacaagta
      181 ttcaagtgat tgttaaagag ggaggcctga agttgattca gatccaagac aatggcaccg
      241 ggatcaggaa agaagatctg gatattgtat gtgaaaggtt cactactagt aaactgcagt
      301 cctttgagga tttagccagt atttctacct atggctttcg aggtgaggct ttggccagca
      361 taagccatgt ggctcatgtt actattacaa cgaaaacagc tgatggaaag tgtgcataca
      421 gagcaagtta ctcagatgga aaactgaaag cccctcctaa accatgtgct ggcaatcaag
      481 ggacccagat cacggtggag gacctttttt acaacatagc cacgaggaga aaagctttaa
      541 aaaatccaag tgaagaatat gggaaaattt tggaagttgt tggcaggtat tcagtacaca
      601 atgcaggcat tagtttctca gttaaaaaac aaggagagac agtagctgat gttaggacac
      661 tacccaatgc ctcaaccgtg gacaatattc gctccgtctt tggaaatgct gttagtcgag
      721 aactgataga aattggatgt gaggataaaa ccctagcctt caaaatgaat ggttacatat
      781 ccaatgcaaa ctactcagtg aagaagtgca tcttcttact cttcatcaac catcgtctgg
      841 tagaatcaac ttccttgaga aaagccatag aaacagtgta tgcagcctat ttgcccaaaa
      901 acacacaccc attcctgtac ctcagtttag aaatcagtcc ccagaatgtg gatgttaatg
      961 tgcaccccac aaagcatgaa gttcacttcc tgcacgagga gagcatcctg gagcgggtgc
     1021 agcagcacat cgagagcaag ctcctgggct ccaattcctc caggatgtac ttcacccaga
     1081 ctttgctacc aggacttgct ggcccctctg gggagatggt taaatccaca acaagtctga
     1141 cctcgtcttc tacttctgga agtagtgata aggtctatgc ccaccagatg gttcgtacag
     1201 attcccggga acagaagctt gatgcatttc tgcagcctct gagcaaaccc ctgtccagtc
     1261 agccccaggc cattgtcaca gaggataaga cagatatttc tagtggcagg gctaggcagc
     1321 aagatgagga gatgcttgaa ctcccagccc ctgctgaagt ggctgccaaa aatcagagct
     1381 tggaggggga tacaacaaag gggacttcag aaatgtcaga gaagagagga cctacttcca
     1441 gcaaccccag aaagagacat cgggaagatt ctgatgtgga aatggtggaa gatgattccc
     1501 gaaaggaaat gactgcagct tgtacccccc ggagaaggat cattaacctc actagtgttt
     1561 tgagtctcca ggaagaaatt aatgagcagg gacatgaggt tctccgggag atgttgcata
     1621 accactcctt cgtgggctgt gtgaatcctc agtgggcctt ggcacagcat caaaccaagt
     1681 tataccttct caacaccacc aagcttagtg aagaactgtt ctaccagata ctcatttatg
     1741 attttgccaa ttttggtgtt ctcaggttat cggagccagc accgctcttt gaccttgcca
     1801 tgcttgcctt agatagtcca gagagtggct ggacagagga agatggtccc aaagaaggac
     1861 ttgctgaata cattgttgag tttctgaaga agaaggctga gatgcttgca gactatttct
     1921 ctttggaaat tgatgaggaa gggaacctga ttggattacc ccttctgatt gacaactatg
     1981 tgcccccttt ggagggactg cctatcttca ttcttcgact agccactgag gtgaattggg
     2041 acgaagaaaa ggaatgtttt gaaagcctca gtaaagaatg cgctatgttc tattccatcc
     2101 ggaagcagta catatctgag gagtcgaccc tctcaggcca gcagagtgaa gtgcctggct
     2161 ccattccaaa ctcctggaag tggactgtgg aacacattgt ctataaagcc ttgcgctcac
     2221 acattctgcc tcctaaacat tttcacagaag atggaaatat cctgcagctt gctaacctgc
     2281 ctgatctata caaagtcttt gagaggtgtt aaatatggtt atttatgcac tgtgggatgt
     2341 gttcttcttt ctctgtattc cgatacaaag tgttgtatca aagtgtgata tacaaagtgt
     2401 accaacataa gtgttggtag cacttaagac ttatacttgc cttctgatag tattcctttt
     2461 tacacagtgg attgattata aataaataga tgtgtcttaa cat
//
```

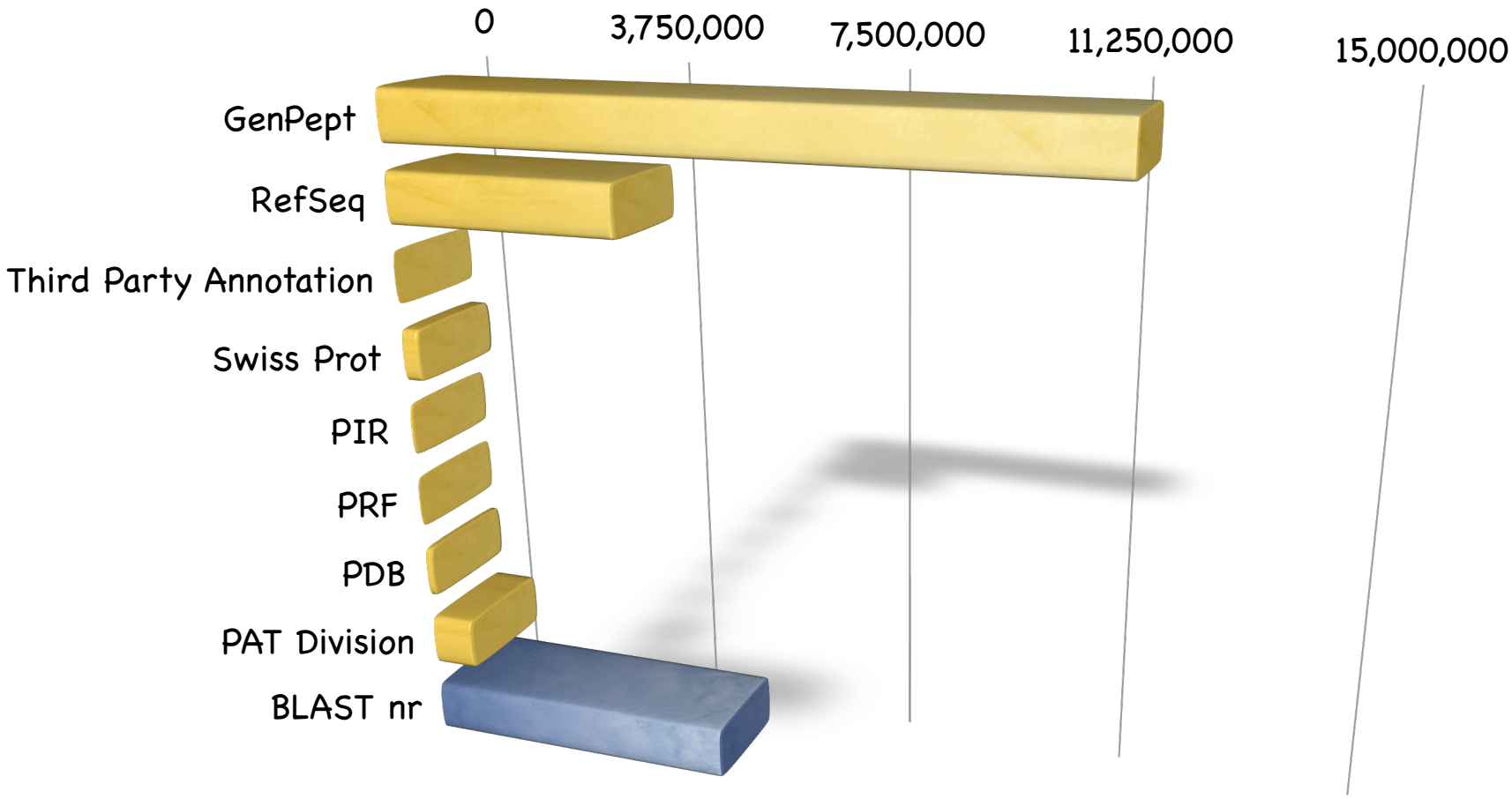
the sequence is the data

Primary vs. Derivative Databases



Derivative Databases

Entrez Protein



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="MLH1"
     CDS             22..2292
                     /gene="MLH1"
                     /note="homo
                     Number P14242), S. cerevisiae MLH1 (GenBank Accession
                     Number U07187), E. col MUTL (Swiss-Prot Accession Number
                     P23367), Salmonella t imurium MUTL (Swiss-Prot Accession
                     Number P14161) and S ococcus pneumoniae (Swiss-Prot
                     Accession Number P141 "
                     /codon_start=1
                     /product="DNA mismatc epair protein homolog"
                     /protein_id="AAC50285
                     /db_xref="GI:463989"
                     /translation="MSFVAGV LDETVVNRIAAGEVIQRPANAIKEMIENCLDAKS
                     TSIQVIVKEGGLKLIQIQDNG RKEDLDIVCERFTTSKLQSFEDLASISTYGFRGE
                     ALASISHVAHVTITTKTADGK RASYSDGKLKAPPKPCAGNQGTQITVEDLFYNIA
                     TRRKALKNPSEEYGKILEVVGRYSVHNAGISFSVKKQGETVADVRTLPNASTVDNIRS
```

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

RefSeq

- The goal is to provide the best single collection of sequence information for each major organism.
  - chromosome, organelle, or plasmid
  - linked by residue to transcripts, translated proteins, and mature peptide product.
  - known and predicted
  - reviewed
  - best view from available data

RefSeq

- DDBJ/EMBL/GenBank remains the primary sequence archive while RefSeq is a summary and synthesis based on that essential primary data.

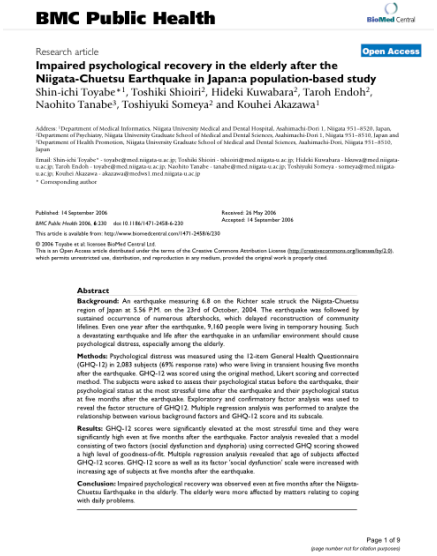
SCIENTIFIC AMERICAN

JUNE 1989
$2.95

M.I.T.'s R for building a new industrial America.
Strange fossil creatures from an ancient sea.
How a crystal lattice channels electrons and positrons.

***An earthquake** waiting to happen? This sharply folded terrain in the foothills of the Andes could conceal a dangerous fault.*

vs



BMC Public Health

BioMed Central

Research article

Open Access

Impaired psychological recovery in the elderly after the Niigata-Chuetsu Earthquake in Japan:a population-based study

Shin-ichi Toyabe*[1], Toshiki Shioiri[2], Hideki Kuwabara[2], Taroh Endoh[2], Naohito Tanabe[3], Toshiyuki Someya[2] and Kouhei Akazawa[1]

Address: [1]Department of Medical Informatics, Niigata University Medical and Dental Hospital, Asahimachi-Dori 1, Niigata 951–8520, Japan, [2]Department of Psychiatry, Niigata University Graduate School of Medical and Dental Sciences, Asahimachi-Dori 1, Niigata 951–8510, Japan and [3]Department of Health Promotion, Niigata University Graduate School of Medical and Dental Sciences, Asahimachi-Dori, Niigata 951–8510, Japan

Email: Shin-ichi Toyabe* - toyabe@med.niigata-u.ac.jp; Toshiki Shioiri - tshioiri@med.niigata-u.ac.jp; Hideki Kuwabara - hkuwa@med.niigata-u.ac.jp; Taroh Endoh - toyabe@med.niigata-u.ac.jp; Naohito Tanabe - tanabe@med.niigata-u.ac.jp; Toshiyuki Someya - someya@med.niigata-u.ac.jp; Kouhei Akazawa - akazawa@medwa1.med.niigata-u.ac.jp

* Corresponding author

Published: 14 September 2006
BMC Public Health 2006, 6:230 doi:10.1186/1471-2458-6-230

Received: 26 May 2006
Accepted: 14 September 2006

This article is available from: http://www.biomedcentral.com/1471-2458/6/230

© 2006 Toyabe et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: An earthquake measuring 6.8 on the Richter scale struck the Niigata-Chuetsu region of Japan at 5.56 P.M. on the 23rd of October, 2004. The earthquake was followed by sustained occurrence of numerous aftershocks, which delayed reconstruction of community lifelines. Even one year after the earthquake, 9,160 people were living in temporary housing. Such a devastating earthquake and life after the earthquake in an unfamiliar environment should cause psychological distress, especially among the elderly.

Methods: Psychological distress was measured using the 12-item General Health Questionnaire (GHQ-12) in 2,083 subjects (69% response rate) who were living in transient housing five months after the earthquake. GHQ-12 was scored using the original method, Likert scoring and corrected method. The subjects were asked to assess their psychological status before the earthquake, their psychological status at the most stressful time after the earthquake and their psychological status at five months after the earthquake. Exploratory and confirmatory factor analysis was used to reveal the factor structure of GHQ12. Multiple regression analysis was performed to analyze the relationship between various background factors and GHQ-12 score and its subscale.

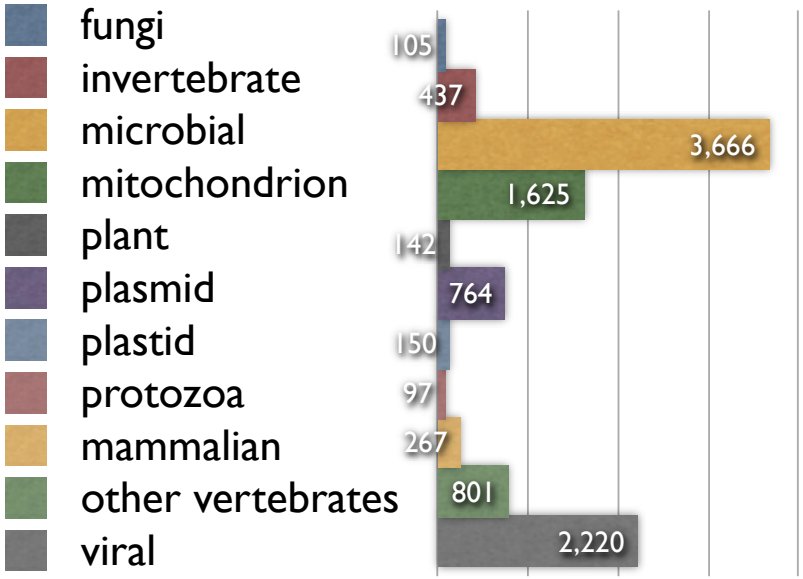
Results: GHQ-12 scores were significantly elevated at the most stressful time and they were significantly high even at five months after the earthquake. Factor analysis revealed that a model consisting of two factors (social dysfunction and dysphoria) using corrected GHQ scoring showed a high level of goodness-of-fit. Multiple regression analysis revealed that age of subjects affected GHQ-12 scores. GHQ-12 score as well as its factor 'social dysfunction' scale were increased with increasing age of subjects at five months after the earthquake.

Conclusion: Impaired psychological recovery was observed even at five months after the Niigata-Chuetsu Earthquake in the elderly. The elderly were more affected by matters relating to coping with daily problems.

Page 1 of 9
(page number not for citation purposes)

RefSeq

- includes species ranging from viral to microbial to eukaryotic, 7000+ species
- organisms with complete & incomplete genomes
- does not include all species
  - ✓ common research organisms, mouse, human, yeast, fly, plants, ...



*refseq release 33

RefSeq Accession Numbers*

- prefix indicates the molecule type.

Molecule Type	Accession Prefix
protein	NP_; XP_; ZP_; AP_; YP_;
rna	NM_; NR_; XM_; XR_
genomic	NC_; NG_; NT_; NW_; NZ_; NS_; AC_

*The underscore ("_") is the primary distinguishing feature of a RefSeq accession

RefSeq Accession Numbers

- mRNAs and Proteins

NM_123456	Curated mRNA
NP_123456	Curated Protein
NR_123456	Curated nc RNA
XM_123456	Predicted mRNA
XP_123456	Predicted Protein
XR_123456	Predicted nc RNA

- Genomic Records

NG_123456 Reference Genomic Sequence

- Chromosome

NC_123455 Microbial replicons, organelle, genomes, human chromosomes

- Assemblies

NT_123456	Contig
NW_123456	WGS Supercontig

Other NCBI Databases

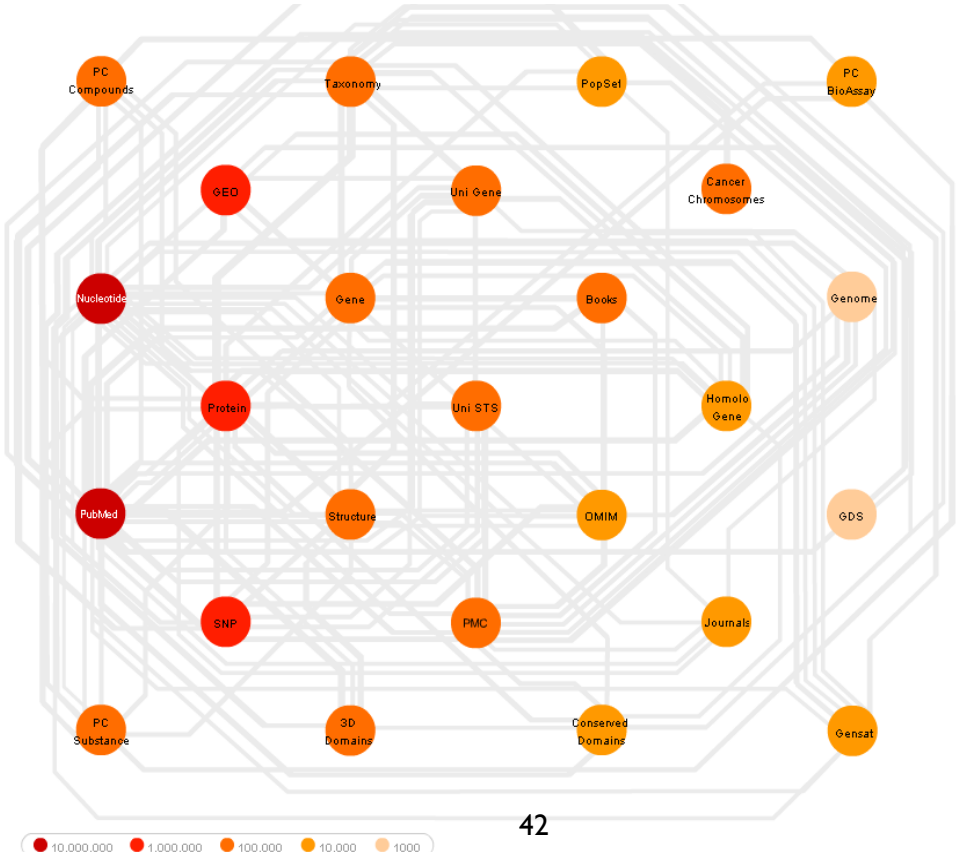
Structure:	imported structures (PDB)	Cn3D viewer, NCBI curation
CDD:	conserved domain database	Protein families (COGs and KOGs); Single domains (PFAM, SMART, CD)
dbSNP:	nucleotide polymorphism	variation data
Gene:	gene records	unified searchable database of genes, replaces locuslink
HomoloGene:	homologs	neighboring function for Gene

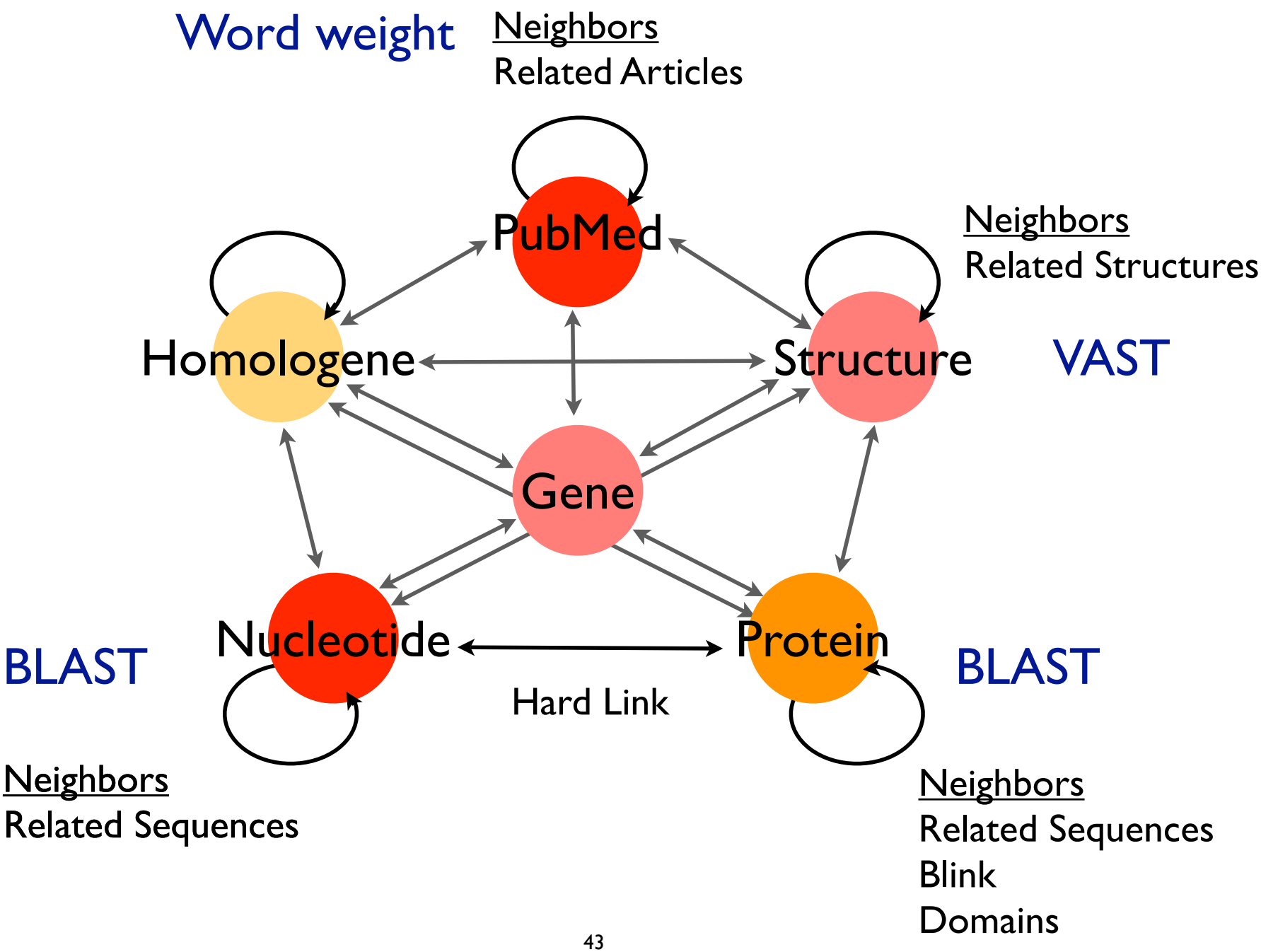
http://www.ncbi.nih.gov/Database/datamodel

NCBI Model of Entrez Databases Search Entrez for Go

The diagram shows the Entrez databases and the connections between them. Each database is represented by a colored circle, where the color indicates the approximate number of records in the database. Mouse over a circle to see which databases are linked to the one selected, and how many links exist between those databases.

This diagram requires Flash for viewing.




Word weight
Neighbors
Related Articles
PubMed
Neighbors
Related Structures
Homologene
Structure
VAST
Gene
BLAST
Nucleotide
Protein
BLAST
Hard Link
Neighbors
Related Sequences
Neighbors
Related Sequences
Blink
Domains

Neighbors in Entrez

1: rs709932 [Homo sapiens]CGAP-GAI, ILLUMINA, ILLUMINA, ILLUMINA, ILLUMINA, LEE, TSC-CSHL

Links

SNP

1: GDS596 record | GPL96 211298_s_at [Homo sapiens]

158 samples Profile Neighbors, Sequence Neighbors, Links

Annotation: ALB: albumin DKFZp779N1935, PRO0883, PRO0903, PRO1341

Reporter: AF116645

Large-scale analysis of the human

GEO

1: MLH1

Order cDNA clone, Links

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

Other Aliases: COCA2, FCC2, HNPCC, HNPCC2, MGC5172, hMLH1

Other Designations: DNA mismatch repair protein Mlh1; MutL protein homolog 1

Location: 3p21.3

Gene

1: Plotz G, Welsch C, Giron-Monzon L, Friedhoff P, Albrecht M, Piiper A, Raedle J.

Related Articles, Links

PubMed

1: NP_000240. Reports MutL protein homo...[gi:4557757]

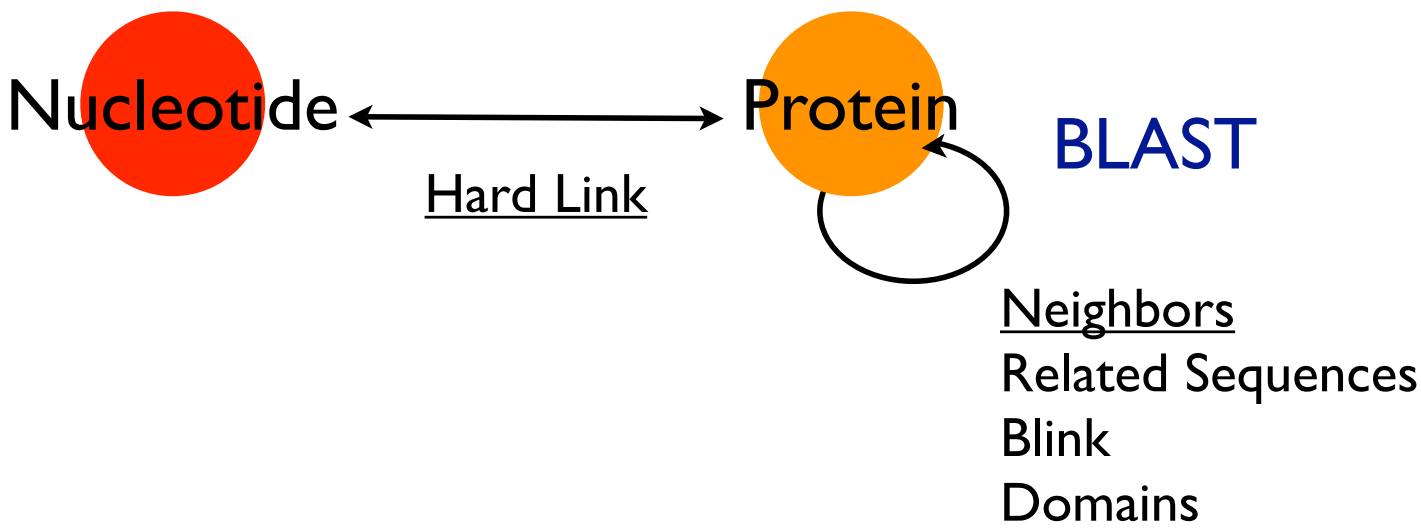
BLink, Conserved Domains, Links

Comment Features Sequence

```
LOCUS       NP_000240                 756 aa            linear   PRI 22-APR-2007
DEFINITION  MutL protein homolog 1 [Homo sapiens].
ACCESSION   NP_000240
VERSION     NP_000240.1  GI:4557757
DBSOURCE    REFSEQ: accession NM_000249.2
```

Protein

Entrez - Linking Data



Blink & Domains

Neighbors: BLAST Link
pre-computed BLAST

1: NP_000240. Reports MutL protein homo...[gi:4557757]

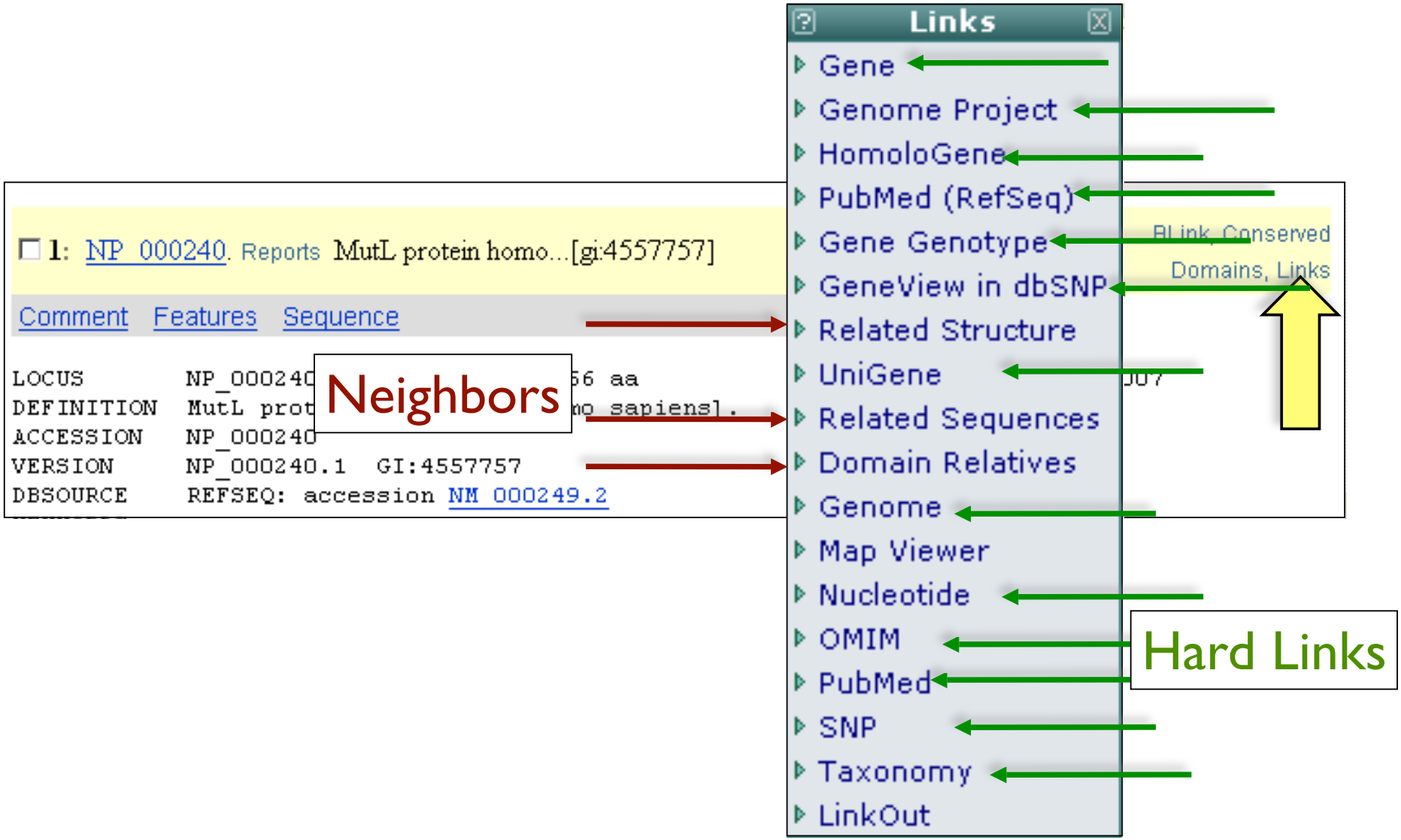
BLink, Conserved Domains, Links

Comment Features Sequence

Neighbors:
pre-computed CDD search

```
LOCUS       NP_000240                                    -APR-2007
DEFINITION  MutL protein homolog 1 [
ACCESSION   NP_000240
VERSION     NP_000240.1  GI:4557757
DBSOURCE    REFSEQ: accession NM_000249.2
```

Links


Links
Gene
Genome Project
HomoloGene
PubMed (RefSeq)
Gene Genotype
GeneView in dbSNP
Related Structure
UniGene
Related Sequences
Domain Relatives
Genome
Map Viewer
Nucleotide
OMIM
PubMed
SNP
Taxonomy
LinkOut
1: NP_000240. Reports MutL protein homo...[gi:4557757]
Comment Features Sequence
LOCUS NP_000240
DEFINITION MutL prot
ACCESSION NP_000240
VERSION NP_000240.1 GI:4557757
DBSOURCE REFSEQ: accession NM_000249.2
Neighbors
Hard Links
BLink, Conserved Domains, Links

Sequence Databases

GUIDED TOUR: Retrieving Data

Laboratory Bioinformatics Scenario:

You've just read about some interesting genes and now you want to find out more...

transactions
www.biochemsoctrans.org

Humanizing mismatch repair in yeast: towards effective identification of hereditary non-polyposis colorectal cancer alleles

P.M.R. Aldred and R.H. Borts[1]
Department of Genetics, University of Leicester, Adrian Building, University Road, Leicester LE1 7RH, U.K.

Abstract

The correction of replication errors is an essential component of genetic stability. This is clearly demonstrated in humans by the observation that mutations in mismatch repair genes lead to HNPCC (hereditary non-polyposis colorectal cancer). This disease accounts for as many as 2–3% of colon cancers. Of these, most of them are in the two central components of mismatch repair, *MLH1* (mutL homologue 1) and *MSH2* (mutS homologue 2). *MLH1* and *MSH2* function as a complex with two other genes *PMS2* and *MSH6*. Mismatch repair genes, and the mechanism that ensures that incorrectly paired bases are removed, are conserved from prokaryotes to human. Thus yeast can serve as a model organism for analysing mutations/polymorphisms found in human mismatch repair genes for their effect on post-replicative repair. To date, this has predominantly been accomplished by making the analogous mutations in yeast genes. However, this approach is only useful for the most highly conserved regions. Here, we discuss some of the benefits and technical difficulties involved in expressing human genes in yeast. Modelling human mismatch repair in yeast will allow the assessment of any functional effect of novel polymorphisms found in patients diagnosed with colon cancers.

Mismatch repair

The mismatch repair system serves to correct errors that occur during DNA replication. These errors can take the form of misincorporated nucleotides that result in mispaired bases or insertion/deletion loops that can result from replication slippage at polynucleotide tracts [1,2]. The mismatch repair process and therefore an increase in mutation rate or 'mutator' phenotype. As yMlh1p and yMsh2p are involved in the correction of multiple types of mismatch, deletion or mutation of these genes has a greater effect on mutation rate than the equivalent disruption of yMsh6p, which is involved in only one form of mismatch repair (Figure 2).

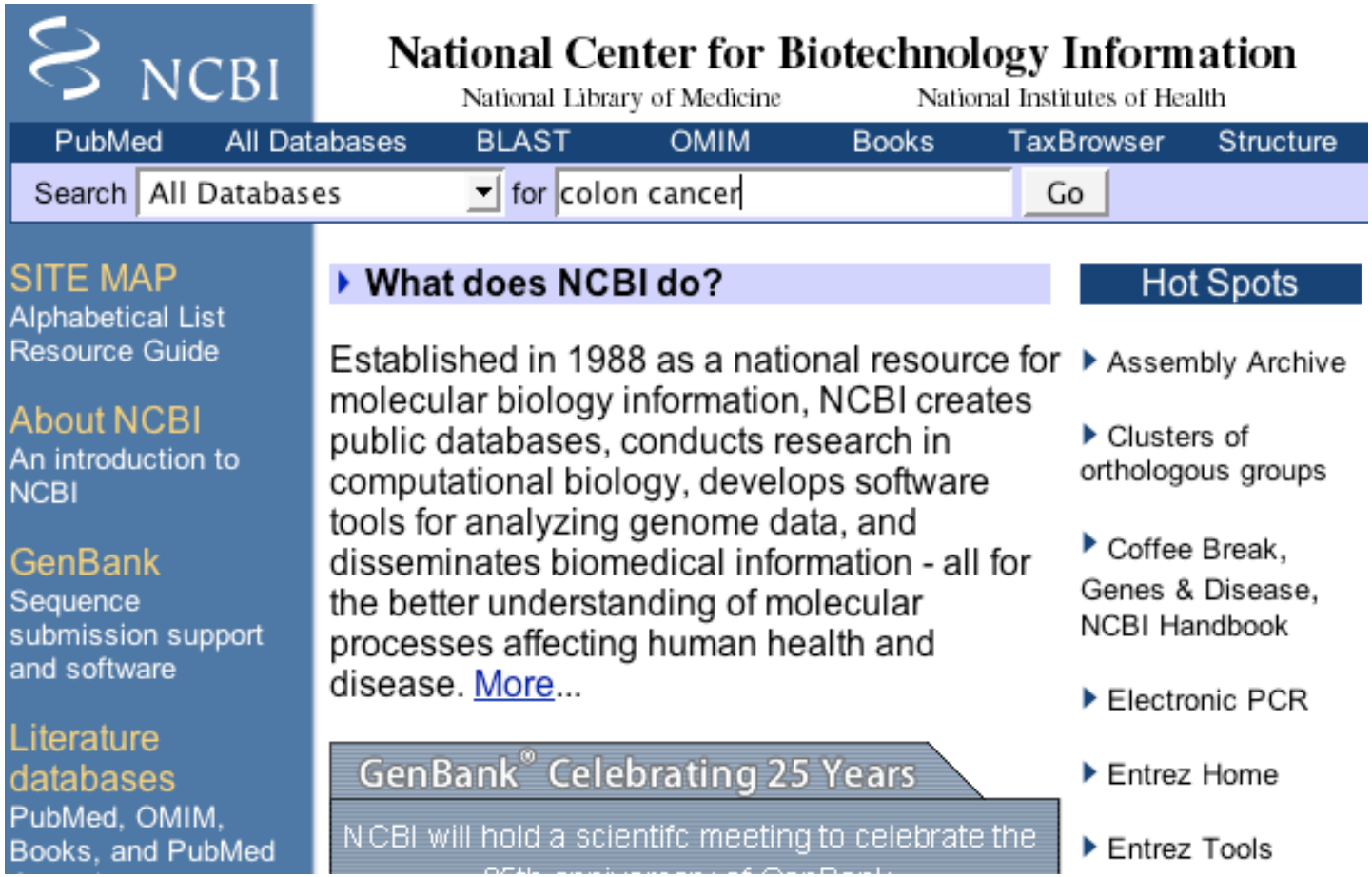
Database searching with Entrez

- **Scenario Summary:** Let's find out more about the genes involved in colon cancer

✓ Using limits and field restriction to find human MutL homolog - MLH1

✓ Linking and neighboring with MLH1

Start with a search for "colon cancer"



NCBI

Entrez, The Life Sciences Search Engine

HOME | SEARCH | SITE MAP | PubMed | All Databases | Human Genome | GenBank | Map Viewer | BLAST

Search across databases colon cancer GO Clear Help

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

58219	**PubMed:** biomedical literature citations and abstracts	894	**Books:** online books
7197	**PubMed Central:** free, full text journal articles	374	**OMIM:** online Mendelian Inheritance in Man
7	**Site Search:** NCBI web and FTP sites	none	**OMIA:** online Mendelian Inheritance in Animals

19529	**CoreNucleotide:** Core subset of nucleotide sequence records	2	**dbGaP:** genotype and phenotype
1156	**EST:** Expressed Sequence Tag records	160	**UniGene:** gene-oriented clusters of transcript sequences
none	**GSS:** Genome Survey Sequence records	6	**CDD:** conserved protein domain database
940	**Protein:** sequence database	19	**3D Domains:** domains from Entrez Structure
6	**Genome:** whole genome sequences	34	**UniSTS:** markers and mapping data
2	**Structure:** three-dimensional macromolecular structures	2	**PopSet:** population study data sets
none	**Taxonomy:** organisms in GenBank	109008	**GEO Profiles:** expression and molecular abundance profiles
none	**SNP:** single nucleotide polymorphism	83	**GEO DataSets:** experimental sets of GEO data
493	**Gene:** gene-centered information	123	**Cancer Chromosomes:** cytogenetic databases
20	**HomoloGene:** eukaryotic homology groups	4	**PubChem BioAssay:** bioactivity screens of chemical substances

Human Disease Genes

NCBI

MIM *120436
Description
Cloning
Gene Function
Biochemical Features
Gene Structure
Mapping
Molecular Genetics
Animal Model
Allelic Variants
- View List

References
Contributors
Creation Date
Edit History

- Clinical Synopsis
- Gene map

Entrez Gene
N Nomenclature
R RefSeq
G GenBank
P Protein
U UniGene

LinkOut
HNPCC
HGVS
HGMD
GAD

OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University

My NCBI
[Sign In] [Reg

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | OMIM

Search OMIM for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Detailed Show 20 Send to

*120436 GeneTests, Links

MutL, E. COLI, HOMOLOG OF, 1; MLH1

Gene map locus 3p21.3

TEXT

DESCRIPTION

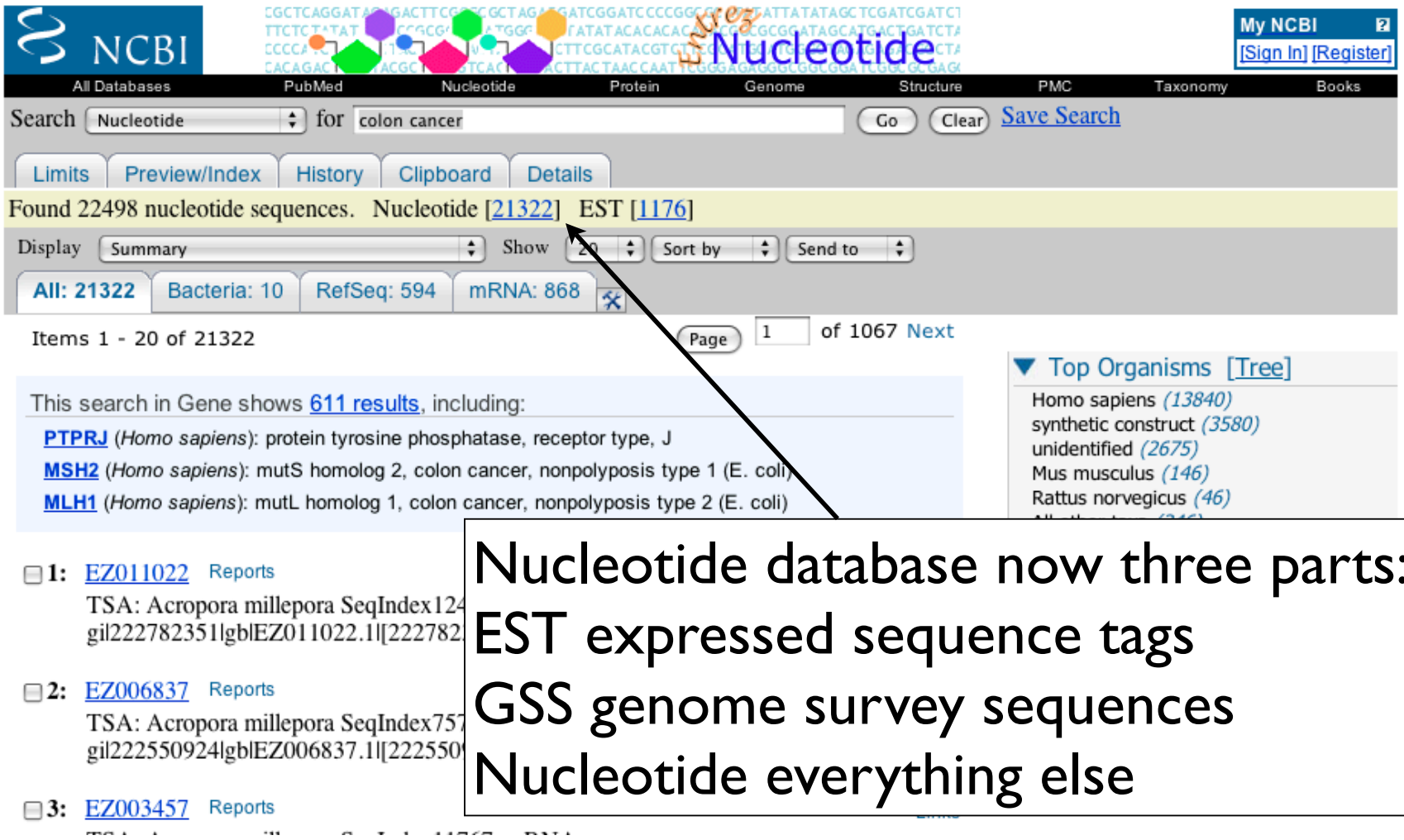
MLH is homologous to the E. coli MutL gene and is involved in DNA mismatch repair. Heterozygous mutations in the MLH1 gene result in hereditary nonpolyposis colorectal cancer-2 (HNPCC2; 609310) (Papadopoulos et al., 1994).

CLONING

After human homologs of the mutS gene of bacteria and yeast were found to have mutations responsible for hereditary nonpolyposis colorectal cancer (HNPCC1; 120435), Papadopoulos et al. (1994) searched for other human mismatch repair (MMR) genes. A survey of EST databases derived from random cDNA clones revealed 3 additional human MMR genes, all related to the bacterial mutL gene. One of these genes was MLH1. The other 2 genes had a slightly greater similarity to the yeast mutL homolog PMS1 and were therefore denoted PMS1 (600258) and PMS2 (600259), respectively.

Genuardi et al. (1998) characterized the normal alternative splicing of the MLH1 gene and reported a number of splice variants that exist in various tissue types. They observed splice variants lacking exons 6/9, 9, 9/10, 9/10/11, 10/11, 12, 16, and 17. The level of

Search Nucleotide



Nucleotide database now three parts:
EST expressed sequence tags
GSS genome survey sequences
Nucleotide everything else

Advanced Search Options

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy

Search Nucleotide for colon cancer Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Tabs

For 22498 nucleotide sequences. Nucleotide [21322] EST [1176]

Dis Summary Show 20 Sort by Send to

A 1322 | Bacteria: 10 | RefSeq: 594 | mRNA: 868

1 - 20 of 21322 Page 1 of 1067 Next

This search in Gene shows 611 results, including:

PTPRJ (*Homo sapiens*): protein tyrosine phosphatase, receptor type, J

MSH2 (*Homo sapiens*): mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)

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

1: EZ011022 Reports Links
TSA: Acropora millepora SeqIndex12410, mRNA sequence
gi|222782351|gb|EZ011022.1|[222782351]

2: EZ006837 Reports Links
TSA: Acropora millepora SeqIndex7572, mRNA sequence
gi|222550924|gb|EZ006837.1|[222550924]

3: EZ003457 Reports Links
TSA: Acropora millepora SeqIndex11767, mRNA sequence
gi|222547544|gb|EZ003457.1|[222547544]

Top Organisms [Tree]
Homo sapiens *(13840)*
synthetic construct *(3580)*
unidentified *(2675)*
Mus musculus *(146)*
Rattus norvegicus *(46)*
All other taxa *(246)*
More...

Recent Activity
Tur
Your browsing activity is empty

NCBI

Entrez Nucleotide

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC

Search CoreNucleotide for colon cancer AND nonpolyposis Go Clear

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI (Cubby)

Related resources
BLAST

Reference sequence project

Search for Genes

Submit to GenBank

Search for full length cDNAs

Field: **Title**

- Use All Fields pull-down menu to specify a field.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available here.

Limited to:

Fields

Title

EC/RN Number
Feature key
Filter
Gene Name
Genome Project
Issue
Journal
Keyword
Modification Date
Organism
Page Number
Primary Accession
Properties
Protein Name
Publication Date
SeqID String
Sequence Length
Substance Name
Text Word
Title

draft TPA patents

Gene Location: Any

Only from: Any

Write to the Help Desk
NCBI | NLM | NIH

colon cancer[Title] AND nonpolyposis[Title]

NCBI

Entrez Nucleotide

My NCBI
Welcome joannealisonfox. [Sign Out]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search CoreNucleotide for colon cancer AND nonpolyposis Go Clear

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI (Cubby)

Related resources
BLAST

Reference sequence project

Search for Genes

Submit to GenBank

Search for full length cDNAs

Field: **Title**

- Use All Fields pull-down menu to specify a field.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available here.

Limited to:

Fields
Title

Exclude
STSs | working draft | TPA | patents

Molecule:
mRNA

Gene Location:
Any

Segmented Sequences:
Any

Only from:
RefSeq

Published in the last:
Any Date

Modified in the last:
Any Date

Write to the Help Desk

colon cancer[Title] AND nonpolyposis[Title] AND biomol_mrna[Properties] AND srcdb_refseq[Properties]

Advanced Search Options

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy

Search Nucleotide for colon cancer Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Tabs

Found 22498 nucl de sequences. Nucleotide [21322] EST [1176]

Display Summary Show 20 Sort by Send to

All: 21322 | Ba a: 10 | RefSeq: 594 | mRNA: 868

Items 1 - 20 of 22 Page 1 of 1067 Next

This search in Gene shows 611 results, including:

PTPRJ (*Homo sapiens*): protein tyrosine phosphatase, receptor type, J

MSH2 (*Homo sapiens*): mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)

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

1: EZ011022 Reports Links
TSA: Acropora millepora SeqIndex12410, mRNA sequence
gi|222782351|gb|EZ011022.1|[222782351]

2: EZ006837 Reports Links
TSA: Acropora millepora SeqIndex7572, mRNA sequence
gi|222550924|gb|EZ006837.1|[222550924]

3: EZ003457 Reports Links
TSA: Acropora millepora SeqIndex11767, mRNA sequence
gi|222547544|gb|EZ003457.1|[222547544]

Top Organisms [Tree]
Homo sapiens *(13840)*
synthetic construct *(3580)*
unidentified *(2675)*
Mus musculus *(146)*
Rattus norvegicus *(46)*
All other taxa *(246)*
More...

Recent Activity
Tur
Your browsing activity is empty

NCBI

Entrez Nucleotide

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy

Search Nucleotide for colon cancer AND nonpolyposis AND human[Organism] Preview Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI (Cubby)

Field: **Title** Limits: **mRNA**, **RefSeq**

- Enter terms and click Preview to see only the number of search results.
- To save search indefinitely, click query # and select Save in My NCBI.
- To combine searches use #search, e.g., #2 AND #3 or click query # for more options.

Search	Most Recent Queries	Time	Result
#43	Search **colon cancer AND nonpolyposis AND human[Organism]** Field: **Title** Limits: **mRNA**, **RefSeq**	17:58:24	2
#40	Search **colon cancer AND nonpolyposis** Field: **Title** Limits: **mRNA**, **RefSeq**	17:58:07	13
#31	Search **colon cancer**	17:53:35	21322

Add Term(s) to Query or View Index:

- Enter a term in the text box; use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within a field.

All Fields Preview Index

Click AND OR NOT to add a term to the query box

Organism
Accession
All Fields
Author
EC/RN Number
Feature key
Filter
Gene Name
Genome Project
Issue
Journal
Keyword
Modification Date
Organism
Page Number
Primary Accession
Properties
Protein Name
Publication Date
SeqID String
Sequence Length

Refining your Search

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Nucleotide for colon cancer AND nonpolyposis AND human[Organism] Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Field: **Title** Limits: **mRNA**, **RefSeq**

Display Summary Show 20 Sort by Send to

All: 2 | Bacteria: 0 | RefSeq: 2 | mRNA: 2

Items 1 - 2 of 2 One page.

This search in Gene shows 9 results, including:

MSH2 (*Homo sapiens*): mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)
MLH1 (*Homo sapiens*): mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
MSH6 (*Homo sapiens*): mutS homolog 6 (E. coli)

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

2: NM_000251 Reports Order cDNA clone, Links
Homo sapiens mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli) (MSH2), mRNA
gi|4557760|ref|NM_000251.1|[4557760]

Recent Activity
Turn Off Clear
colon cancer AND nonpolyp... (2)
colon cancer AND nonpolyp... (13) Nucleotide

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

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

Entrez Nucleotide

My NCBI
Welcome joannealisonfox. [Sign Out]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search CoreNucleotide for colon cancer AND nonpolyposis AND human[Organism] Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Field: **Title** Limits: **mRNA**, **RefSeq**, **RefSeq**

Display Summary Show 20 Sort by Send to

All: 2 | Bacteria: 0 | RefSeq: 2 | mRNA: 2

Items 1 - 2 of 2 One page.

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

2: NM_000251 Reports Order cDNA clone, Links
Homo sapiens mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli) (MSH2), mRNA
gi|4557760|ref|NM_000251.1|[4557760]

About Entrez
Entrez Nucleotide
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut
My NCBI (Cubby)
Related resources
BLAST
Reference sequence project
Search for Genes
Submit to GenBank
Search for full length cDNAs

Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services
Privacy Statement | Freedom of Information Act | Disclaimer

NCBI Entrez Nucleotide

My NCBI [Sign In] [Register]

PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | OMIM | Books

Search CoreNucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display FASTA Show 5 Send to Hide: sequence all but gene, CDS and mRNA features

Range: from begin to end Reverse complemented strand Features: SNP STS Exon + Refresh

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

Comment Features Sequence

```
LOCUS       NM_000249                2524 bp    mRNA    linear   PRI 20-AUG-2007
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
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2524)
  AUTHORS   Perri,F., Cotugno,R., Piepoli,A., Merla,A., Quitadamo,M.,
            Gentile,A., Pilotto,A., Annese,V. and Andriulli,A.
  TITLE     Aberrant DNA methylation in non-neoplastic gastric mucosa of H.
            Pylori infected patients and effect of eradication
  JOURNAL   Am. J. Gastroenterol. 102 (7), 1361-1371 (2007)
   PUBMED   17509026
  REMARK    GeneRIF: While CDH1 methylation seems to be an early event in Hp
            gastritis, MLH1 methylation occurs late along with IM.
REFERENCE   2  (bases 1 to 2524)
  AUTHORS   Bettstetter,M., Dechant,S., Ruemmele,P., Grabowski,M., Keller,G.,
            Holinski-Feder,E., Hartmann,A., Hofstaedter,F. and Dietmaier,W.
  TITLE     Distinction of hereditary nonpolyposis colorectal cancer and
            sporadic microsatellite-unstable colorectal cancer through
            quantification of MLH1 methylation by real-time PCR
  JOURNAL   Clin. Cancer Res. 13 (11), 3221-3228 (2007)
   PUBMED   17545526
  REMARK    GeneRIF: quantitative MLH1 methylation analysis in MSI-H CRC is a
            valuable molecular tool to distinguish between HNPCC and sporadic
            MSI-H CRC
REFERENCE   3  (bases 1 to 2524)
  AUTHORS   Takahashi,M., Shimodaira,H., Andreutti-Zaugg,C., Iggo,R.,
            Kolodner,R.D. and Ishioka,C.
  TITLE     Functional analysis of human MLH1 variants using yeast and in vitro
            mismatch repair assays
  JOURNAL   Cancer Res. 67 (10), 4595-4604 (2007)
   PUBMED   17510385
  REMARK    GeneRIF: The 101 MLH1 variants were examined for the dominant
```

Links
- Gene
- HomoloGene
- Genome
- Genome Project
- Master
- Full text in PMC
- Probe
- Protein
- PubMed
- PubMed (RefSeq)
- Gene Genotype
- GeneView in dbSNP
- Taxonomy
- Related Sequences
- Map Viewer
- OMIM
- GEO Profiles
- SNP
- UniGene
- UniSTS
- LinkOut

NCBI

Entrez Nucleotide

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search CoreNucleotide for colon cancer AND nonpolyposis AND human[Organism] Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI (Cubby)

Related resources
BLAST

Reference sequence project

Search for Genes

Submit to GenBank

Search for full length cDNAs

Field: **Title** Limits: **mRNA**, **RefSeq**, **RefSeq**

Found 2 nucleotide sequences. CoreNucleotide [2]

Display Summary Show 20 Sort by Send to

All: 2 | Bacteria: 0 | RefSeq: 2 | mRNA: 2

Items 1 - 2 of 2 One page.

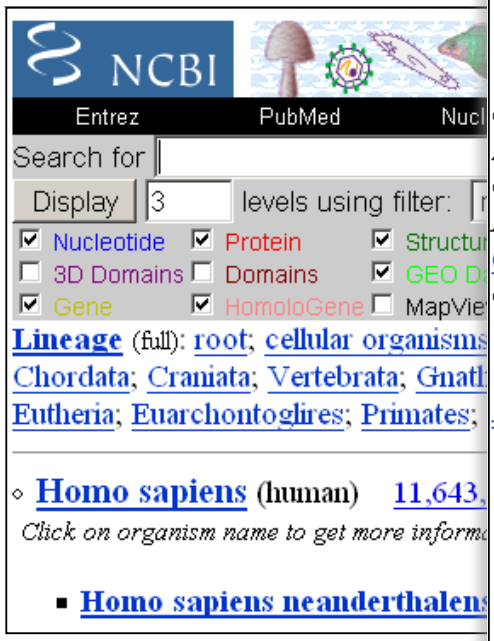
1: NM_000249 Reports
Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), mRNA
gi|28559089|ref|NM_000249.2|[28559089]

2: NM_000251 Reports
Homo sapiens mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli) (MSH2), mRNA
gi|4557760|ref|NM_000251.1|[4557760]

Links
- Full text in PMC
- Gene
- Gene Genotype
- GeneView in dbSNP
- Genome
- Genome Project
- HomoloGene
- Master
- Probe
- Protein
- PubMed
- PubMed (RefSeq)
- Taxonomy
- Related Sequences
- Map Viewer
- OMIM
- GEO Profiles
- SNP
- UniGene
- UniSTS
- LinkOut

Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services
Privacy Statement | Freedom of Information Act | Disclaimer

Taxonomy



Homo sapiens

Taxonomy ID: 9606
Genbank common name: **human**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:
common name: **man**

Lineage(full)
cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homo/Pan/Gorilla group; Homo

Entrez records

Database name	Subtree links	Direct links
Nucleotide	11,643,469	11,642,134
Protein	392,990	392,989
Structure	9,472	9,472
Genome Sequences	51	51
Genome Projects	1	1
Popset	20,878	20,878
SNP	11,870,024	11,870,024
3D Domains	35,848	35,848
Domains	19	19
GEO Datasets	3,525	3,525
GEO Expressions	10,649,715	10,649,715
UniGene	124,179	124,179
UniSTS	322,789	322,789
PubMed Central	3,586	3,586
Gene	38,624	38,624
HomoloGene	20,167	20,167
Taxonomy	2	1

Genome Information

See the NCBI Genome homepage
Go to NCBI genomic BLAST page for Homo sapiens

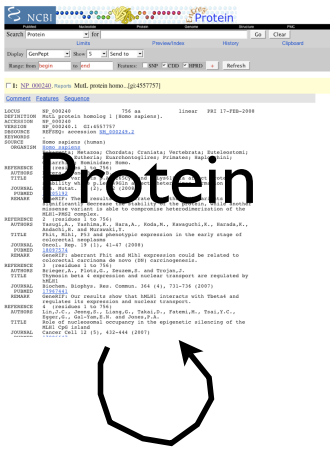
Genome view: 24 chromosomes

Names	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y

All molecular databases

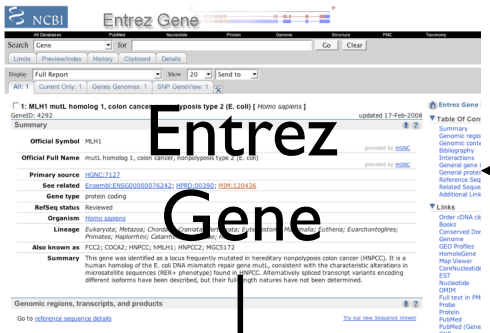
Goal: Find MLH1 homologs

- Tip: Use Entrez Gene as your hub to connect to everything else!



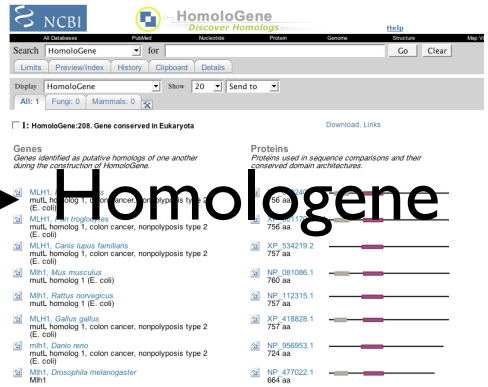
Protein

BLink



Entrez Gene

Other Entrez Databases



Homologene

Gene neighbors

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Nucleotide for colon cancer AND nonpolyposis AND human[Organism] Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Field: **Title** Limits: **mRNA**, **RefSeq**

Display Summary Show 20 Sort by Send to

All: 2 | Bacteria: 0 | RefSeq: 2 | mRNA: 2

Items 1 - 2 of 2 One page.

This search in Gene shows 9 results, including:

MSH2 (*Homo sapiens*): mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)

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

MSH6 (*Homo sapiens*): mutS homolog 6 (E. coli)

1: NM_000249 Reports

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

2: NM_000251 Reports

Homo sapiens mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli) (MS
gi|4557760|ref|NM_000251.1|[4557760]

Recent Activity

Turn Off Clear

colon cancer AND nonpolyp... (2)

colon cancer AND nonpolyp... (13) Nucleotide

Links

- Full text in PMC
- Gene
- Gene Genotype
- GeneView in dbSNP
- Genome
- Genome Project
- HomoloGene
- Master
- Probe
- Protein
- PubMed
- PubMed (RefSeq)
- Taxonomy
- Related Sequences
- Map Viewer
- OMIM
- GEO Profiles
- SNP
- UniGene
- UniSTS
- LinkOut

MLH1 Gene Record

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

GeneID: 4292 updated 10-Apr-2007

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	HGNC:7127
See related	HPRD:0039
Gene type	protein co
RefSeq status	Reviewed
Organism	*Homo sapi*

Genomic regions, transcripts, and products

Go to reference sequence details

NC_000003.10

[37009983 ▶ [37067341 ▶

5' 3'

NM_000249.2 NP_000240.1 CCDS2663.1

GeneRIFs: Gene References Into Function What's a GeneRIF?

1. Results confirmed complete exon skipping for the mutations of MLH1 in hereditary nonpolyposis colorectal cancer patients.
2. hMLH1 may have a role in development of secondary carcinoma in the gastrointestinal tract in patients (stomach and colorectal carcinoma)
3. Inactivation of MLH1 gene is associated with head and neck squamous cell carcinoma tumors and leukoplakia
4. In three adenocarcinomas, microsatellite instability and lack of the MLH1 protein expression were detected.
5. MLH1 is associated with longevity.
6. The identification of residues whose mutation disrupts MutL-MutS interaction and affects mismatch repair activity, suggesting a mechanism by which hereditary mutations in this region can produce a cancer predisposition.
7. These results indicate that an age-related increase of medullary-type tumors in poorly differentiated adenocarcinoma may play an important

See MLH1 in MapViewer

[37383246 ▶

A4

TCEA1P2

Interactions + GO

Interactions

Description .....

Product	Interactant	Other Gene	Complex	Source	P
E2F1 interacts with the MLH1 promoter.					
NC_000003.9	NP_005216.1	E2F1		BIND	
E2F4 interacts with the MLH1 promoter region.					
NC_000003.9	NP_001941.2	E2F4		BIND	
NP_000240.1	NP_000048.1	BLM		HPRD	
MLH1 interacts with BLM.					
NP_000240.1	NP_000048.1	BLM		BIND	
NP_000240.1	NP_009225.1	BRCA1		HPRD	
The exonuclease HEX1 interacts with the mismatch repair protein hMLH1.					
NP_000240.1	NP_003677.3	EXO1		BIND	
The exonuclease hEXO1b interacts with the mismatch repair protein hMLH1.					
NP_000240.1	NP_006018.3	EXO1		BIND	
NP_000240.1	NP_569082.1	EXO1		HPRD	
NP_000240.1	NP_003916.1	MBD4		HPRD	
MLH1 and interacts with MED1.					
NP_000240.1	NP_003916.1	MBD4		BIND	
NP_000240.1	BAA92353.1	MLH3		HPRD	

GeneOntology Provided by GOA

Function	Evidence	
ATP binding	IEA	
contributes_to MutSalpha complex binding	IDA	Pubmed
guanine/thymine mispair binding	IMP	Pubmed
guanine/thymine mispair binding	IEA	
mismatched DNA binding	IEA	
protein binding	IPI	Pubmed
contributes_to single-stranded DNA binding	IDA	Pubmed

Process	Evidence	
DNA damage response, signal transduction resulting in induction of apoptosis	IEA	
cell cycle	IEA	
male meiosis chromosome segregation	IEA	
meiotic recombination	IEA	
mismatch repair	IEA	
mismatch repair	TAS	Pubmed
negative regulation of mitotic recombination	IEA	
negative regulation of progression through cell cycle	IEA	

Component	Evidence	
MutLalpha complex	IEA	
condensed chromosome	IEA	
nucleus	IC	Pubmed
nucleus	IEA	
synaptonemal complex	IEA	

Sequences

NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. Explain

mRNA and Protein(s)

1. NM_000249.2→NP_000240.1 **MutL protein homolog 1**

Source sequence(s)	AU127758,BC006850,U07343
Consensus CDS	CCDS2663.1
Conserved Domains (3)	summary

cd00075	Location:31-122 Blast Score:107	HATPase_c; Histidine kinase-like ATPases; This family includes several ATP-binding proteins for example: histidine kinase, DNA gyrase B, topoisomerases, heat shock protein HSP90, phytochrome-like ATPases and

RefSeqs of Annotated Genomes: Build 36.2

The following sections contain reference sequences that belong to a specific genome build. Explain

Reference assembly

Genomic

1. **NC_000003.10 Reference assembly**

Range	37009983..37067341
Download	GenBank FASTA

2. **NT_022517.17**

Range	36974983..37032341
Download	GenBank FASTA

Alternate assembly (based on Celera assembly)

Genomic

1. **AC_000046.1 Alternate assembly (based on Celera assembly)**

Range	36977744..37035102
Download	GenBank FASTA

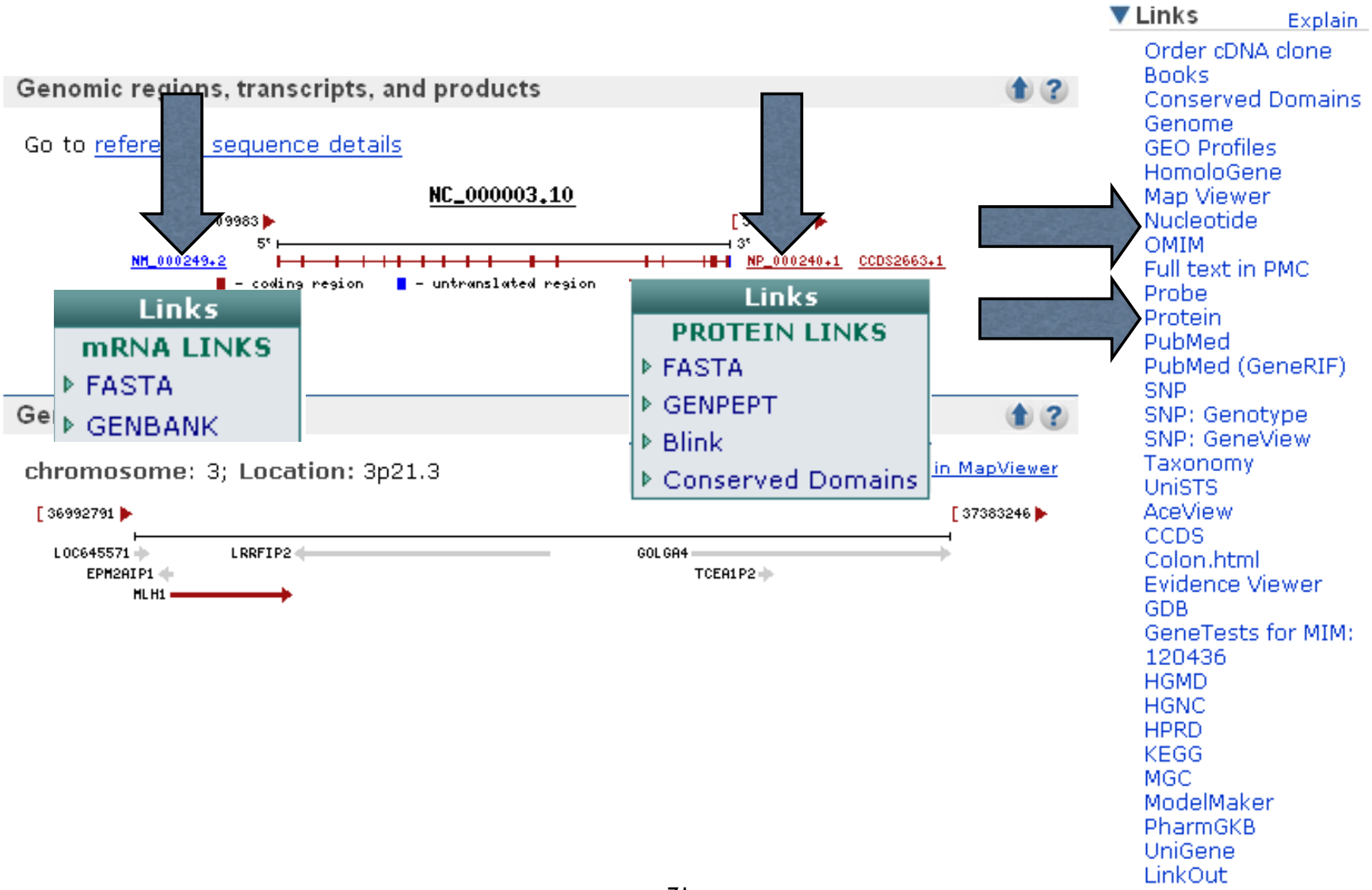
2. **NW_921651.1**

Range	36977744..37035102
Download	GenBank FASTA

Related Sequences

Nucleotide		Protein
Genomic	AC006583.31 (69181..100370, complement)	None
Genomic	AC011816.17 (143145..169313)	None
Genomic	AY217549.1	AAO22994.1
Genomic	AY344475.1	AAQ23474.1
Genomic	AY706914.1	AAU21566.1
Genomic	CH471055.1	EAW64483.1
		EAW64484.1
		EAW64485.1
Genomic	U17839.1	AAA85687.1
Genomic	U17840.1	AAA85687.1
Genomic	U17841.1	AAA85687.1
Genomic	U17842.1	AAA85687.1
Genomic	U17843.1	AAA85687.1
Genomic	U17844.1	AAA85687.1
Genomic	U17845.1	AAA85687.1
Genomic	U17846.1	AAA85687.1
Genomic	U17847.1	AAA85687.1
Genomic	U17848.1	AAA85687.1
Genomic	U17849.1	AAA85687.1
Genomic	U17850.1	AAA85687.1
Genomic	U17851.1	AAA85687.1
Genomic	U17852.1	AAA85687.1
Genomic	U17853.1	AAA85687.1
Genomic	U17854.1	AAA85687.1
Genomic	U17855.1	AAA85687.1
Genomic	U17856.1	AAA85687.1
Genomic	U17857.1	AAA85687.1
Genomic	U40978.1	AAA82079.1
mRNA	AB209848.1	BAD93085.1
mRNA	AF001359.1	AAB58936.1
mRNA	AK222810.1	BAD96530.1
mRNA	AU127758.1	None
mRNA	AY517558.1	AAT44531.1
mRNA	BC006850.1	AAH06850.1
mRNA	BX648844.1	None
mRNA	CR609870.1	None
mRNA	CR617505.1	None
mRNA	DQ648888.1	ABG49483.1
mRNA	DQ648889.1	ABG49484.1
mRNA	DQ648890.1	ABG49485.1
mRNA	DQ648891.1	ABG49486.1
mRNA	DQ648892.1	ABG49487.1
mRNA	DQ648893.1	ABG49488.1
mRNA	S77856.1	AAB34135.1
mRNA	U07343.1	AAC50285.1
mRNA	U07418.1	AAA17374.1

MLH1: Sequence Links



NCBI Entrez Gene

My NCBI
Welcome joannealisonfox. [Sign Out]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books | OMIM

Search Gene for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Full Report Show 20 Send to

All: 1 | Current Only: 1 | Genes Genomes: 1 | SNP GeneView: 1

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

GeneID: 4292 updated 16-Sep-2007

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	HGNC:7127
See related	Ensembl:ENSG00000076242; HPRD:00390; MIM:120436
Gene type	protein coding
RefSeq status	Reviewed
Organism	*Homo sapiens*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Also known as	FCC2; COCA2; HNPCC; hMLH1; HNPCC2; MGC5172
Summary	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. Alternatively spliced transcript variants encoding different isoforms have been described, but their full-length natures have not been determined.

Genomic regions, transcripts, and products

Go to reference sequence details

NC_000003.10

Entrez Gene Home

Table Of Contents

- Summary
- Genomic regions, transcripts...
- Genomic context
- Bibliography
- Interactions
- General gene information
- General protein information
- Reference Sequences
- Related Sequences
- Additional Links

Links Explain

- Order cDNA clone
- Books
- Conserved Domains
- Genome
- GEO Profiles
- HomoloGene
- Map Viewer
- CoreNucleotide
- EST
- Nucleotide
- OMIM
- Full text in PMC
- Probe
- Protein
- PubMed
- PubMed (GeneRIF)
- SNP
- SNP: Genotype

Finding Homologs:

1: HomoloGene:208. Gene conserved in Eukaryota

Download, Links

Genes
Genes identified as putative homologs of one another during the construction of HomoloGene.

Proteins
Proteins used in sequence comparisons and their conserved domain architectures.

NP_000240.1 756 aa
XP_001170433.1 756 aa
XP_534219.2 757 aa
NP_081086.1 760 aa
NP_112315.1 757 aa
XP_418828.1 757 aa
NP_477022.1 664 aa
XP_320342.2 671 aa
XP_307435.2 395 aa
NP_596199.1 684 aa
NP_013890.1 769 aa
XP_453504.1 724 aa
NP_985351.1 771 aa
XP_329015.1 751 aa
NP_567345.2 737 aa
NP_001045457.1

HomoloGene Downloader

Homologene:208. Gene conserved in Eukaryota

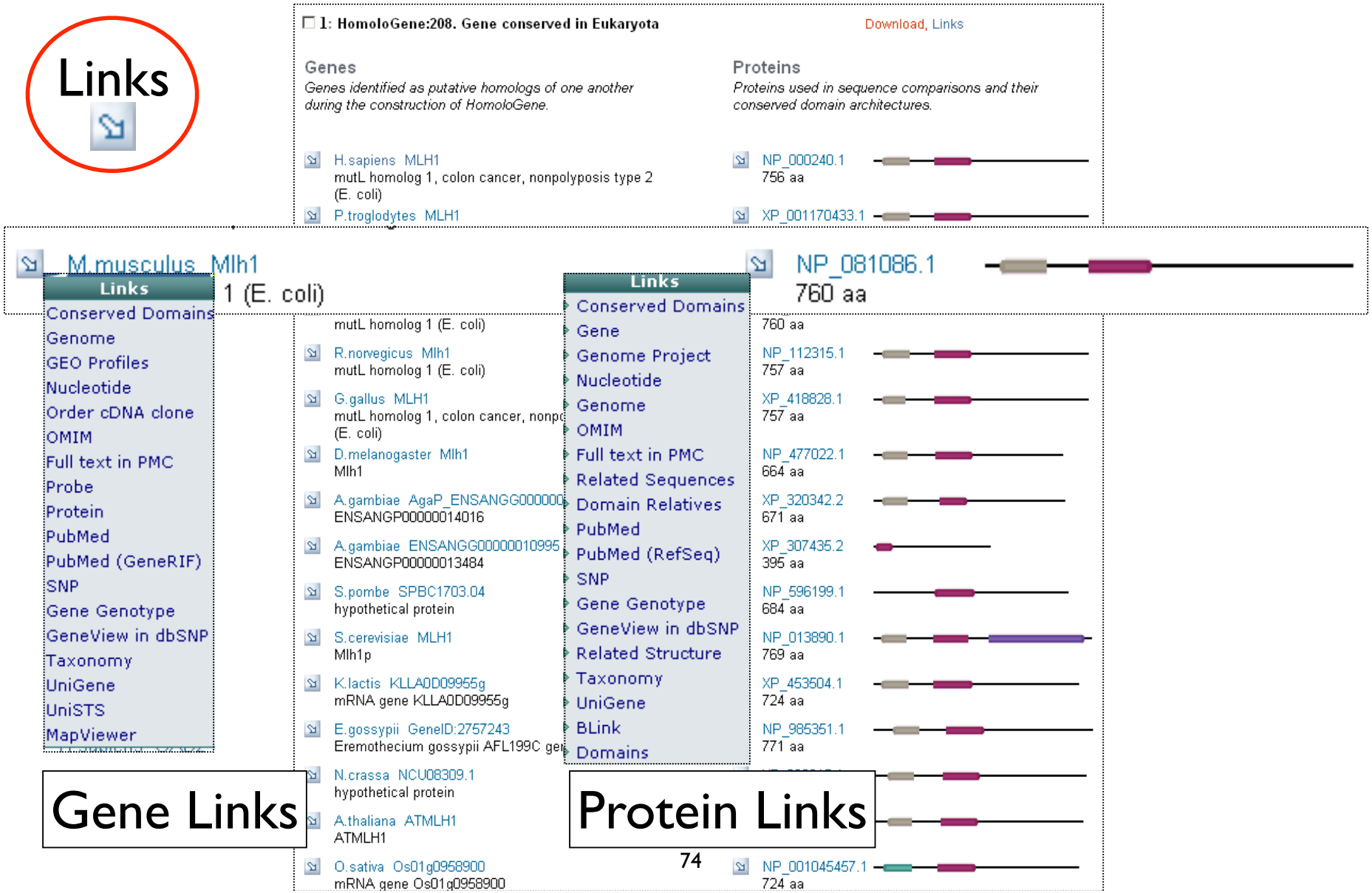
Download | Protein | sequences (in FASTA format)
Include 0 bp upstream of gene
Include 0 bp downstream of gene

Select which sequences should be included
Select All | Unselect All

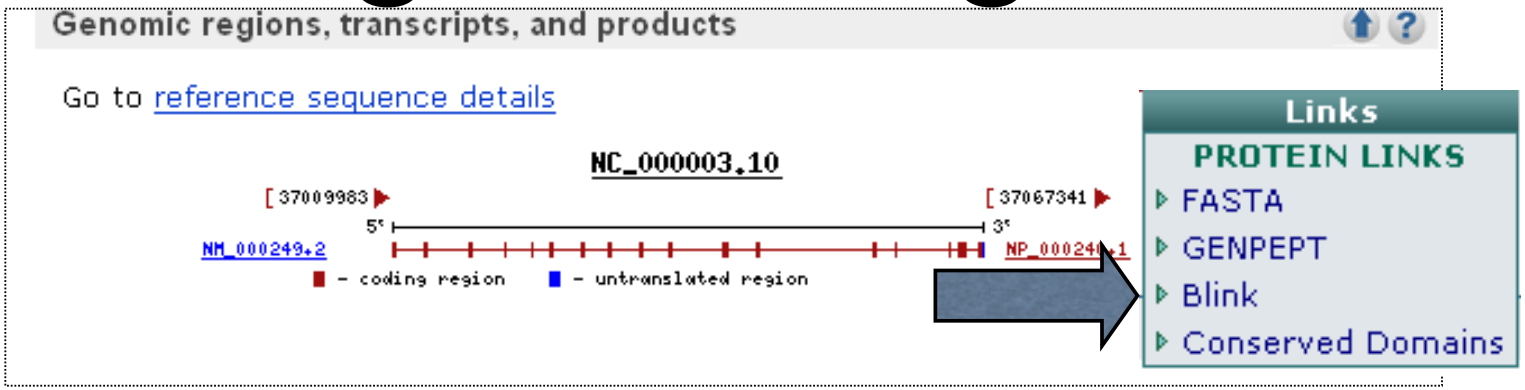
	Species	Gene		
☑	H.sapiens	MLH1	NM_0	
☑	P.troglodytes	MLH1	XM_0	
☑	C.familiaris	LOC477019	XM_5	
☑	M.musculus	Mlh1	NM_0	
☑	R.norvegicus	Mlh1	NM_031053.1	NP_112315.1
☑	G.gallus	MLH1	XM_418828.1	XP_418828.1
☑	D.melanogaster	Mlh1	NM_057674.2	NP_477022.1
☑	A.gambiae	AgaP_ENSANGG00000011527	XM_320342.2	XP_320342.2
☑	A.gambiae	ENSANGG00000010995	XM_307435.2	XP_307435.2
☑	S.pombe	SPBC1703.04	NM_001022118.1	NP_596199.1
☑	S.cerevisiae	MLH1	MLH1_6323819	NP_013890.1
☑	K.lactis	KLLA0D09955g	XM_453504.1	XP_453504.1

Protein
mRNA
Genomic

HomoloGene Cluster



Finding Homologs 2: BLink



BLink: BLAST Link

BLINK *precomputed BLAST*

Home | Taxonomy Report | Multiple Alignment | Blast | Help

My NCBI [Sign In] [Register]

Pre-computed BLAST results for: gi|4557757|ref|NP_000240.1 **MutL protein homolog 1 [Homo sapiens]**

Matching gis: 33738032;13905126;155685496;157928134;157928839;53932122;463989;91132884;155119205;730028;741682;1079787;119584889;27805155

Total (score > 100) : 4528 hits in 4468 proteins in 1318 species

Selected: 4528 hits in 4468 proteins in 1318 species Filter: Min Score: 100 |

Other views (Reports): Taxonomy report | Multiple Alignment | Blast

Reset all filters

Choose Display Options

40 Archaea | 2479 Bacteria | 443 Metazoa | 326 Fungi | 60 Plants | 0 Viruses | 1180 The Others reset selection

Results: 1 - 100 Next Page Last

% hits

756 aa

blink

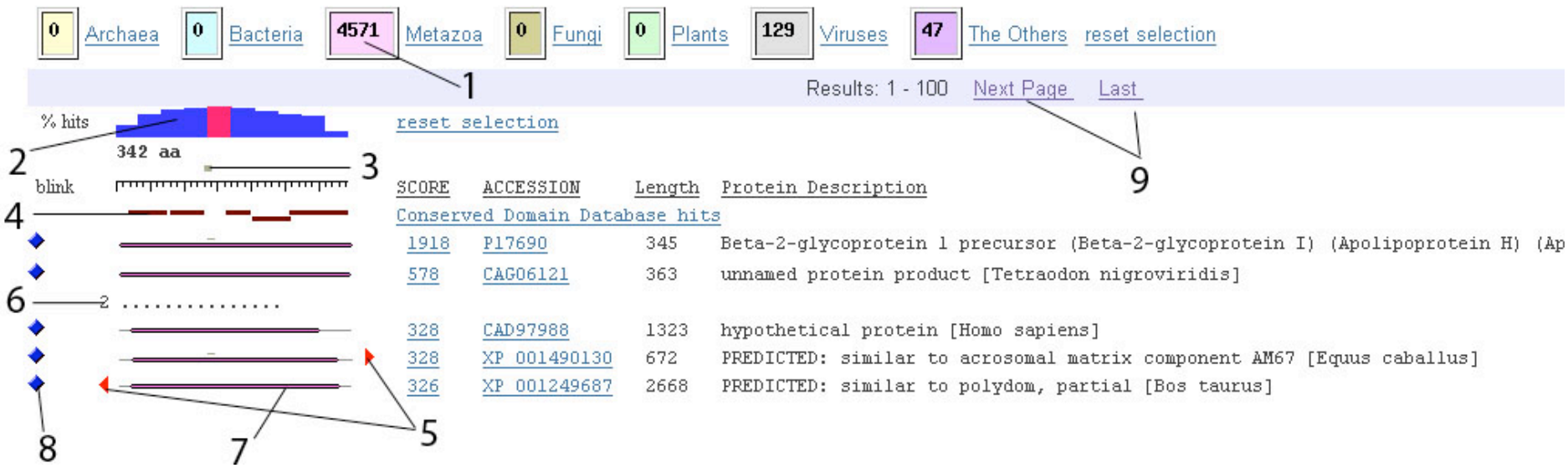
reset selection

SCORE	ACCESSION	Length	Protein Description
Conserved Domain Database hits			
3869	AAH06850	756	MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [Homo sap
3869	ABW03363	756	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [syntheti
3869	ABW03705	756	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [syntheti
3869	AAC50285	756	DNA mismatch repair protein homolog [Homo sapiens]
3869	P40692	756	RecName: Full=DNA mismatch repair protein Mlh1; AltName: Full=MutL pr
3869	gi|741682	756	DNA mismatch repair protein
3869	AAA82079	756	DNA mismatch repair protein homolog
3869	EAW64485	756	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli), isoform
3869	AAO22994	756	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [Homo sap
3869	AAQ02400	757	mutL-like 1, colon cancer, nonpolyposis type 2 [synthetic construct]

BLINK

- tool for exploring similar protein sequences by accessing precomputed BLAST searches
  - for every protein in Entrez against non-redundant (nr) protein database

BLINK precomputed BLAST



new and improved!
new display, previously limited to only 200 hits, now includes all hits

Sample Questions that can be answered with BLink

1. What protein sequences are similar to an Entrez protein sequence of interest, and what is the position and BLAST score of each hit? (see All Hits)

2. What are all the organisms to which a query sequence gets hits? Display the best hit to each organism? (see Best Hits)

3. What is the taxonomy tree structure of the set of organisms to which hits were found? (see TaxonomyReport)

4. What protein sequences with known 3-D structures are similar to the query sequence?

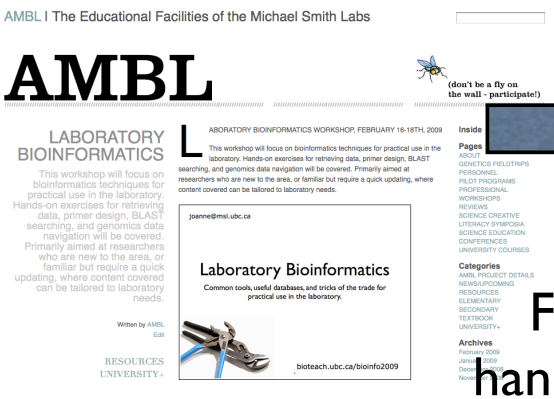
5. What domains are present in the query sequence?

Sequence Databases

PRACTICAL EXERCISES: Navigating Links, Retrieving Data with Entrez, and Advanced Tips & Tricks for Searching PubMed

I am studying the regulation of cancer genes and would like to retrieve all human sequence records associated with cancer that contain a promoter region.

navigate to:
bioteach.ubc.ca/bioinfo2009



Let's compare our results

Follow step-by-step instructions in handout and use links on the workshop website to complete the practical exercise

Use the preview tab and feature keys

Strategy #1: search nt

Strategy #2: search entrez gene

Check your History

Search	Most Recent Queries	Result
#5	Search #3 NOT #1 (unique hits from Approach B: Entrez Gene to CoreNucleotide)	329
#4	Search #1 NOT #3 (unique hits from Approach A: straight to Entrez CoreNucleotide search)	214
#3	Search #2 AND promoter[Feature key] (limit Approach B search to records with promoter annotated)	380
#2	CoreNucleotide Links for Gene (Search human[Organism] AND cancer[Text Word] AND gene_nucleotide[Filter]) (Approach B: Entrez gene follow link to CoreNucleotide)	65604
#1	Search human[Organism] AND cancer[Text Word] AND promoter[Feature key] (Approach A: Entrez CoreNucleotide search)	265

Advanced Tips & Tricks for Searching PubMed

My NCBI

Bookshelf

- Advanced Tabs - Limits; Preview/Index; History
- Entrez Gene RIF - reference into function sets
- Save collections with your MyNCBI account
- Search the NCBI Bookshef

NCBI

PubMed www.pubmed.gov

A service of the National Library of Medicine and the National Institutes of Health

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Search PubMed for cancer AND carrots Preview Go Clear

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI

Related Resources
Order Documents

- Search History will be lost after eight hours of inactivity.
- Search numbers may not be continuous; all searches are represented.
- To save search indefinitely, click query # and select Save in My NCBI.
- To combine searches use #search, e.g., #2 AND #3 or click query # for more options.

Search	Most Recent Queries	Time	Result
#22	Search **cancer AND carrots**	17:18:07	115
#21	Search **carrots**	17:17:56	1419
#20	Search **cancer**	17:17:48	1957409

Clear History

New PubMed display search: TPH1

All: 128 Review: 11

Gene Information

TPH1 tryptophan hydroxylase 1 [Homo sapiens]

This gene encodes a member of the pterin-dependent aromatic acid hydroxylase family. The encoded protein catalyzes the f... More

Location: 11p15.3-p14

tph1 in Homo sapiens | Mus musculus | Rattus norvegicus | All 12 Gene records

Gene

Items 1 - 20 of 128

Page 1 of 7 Next

1: Dopamine-melatonin neurons in the avian hypothalamus and their role as photoperiodic clocks.
El Halawani ME, Kang SW, Leclerc B, Kosonsiriluk S, Chaiseha Y.
Gen Comp Endocrinol. 2008 Dec 11. [Epub ahead of print]
PMID: 19114045 [PubMed - as supplied by publisher]
Related Articles

2: Resequencing of serotonin-related genes and association of tagging SNPs to citalopram response.
Peters EJ, Slager SL, Jenkins GD, Reinalda MS, Garriock HA, Shyn SI, Kraft JB, McGrath PJ, Hamilton SP.
Pharmacogenet Genomics. 2009 Jan;19(1):1-10.
PMID: 19077664 [PubMed - as supplied by publisher]
Related Articles

3: Lrp5 controls bone formation by inhibiting serotonin synthesis in the duodenum.
Yadav VK, Ryu JH, Suda N, Tanaka KF, Gingrich JA, Schütz G, Glorieux FH, Chiang CY, Zajac JD, Insogna KL, Mann JJ, Hen R, Ducy P, Karsenty G.
Cell. 2008 Nov 28;135(5):825-37.
PMID: 19041748 [PubMed - indexed for MEDLINE]
Related Articles

4: Serotonin genes and gene-gene interactions in borderline personality disorder in a matched case-control study.
Ni X, Chan D, Chan K, McMain S, Kennedy JL.
Prog Neuropsychopharmacol Biol Psychiatry. 2008 Nov 12. [Epub ahead of print]
PMID: 19032968 [PubMed - as supplied by publisher]
Related Articles

Also try:

- tph1 tph2
- tph1 knockout
- tph1 gene
- tph1 polymorphism
- tph1 depression

Titles with your search terms

- No association of TPH1 218A/C polymorphism with treatment response and ir [Neuropsychobiology. 2007]
- Stress upregulates TPH1 but not TPH2 mRNA in the rat dorsal raphe nucleus: ide [Cell Mol Neurobiol. 2008]
- TPH2 and TPH1: association of variants and interactions with heroin addiction. [Behav Genet. 2008]

» See all...

Recent Activity

Turn Off Clear

TPH1 (128)

The medical treatment of obsessive-compulsive disorder and anxiety.

clomipramine (3295) PubMed

Maylandia zebra M...[gi:193992698]

(Cichlidae) AND "Maylandi... (105438) Nucleotide

The Abstract plus page

1: PLoS ONE. 2008;3(10):e3301. Epub 2008 Oct 15.

Links

Genetic disruption of both tryptophan hydroxylase genes dramatically reduces serotonin and affects behavior in models sensitive to antidepressants.

Savelieva KV, Zhao S, Pogorelov VM, Rajan I, Yang Q, Cullinan E, Lanthorn TH.

Lexicon Pharmaceuticals Incorporated, The Woodlands, TX, USA. ksavelieva@lexpharma.com

The neurotransmitter serotonin (5-HT) plays an important role in both the peripheral and central nervous systems. The biosynthesis of serotonin is regulated by two rate-limiting enzymes, tryptophan hydroxylase-1 and -2 (TPH1 and TPH2). We used a gene-targeting approach to generate mice with selective and complete elimination of the two known TPH isoforms. This resulted in dramatically reduced central 5-HT levels in Tph2 knockout (TPH2KO) and Tph1/Tph2 double knockout (DKO) mice; and substantially reduced peripheral 5-HT levels in DKO, but not TPH2KO mice. Therefore, differential expression of the two isoforms of TPH was reflected in corresponding depletion of 5-HT content in the brain and periphery. Surprisingly, despite the prominent and evolutionarily ancient role that 5-HT plays in both vertebrate and invertebrate physiology, none of these mutations resulted in an overt phenotype. TPH2KO and DKO mice were viable and normal in appearance. Behavioral alterations in assays with predictive validity for antidepressants were among the very few phenotypes uncovered. These behavioral changes were subtle in the TPH2KO mice; they were enhanced in the DKO mice. Herein, we confirm findings from prior descriptions of TPH1 knockout mice and present the first reported phenotypic evaluations of Tph2 and Tph1/Tph2 knockout mice. The behavioral effects observed in the TPH2 KO and DKO mice strongly confirm the role of 5-HT and its synthetic enzymes in the etiology and treatment of affective disorders.

PMID: 18923670 [PubMed - indexed for MEDLINE]

PMCID: PMC2565062

Related Articles

- Late developmental stage-specific role of tryptophan hydroxylase 1 in brain serotonin levels. [J Neurosci. 2006]
- Tryptophan hydroxylase 1 knockout and tryptophan hydroxylase 2 polymo [Am J Physiol Lung Cell Mol Physiol. 2007]
- Deficiency of brain 5-HT synthesis but serotonergic neuron formation in Tph2 knockout mice. [J Neural Transm. 2008]
- Review [Abnormal cardiac activity in mice in the absence of peripheral serotonin synthesis] [J Soc Biol. 2004]
- Review Developmental role of tryptophan hydroxylase in the nervous system. [Mol Neurobiol. 2007]

» See Reviews... | » See All...

Recent Activity

Turn Off Clear

- Genetic disruption of both tryptophan hydroxylase genes dramatically reduces serotonin and...
- Crystal structure of tryptophan hydroxylase with bound amino acid substrate.
- Related Reviews for PubMe... (41)
- Deficiency of brain 5-HT synthesis but serotonergic neuron formation in Tph2 knockout mice...
- Modulation of peripheral serotonin levels by novel tryptophan hydroxylase inhibitors for t...

NCBI

National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

PubMed | All Databases | BLAST | OMIM | Books | TaxBrowser | Structure

Search All Databases for Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

Genomic biology
The human genome, whole genomes, and related resources

Tools
Data mining

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

New **Protein Clusters**
Entrez Protein Clusters database

The new Entrez Protein Clusters database is a collection of Reference Sequence (RefSeq) proteins, from the complete genomes of prokaryotes, plasmids, and organelles, that have been grouped and annotated based on sequence similiarity and protein function. Click here to find out more about the Protein Clusters database.

1 Billion Live Traces

The Trace Archive of sequencing traces has reached 1 billion live traces from over 480 organisms. For more information about the Trace Archive database click here.

PubMed Central

Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources
- Influenza Virus Resource
- Map Viewer
- dbMHC
- Mouse genome resources
- My NCBI

NCBI

Bookshelf

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC

Search Books for mlh1 Go Clear

Limits | Preview/Index | History | Clipboard | Details

- About Entrez
- Books
- Overview
- Using the books
- Information for authors and publishers
- Contact us
- Mailing list
- Project background
- FAQ
- My NCBI
- Privacy Policy

The **Bookshelf** is a growing collection of biomedical books that can be searched directly by typing a concept into the textbox above and selecting "Go". Try one of these searches:

- cell cycle control
- immunodeficiency
- protein evolution

New on the Bookshelf:

Animal Models of Cognitive Impairment
Frontiers in Neuroscience Series
Levin, Edward D.; Buccafusco, Jerry J., editors
Boca Raton (FL): CRC Press, Taylor & Francis Group; c2006

Baculovirus Molecular Biology [Internet]
Rohrmann, George F.
Bethesda (MD): National Library of Medicine (US), NCBI; 2008

Defining and Defeating the Intolerable Burden of Malaria III: Progress and Perspectives
Breman, Joel G.; Alilio, Martin S.; and White, Nicholas J.; editors
Northbrook (IL): The American Society of Tropical Medicine and Hygiene; 2007

The Intolerable Burden of Malaria: A New Look at the Numbers
Breman, Joel G.; Egan, Andréa; Keusch, Gerald T., editors
Northbrook (IL): The American Society of Tropical Medicine and Hygiene; c2001

NCBI **Bookshelf**

All Databases | PubMed | Nucleotide | Protein | Genome | Structure

Search Books for mlh1 Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Books Show 20 Send to

All: 204 | Figures: 8

183 items in **Health Services/Technology Assessment Text (HSTAT)**
Bethesda (MD): National Library of Medicine (US); 2003-2008

17 items in **GeneReviews**
Pagon, Roberta A., Editor-in-chief; Bird, Thomas C.; Dolan, Cynthia R.; Smith, Richard J.H.; Stephens, Karen; Associate editors.
Seattle (WA): University of Washington; c1993-2008

13 items in **Cancer Medicine**
Kufe, Donald W.; Pollock, Raphael E.; Weichselbaum, Ralph R.; Bast, Robert C., Jr.; Gansler, Ted S.; Holland, James F.; Frei III, Emil, editors.
Hamilton (Canada): BC Decker Inc.; c2003

10 items in **Madame Curie Bioscience Database**
Chapters taken from the Madame Curie Bioscience Database (formerly, Eurekah Bioscience Database)
Eurekah.com and Landes Bioscience and Springer Science+Business Media; c2009

Navigation

About this book

Part II Scientific Foundation, Section 1: Cancer Biology

7. Tumor-Suppressor Genes

Genetic Basis for Tumor Development

Somatic Cell Genetic Studies of Tumorigenesis

Retinoblastoma—A Paradigm for Tumor-Suppressor Gene Function

The *p53* Gene

The *INK4A* Locus and the *p16* ^{INK4A} and *p19* ^{ARF} Genes

The *APC* Gene

BRCA1 and *BRCA2* Genes

WT1 Gene

NF1 and *NF2* Genes

VHL Gene

DNA Repair Pathway Genes

Candidate Tumor-Suppressor Genes

Summary

References

Figures

Figure 7-10. Mismatch repair pathway in human...

DNA Repair Pathway Genes

At the outset of the chapter, tumor-suppressor genes were defined as those genes inactivated by germ line or somatic mutations in cancer. It was also emphasized that DNA damagerecognition and repair genes constitute a subset of the tumor-suppressor gene class, because they are affected by inactivating mutations in cancer. Whereas tumor-suppressor genes such as *RB1*, *p53*, *APC*, and *INK4a* appear to have active roles in regulating cell growth and/or apoptosis, the DNA damage-recognition and repair genes can arguably be viewed as having more passive roles in processes controlling growth. Distinguishing between what constitutes a growth-regulating tumor-suppressor gene versus a DNA repair-type tumor-suppressor gene may be difficult because some tumor-suppressor genes, including perhaps *p53*, *BRCA1*, and *BRCA2*, may ultimately be established to have functions in both growth control and DNA repair. Nevertheless, based on present data, there is a reasonable basis to suggest that loss-of-function mutations in both alleles of certain DNA repair pathway genes, such as the DNA mismatch repair genes, probably do not directly alter cell growth. Rather, inactivation of DNA mismatch repair activity likely contributes to cancer via an increased frequency of mutations in other cellular genes, particularly genes that are rate determining in tumor development.

Several recessive cancer predisposition syndromes resulting from inactivation of genes that function in DNA damage recognition and repair have been well described, including ataxia-telangiectasia (AT), Bloom syndrome, xeroderma pigmentosum, and Fanconi anemia. In each case, the specific cancer types and DNA-damaging agents that increase cancer risk are essentially distinct. Although AT heterozygotes may have a subtly increased risk of breast cancer,[264] in other recessive cancer syndromes, only homozygotes appear to have a clearly increased cancer risk. This observation contrasts sharply with the picture in the dominant cancer predisposition syndromes discussed earlier (eg, inherited retinoblastoma, familial adenomatous polyposis, NF1, and NF2), where heterozygotes have a clearly elevated cancer risk. Furthermore, as discussed earlier, the basis for increased cancer risk in an individual with a dominant cancer syndrome attributable to a germ line tumor-suppressor mutation (eg, *RB1* or *APC* mutation) is that cancers arise following inactivation of the remaining normal copy of the gene by a second "hit" in somatic cells (ie, the Knudson hypothesis). Therefore, it seems reasonable to argue that second "hits" in tumor-suppressor genes of the type that underlie dominant cancer syndromes must have considerably more potent effects on initiating cancer development than second "hits" in tumor-suppressor genes of the type that underlie recessive cancer syndromes.

In light of these considerations and because recessive cancer syndromes are quite rare, our discussion of the role of

encouraged. ↑ TOP

Navigation

About this book

Part II Scientific Foundation, Section 1: Cancer Biology

7. Tumor-Suppressor Genes

Genetic Basis for Tumor Development

Somatic Cell Genetic Studies of Tumorigenesis

Retinoblastoma—A Paradigm for Tumor-Suppressor Gene Function

The *p53* Gene

The *INK4A* Locus and the *p16* ^{INK4A} and *p19* ^{ARF} Genes

The *APC* Gene

BRCA1 and *BRCA2* Genes

WT1 Gene

NF1 and *NF2* Genes

VHL Gene

DNA Repair Pathway Genes

Candidate Tumor-Suppressor Genes

Summary

References

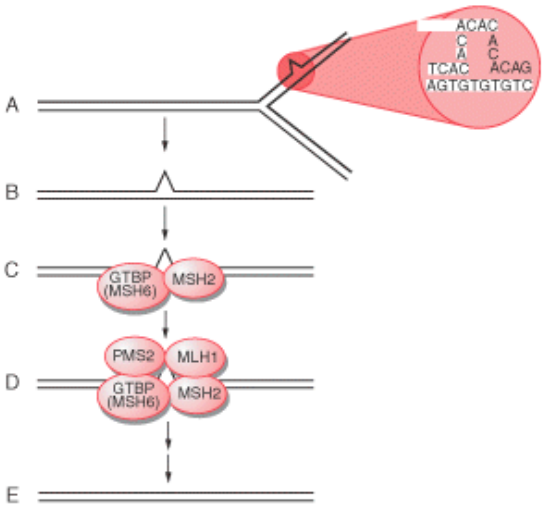


Figure 7-10. Mismatch repair pathway in human cells. **A** and **B,** During DNA replication, DNA mismatches may arise, such as from strand slippage (shown) or misincorporation of bases (not shown). **C,** The mismatch is recognized by MutS homologs, perhaps most often MSH2 and GTBP/MSH6, although another MutS homolog, MSH3, may substitute for GTBP/MSH6 in some cases. **D** and **E,** MutL homologs, such as MLH1 and PMS2, are recruited to the complex and the mismatch is repaired through the action of a number of proteins, including an exonuclease, helicase, DNA polymerase, and ligase. (Modified and reproduced with permission from Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. Cell 1996;87:159–70.)

You can make up your own examples, to search

Pubmed…

the Bookshelf…

and sign up for MyNCBI

NCBI Home PubMed GenBank BLAST Sign In | My NCBI

My NCBI

A division of the *National Library of Medicine* at the *National Institutes of Health*

Table of Contents

- My NCBI Home
- My Saved Data
- Search Filters
- Preferences
- About My NCBI

Register

Username:

Password:

Repeat Password:

- Usernames must be 3 or more letters, numbers or underscores
- Usernames, passwords and security question answers are case sensitive
- Passwords must be 6 or more characters
- Passwords must match

☐ Keep me signed in

☐ Remember my username

Security Question: Pet's Name

Answer:

VE3 5M

Please type the five characters you see above.

You can provide an e-mail address (optional).

Register

Or cancel and return home

Help Desk | Copyright | Disclaimer | Privacy | Accessibility | Contact

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

NCBI Home PubMed GenBank BLAST Sign In | My NCBI

My NCBI

A division of the *National Library of Medicine* at the *National Institutes of Health*

Table of Contents

- My NCBI Home
- My Saved Data
- Search Filters
- Preferences
- About My NCBI

Welcome to My NCBI

Use My NCBI to save your searches and data, and to set NCBI Web site preferences *About My NCBI...*

Sign into My NCBI

Username

Password

Keep me signed in

Remember my username

Sign In

- Register for an account
- I forgot my username
- I forgot my password
- About automatic sign in

See more sign in options for My NCBI partner organizations.

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

NCBI Home PubMed GenBank BLAST joannealisonfox | Sign Out | My NCBI

My NCBI

A division of the *National Library of Medicine* at the *National Institutes of Health*

Table of Contents

- My NCBI Home
- My Saved Data
- Search Filters
- Preferences
- About My NCBI

My NCBI Home » Saved Data

My Saved Data

Bibliographies

My Bibliography	6 Items
Other Citations	Not Created

Saved Searches (Manage)

cancer-omim	OMIM
Fox JA (full text free fu...	PubMed

Collections (Manage)

PTEN test search - 2 item...	PubMed, 2 Items
bootcamp collection	PubMed, 5 Items
488 items - pten generif	PubMed, 488 Items
cancer and carrot*	PubMed, 7 Items

Help Desk | Copyright | Disclaimer | Privacy | Accessibility | Contact

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

NCBI

PubMed www.pubmed.gov

A service of the U.S. National Library of Medicine and the National Institutes of Health

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Search PubMed for Go Clear Advanced Search

Limits | Preview/Index | History | Clipboard | Details

About Entrez
Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI

Related Resources
Order Documents
NLM Mobile
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

To get started with PubMed, enter one or more search terms.

Search terms may be topics, authors or journals.

The NIH Public Access Policy May Affect You

Does NIH fund your work?

Then your manuscript must be made available in PubMed Central

How?

If you publish in one of these journals, they will take care of the whole process.

If you publish *anywhere else*, deposit the manuscript in PubMed Central via one of the options described at publicaccess.nih.gov.

Note: Other funding organizations, including **HHMI**, **Wellcome Trust** and the **MRC** also require papers to be made freely available through PMC.

search with a gene name of interest to you

PubMed is a service of the U.S. National Library of Medicine that includes over 18 million citations from MEDLINE and other life science journals for biomedical articles back to 1948. PubMed includes links to full text articles and other related resources.

Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services

Bibliography

Related Articles in PubMed

PubMed links

GeneRIFs: Gene References Into Function

What's a Gene

1. the consequence of PTEN loss and Akt2 overexpression function synergistically to promote metastasis
2. Reduced PTEN expression was detected in more than one third of ovarian clear cell adenocarcinoma cases. Neither PTEN promoter methylation nor LOH at 10q23 locus is significantly related to PTEN inactivation and is not an adverse prognostic factor in OCCA.
3. Total PTEN was absent in 33.3% of ameloblastomas, while its stabilized, phosphorylated(ser380 / thr382 / thr383) form was absent in 83.3% of tumors.
4. report a statistically significant lower expression intensity of PTEN and HePTP and higher nuclear SHP2 expression
5. PTEN posttranslational inactivation and hyperacivation of the PI3K/Akt pathway sustain primary T cell leukemia.
6. coexpression of PTEN and AR should be undertaken to validate this pilot study and the utility of these biomarkers in routine histopathologic workup of patients with PC
7. Observational study and meta-analysis of gene-disease association. (HuGE Navigator)
8. im
thr

Submit:

Follow link from PubMed to Entrez Gene

GeneRIFs are intended to facilitate access to publications documenting experiments that add to our understanding of a gene and its function.

NCBI

PubMed www.pubmed.gov

A service of the U.S. National Library of Medicine and the National Institutes of Health

My NCBI
Welcome joannealisonfox. [Sign Out]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Search PubMed for Go Clear Advanced Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort By Send to

Send to
Text
File
Printer
Clipboard
Collections
E-mail
Order

All: 493 | Review: 20

Items 1 - 20 of 493

Page 1 of 25 Next

1: Genetic Variations in the PI3K/PTEN/AKT/mTOR Pathway Are Associated With Clinical Outcomes in Esophageal Cancer Patients Treated With Chemotherapy.
Hildebrandt MA, Yang H, Hung MC, Izzo JG, Huang M, Lin J, Ajani JA, Wu X.
J Clin Oncol. 2009 Jan 21. [Epub ahead of print]
PMID: 19164214 [PubMed - as supplied by publisher]
Related Articles

2: PTEN polymorphisms and the risk of esophageal carcinoma and gastric cardiac carcinoma in a high incidence region of China.
Ge H, Cao YY, Chen LQ, Wang YM, Chen ZF, Wen DG, Zhang XF, Guo W, Wang N, Li Y, Zhang JH.
Dis Esophagus. 2008;21(5):409-15.
PMID:
Related

3: Akt2 o
metas
Rycha
Proc Natl Acad Sci U S A. 2008 Dec 23;105(51):20315-20. Epub 2008 Dec 15.
PMID: 19075230 [PubMed - indexed for MEDLINE]
Related Articles

Recent Activity
Turn Off Clear

PubMed Links for Gene (Se... (493) PubMed

PTEN phosphatase and tensin homolog [Homo sapiens]

pten (3788) PubMed

mlh1 AND cmed6[book] (13) Books

Toward a confocal subcellular atlas of the human proteome.

Save your PubMed results to your MyNCBI collections

Credits

- Materials for this presentation have been adapted from the following sources:

  NCBI HelpDesk - Field Guide Course Materials

  Bioinformatics: A practical guide to the analysis of genes and proteins

- Questions? Please contact:

Dr. Joanne Fox
Michael Smith Laboratories
joanne@msl.ubc.ca

AMBL I The Educational Facilities of the Michael Smith Labs

AMBL

(don't be a fly on the wall - participate!)

LABORATORY BIOINFORMATICS

This workshop will focus on bioinformatics techniques for practical use in the laboratory. Hands-on exercises for retrieving data, primer design, BLAST searching, and genomics data navigation will be covered. Primarily aimed at researchers who are new to the area, or familiar but require a quick updating, where content covered can be tailored to laboratory needs.

Written by AMBL
Edit

RESOURCES
UNIVERSITY+

LABORATORY BIOINFORMATICS WORKSHOP, FEBRUARY 16-18TH, 2009

This workshop will focus on bioinformatics techniques for practical use in the laboratory. Hands-on exercises for retrieving data, primer design, BLAST searching, and genomics data navigation will be covered. Primarily aimed at researchers who are new to the area, or familiar but require a quick updating, where content covered can be tailored to laboratory needs.

joanne@msl.ubc.ca

Laboratory Bioinformatics

Common tools, useful databases, and tricks of the trade for practical use in the laboratory.

bioteach.ubc.ca/bioinfo2009

Inside

Pages
ABOUT
GENETICS FIELDTRIPS
PERSONNEL
PILOT PROGRAMS
PROFESSIONAL
WORKSHOPS
REVIEWS
SCIENCE CREATIVE
LITERACY SYMPOSIA
SCIENCE EDUCATION
CONFERENCES
UNIVERSITY COURSES

Categories
AMBL PROJECT DETAILS
NEWS/UPCOMING
RESOURCES
ELEMENTARY
SECONDARY
TEXTBOOK
UNIVERSITY+

Archives
February 2009
January 2009
December 2008
November 2008

Let's start at 9:00am

BLAST background, guided tour & practical exercises

joanne@msl.ubc.ca

BLink: BLAST Link

BLINK *precomputed BLAST*

Home | Taxonomy Report | Multiple Alignment | Blast | Help

My NCBI [Sign In] [Register]

Pre-computed BLAST results for: gi|4557757|ref|NP_000240.1 **MutL protein homolog 1 [Homo sapiens]**

Matching gis: 33738032;13905126;155685496;157928134;157928839;53932122;463989;91132884;155119205;730028;741682;1079787;119584889;27805155

Total (score > 100) : 4528 hits in 4468 proteins in 1318 species

Selected: 4528 hits in 4468 proteins in 1318 species Filter: Min Score: 100 |

Other views (Reports): Taxonomy report | Multiple Alignment | Blast

Reset all filters

Choose Display Options

40 Archaea | 2479 Bacteria | 443 Metazoa | 326 Fungi | 60 Plants | 0 Viruses | 1180 The Others reset selection

Results: 1 - 100 Next Page Last

% hits

756 aa

blink

reset selection

SCORE	ACCESSION	Length	Protein Description
Conserved Domain Database hits			
3869	AAH06850	756	MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [Homo sap
3869	ABW03363	756	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [syntheti
3869	ABW03705	756	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [syntheti
3869	AAC50285	756	DNA mismatch repair protein homolog [Homo sapiens]
3869	P40692	756	RecName: Full=DNA mismatch repair protein Mlh1; AltName: Full=MutL pr
3869	gi|741682	756	DNA mismatch repair protein
3869	AAA82079	756	DNA mismatch repair protein homolog
3869	EAW64485	756	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli), isoform
3869	AAO22994	756	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [Homo sap
3869	AAQ02400	757	mutL-like 1, colon cancer, nonpolyposis type 2 [synthetic construct]

BLAST

Finding Function By Sequence Similarity

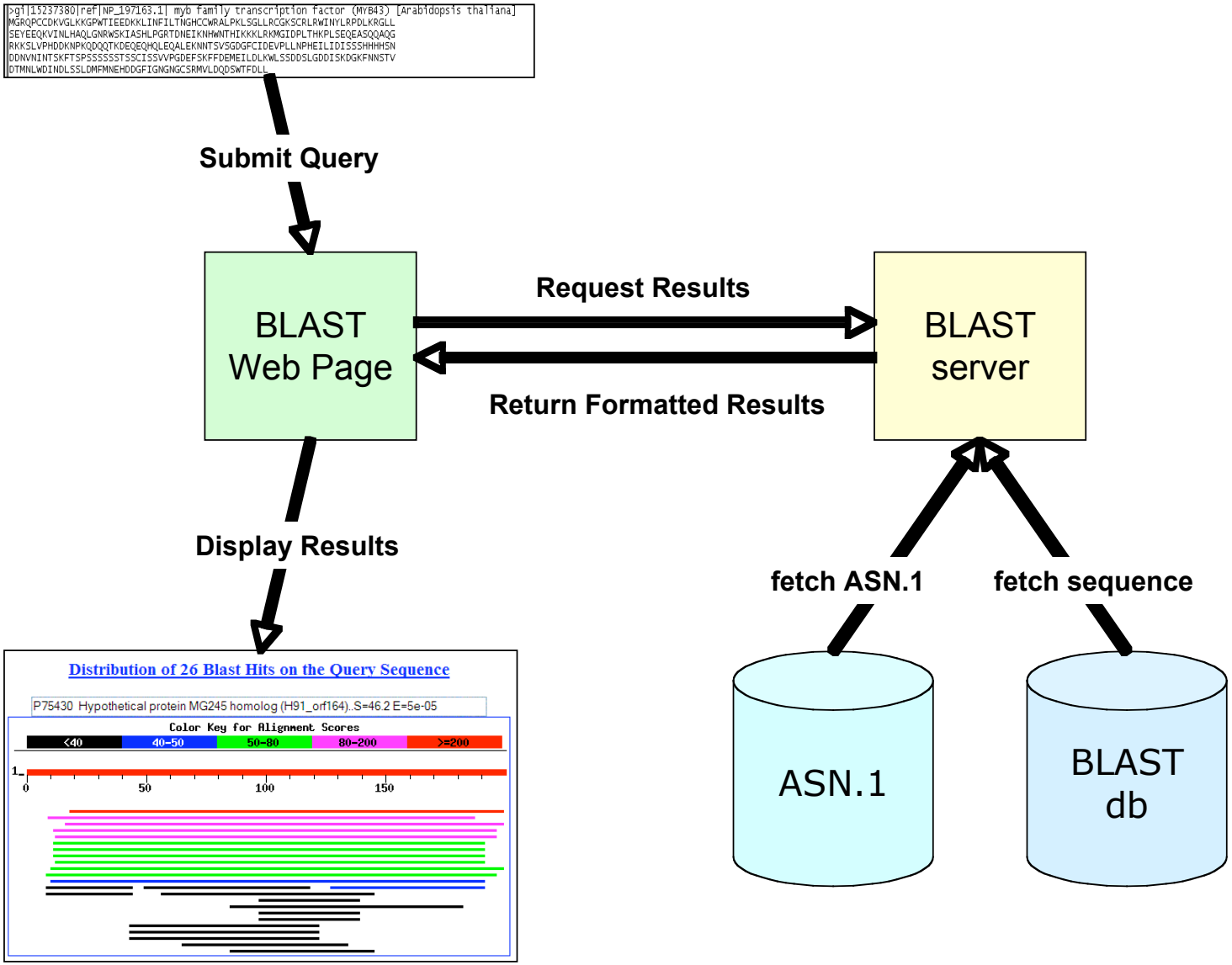
Concepts of Sequence Similarity Searching

- The premise:

  One sequence by itself is not informative; it must be analyzed by comparative methods against existing sequence databases to develop hypothesis concerning relatives and function.

The BLAST algorithm

- The BLAST programs (Basic Local Alignment Search Tools) are a set of sequence comparison algorithms introduced in 1990 that are used to search sequence databases for optimal local alignments to a query.
  - Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
  - Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." NAR 25:3389-3402.


>gi|15237380|ref|NP_197163.1| myb family transcription factor (MYB43) [Arabidopsis thaliana]
MGRQPCCDKVGLKKGPWTIEEDKKLINFILTNGHCCWRALPKLSGLLRCGKSCRLRWINYLRPDLKRGLL
SEYEEQKVINLHAQLGNRWSKIASHLPGRTDNEIKNHWNTHIKKKLRKMGIDPLTHKPLSEQEASQQAQG
RKKSLVPHDDKNPKQDQQTKDEQEQHQLEQALEKNNTSVSGDGFCIDEVPLLNPHEILIDISSSHHHHSN
DDNVNINTSKFTSPSSSSSSTSSCISSVVPGDEFSKFFDEMEILDLKWLSSDDSLGDDISKDGKFNNSTV
DTMNLWDINDLSSLDMFMNEHDDGFIGNGNGCSRMVLDQDSWTFDLL
Submit Query
BLAST Web Page
Request Results
Return Formatted Results
BLAST server
Display Results
Distribution of 26 Blast Hits on the Query Sequence
P75430 Hypothetical protein MG245 homolog (H91_orf164)..S=46.2 E=5e-05
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1
0
50
100
150
fetch ASN.1
fetch sequence
ASN.1
BLAST db

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

Basic **L**ocal **A**lignment **S**earch **T**ool

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- www, standalone, and network clients

BLAST programs

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

more BLAST programs

Program		Notes
Megablast	Contiguous	Nearly identical sequences
	Discontiguous	Cross-species comparison
Position Specific	PSI-BLAST	Automatically generates a position specific score matrix (PSSM)
	RPS-BLAST	Searches a database of PSI-BLAST PSSMs

nucleotide only

protein only

BLAST Algorithm

- Scoring of matches done using scoring matrices
- Sequences are split into words (default n=3)
  - Speed, computational efficiency
- BLAST algorithm extends the initial "seed" hit into an HSP
  - HSP = high scoring segment pair = Local optimal alignment

Sequence Similarity Searching – The statistics are important

Discriminating between real and artifactual matches is done using an estimate of probability that the match might occur by chance.

We'll talk more about the meaning of the scores (S) and e-values (E) that are associated with BLAST hits

Where does the score (S) come from?

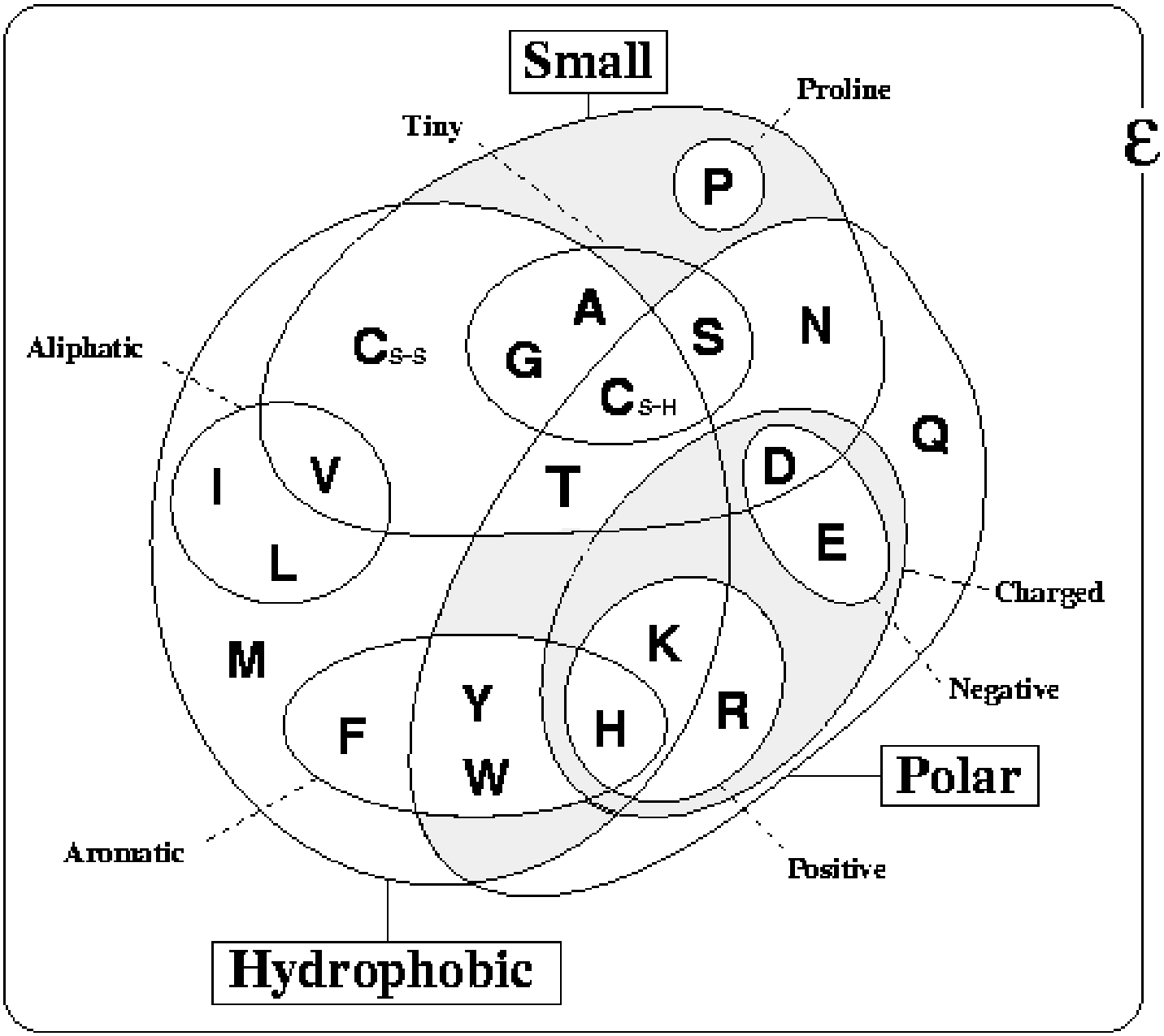
- The quality of each pair-wise alignment is represented as a score and the scores are ranked.
- Scoring matrices are used to calculate the score of the alignment base by base (DNA) or amino acid by amino acid (protein).
- The alignment score will be the sum of the scores for each position.

What's a scoring matrix?

- Substitution matrices are used for amino acid alignments.
  - each possible residue substitution is given a score
- A simpler unitary matrix is used for DNA pairs (+1 for match, -2 mismatch)

	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	[illegible]
G	0	-3	-1	-2	-3	6	[illegible]
H	-2	-3	-1	0	[illegible]	[illegible]	[illegible]

BLOSUM 62


Small
Proline
Tiny
ε
P
A
S
N
Aliphatic
Cs-s
G
Cs-H
Q
I
V
T
D
E
L
Charged
M
K
Negative
Y
R
F
H
W
Polar
Aromatic
Positive
Hydrophobic

BLOSUM vs PAM

BLOSUM 45	**BLOSUM 62**	**BLOSUM 90**
PAM 250	**PAM 160**	**PAM 100**
More Divergent		***Less Divergent***

- BLOSUM 62 is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.

What do the Score and the e-value really mean?

- The quality of the alignment is represented by the Score (S).

  The score of an alignment is calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (PAM, BLOSUM) whereas gap scores are assigned empirically .

- The significance of each alignment is computed as an E value (E).

  Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

Notes on E-values

- Low E-values suggest that sequences are homologous
  - Can't show non-homology
- Statistical significance depends on both the size of the alignments and the size of the sequence database
  - Important consideration for comparing results across different searches
  - E-value increases as database gets bigger
  - E-value decreases as alignments get longer

Homology: Some Guidelines

- Similarity can be indicative of homology
- Generally, if two sequences are significantly similar over entire length they are likely homologous
- Low complexity regions can be highly similar without being homologous
- Homologous sequences not always highly similar

Suggested BLAST cutoffs

Take Home Message:
Always look at your alignments

- Source: Chapter 11 – Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins
- For nucleotide based searches, one should look for hits with E-values of 10-6 or less and sequence identity of 70% or more
- For protein based searches, one should look for hits with E-values of 10-3 or less and sequence identity of 25% or more

BLAST Algorithm

- Scoring of matches done using scoring matrices
- Sequences are split into words (default n=3)
  - Speed, computational efficiency
- BLAST algorithm extends the initial "seed" hit into an HSP
  - HSP = high scoring segment pair = Local optimal alignment

How Does BLAST Really Work?

- The BLAST programs improved the overall speed of searches while retaining good sensitivity (important as databases continue to grow) by breaking the query and database sequences into fragments ("words"), and initially seeking matches between fragments.
- Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S".

BLAST Algorithm

Query Word (W = 3)

TLSHAWRLSNETDKRPFIETAERL**RDQ**HKKDYPEYKYQPRRRKNGKPGSSSEADAHSE

Determine neighborhood

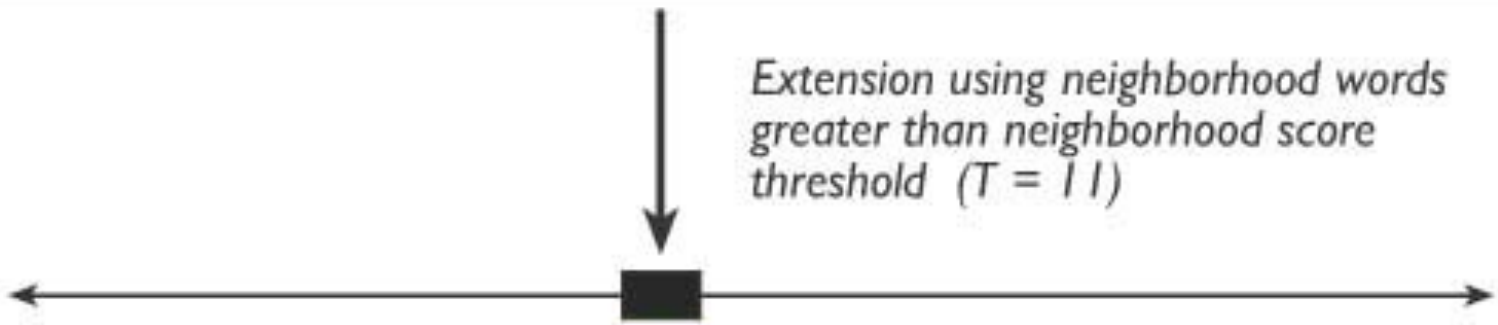
RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...

How Does BLAST Really Work?

- The BLAST programs improved the overall speed of searches while retaining good sensitivity (important as databases continue to grow) by breaking the query and database sequences into fragments ("words"), and initially seeking matches between fragments.
- Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S".

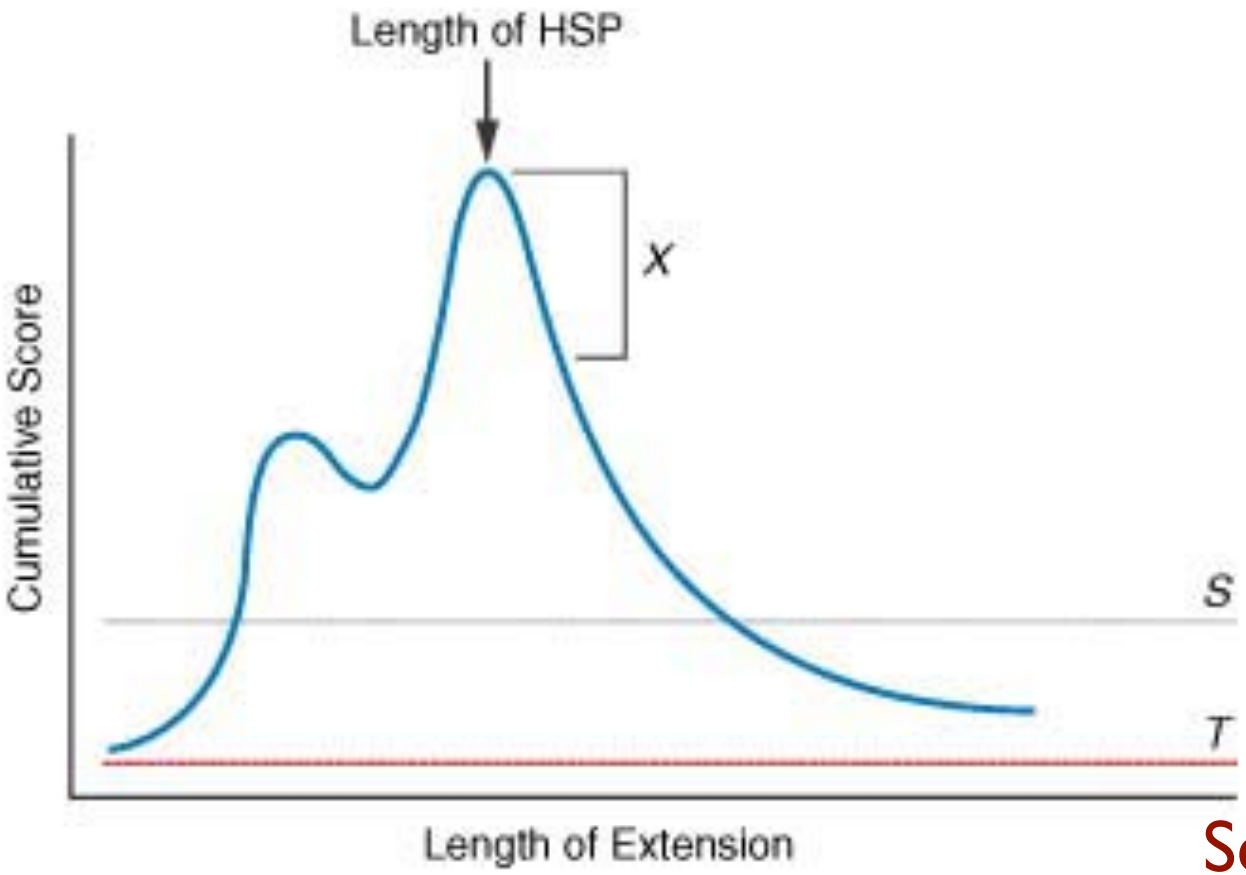
BLAST Algorithm

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...




```
Query: 1   TLSHAWRLSNETDKRPFIETAERLRDQHKKDYPEYKYQPRRRKNGKPGSSSEADAHSE 58
           TL   WRL N  +KRPF+E AERLR+QHKKD+P+YKYQPRRRK+ K G S   D   +
Sbjct: 140 TLESGWRLENPGEKRPFVEGAERLREQHKKDHPDYKYQPRRRKSVKNGQSEPEDGSEQ 197
```

Extending the High Scoring Segment Pair (HSP)



Minimum Score (S)

Neighborhood Score Threshold (T)

```
>gb|AAL08419.1|  PTEN [Takifugu rubripes]
Length=412

 Score =  197 bits (501),  Expect = 2e-49, Method: Composition-based stats.
 Identities = 95/100 (95%), Positives = 98/100 (98%), Gaps = 0/100 (0%)

Query  2    IVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  61
            +VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI
Sbjct  8    MVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  67

Query  62   YNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN  101
            YNLCAERHYD AKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct  68   YNLCAERHYDAAKFNCRVAQYPFEDHNPPQLELIKPFCED  107


 Score = 83.6 bits (205),  Expect = 4e-15, Method: Composition-based stats.
 Identities = 60/103 (58%), Positives = 68/103 (66%), Gaps = 32/103 (31%)

Query  99   KQNKMLKKDKMFHFWVNTFFIPGPEEV--------------------------------D  126
            KQNKM+KKDKMFHFWVNTFFIPGPEE                                  +
Sbjct  260  KQNKMMKKDKMFHFWVNTFFIPGPEESRDKLENGAVNNADSQQGVPAPGQGQPQSAECRE  319

Query  127  NDKEYLVLTLTkndldkankdkanRYFSPNFKVKLYFTKTVEE  169
            +D++YL+LTL+KND DKANKDKANRYFSPNFKVKL F+KTVEE
Sbjct  320  SDRDYLILTLSKNDRDKANKDKANRYFSPNFKVKLCFSKTVEE  362


>gb|AAH93110.1|  Ptenb protein [Danio rerio]
Length=289

 Score =  197 bits (500),  Expect = 2e-49, Method: Composition-based stats.
 Identities = 95/99 (95%), Positives = 98/99 (98%), Gaps = 0/99 (0%)

Query  3    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIY  62
            VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHK+HYKIY
Sbjct  9    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKDHYKIY  68

Query  63   NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFKQN  101
            NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPF ++
Sbjct  69   NLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCED  107
```

BLAST Algorithm

- Scoring of matches done using scoring matrices
- Sequences are split into words (default n=3)
  - Speed, computational efficiency
- BLAST algorithm extends the initial "seed" hit into an HSP
  - HSP = high scoring segment pair = Local optimal alignment

Credits

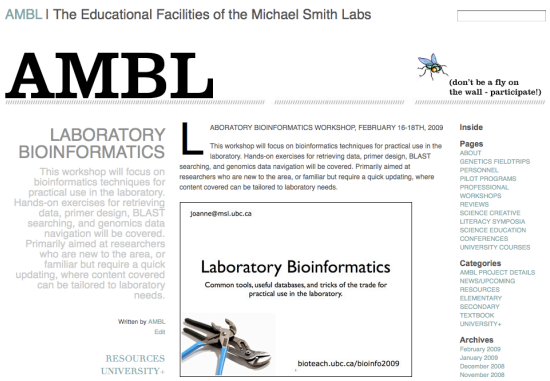
- Materials for this presentation have been adapted from the following sources:

  NCBI HelpDesk - Field Guide Course Materials

  Bioinformatics: A practical guide to the analysis of genes and proteins

- Questions? Please contact:

Dr. Joanne Fox
Michael Smith Laboratories
joanne@msl.ubc.ca



AMBL I The Educational Facilities of the Michael Smith Labs

AMBL

(don't be a fly on the wall - participate!)

LABORATORY BIOINFORMATICS

This workshop will focus on bioinformatics techniques for practical use in the laboratory. Hands-on exercises for retrieving data, primer design, BLAST searching, and genomics data navigation will be covered. Primarily aimed at researchers who are new to the area, or familiar but require a quick updating, where content covered can be tailored to laboratory needs.

Written by AMBL
Edit

RESOURCES
UNIVERSITY+

LABORATORY BIOINFORMATICS WORKSHOP, FEBRUARY 16-18TH, 2009

This workshop will focus on bioinformatics techniques for practical use in the laboratory. Hands-on exercises for retrieving data, primer design, BLAST searching, and genomics data navigation will be covered. Primarily aimed at researchers who are new to the area, or familiar but require a quick updating, where content covered can be tailored to laboratory needs.

joanne@msl.ubc.ca

Laboratory Bioinformatics

Common tools, useful databases, and tricks of the trade for practical use in the laboratory.

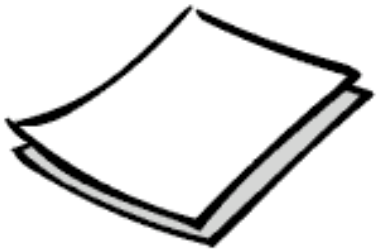
bioteach.ubc.ca/bioinfo2009

Inside

Pages
ABOUT
GENETICS FIELDTRIPS
PERSONNEL
PILOT PROGRAMS
PROFESSIONAL
WORKSHOPS
REVIEWS
SCIENCE CREATIVE
LITERACY SYMPOSIA
SCIENCE EDUCATION
CONFERENCES
UNIVERSITY COURSES

Categories
AMBL PROJECT DETAILS
NEWS/UPCOMING
RESOURCES
ELEMENTARY
SECONDARY
TEXTBOOK
UNIVERSITY+

Archives
February 2009
January 2009
December 2008
November 2008



Let's start at 9:00am

BLAST background, guided tour & practical exercises

joanne@msl.ubc.ca