joanne@msl.ubc.ca

wireless login:
mslguest
4myguest

Laboratory Bioinformatics

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

bioteach.ubc.ca/bioinfo2010

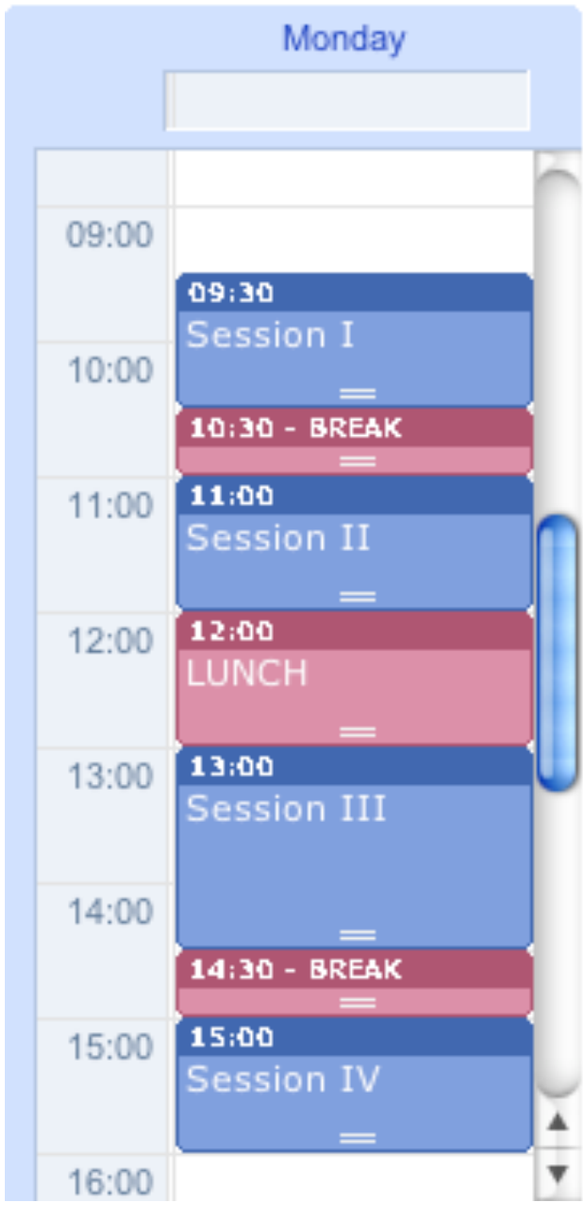
Workshop Schedule

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

- wireless login

  mslguest

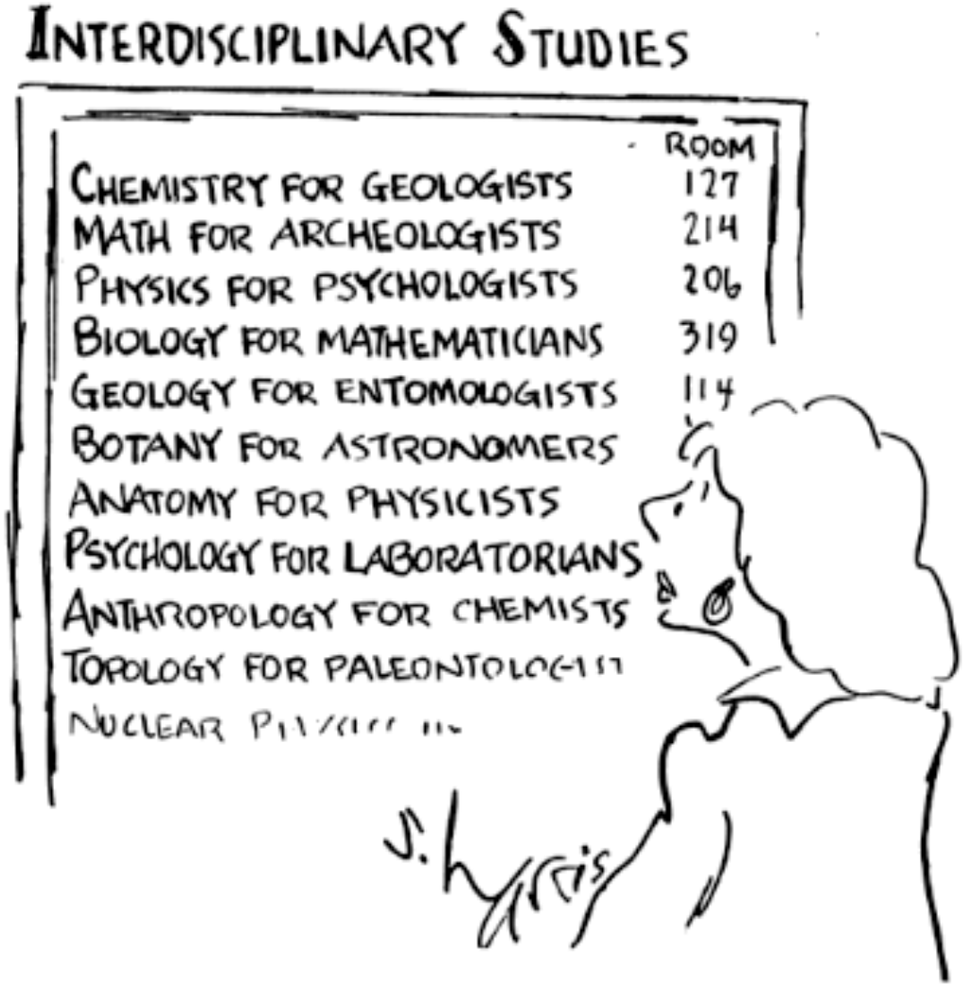
  4myguest



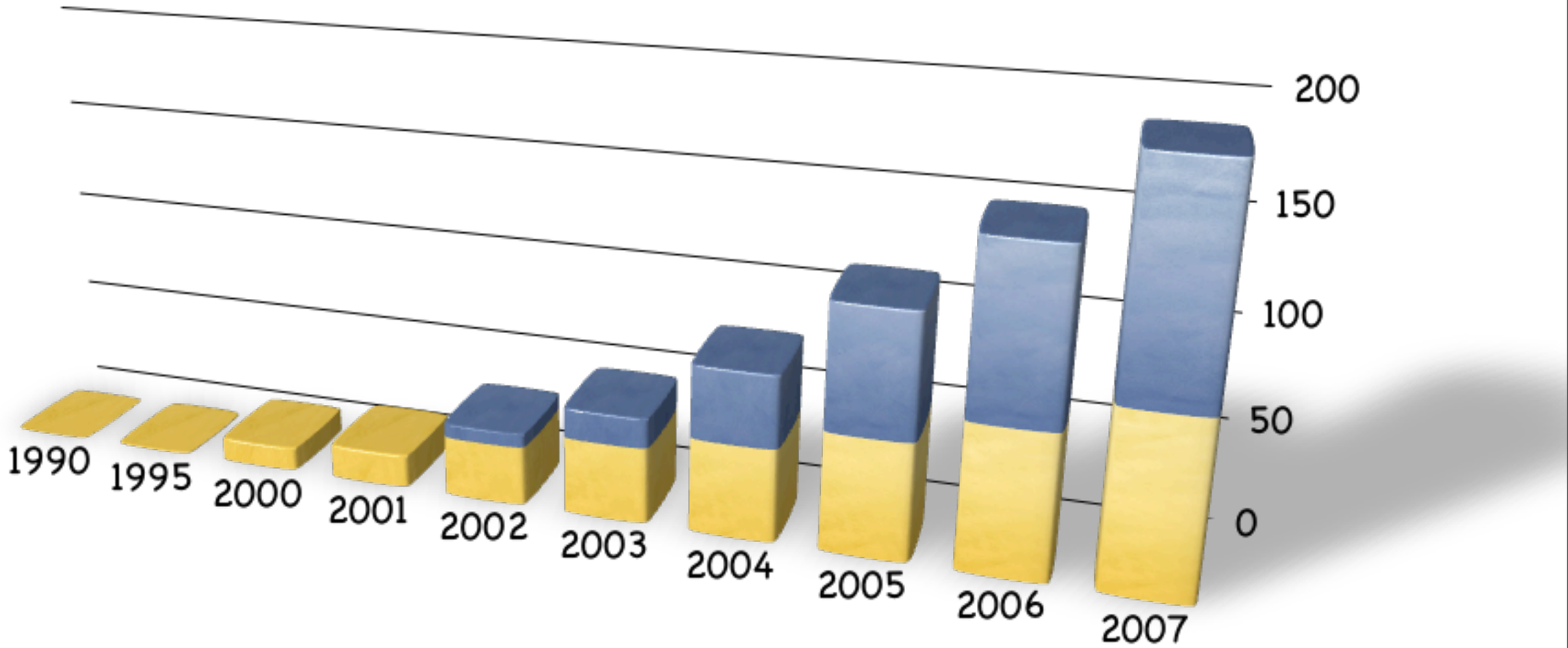
Today's Plan

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

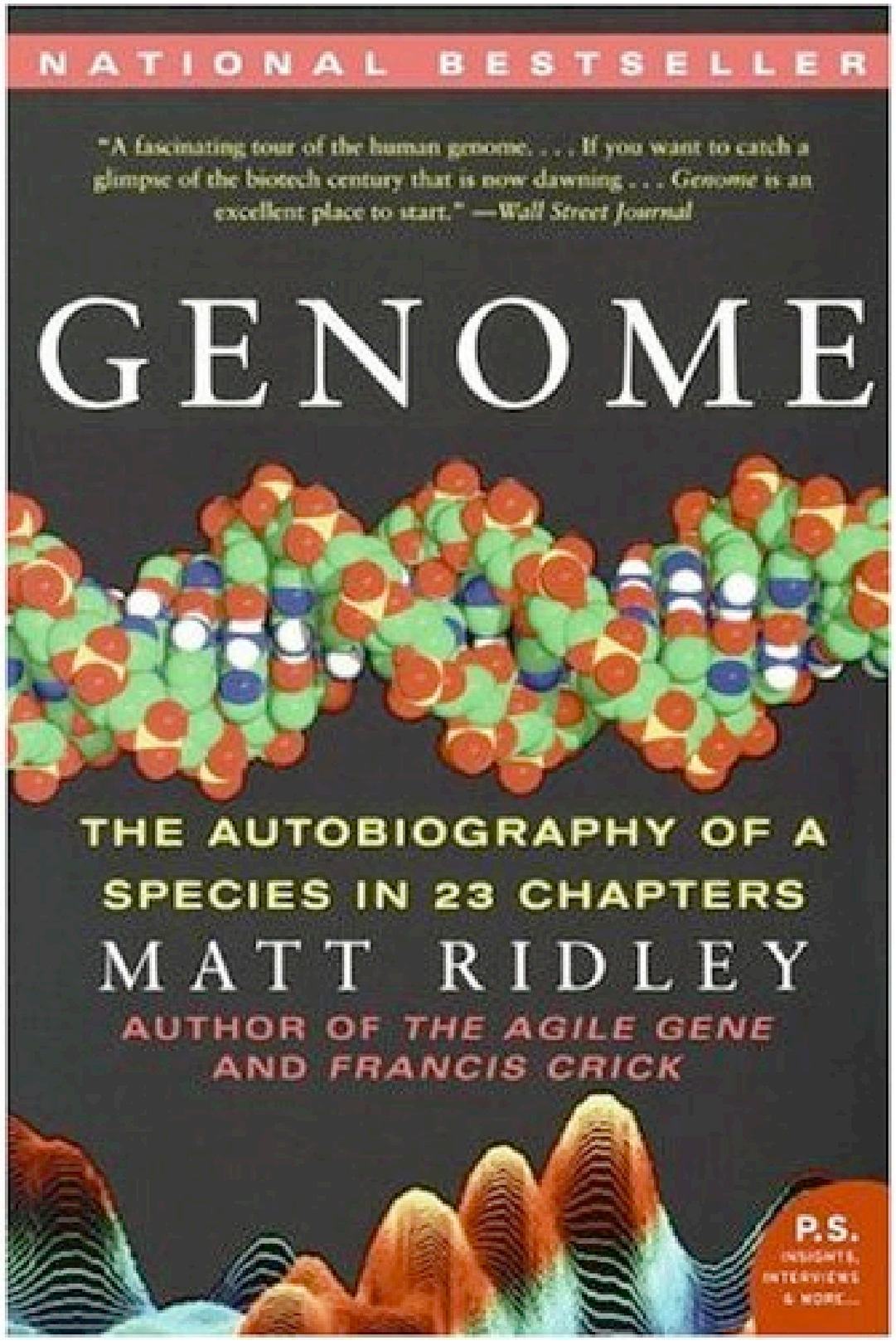
Bioinformatics for Biologists



Growth of GenBank

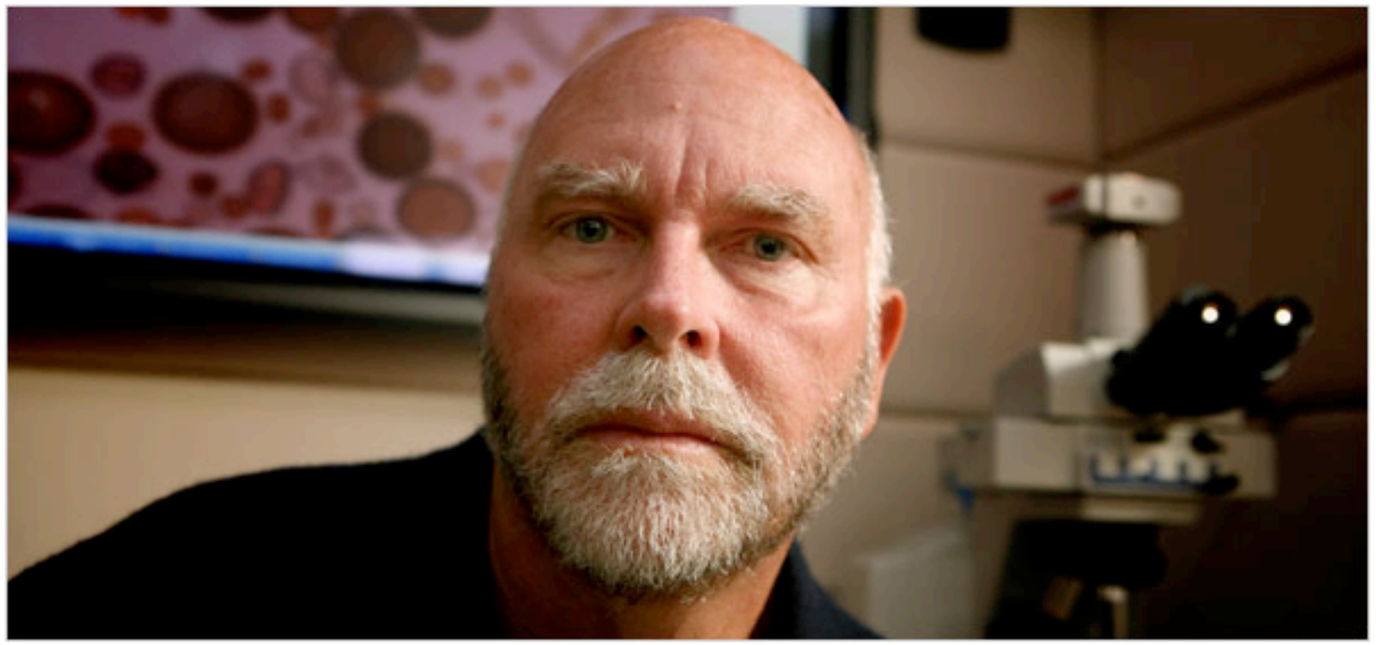


In 2005, International sequence databases exceed 100 gigabases


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



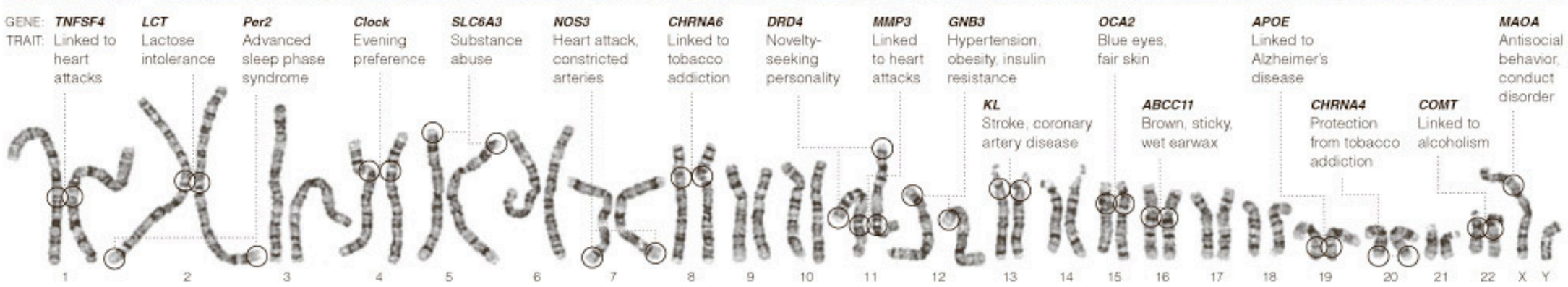
Thor Swift for The New York Times

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.

By NICHOLAS WADE

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.

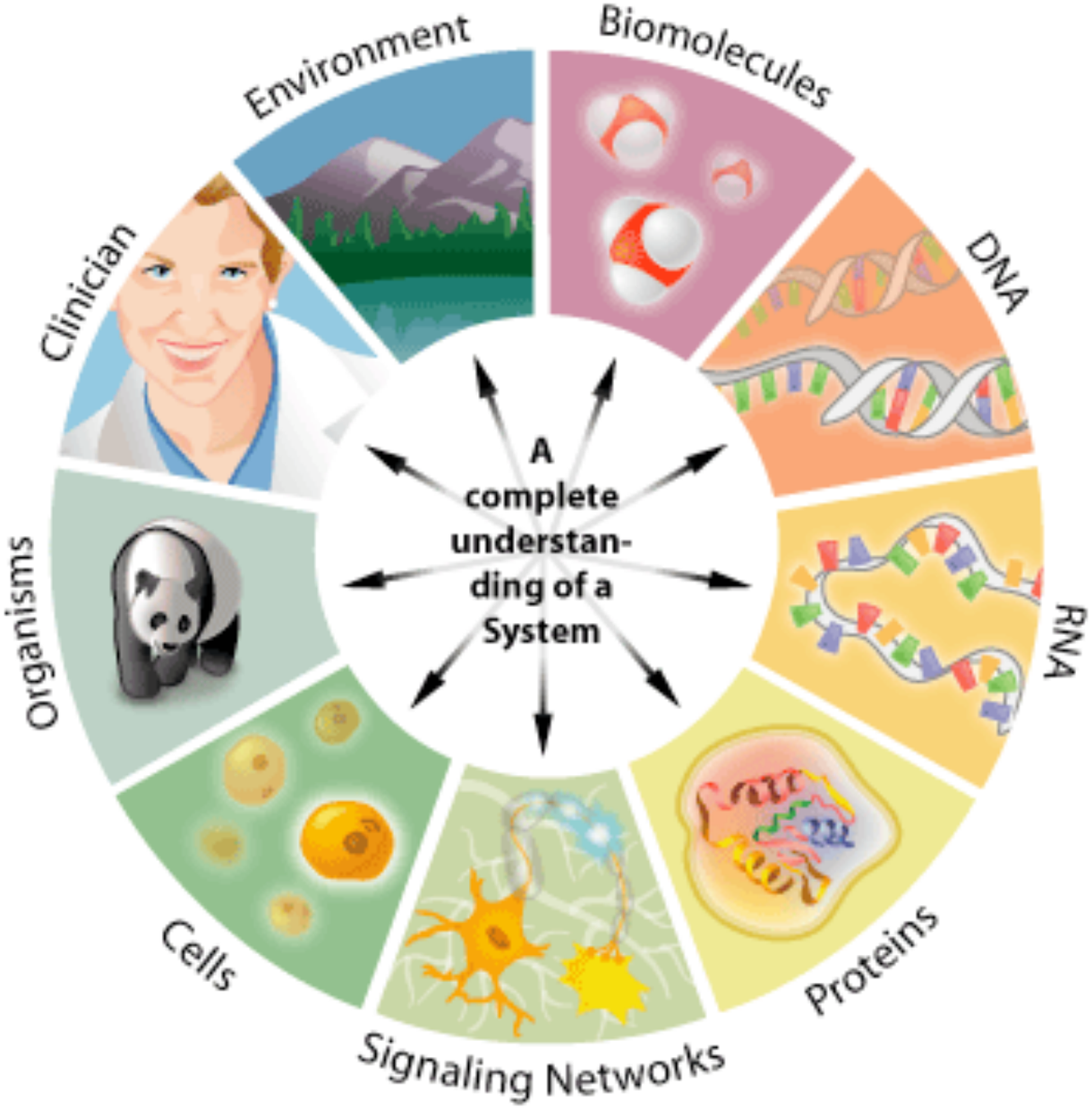


Sources: PLoS Biology; GeneCards

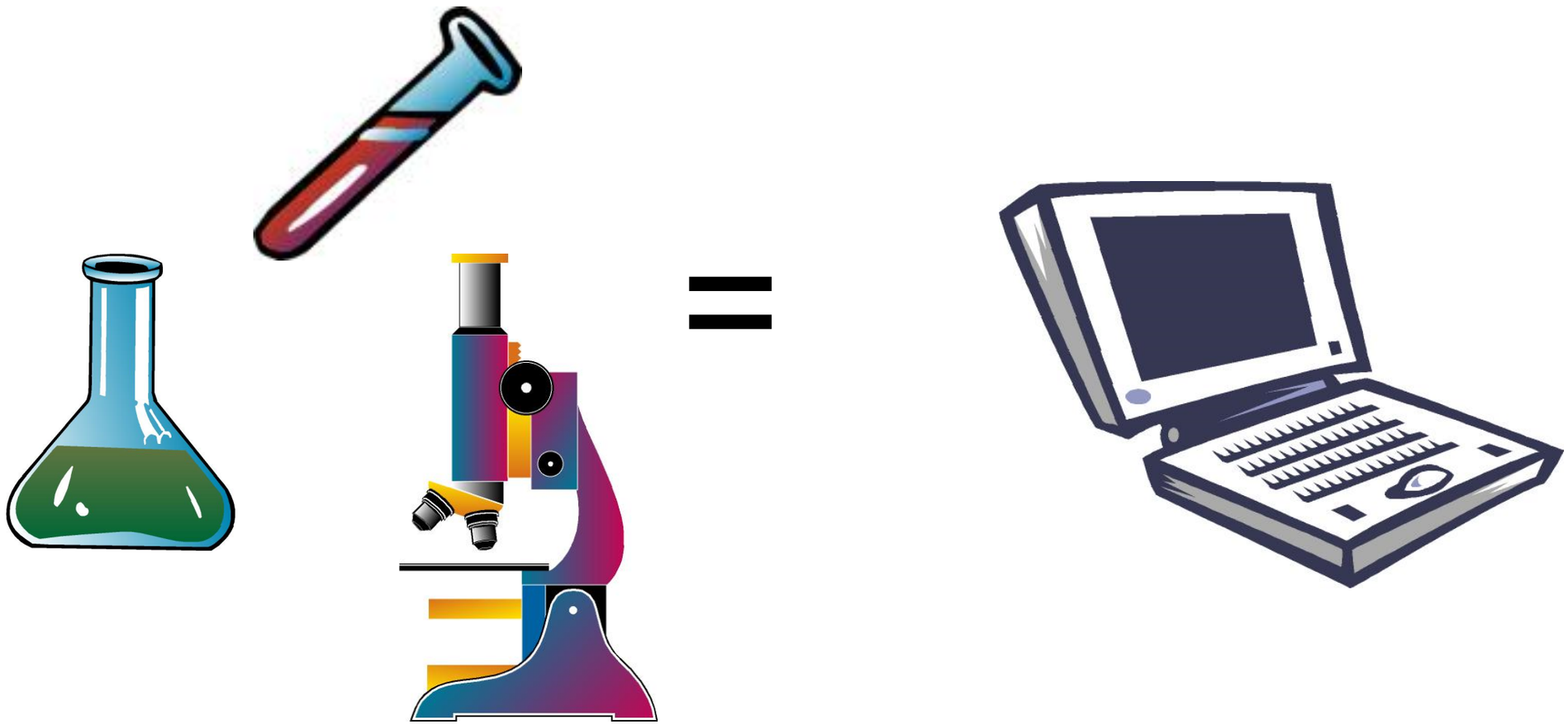
Genes highlighted above are linked to specific traits but do not mean the conditions are inevitable.

THE NEW YORK TIMES

What is Bioinformatics?

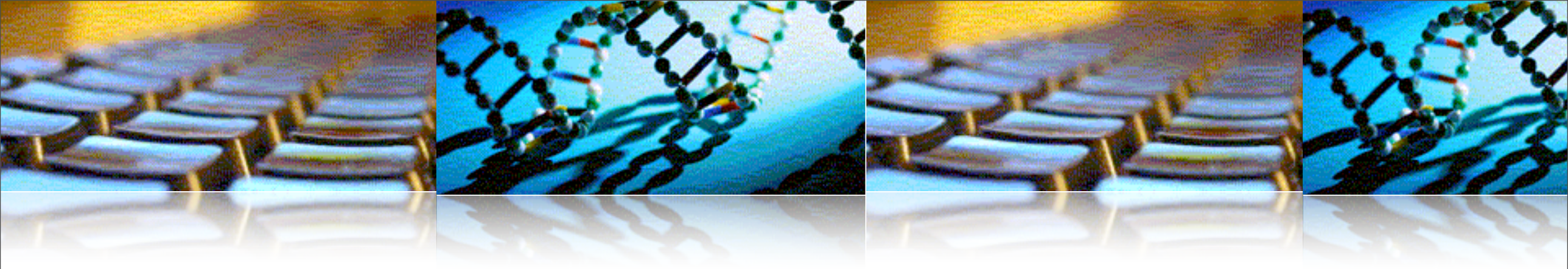


Laboratory Bioinformatics



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.

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?

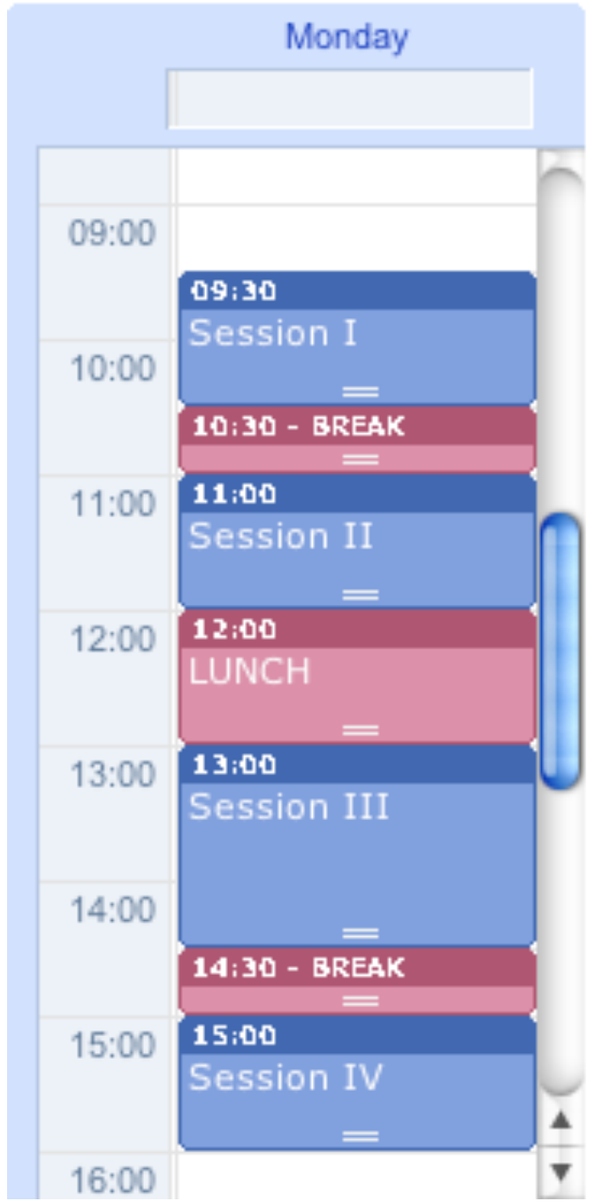
Overview of Topics*

✓ 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

NCBI
National Center for Biotechnology Information

Search All Databases

Search Clear

Resources

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

Welcome to NCBI

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

More about the NCBI | Mission | Organization | Research | RSS

Genome

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

II 1 2 3 4

How To...

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

See all ...

Popular Resources

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

NCBI News

OMIM's new look, Epigenomics in April — 10 May 2010
NCBI News
The April NCBI News issue is now available.

NIH Roadmap Epigenomics Project data in GEO database — 22 Apr 2010
GEO's Roadmap Epigenomics Project Data Listings page allows ...

March News issue available — 09 Apr 2010

New NCBI Site Guide

Resources

How To...

adapted from NCBI News, November 2009

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)

DDBJ

EMBL

CIB

getentry

SRS

EBI

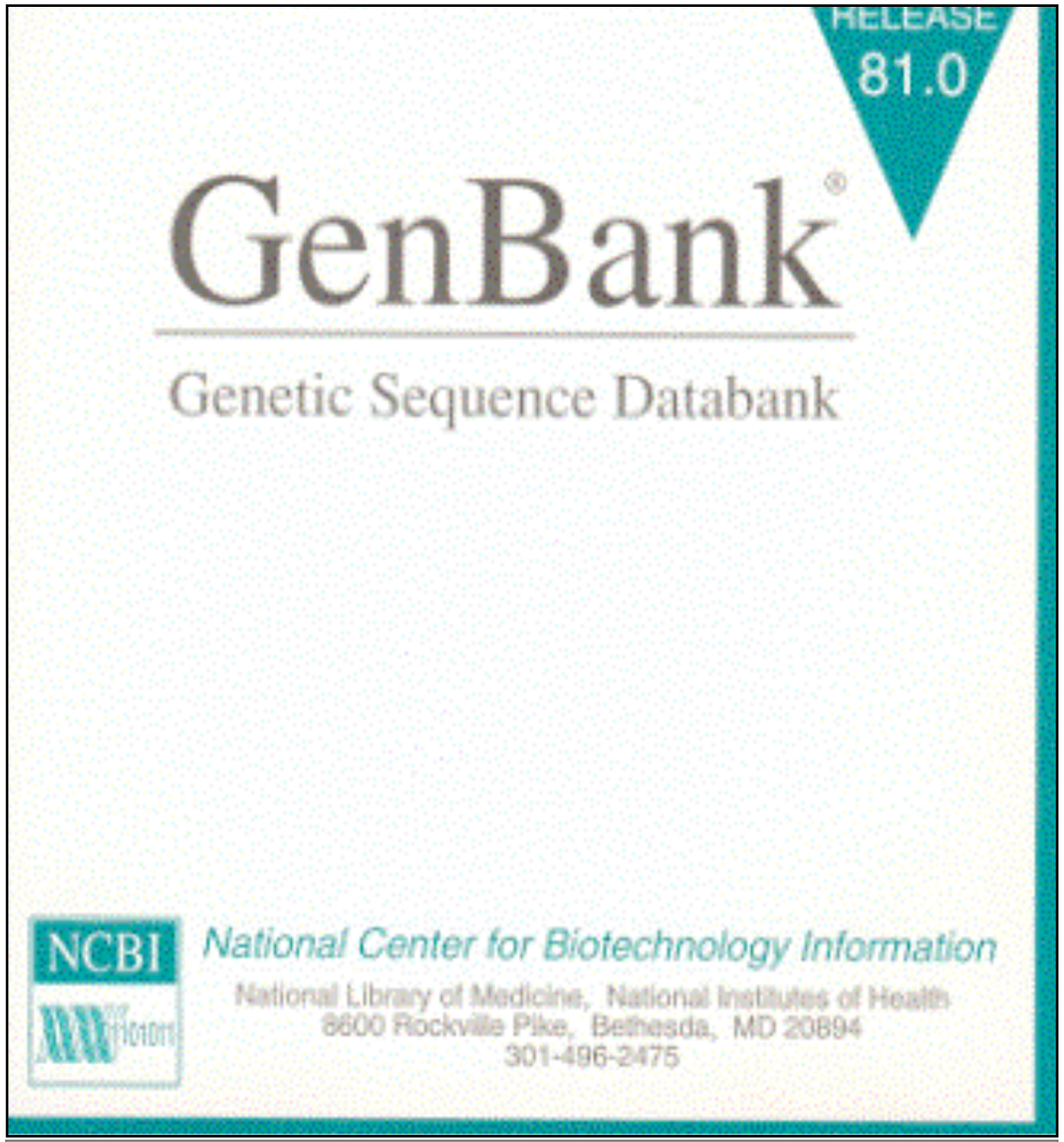
International Sequence Database Collaboration

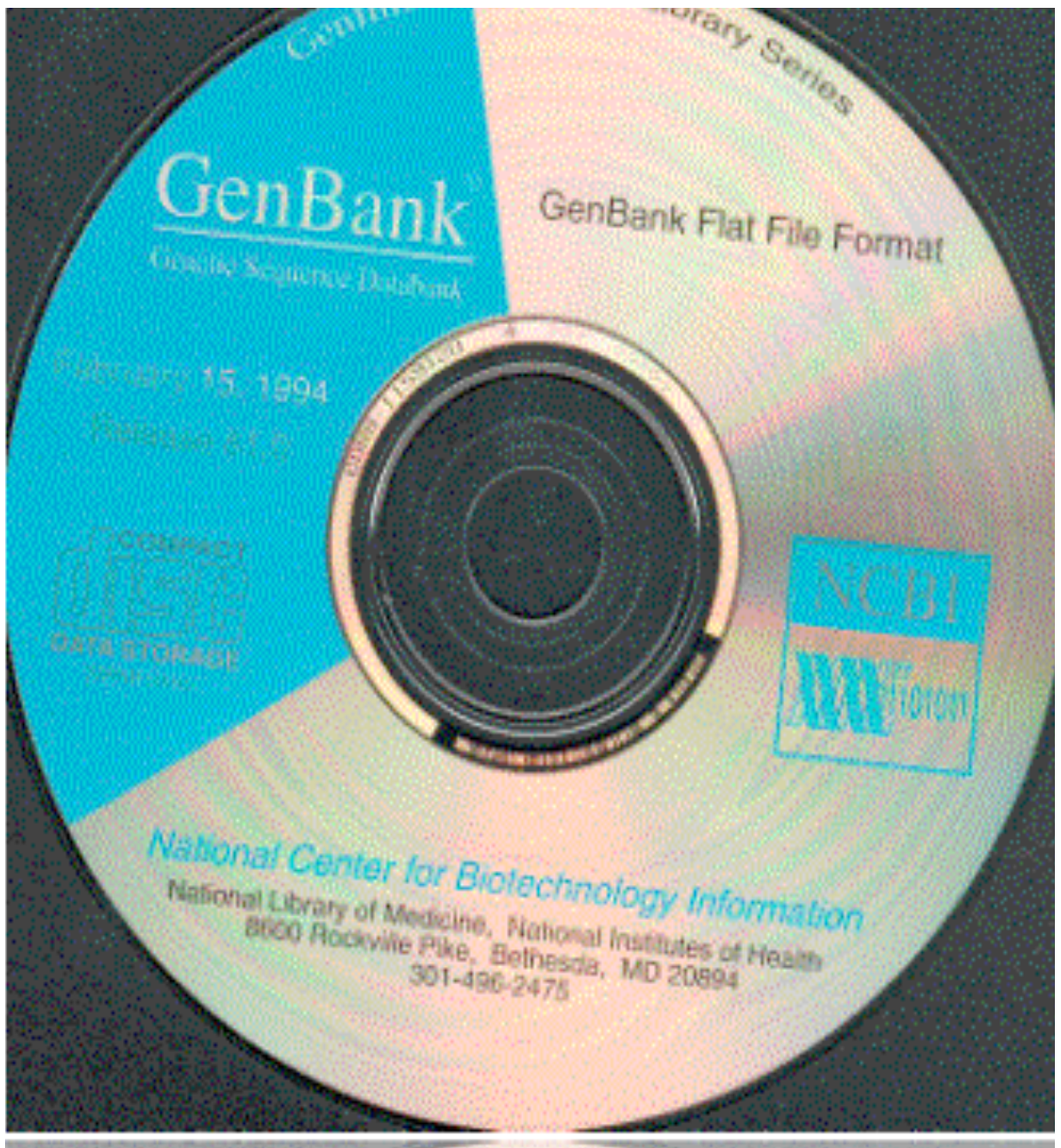
- submit anywhere
- daily updates

GenBank

Entrez

NCBI





GenBank: NCBI's Primary Sequence Database

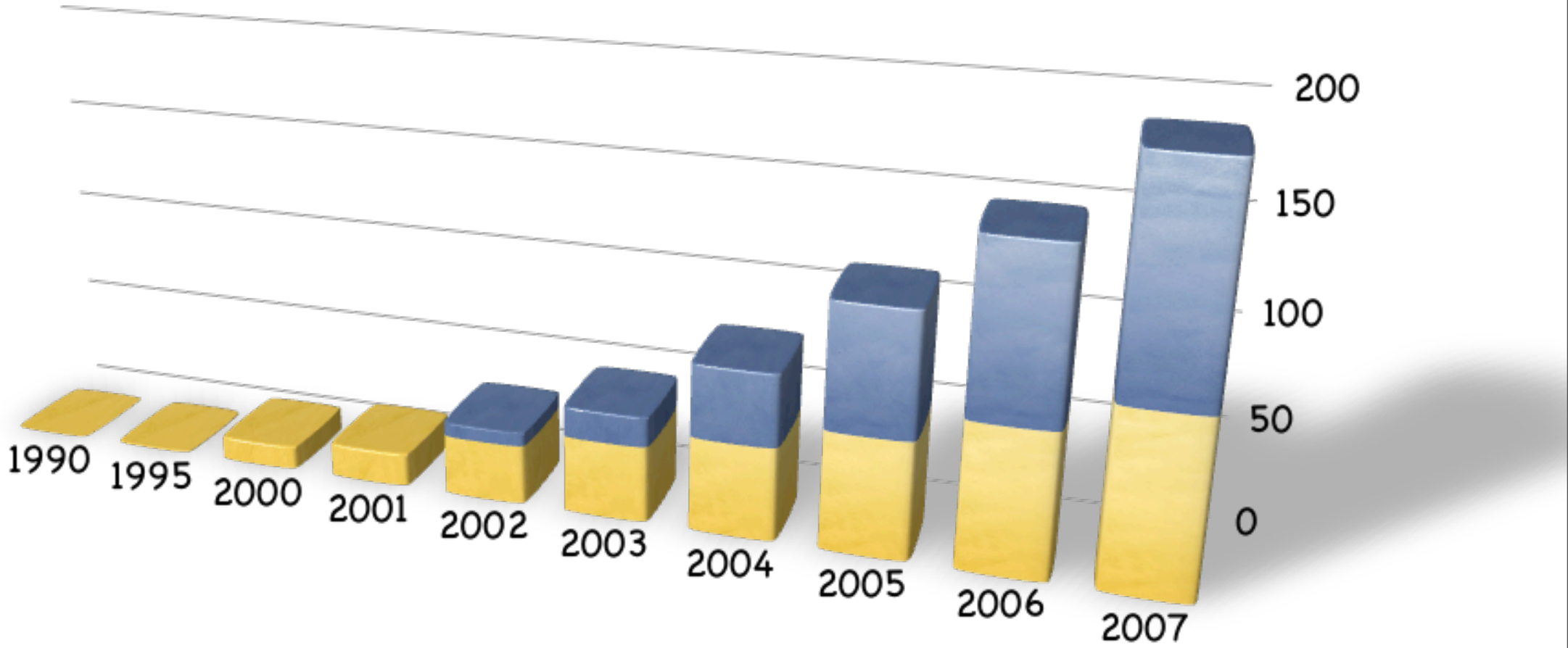
ftp://ftp.ncbi.nih.gov/genbank/

Release 177	April 2010
177,473,850	Records
279,884,898,285*	Total Bases

*includes WGS

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

Growth of GenBank



Current Release 177

Doubling time 12-14 months

GenBank WGS

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
            Fraser,C.M., Adams,M.D., Venter,J.C.                    .M.,
            Watson,P., Lynch,H.T., Peltomaki,P.,                    .,
            Kinzler,K.W. and Vogelstein,B.
  TITLE     Mutation of a mutL homolog in heredit
  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

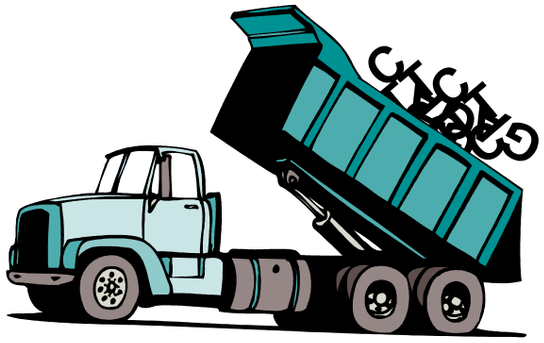
```
FEATURES             Location/Qualifiers
     source          1..2503
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="3"
                     /map="p21"
                     /tissue_type="gall bladder"
                     /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 ttcacagaag atggaaatat cctgcagctt gctaacctgc
     2281 ctgatctata caaagtcttt gagaggtgtt aaatatggtt atttatgcac tgtgggatgt
     2341 gttcttcttt ctctgtattc cgatacaaag tgttgtatca aagtgtgata tacaaagtgt
     2401 accaacataa gtgttggtag cacttaagac ttatacttgc cttctgatag tattccttta
     2461 tacacagtgg attgattata aataaataga tgtgtcttaa cat
//
```

the sequence is the data

Primary vs. Derivative Databases

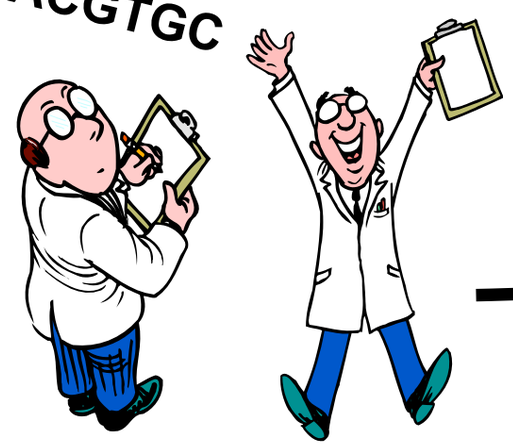


Sequencing Centers

TATAGCCG

ACGTGC

GenBank



ACGTGC

Labs

Algorithms

Updated continually by NCBI

RefSeq: Annotation Pipeline

RefSeq: Entrez Gene and Genomes Pipelines

TATAGCCG
AGCTCCGATA
CCGATGACAA

Curators

Derivative Databases

GenPept

- GenBank CDS translations

```
FEATURES             Location/Qualifiers
     source          1..2484
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="3"
                     /map="3p22-p23"
     gene            1..2484
                     /gene="MLH1"
     CDS             22..2292
                     /gene="MLH1"
                     /note="homol
                     Number P14242), S. cerevisiae MLH1 (GenBank Accession
                     Number U07187), E. coli MUTL (Swiss-Prot Accession Number
                     P23367), Salmonella typhimurium MUTL (Swiss-Prot Accession
                     Number P14161) and Streptococcus pneumoniae (Swiss-Prot
                     Accession Number 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.

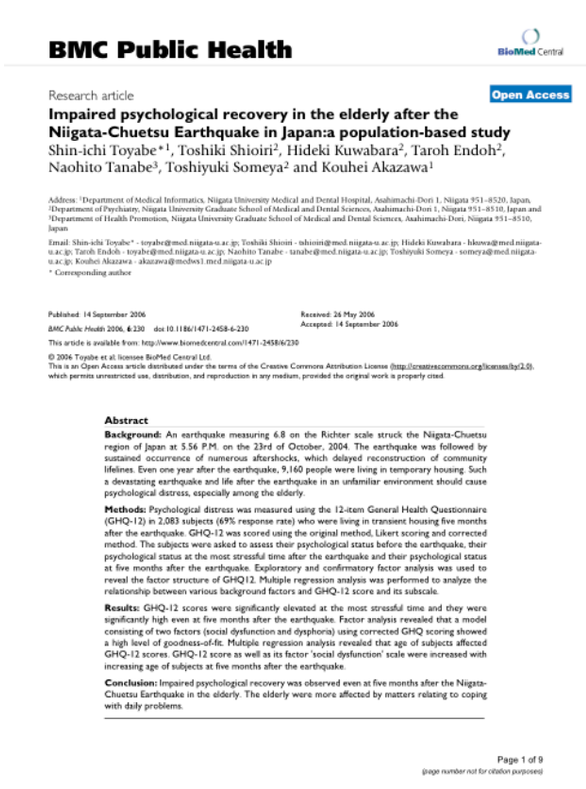

SCIENTIFIC AMERICAN

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@medsv1.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

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.

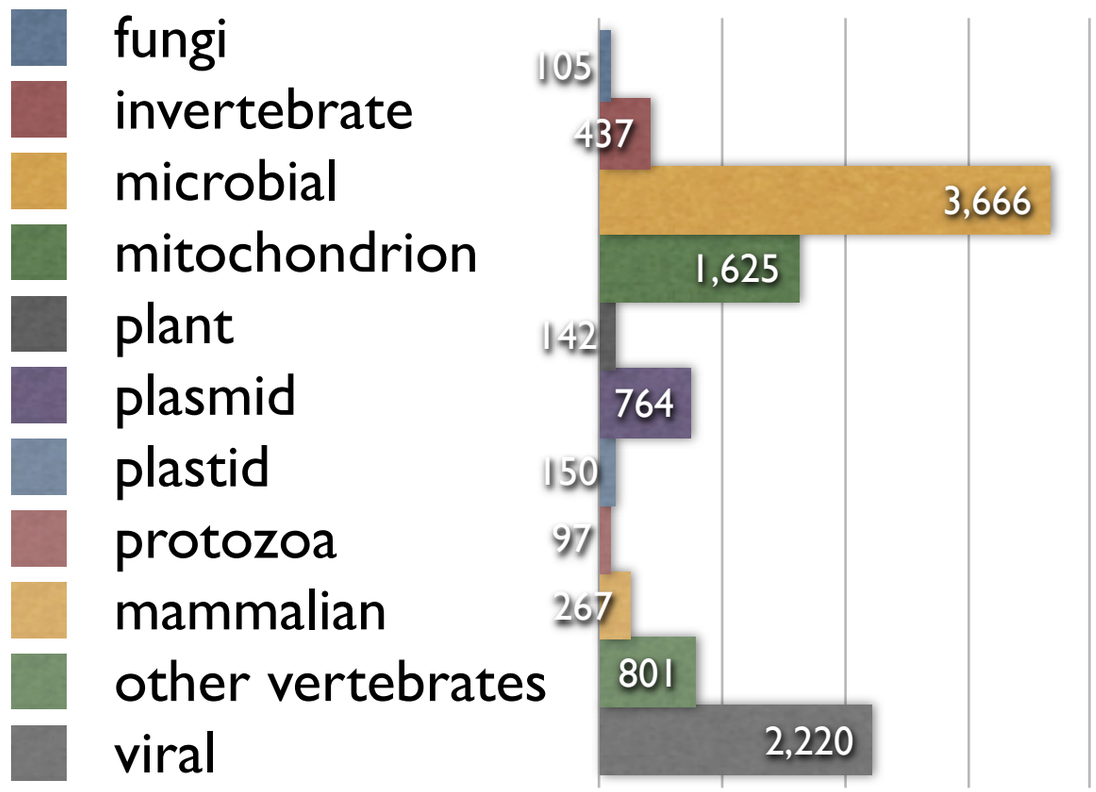
RefSeq

- includes species ranging from viral to microbial to eukaryotic, 10,000+ species (current release RefSeq 41)

- 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

Table 1. The Entrez Databases

(# records as of 8/14/2009)

★ Nucleotide (78 783 103)
EST (62 838 170)
PubChem Substance (61 056 228)
SNP (59 806 469)
GEO Profiles (42 751 725)
Protein (28 475 324)
GSS (25 787 403)
PubChem Compound (25 668 433)
★ PubMed (19 076 621)
Probe (10 187 129)
☆ Gene (6 261 420)
UniGene (3 645 645)
PubMed Central (1 834 865)
NLM Catalog 1 (394 522)
Taxonomy (525 252)
UniSTS (524 629)
Protein Clusters (413 052)
3D Domains (280 897)
☆ Books (237 535)

MeSH (211 794)
Cancer Chromosomes (134 570)
☆ Homologene (123 767)
PopSet (101 569)
Biosystems (96 559)
GENSAT (91 458)
dbGaP (62 335)
☆ Structure (59 329)
CDD (34 735)
Journals (23 939)
GEO Datasets (21 358)
OMIM (20 548)
Site Search (25 070)
Genome (10 777)
☆ SRA (6562)
Projects (5234)
OMIA (2599)
PubChem Bioassay (1691)
Peptidome (79)

adapted from Sayers et al NAR 2010

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)
SRA:	sequence read archive	next generation sequencing data
Gene:	gene records	unified searchable database of genes
HomoloGene:	homologs	neighboring function for Gene

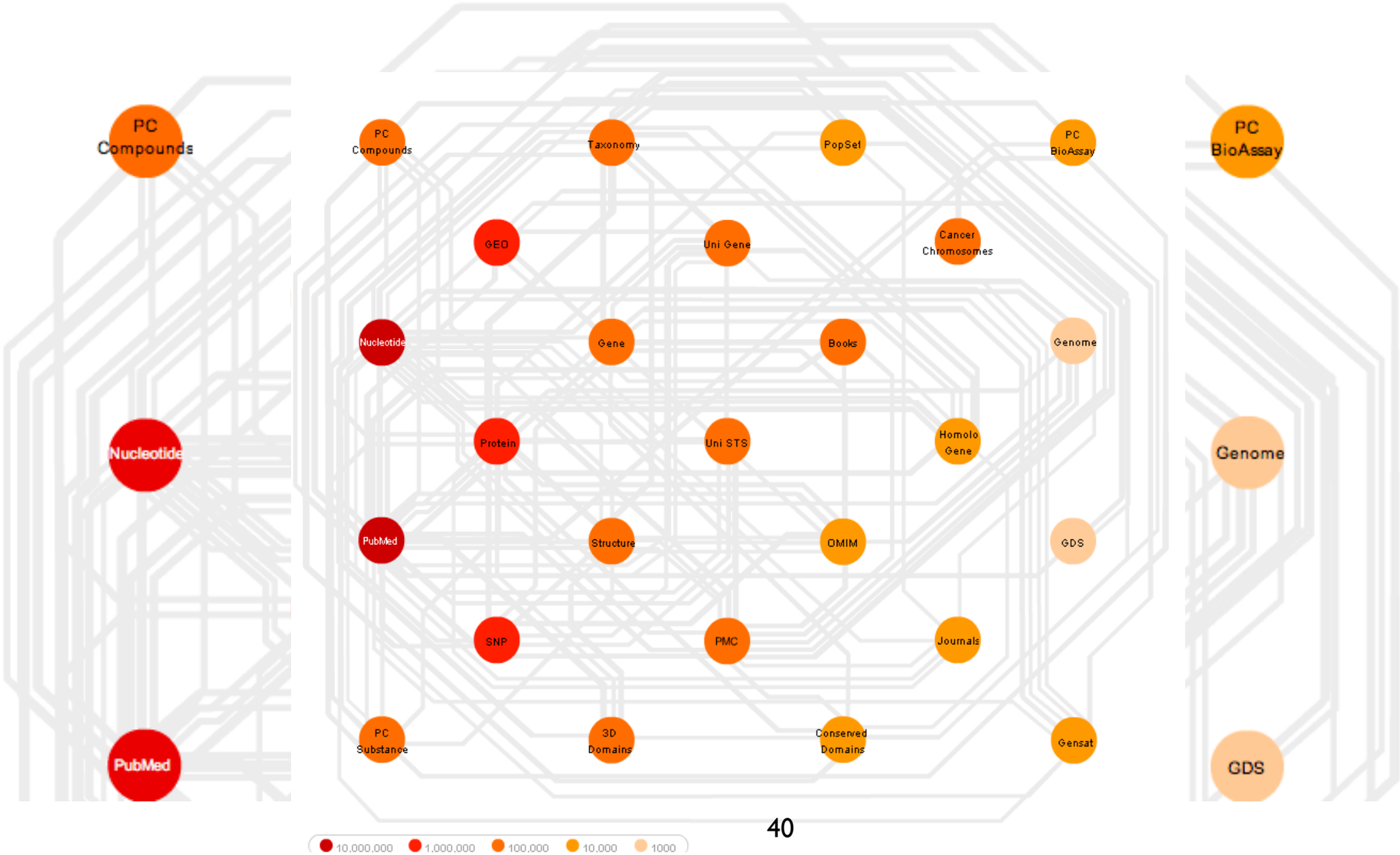
http://www.ncbi.nih.gov/Database/datamodel

NCBI Model of Entrez Databases

Search Entrez for Go

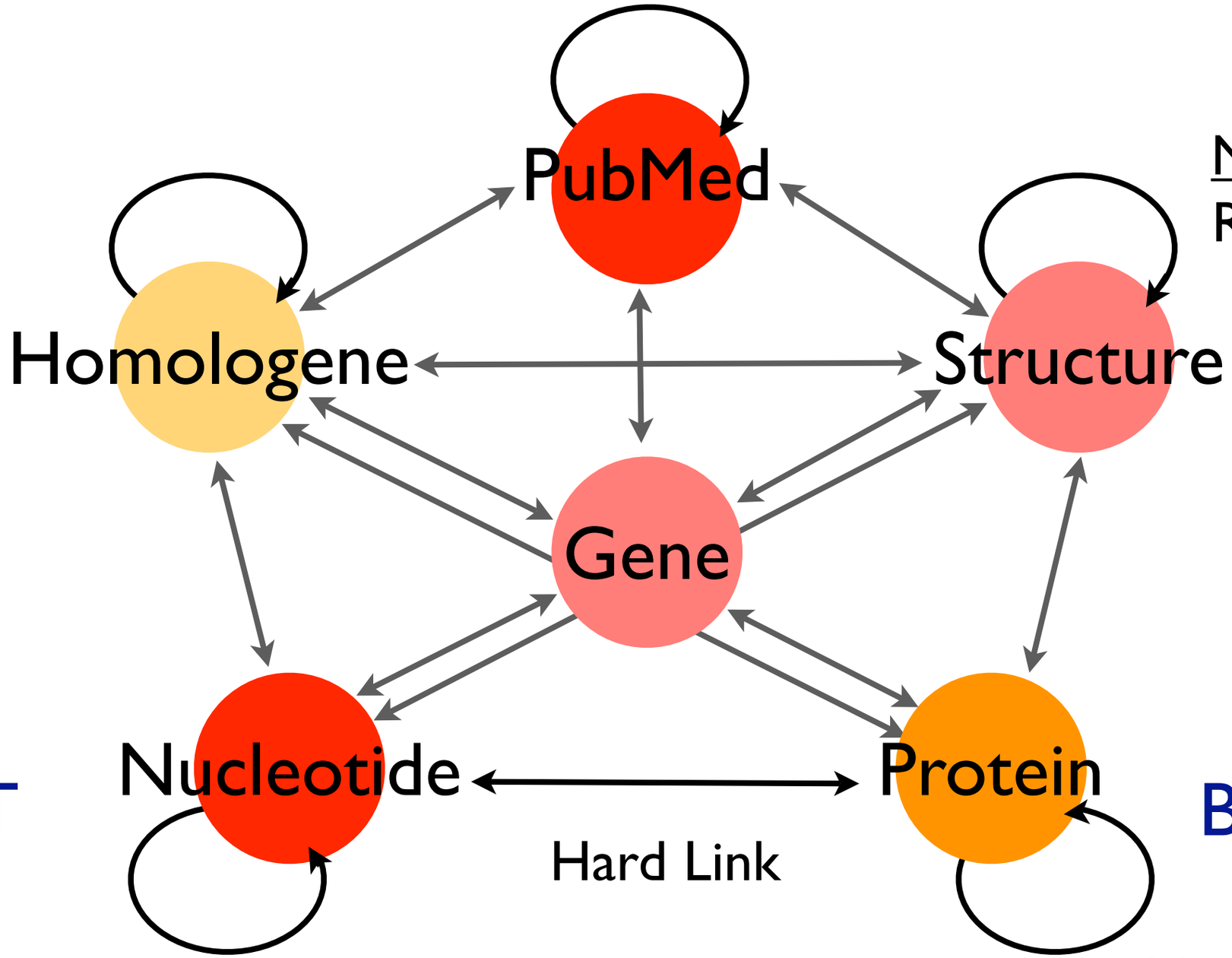
The diagram shows the Entrez databases and the connections between them. Each database is represented by a colored circle, where the color indicates the approximate number of records in the database. Mouse over a circle to see which databases are linked to the one selected, and how many links exist between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

This diagram requires Flash for viewing.



Word weight

Neighbors
Related Articles



Neighbors
Related Structures

VAST

BLAST

Neighbors
Related Sequences

BLAST

Neighbors
Related Sequences
Blink
Domains

NEW! NCBI Discovery Initiatives

- Sensors
- Database "ADs"

All: 14 | Review: 4

We found 3 articles in Nature 2001 by Lander:

Linkage disequilibrium in the human genome. Reich DE et al. *Nature*. (2001)

A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Sachidanandam R et al. *Nature*. (2001)

Initial sequencing and analysis of the human genome. Lander ES et al. *Nature*. (2001)

Items 1 - 14 of 14

1: Automation, parallelism, and robotics for proteomics.
Alterovitz G, Liu J, Chow J, Ramoni MF.
Proteomics. 2006 Jul;6(14):4016-22. Review.
PMID: 16786489 [PubMed - indexed for MEDLINE]
Related Articles

Are you looking for a sequence?
Result for term X51362 found in the Nucleotide database
- Human mRNA for dopamine D2 receptor [Homo sapiens]

Items 1 - 2 of 2 One page.

1: Sequence specific binding of cytosolic proteins to a 12 nucleotide sequence in the 5' untranslated region of FMR1 mRNA.
Iber H.
Biochim Biophys Acta. 1996 Dec 11;1309(3):167-73.
PMID: 8982249 [PubMed - indexed for MEDLINE]
Related Articles

2: Human retina D2 receptor cDNAs have multiple polyadenylation sites and differ from a pituitary clone at the 5' non-coding region.
Robakis NK, Mohamadi M, Fu DY, Sambamurti K, Refolo LM.
Nucleic Acids Res. 1990 Mar 11;18(5):1299. No abstract available.
PMID: 2138729 [PubMed - indexed for MEDLINE]
Related Articles Free article in PMC | at journal site

NCBI Reference Sequence: NM_001133.2

Homo sapiens afamin (AFM), mRNA

Comment Features Sequence

```
LOCUS       NM_001133               1997 bp    mRNA    linear   PRI 05-MAR-2010
DEFINITION  Homo sapiens afamin (AFM), mRNA.
ACCESSION   NM_001133
VERSION     NM_001133.2  GI:27754774
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 1997)
  AUTHORS   Dieplinger,H., Ankerst,D.P., Burges,A., Lenhard,M., Lingenhel,A.,
            Fineder,L., Buchner,H. and Stieber,P.
  TITLE     Afamin and apolipoprotein A-IV: novel protein markers for ovarian
            cancer
  JOURNAL   Cancer Epidemiol. Biomarkers Prev. 18 (4), 1127-1133 (2009)
   PUBMED   19336561
  REMARK    GeneRIF: Reduced Afamin expression is associated with ovarian
            cancer.
REFERENCE   2  (bases 1 to 1997)
  AUTHORS   Kratzer,I., Bernhart,E., Wintersperger,A., Hammer,A., Waltl,S.,
            Malle,E., Sperk,G., Wietzorrek,G., Dieplinger,H. and Sattler,W.
  TITLE     Afamin is synthesized by cerebrovascular endothelial cells and
            mediates alpha-tocopherol transport across an in vitro model of the
            blood-brain barrier
  JOURNAL   J. Neurochem. 108 (3), 707-718 (2009)
   PUBMED   19046407
  REMARK    GeneRIF: afamin might be a new family member of binding/transport
            proteins contributing to alpha-tocopherol homeostasis at the
            blood-brain barrier
REFERENCE   3  (bases 1 to 1997)
  AUTHORS   Ramachandran,P., Boontheung,P., Xie,Y., Sondej,M., Wong,D.T. and
            Loo,J.A.
  TITLE     Identification of
            glycoprotein captu
  JOURNAL   J. Proteome Res.
   PUBMED   16740002
REFERENCE   4  (bases 1 to 19
  AUTHORS   Hu,Y., Malone,J.P
  TITLE     Comparative protec
            variation in humar
```

Change Region Shown
Customize View

Analyze This Sequence
- Run BLAST
- Pick Primers

Articles about the AFM gene
- Afamin and apolipoprotein A-IV: novel protein mar [Cancer Epidemiol Biomarkers Prev. 2009]
- Afamin is synthesized by cerebrovascular endothelial cells and medi [J Neurochem. 2009]
- Identification of N-linked glycoproteins in human saliva by glycoprotein c [J Proteome Res. 2006]

» See all...

RefSeq Protein Product
See the reference protein sequence for afamin precursor (NP_001124.1).

More about the AFM gene
This gene is a member of the albumin gene family, which is comprised of four genes that localize to chromosome 4 in a tandem arrangement. Th...
Also Known As: ALB2, ALBA, ALF, MGC125...

Homologs of the AFM gene
...erved in chimpanzee, dog,

...record

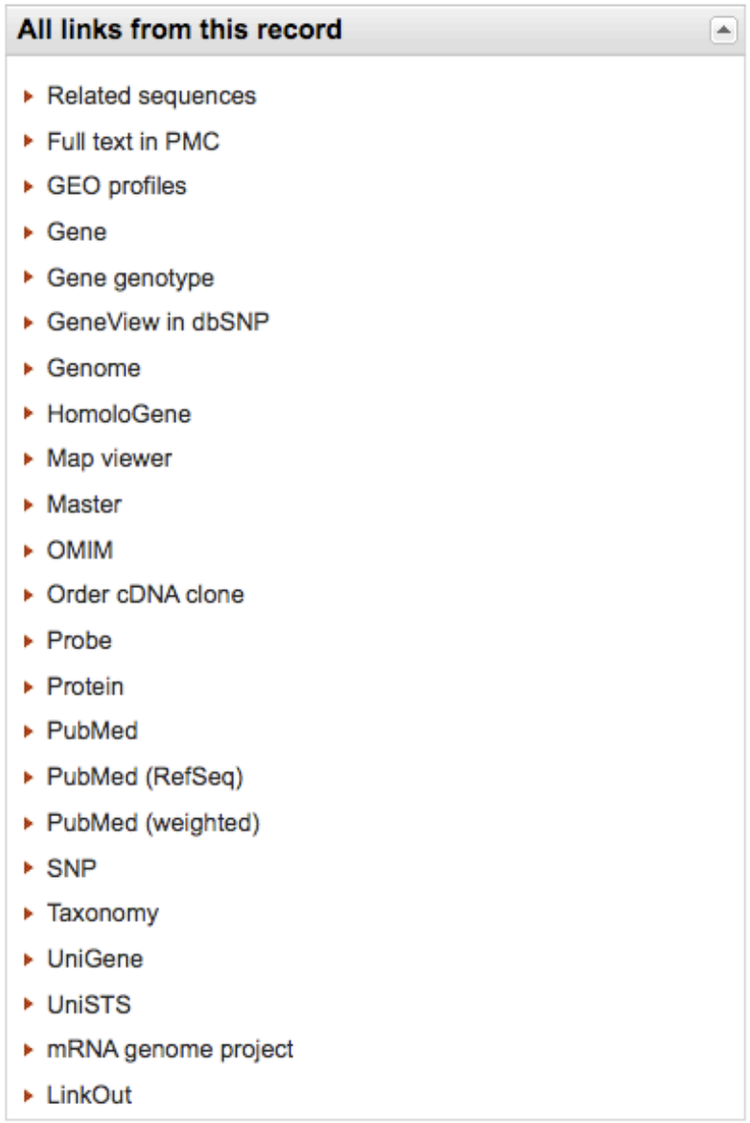
RefSeq Protein Product
See the reference protein sequence for afamin precursor (NP_001124.1).

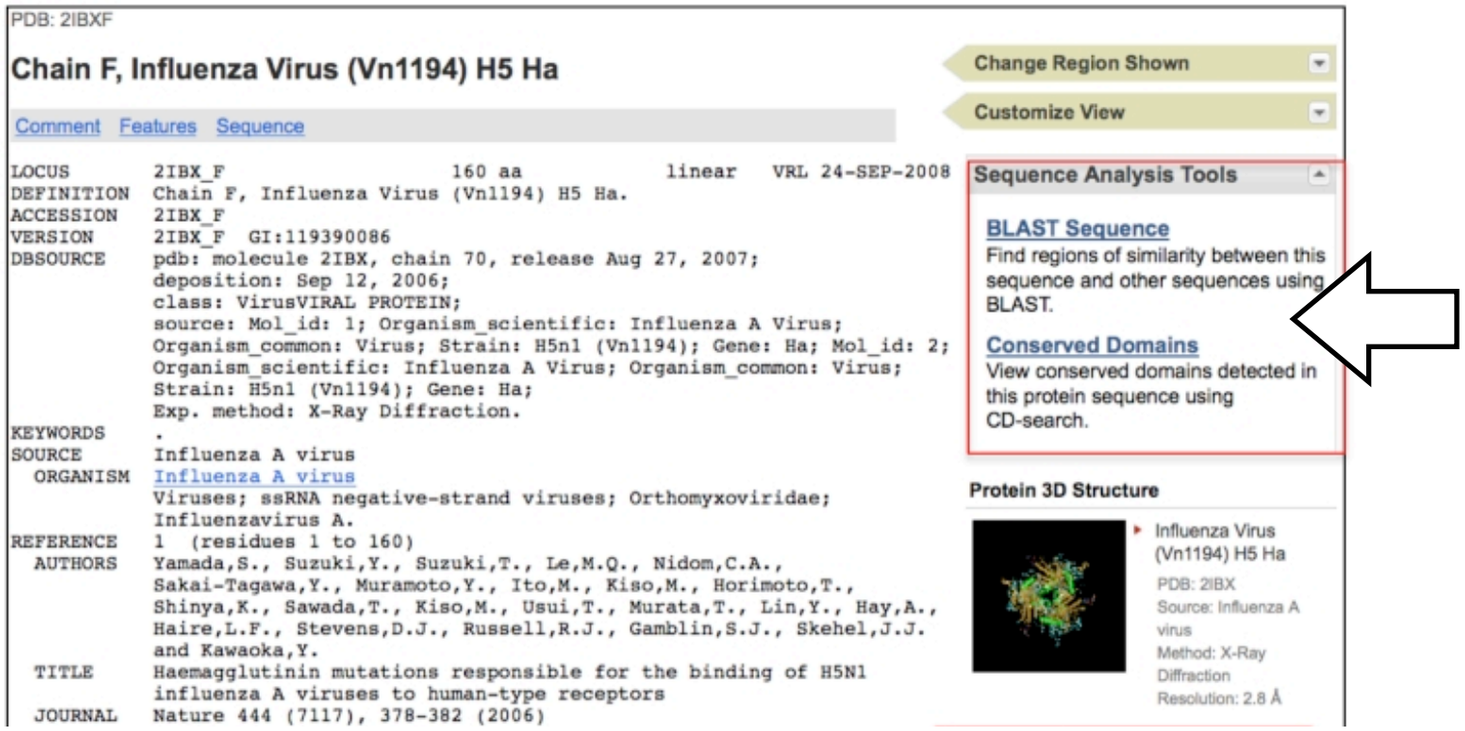
More about the AFM gene
This gene is a member of the albumin gene family, which is comprised of four genes that localize to chromosome 4 in a tandem arrangement. Th...
Also Known As: ALB2, ALBA, ALF, MGC125...

Homologs of the AFM gene
The AFM gene is conserved in chimpanzee, dog, cow, mouse, and rat.

NEW! NCBI Discovery Initiatives

- Easier Access to Links
- Analysis Tools





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...

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).

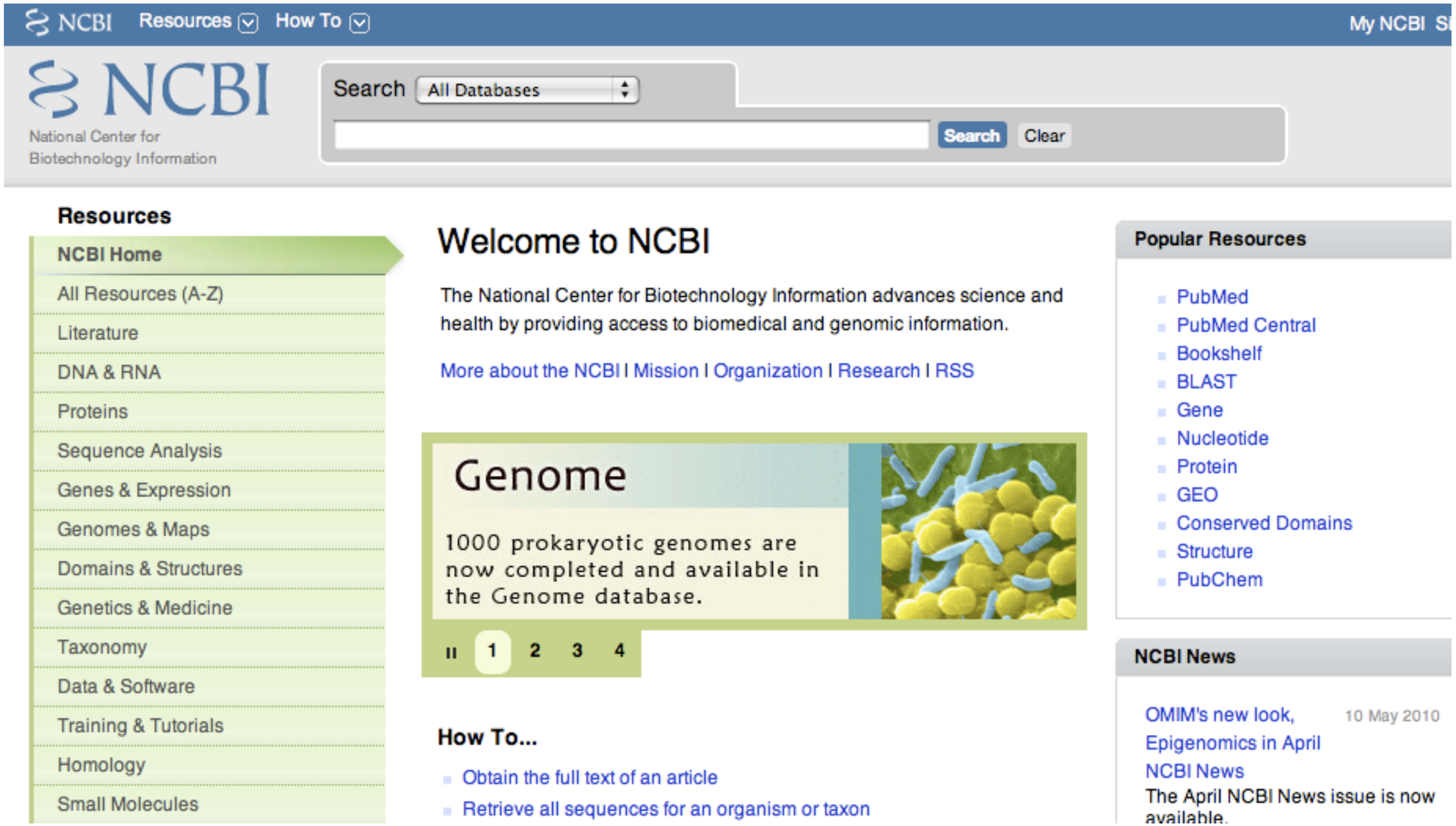
Database searching with Entrez

- **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

✓ Using NCBI's Discovery Components to explore links & neighbors of 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

85440	**PubMed:** biomedical literature citations and abstracts	1018	**Books:** online books
15256	**PubMed Central:** free, full text journal articles	464	**OMIM:** online Mendelian Inheritance in Man
11	**Site Search:** NCBI web and FTP sites	1	**OMIA:** online Mendelian Inheritance in Animals

21768	**Nucleotide:** Core subset of nucleotide sequence records	152	**dbGaP:** genotype and phenotype
1161	**EST:** Expressed Sequence Tag records	200	**UniGene:** gene-oriented clusters of transcript sequences
none	**GSS:** Genome Survey Sequence records	13	**CDD:** conserved protein domain database
1314	**Protein:** sequence database	22	**3D Domains:** domains from Entrez Structure
8	**Genome:** whole genome sequences	34	**UniSTS:** markers and mapping data
28	**Structure:** three-dimensional macromolecular structures	1	**PopSet:** population study data sets
none	**Taxonomy:** organisms in GenBank	86766	**GEO Profiles:** expression and molecular abundance profiles
16	**SNP:** single nucleotide polymorphism	157	**GEO DataSets:** experimental sets of GEO data
11	**dbVar:** Genomic structural variation	162	**Cancer Chromosomes:** cytogenetic databases
869	**Gene:** gene-centered information	367	**PubChem BioAssay:** bioactivity screens of chemical substances

Human Disease Genes

NCBI

OMIM
Online Mendelian Inheritance in Man

Johns Hopkins University

My NCBI [Sign In] [Register]

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

All: 1 OMIM UniSTS: 0 OMIM dbSNP: 1

MIM *120436

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

MGI, GeneTests, Links

Gene map locus: 3p21.3

Clinical Synopsis

Description

Back to Top

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

Back to Top

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

Table of Contents

Links

Selected Gene Related Links
- Entrez Gene
- Nomenclature

Search Nucleotide

NCBI Entrez Nucleotide

My NCBI [Sign In] [Register]

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

Search Nucleotide for colon cancer Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Found 22929 nucleotide sequences. Nucleotide [21768] EST [1161]

Display Summary Show 20 Sort By Send to

All: 21768 | Bacteria: 11 | INSDC (GenBank): 21077 | RefSeq: 691 | mRNA: 955

This search in Gene shows 789 results, including:

- **PTPRJ** (*Homo sapiens*): protein tyrosine phosphatase, receptor type, J
- **MLH3** (*Homo sapiens*): mutL homolog 3 (E. coli)
- **MSH2** (*Homo sapiens*): mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)

Items 1 - 20 of 21768 Page 1 of 1089 Next

1. PREDICTED: Callithrix jacchus serologically defined colon cancer antigen 8 (SDCCAG8), mRNA
   2,453 bp linear mRNA
   XM_002760861.1 GI:296230841

2. PREDICTED: Callithrix jacchus mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), mRNA
   2,578 bp linear mRNA
   XM_002759730.1 GI:296228348

Top Organisms [Tree]

- Homo sapiens *(14022)*
- synthetic construct *(3612)*
- unidentified *(2719)*
- Mus musculus *(151)*
- Rattus norvegicus *(48)*
- All other taxa *(305)*

More...

Recent activity

Turn Off Clear

colon cancer (21768) Nucleotide

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

» See more...

Advanced Search Options

NCBI Entrez Nucleotide

My NCBI [Sign In] [Register]

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

Search Nucleotide for colon cancer Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Tabs

Found 22929 nucleotide sequences. Nucleotide [21768] EST [1161]

Summary Show 20 Sort By Send to

21768 | Bacteria: 11 | INSDC (GenBank): 21077 | RefSeq: 691 | mRNA: 955

search in Gene shows 789 results, including:

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

MLH3 (*Homo sapiens*): mutL homolog 3 (E. coli)

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

Items 1 - 20 of 21768 Page 1 of 1089 Next

1. PREDICTED: Callithrix jacchus serologically defined colon cancer antigen 8 (SDCCAG8), mRNA
   2,453 bp linear mRNA
   XM_002760861.1 GI:296230841

2. PREDICTED: Callithrix jacchus mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), mRNA
   2,578 bp linear mRNA
   XM_002759730.1 GI:296228348

Top Organisms [Tree]

Homo sapiens *(14022)*
synthetic construct *(3612)*
unidentified *(2719)*
Mus musculus *(151)*
Rattus norvegicus *(48)*
All other taxa *(305)*
More...

Recent activity

Turn Off Clear

colon cancer (21768) Nucleotide

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

» See more...

NCBI

Entrez Nucleotide

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

Search Nucleotide 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

Related resources

BLAST

Reference sequence project

Search for Genes

Submit to GenBank

Search for full length cDNAs

- Use All Fields pull-down menu to specify a field.
- Boolean operators AND, OR, NOT must be in upper case.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available here.

Limited to:

All Fields

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

ng draft TPA patents

Gene Location: Any

nces:

Only from: Any

st:

st:

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

NCBI Entrez Nucleotide

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search Nucleotide for colon cancer Go Clear Save Search

Limits Preview/Index History Clipboard Details

Tabs

Found 22929 nucl de sequences. Nucleotide [21768] EST [1161]

Display Summary Show 20 Sort By Send to

All: 21768 B ia: 11 INSDC (GenBank): 21077 RefSeq: 691 mRNA: 955

This search in Gene shows 789 results, including:

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

MLH3 (*Homo sapiens*): mutL homolog 3 (E. coli)

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

Items 1 - 20 of 21768 Page 1 of 1089 Next

1. PREDICTED: Callithrix jacchus serologically defined colon cancer antigen 8 (SDCCAG8), mRNA
2,453 bp linear mRNA
XM_002760861.1 GI:296230841

2. PREDICTED: Callithrix jacchus mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), mRNA
2,578 bp linear mRNA
XM_002759730.1 GI:296228348

Top Organisms [Tree]

Homo sapiens *(14022)*
synthetic construct *(3612)*
unidentified *(2719)*
Mus musculus *(151)*
Rattus norvegicus *(48)*
All other taxa *(305)*
More...

Recent activity

Turn Off Clear

colon cancer (21768) Nucleotide

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

» See more...

NCBI

Entrez Nucleotide

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

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

Limits | Preview/Index | History | Clipboard | Details

Field: **Title** Limits: **Molecule: mRNA, Only from: 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	Ti
#18	Search **colon cancer AND nonpolyposis** Field: **Title** Limits: **Molecule: mRNA, Only from: RefSeq**	16:4
#15	Search **colon cancer AND nonpolyposis** Field: **Title**	16:4
#30	Search **colon cancer** Field: **Title** Limits: **Molecule: mRNA, Only from: RefSeq**	16:4

About Entrez

Entrez Nucleotide

Help | FAQ

Entrez Tools

Check sequence

revision history

LinkOut

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

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.

Organism Preview Index

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

Refining your Search

NC
All Databa
Search Nucle

CBI
In] [Register]

Limits | Preview/Index | History | Clipboard | Details

Found 5 nucleotide sequences. Nucleotide [5]

Display Summary | Show 20 | Sort By | Send to

All: 5 | Bacteria: 0 | INSDC (GenBank): 0 | RefSeq: 5 | mRNA: 5

Items 1 - 5 of 5 One page.

1. Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), transcript variant 4, mRNA
2,386 bp linear mRNA
NM_001167619.1 GI:263191732

2. Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), transcript variant 3, mRNA
2,473 bp linear mRNA
NM_001167618.1 GI:263191712

3. Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), transcript variant 1, mRNA
2,662 bp linear mRNA
NM_000249.3 GI:263191547

4. Homo sapie
3,145 bp lin
NM_000251

5. Homo sapie
2, mRNA
2,473 bp linear mRNA

Recent activity

Turn Off Clear

- colon cancer[Title] AND n... (5)
- colon cancer AND nonpolyp... (10)
- colon cancer AND nonpolyp... (39)
- colon cancerANDnonpolypos... (0)
- colon%20cancer%20AND%20no... (0) Nucleotide

» See more...

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 [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

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

Limits Preview/Index History Clipboard Details

Found 5 nucleotide sequences. Nucleotide [5]

Display Summary Show 20 Sort By Send to

All: 5 Bacteria: 0 INSDC (GenBank): 0 RefSeq: 5 mRNA: 5

Items 1 - 5 of 5 One page.

1. Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), transcript variant 4, mRNA
2,386 bp linear mRNA
NM_001167619.1 GI:263191732

2. Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), transcript variant 3, mRNA
2,473 bp linear mRNA
NM_001167618.1 GI:263191712

3. Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), transcript variant 1, mRNA
2,662 bp linear mRNA
NM_000249.3 GI:263191547

4. Homo sapiens mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli) (MSH2), mRNA
3,145 bp linear mRNA
NM_000251.1 GI:4557760

5. Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), transcript variant 2, mRNA
2,473 bp linear mRNA

Recent activity

Turn Off Clear

colon cancer[Title] AND n... (5)
colon cancer AND nonpolyp... (10)
colon cancer AND nonpolyp... (39)
colon cancerANDnonpolypos... (0)
colon%20cancer%20AND%20no... (0) Nucleotide

» See more...

NCBI Entrez Nucleotide

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Format: **GenBank** FASTA Graphics More Formats▼ Download▼ Save▼ Links▼

NCBI Reference Sequence: NM_000249.3

Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), transcript variant 1, mRNA

Comment Features Sequence

```
LOCUS       NM_000249               2662 bp    mRNA    linear   PRI 16-MAY-2010
DEFINITION  Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E.
            coli) (MLH1), transcript variant 1, mRNA.
ACCESSION   NM_000249
VERSION     NM_000249.3  GI:263191547
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 2662)
  AUTHORS   Alvarez,K., Hurtado,C., Hevia,M.A., Wielandt,A.M., de la Fuente,M.,
            Church,J., Carvallo,P. and Lopez-Kostner,F.
  TITLE     Spectrum of MLH1 and MSH2 mutations in Chilean families with
            suspected Lynch syndrome
  JOURNAL   Dis. Colon Rectum 53 (4), 450-459 (2010)
   PUBMED   20305446
  REMARK    GeneRIF: 21 Chilean families with Lynch syndrome showed 6 mutations
            in MLH1.
REFERENCE   2  (bases 1 to 2662)
  AUTHORS   Vasen,H.F., Abdirahman,M., Brohet,R., Langers,A.M.,
            Kleibeuker,J.H., Kouwen,M.V., Koornstra,J.J., Boot,H., Cats,A.,
            Dekker,E., Sanduleanu,S., Poley,J.W., Hardwick,J.C., Cappel,W.H.,
            Jong,A.E., Tan,T.G., Jacobs,M., Mohamed,F.A., Boer,S.Y.,
            Meeberg,P.C., Verhulst,M.L., Salemans,J.M., Bentem,N.V.,
            Westerveld,B.D., Vecht,J. and Nagengast,F.M.
```

Change Region Shown

Customize View

Analyze This Sequence

- Run BLAST
- Pick Primers

Articles about the MLH1 gene

- [Mismatch repair gene hMLH1 A655G/A [Zhonghua Wei Chang Wai Ke Za Zhi. 2010]
- Spectrum of MLH1 and MSH2 mutations in Chilean famil [Dis Colon Rectum. 2010]
- Prognostic relevance of MLH1 and MSH2 mutations in hereditary nor [Tumori. 2009]

» See all...

RefSeq Alternative Splicing

See 4 reference mRNA sequence splice variants for the MLH1 gene.

RefSeq Protein Product

See the reference protein sequence for DNA mismatch repair protein Mlh1 isoform 1 (NP_000240.1).

More about the MLH1 gene

This gene was identified as a locus frequently mutated in hereditary nonpolyposis colon cancer (HNPCC). It is a

disease association. (HuGE

prior to print

li,I., Radice,P.,
Fornasarig,M., Santarosa,M.,
son,M.P., Lucci-Cordisco,E.,
V., Cama,A., Curia,M.C., de
. and Bertario,L.
mutations in hereditary
nts

disease association. (HuGE

33. Pedroni, Maurizio

es,N.C., Chang,S.Y.,
H.J., Mecklin,J.P.,

ecombination in hereditary

Sibert,L., Moreau,V. and

r RNA in human normal cells

, Narc

association between
sms: the Muir-Torre syndrome

All links from this record

- Related sequences
- Components(Core)
- Components(EST)
- Full text in PMC
- Gene
- GeneView in dbSNP
- HomoloGene
- Master
- OMIM
- Probe
- Protein
- PubMed
- PubMed (RefSeq)
- PubMed (weighted)
- SNP
- Taxonomy
- UniGene
- UniSTS

NCBI Taxonomy entries and classification information for the source organisms of the current set of records.

Taxonomy

NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search for [ ] as [complete name] [x] lock [Go] [Clear]

Display [3] levels using filter: [none]

[ ] Nucleotide [ ] Nucleotide EST [ ] Nucleotide GS
[ ] Popset [ ] SNP [ ] 3D Domains
[ ] UniSTS [ ] PubMed Central [ ] Gene
[ ] LinkOut [ ] BLAST [ ] TRACE

Lineage (full): root; cellular organisms; Eukar
Deuterostomia; Chordata; Craniata; Vertebrat
Amniota; Mammalia; Theria; Eutheria; Euarch
Hominidae; Homininae; Homo

- **Homo sapiens** (human) *Click on organism*
  - **Homo sapiens neanderthalensis** (Nea

All molecular databases

NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search for [ ] as [complete name] [x] lock [Go] [Clear]

Display [3] levels using filter: [none]

Homo sapiens

Taxonomy ID: 9606
Genbank common name: **human**
Inherited blast name: **primates**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:
common name: **man**
authority: **Homo sapiens Linnaeus, 1758**

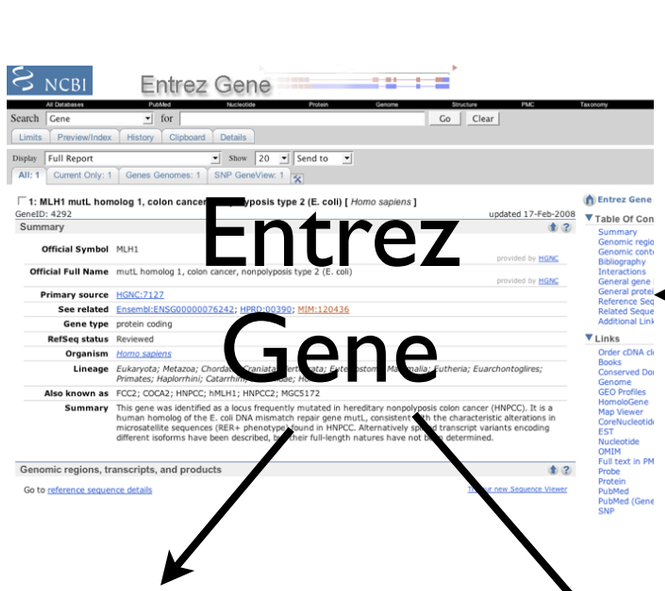
Lineage(full)
cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes;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

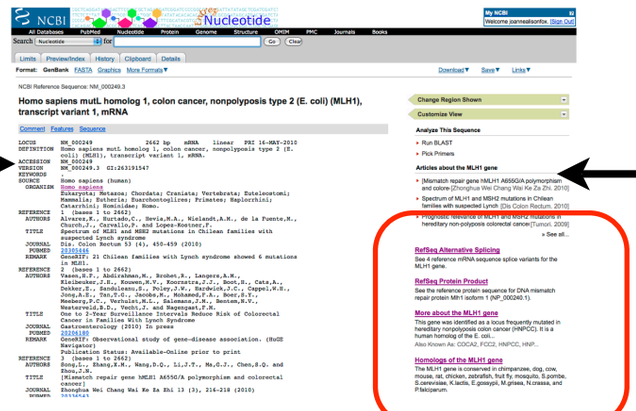
Entrez records

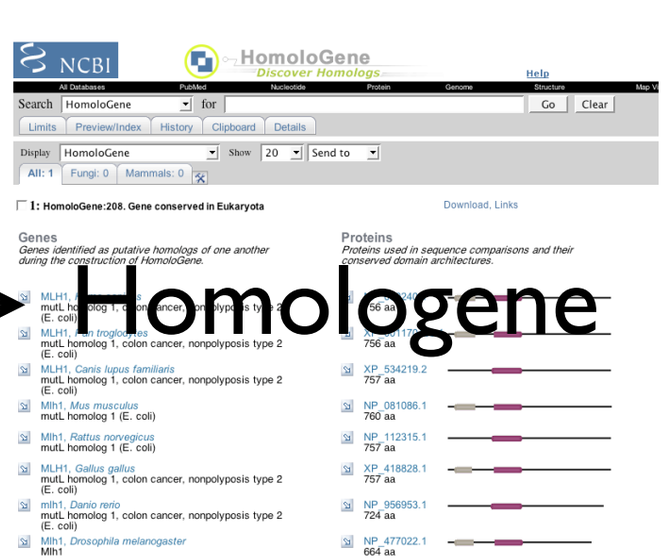
Database name	Subtree links	Direct links
Nucleotide	16,894,199	16,894,174
Nucleotide EST	8,301,471	8,301,471
Nucleotide GSS	1,293,831	1,292,505
Protein	529,402	529,306
Structure	15,545	15,545
Genome Sequences	75	74
Genome Projects	32	32
Popset	18,133	18,133
SNP	29,585,299	29,585,299
3D Domains	61,175	61,175
Domains	8	8
GEO Datasets	9,403	9,403
GEO Expressions	17,689,684	17,689,684
UniGene	123,200	123,200
UniSTS	327,522	327,522
PubMed Central	7,771	7,768

Goal: Investigate MLH1 - function & homologs

- Tip: Use Database Adverts in sidebar of nucleotide entry to navigate to other databases



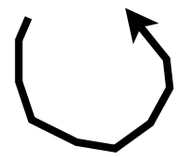




Gene RIF (reference into function)

Other Entrez Databases

Gene neighbors



NCBI Entrez Nucleotide

My NCBI
Welcome joannealisonfox. [Sign Out]

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

Search Nucleotide for [ ] Go Clear

Limits | Preview/Index | History | Clipboard | Details

Format: **GenBank** FASTA Graphics More Formats▼ Download▼ Save▼ Links▼

NCBI Reference Sequence: NM_000249.3

Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), transcript variant 1, mRNA

Comment Features Sequence

```
LOCUS       NM_000249               2662 bp    mRNA    linear   PRI 16-MAY-2010
DEFINITION  Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E.
            coli) (MLH1), transcript variant 1, mRNA.
ACCESSION   NM_000249
VERSION     NM_000249.3  GI:263191547
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 2662)
  AUTHORS   Alvarez,K., Hurtado,C., Hevia,M.A., Wielandt,A.M., de la Fuente,M.,
            Church,J., Carvallo,P. and Lopez-Kostner,F.
  TITLE     Spectrum of MLH1 and MSH2 mutations in Chilean families with
            suspected Lynch syndrome
  JOURNAL   Dis. Colon Rectum 53 (4), 450-459 (2010)
   PUBMED   20305446
  REMARK    GeneRIF: 21 Chilean families with Lynch syndrome showed 6 mutations
            in MLH1.
REFERENCE   2  (bases 1 to 2662)
  AUTHORS   Vasen,H.F., Abdirahman,M., Brohet,R., Langers,A.M.,
            Kleibeuker,J.H., Kouwen,M.V., Koornstra,J.J., Boot,H., Cats,A.,
            Dekker,E., Sanduleanu,S., Poley,J.W., Hardwick,J.C., Cappel,W.H.,
            Jong,A.E., Tan,T.G., Jacobs,M., Mohamed,F.A., Boer,S.Y.,
            Meeberg,P.C., Verhulst,M.L., Salemans,J.M., Bentem,N.V.,
            Westerveld,B.D., Vecht,J. and Nagengast,F.M.
  TITLE     One to 2-Year Surveillance Intervals Reduce Risk of Colorectal
            Cancer in Families With Lynch Syndrome
  JOURNAL   Gastroenterology (2010) In press
   PUBMED   20206180
  REMARK    GeneRIF: Observational study of gene-disease association. (HuGE
            Navigator)
            Publication Status: Available-Online prior to print
REFERENCE   3  (bases 1 to 2662)
  AUTHORS   Song,L., Zhang,X.M., Wang,D.Q., Li,J.T., Ma,G.J., Chen,S.Q. and
            Zhou,J.N.
  TITLE     [Mismatch repair gene hMLH1 A655G/A polymorphism and colorectal
            cancer]
  JOURNAL   Zhonghua Wei Chang Wai Ke Za Zhi 13 (3), 216-218 (2010)
   PUBMED   20336543
```

GENE ⇨

Change Region Shown

Customize View

Analyze This Sequence

- Run BLAST
- Pick Primers

Articles about the MLH1 gene

- [Mismatch repair gene hMLH1 A655G/A polymorphism and colore [Zhonghua Wei Chang Wai Ke Za Zhi. 2010]
- Spectrum of MLH1 and MSH2 mutations in Chilean families with suspected Lynch [Dis Colon Rectum. 2010]
- Prognostic relevance of MLH1 and MSH2 mutations in hereditary non-polyposis colorectal cance [Tumori. 2009]

» See all...

RefSeq Alternative Splicing

See 4 reference mRNA sequence splice variants for the MLH1 gene.

RefSeq Protein Product

See the reference protein sequence for DNA mismatch repair protein Mlh1 isoform 1 (NP_000240.1).

More about the MLH1 gene

This gene was identified as a locus frequently mutated in hereditary nonpolyposis colon cancer (HNPCC). It is a human homolog of the E. coli...

Also Known As: COCA2, FCC2, HNPCC, HNP...

Homologs of the MLH1 gene

The MLH1 gene is conserved in chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, fruit fly, mosquito, S.pombe, S.cerevisiae, K.lactis, E.gossypii, M.grisea, N.crassa, and P.falciparum.

MLH1 Gene Record

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

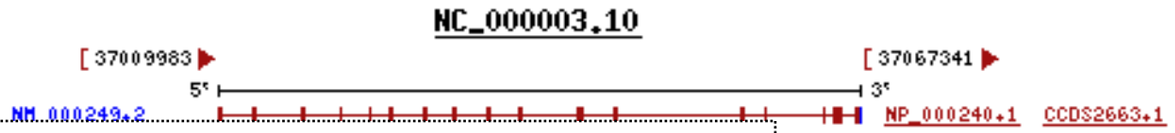
GeneID: 4292 updated 10-Apr-2007

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



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

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

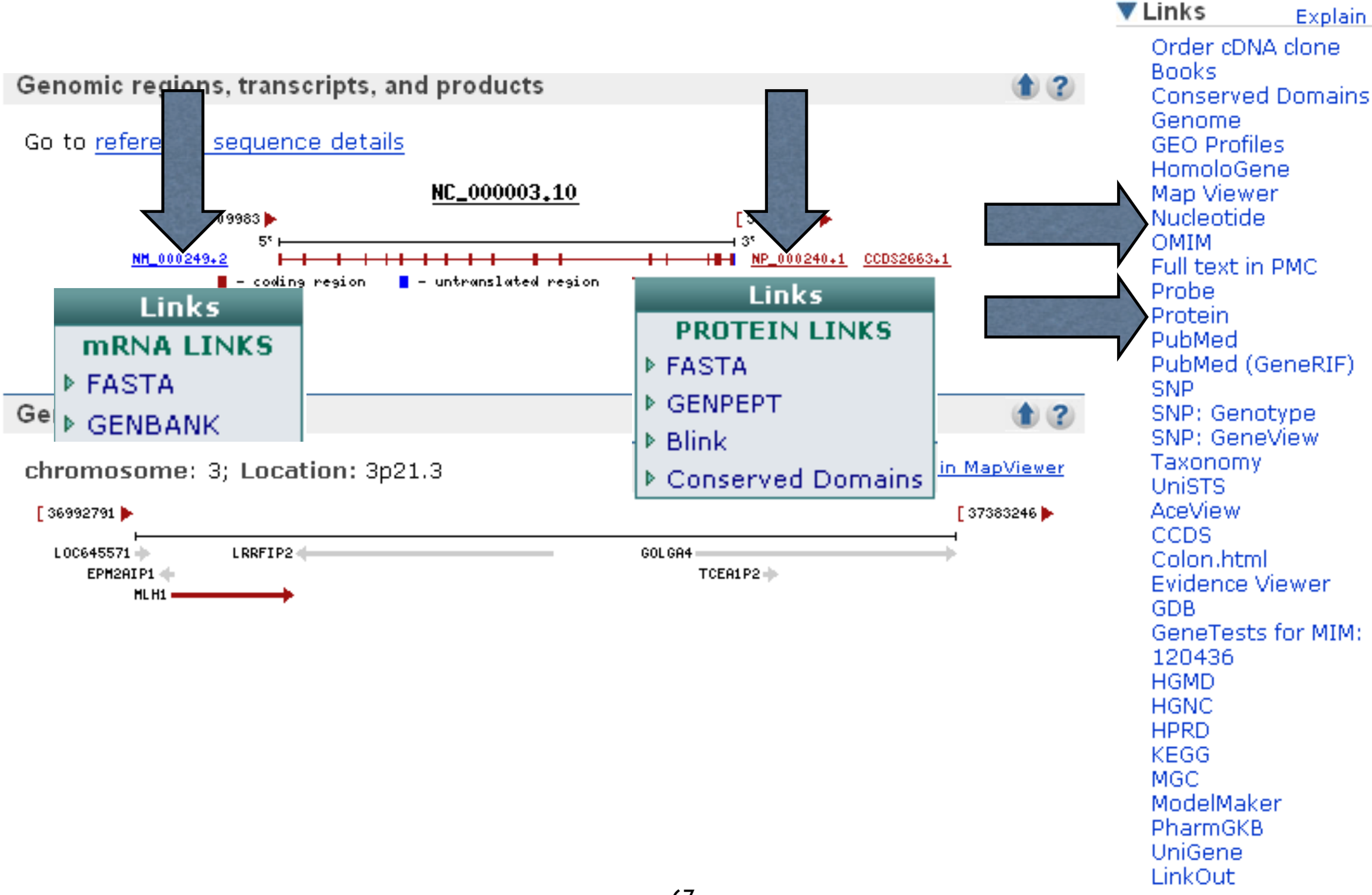
2. NW_921651.1

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

Links Explain

NCBI Entrez Nucleotide

My NCBI
Welcome joannealisonfox. [Sign Out]

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

Search Nucleotide for [ ] Go Clear

Limits | Preview/Index | History | Clipboard | Details

Format: GenBank FASTA Graphics More Formats▼ Download▼ Save▼ Links▼

NCBI Reference Sequence: NM_000249.3

Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), transcript variant 1, mRNA

Comment Features Sequence

```
LOCUS       NM_000249               2662 bp    mRNA    linear   PRI 16-MAY-2010
DEFINITION  Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E.
            coli) (MLH1), transcript variant 1, mRNA.
ACCESSION   NM_000249
VERSION     NM_000249.3  GI:263191547
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 2662)
  AUTHORS   Alvarez,K., Hurtado,C., Hevia,M.A., Wielandt,A.M., de la Fuente,M.,
            Church,J., Carvallo,P. and Lopez-Kostner,F.
  TITLE     Spectrum of MLH1 and MSH2 mutations in Chilean families with
            suspected Lynch syndrome
  JOURNAL   Dis. Colon Rectum 53 (4), 450-459 (2010)
   PUBMED   20305446
  REMARK    GeneRIF: 21 Chilean families with Lynch syndrome showed 6 mutations
            in MLH1.
REFERENCE   2  (bases 1 to 2662)
  AUTHORS   Vasen,H.F., Abdirahman,M., Brohet,R., Langers,A.M.,
            Kleibeuker,J.H., Kouwen,M.V., Koornstra,J.J., Boot,H., Cats,A.,
            Dekker,E., Sanduleanu,S., Poley,J.W., Hardwick,J.C., Cappel,W.H.,
            Jong,A.E., Tan,T.G., Jacobs,M., Mohamed,F.A., Boer,S.Y.,
            Meeberg,P.C., Verhulst,M.L., Salemans,J.M., Bentem,N.V.,
            Westerveld,B.D., Vecht,J. and Nagengast,F.M.
  TITLE     One to 2-Year Surveillance Intervals Reduce Risk of Colorectal
            Cancer in Families With Lynch Syndrome
  JOURNAL   Gastroenterology (2010) In press
   PUBMED   20206180
  REMARK    GeneRIF: Observational study of gene-disease association. (HuGE
            Navigator)
            Publication Status: Available-Online prior to print
REFERENCE   3  (bases 1 to 2662)
  AUTHORS   Song,L., Zhang,X.M., Wang,D.Q., Li,J.T., Ma,G.J., Ch
            Zhou,J.N.
  TITLE     [Mismatch repair gene hMLH1 A655G/A polymorphism and colorectal
            cancer]
  JOURNAL   Zhonghua Wei Chang Wai Ke Za Zhi 13 (3), 216-218 (2010)
   PUBMED   20336543
```

Homologene ⇨

Change Region Shown

Customize View

Analyze This Sequence

- Run BLAST
- Pick Primers

Articles about the MLH1 gene

- [Mismatch repair gene hMLH1 A655G/A polymorphism and colore [Zhonghua Wei Chang Wai Ke Za Zhi. 2010]
- Spectrum of MLH1 and MSH2 mutations in Chilean families with suspected Lynch [Dis Colon Rectum. 2010]
- Prognostic relevance of MLH1 and MSH2 mutations in hereditary non-polyposis colorectal cancer [Tumori. 2009]

» See all...

RefSeq Alternative Splicing

See 4 reference mRNA sequence splice variants for the MLH1 gene.

RefSeq Protein Product

See the reference protein sequence for DNA mismatch repair protein Mlh1 isoform 1 (NP_000240.1).

More about the MLH1 gene

This gene was identified as a locus frequently mutated in hereditary nonpolyposis colon cancer (HNPCC). It is a human homolog of the E. coli...

Also Known As: COCA2, FCC2, HNPCC, HNP...

Homologs of the MLH1 gene

The MLH1 gene is conserved in chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, fruit fly, mosquito, S.pombe, S.cerevisiae, K.lactis, E.gossypii, M.grisea, N.crassa, and P.falciparum.

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.

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 ge

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

n cancer, nonpolyposis type 2

g 1

019
n homolog 1

oli)

ancer, nonpolyposis type 2

NGG00000011527

0000010995
484

.04

55g
09955g

757243
pii AFL199C gene

.1

70
900

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 Cluster

Links

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.

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

NP_000240.1
756 aa

P.troglodytes MLH1

XP_001170433.1

M.musculus Mlh1
1 (E. coli)

NP_081086.1
760 aa

mutL homolog 1 (E. coli)

760 aa

R.norvegicus Mlh1
mutL homolog 1 (E. coli)

NP_112315.1
757 aa

G.gallus MLH1
mutL homolog 1, colon cancer, nonpc
(E. coli)

XP_418828.1
757 aa

D.melanogaster Mlh1
Mlh1

NP_477022.1
664 aa

A.gambiae AgaP_ENSANGG000000
ENSANGP00000014016

XP_320342.2
671 aa

A.gambiae ENSANGG00000010995
ENSANGP00000013484

XP_307435.2
395 aa

S.pombe SPBC1703.04
hypothetical protein

NP_596199.1
684 aa

S.cerevisiae MLH1
Mlh1p

NP_013890.1
769 aa

K.lactis KLLA0D09955g
mRNA gene KLLA0D09955g

XP_453504.1
724 aa

E.gossypii GeneID:2757243
Eremothecium gossypii AFL199C ger

NP_985351.1
771 aa

N.crassa NCU08309.1
hypothetical protein

A.thaliana ATMLH1
ATMLH1

O.sativa Os01g0958900
mRNA gene Os01g0958900

NP_001045457.1
724 aa

Links
- Conserved Domains
- Genome
- GEO Profiles
- Nucleotide
- Order cDNA clone
- OMIM
- Full text in PMC
- Probe
- Protein
- PubMed
- PubMed (GeneRIF)
- SNP
- Gene Genotype
- GeneView in dbSNP
- Taxonomy
- UniGene
- UniSTS
- MapViewer

Gene Links

Links
- Conserved Domains
- Gene
- Genome Project
- Nucleotide
- Genome
- OMIM
- Full text in PMC
- Related Sequences
- Domain Relatives
- PubMed
- PubMed (RefSeq)
- SNP
- Gene Genotype
- GeneView in dbSNP
- Related Structure
- Taxonomy
- UniGene
- BLink
- Domains

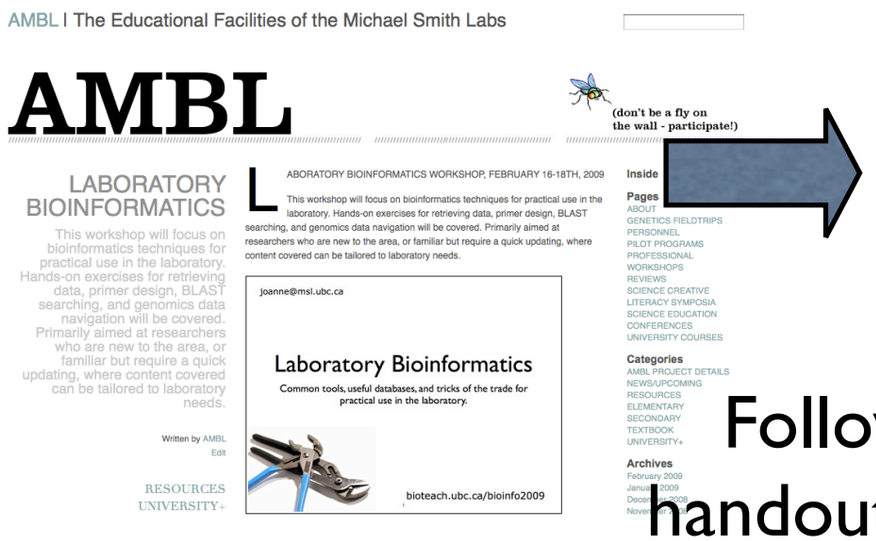
Protein Links

Sequence Databases

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

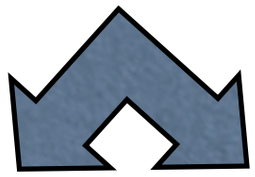
I am studying the regulation of cancer genes and would like to retrieve all human sequence records associated with cancer that contain a promoter region.

navigate to:
bioteach.ubc.ca/bioinfo2010



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)	368
#4	Search #1 NOT #3 (unique hits from Approach A: straight to Entrez CoreNucleotide search)	210
#3	Search #2 AND promoter[Feature key] (limit Approach B search to records with promoter annotated)	421
#2	CoreNucleotide Links for Gene (Search human[Organism] AND cancer[Text Word] AND gene_nucleotide[Filter]) (Approach B: Entrez gene follow link to CoreNucleotide)	87258
#1	Search human[Organism] AND cancer[Text Word] AND promoter[Feature key] (Approach A: Entrez CoreNucleotide search)	263

Advanced Tips & Tricks for Searching PubMed

My NCBI

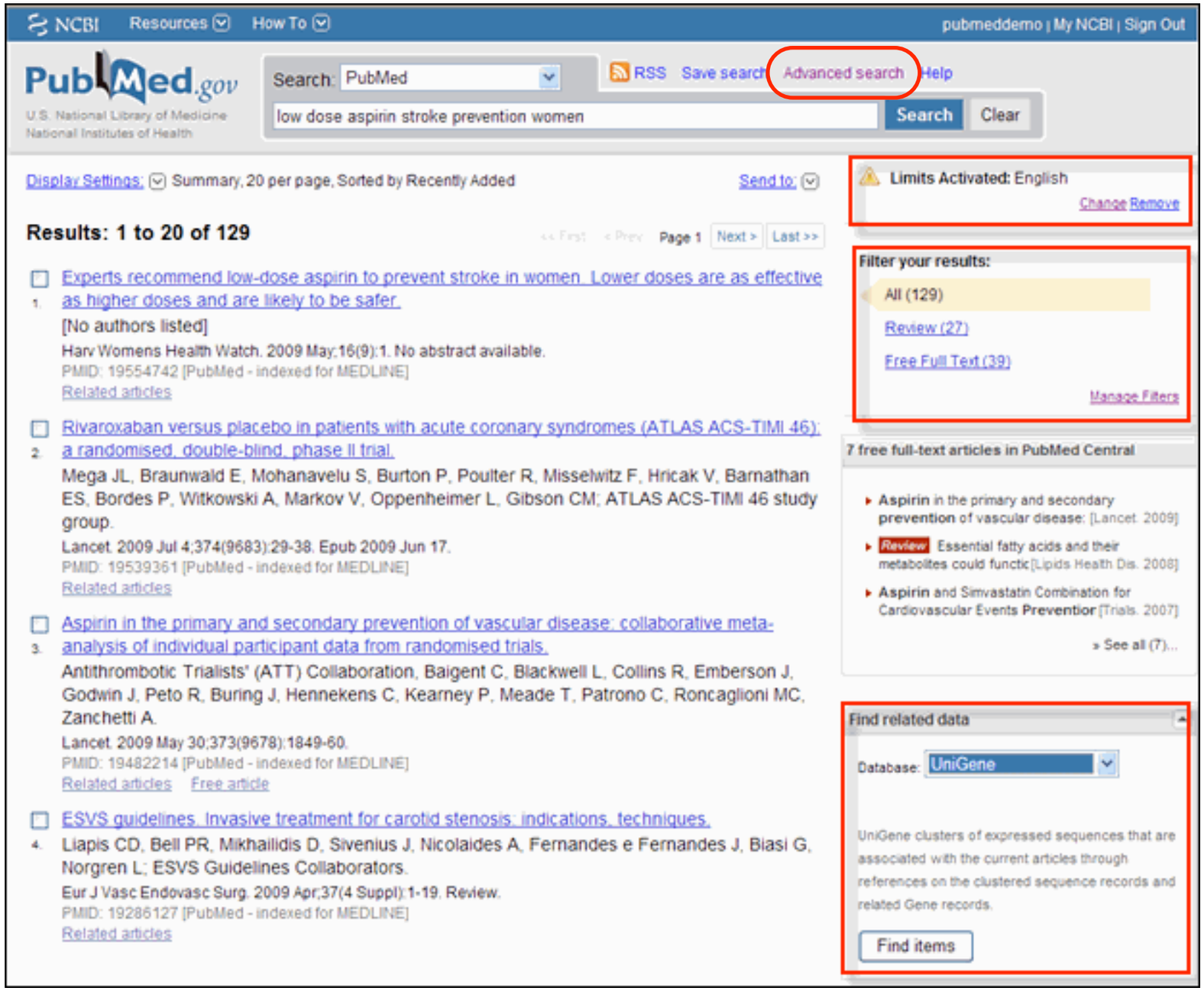
Bookshelf

- Advanced Features - Limits; Related Data; RSS; Send to options; Citation Sensor; Clipboard
- Entrez Gene RIF - reference into function sets
- Save collections with your MyNCBI account
- Search the NCBI Bookshef

PubMed Advanced Searching; Limits & Related Data

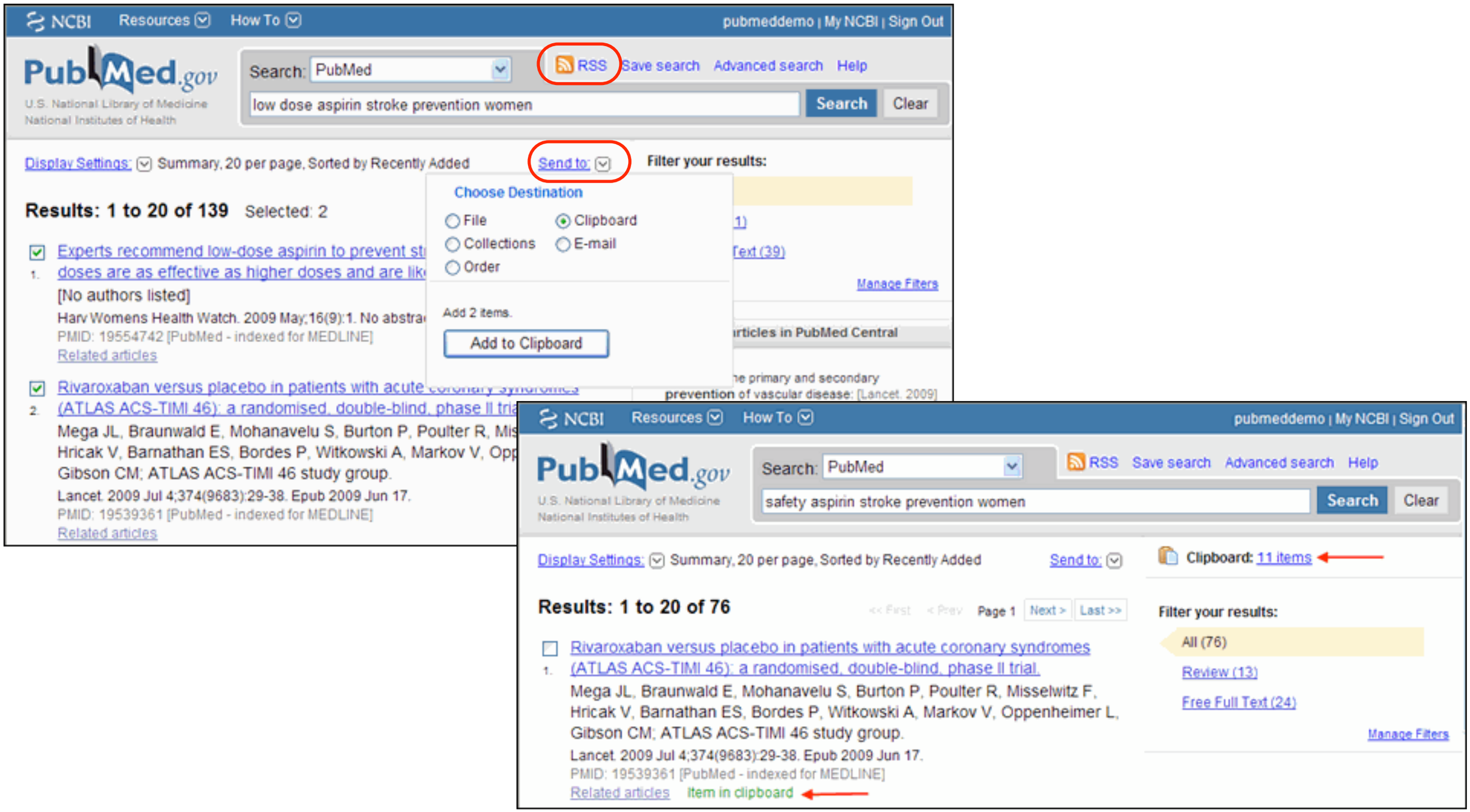
Example Search:

low dose aspirin stroke prevention women



adapted from NLM Technical Bulletin. 2009 Sep-Oct

PubMed Advanced Searching; Send to (destination options)



adapted from NLM Technical Bulletin. 2009 Sep-Oct

PubMed

PubMed comprises more than 19 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

Using PubMed

- PubMed Quick Start Guide
- Full Text Articles
- PubMed FAQs
- PubMed Tutorials
- New and Noteworthy

PubMed Tools

- Single Citation Matcher
- Batch Citation Matcher
- Clinical Queries
- Topic-Specific Queries

More Resources

- MeSH Database
- Journals Database
- Clinical Trials
- E-Utilities
- LinkOut

Example Search:

Lander 2001 Nature

All: 14 | Review: 4

We found 3 articles in Nature 2001 by Lander:

Linkage disequilibrium in the human genome. Reich DE et al. *Nature.* (2001)

A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Sachidanandam R et al. *Nature.* (2001)

Initial sequencing and analysis of the human genome. Lander ES et al. *Nature.* (2001)

Items 1 - 14 of 14

1: Automation, parallelism, and robotics for proteomics.
Alterovitz G, Liu J, Chow J, Ramoni MF.
Proteomics. 2006 Jul;6(14):4016-22. Review.
PMID: 16786489 [PubMed - indexed for MEDLINE]
Related Articles

adapted from NCBI News, June 2009

NCBI
National Center for Biotechnology Information

Search All Databases

Search Clear

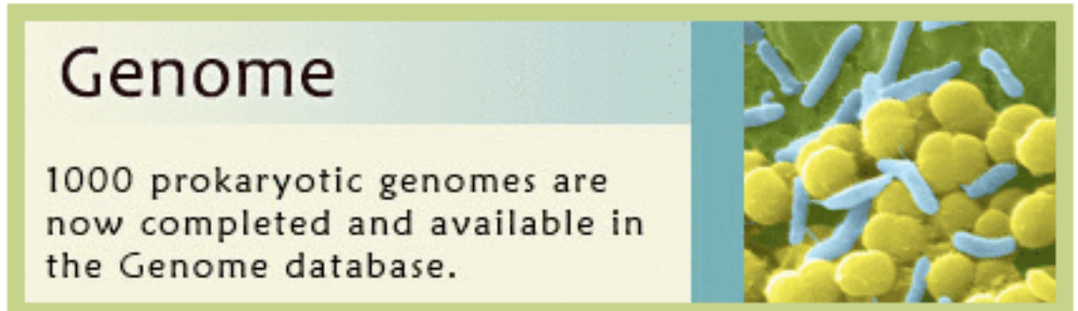
Resources

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

Welcome to NCBI

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

More about the NCBI | Mission | Organization | Research | RSS



How To...

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

See all ...

NLM/NCBI H1N1 Flu Resources

Popular Resources

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

NCBI News

OMIM's new look, Epigenomics in April — 10 May 2010

NCBI News
The April NCBI News issue is now available.

NIH Roadmap Epigenomics Project data in GEO database — 22 Apr 2010

GEO's Roadmap Epigenomics Project Data Listings page allows ...

March News issue available — 09 Apr 2010

Includes My NCBI, E-Utility, and

NCBI

Bookshelf

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

Search Books for MLH1 Go Clear

Limits | Preview/Index | History | Clipboard | Details

- About Entrez
- Books
- Overview
- Using the books
- Information for authors and publishers
- What's new
- 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:

Comparative Oncology
Baba, Alecsandru Ioan; Câtoi, Cornel
Bucharest: The Publishing House of the Romanian Academy; 2007

Familial Cancer Syndromes [Internet]
Riegert-Johnson, Douglas L.; Boardman, Lisa A.; Hefferon, Timothy; Spurck, Lauren, editors
Bethesda (MD): National Library of Medicine (US), NCBI; 2009

Health, United States, 2008
Atlanta (GA): Centers for Disease Control and Prevention; 2008

The National Academies Collection: Reports funded by National Institutes of Health
Washington (DC): The National Academies Press; c2004-2010

National Institute for Health and Clinical Excellence: Guidance
London: National Institute for Health and Clinical Excellence; 2009

Patient Safety and Quality: An Evidence-Based Handbook for Nurses
Hughes, Ronda G., editor
Rockville (MD): Agency for Healthcare Research and Quality (US); 2008

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: 294 | 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...

Search

Go

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

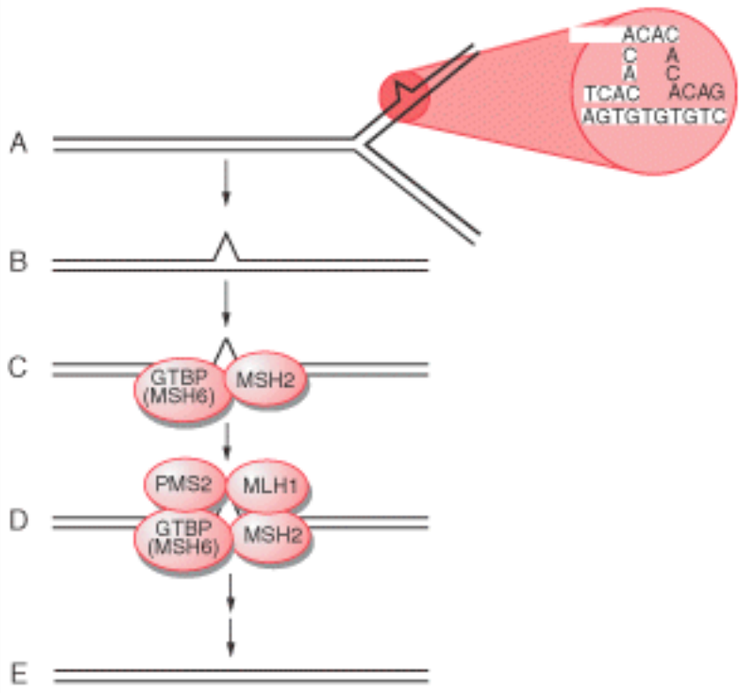


Figure 7-10. Mismatch repair pathway in human cells. **A** and **B,** During DNA replication, DNA mismatches may arise, such as from strand slippage (shown) or misincorporation of bases (not shown). **C,** The mismatch is recognized by MutS homologs, perhaps most often MSH2 and GTBP/MSH6, although another 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.)

Search
Go
This book All books
PubMed

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:

Please type the five characters you see above.

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

Register

Or cancel and return home

My NCBI

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

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

My NCBI

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

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

NCBI Resources How To My NCBI Sign In

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed Limits Advanced search Help

Search Clear

PubMed

PubMed comprises more than 19 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citat
text content from PubMed Central and publisher web si

search with a gene name of interest to you

Using PubMed

- PubMed Quick Start Guide
- Full Text Articles
- PubMed FAQs
- PubMed Tutorials
- New and Noteworthy

PubMed Tools

- Single Citation Matcher
- Batch Citation Matcher
- Clinical Queries
- Topic-Specific Queries

More Re

- MeSH Data
- Journals Database
- Clinical Trials
- E-Utilities
- LinkOut

You are here: NCBI > Literature > PubMed

Write to the Help Desk

GETTING STARTED
NCBI Help Manual
NCBI Handbook
Training & Tutorials

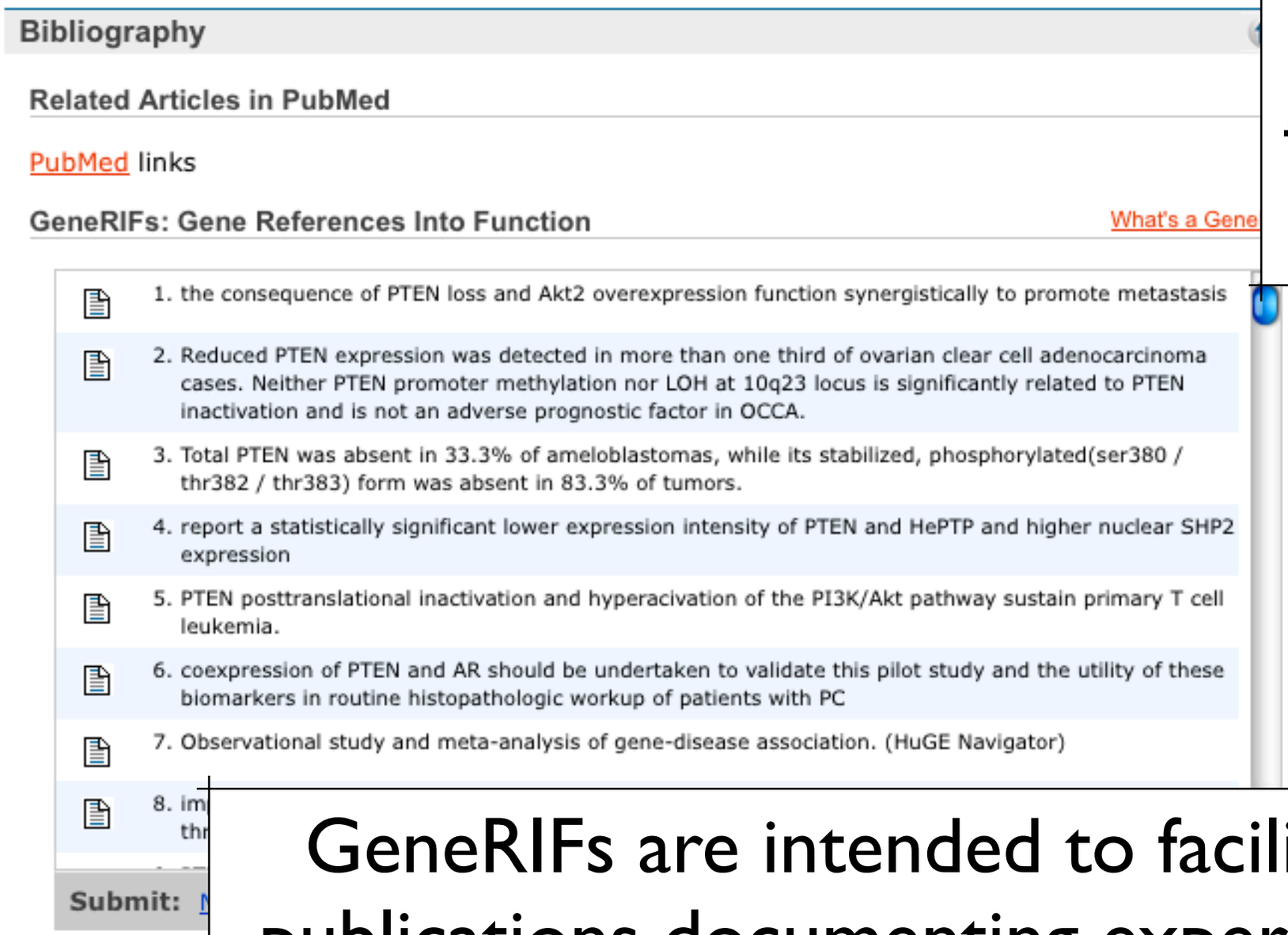
RESOURCES
Literature
DNA & RNA
Proteins
Sequence Analysis

POPULAR
PubMed
PubMed Central
Bookshelf
BLAST

FEATURED
GenBank
Reference Sequences
Map Viewer
Genome Projects

NCBI INFORMATION
About NCBI
Research at NCBI
NCBI Newsletter
NCBI FTP Site

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 Resources How To

joannealisonfox My NCBI Sign Out

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed

Limits Advanced search Help

Search Clear

Display Settings: Summary, 20 per page, Sorted by Recently Added

Send to:

Filter your results:

Choose Destination

File Clipboard
Collections E-mail
Order

Your collection was saved. Edit your collection.

Manage Filters

Results: 1 to 20 of 846

<< First < Pr

1. Depletion of DNMT3A suppressed cell proliferation and restored PTEN in hepatocellular carcinoma cell.
Zhao Z, Wu Q, Cheng J, Qiu X, Zhang J, Fan H.
J Biomed Biotechnol. 2010;2010:737535. Epub 2010 May 12.
PMID: 20467490 [PubMed - in process] **Free PMC Article** Free text
Related citations

2. Upregulation of PTEN in glioma cells by cord blood mesenchymal stem cells inhibits migration via downregulation of the PI3K/Akt pathway.
Dasari VR, Kaur K, Velpula KK, Gujrati M, Fassett D, Klopfenstein JD, Dinh DH, Rao JS.
PLoS One. 2010 Apr 26;5(4):e10350.
PMID: 20436671 [PubMed - in process] **Free PMC Article** Free text
Related citations

3. Activation of the A
Xu J, Zhou JY, We
PLoS One. 2010 Apr
PMID: 20419107 [Pu
Related citations

4. Targeting the Androgen Receptor by Taxol in Castration-Resistant Prostate Cancer.
Jiang J, Huang H.
Mol Cell Pharmacol. 2010 Jan 1;2(1):1-5.
PMID: 20419056 [PubMed] **Free PMC Article** Free text

Save your PubMed results to your MyNCBI collections

Find related data

Database: Select

Find items

Recent activity

Turn On Clear

Activity recording is turned off.

Turn recording back on

Credits

- Materials for this presentation have been adapted from a variety of sources including:

  NCBI HelpDesk - Field Guide Course Materials

  NLM Technical Bulletins

  NCBI News available @ the NCBI Bookshelf

- Questions? Please contact:

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

AMBL

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

LABORATORY BIOINFORMATICS

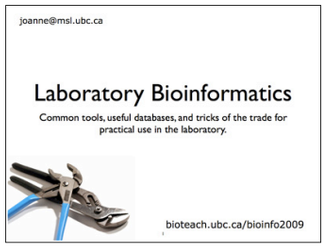
This workshop will focus on bioinformatics techniques for practical use in the laboratory. Hands-on exercises for retrieving data, primer design, BLAST searching, and genomics data navigation will be covered. Primarily aimed at researchers who are new to the area, or familiar but require a quick updating,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



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

BLAST

Finding Function By Sequence Similarity

Concepts of Sequence Similarity Searching

- The premise:

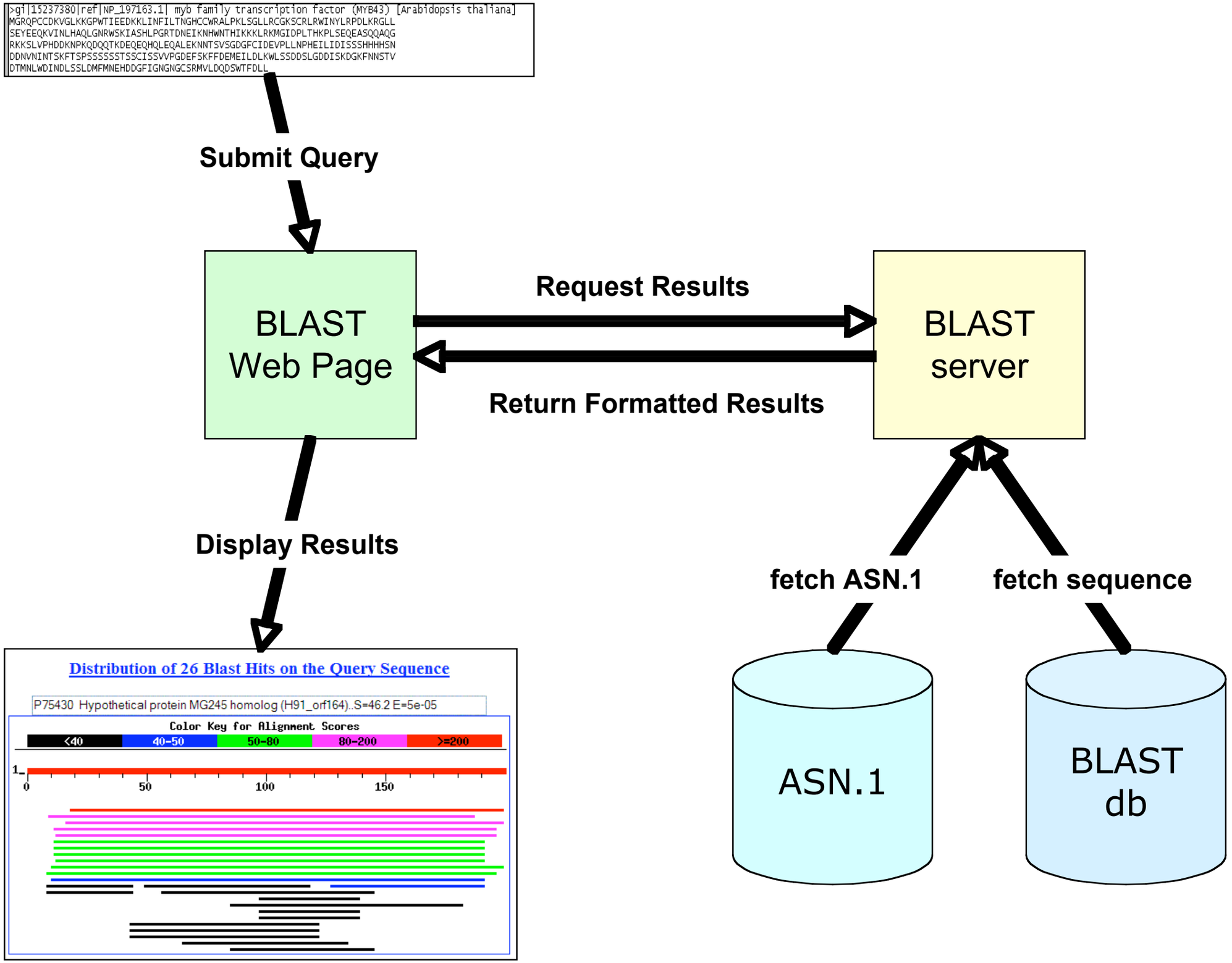
  One sequence by itself is not informative; it must be analyzed by comparative methods against existing sequence databases to develop hypothesis concerning relatives and function.

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
fetch ASN.1
fetch sequence
ASN.1
BLAST db
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

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

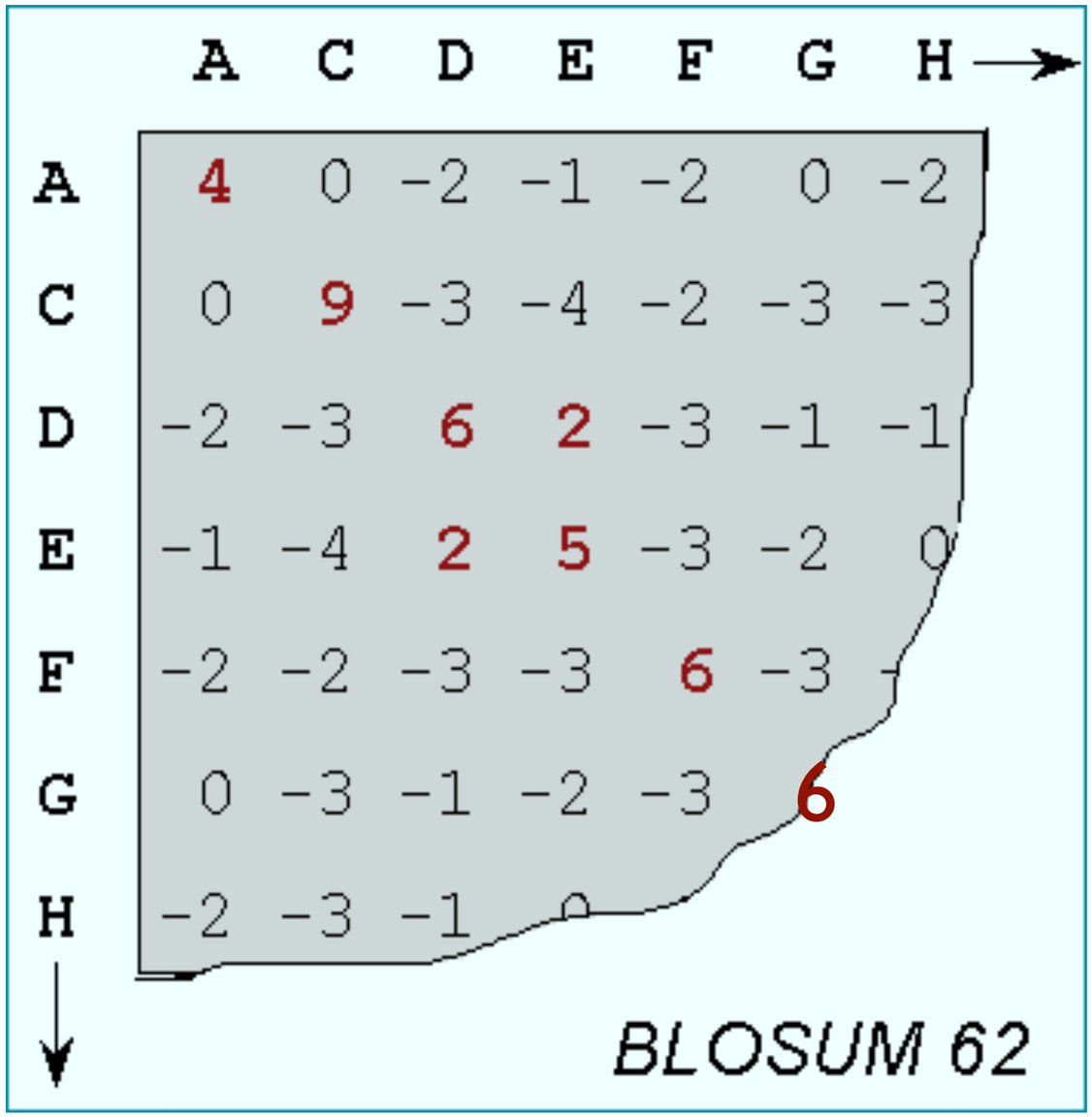
Where does the score (S) come from?

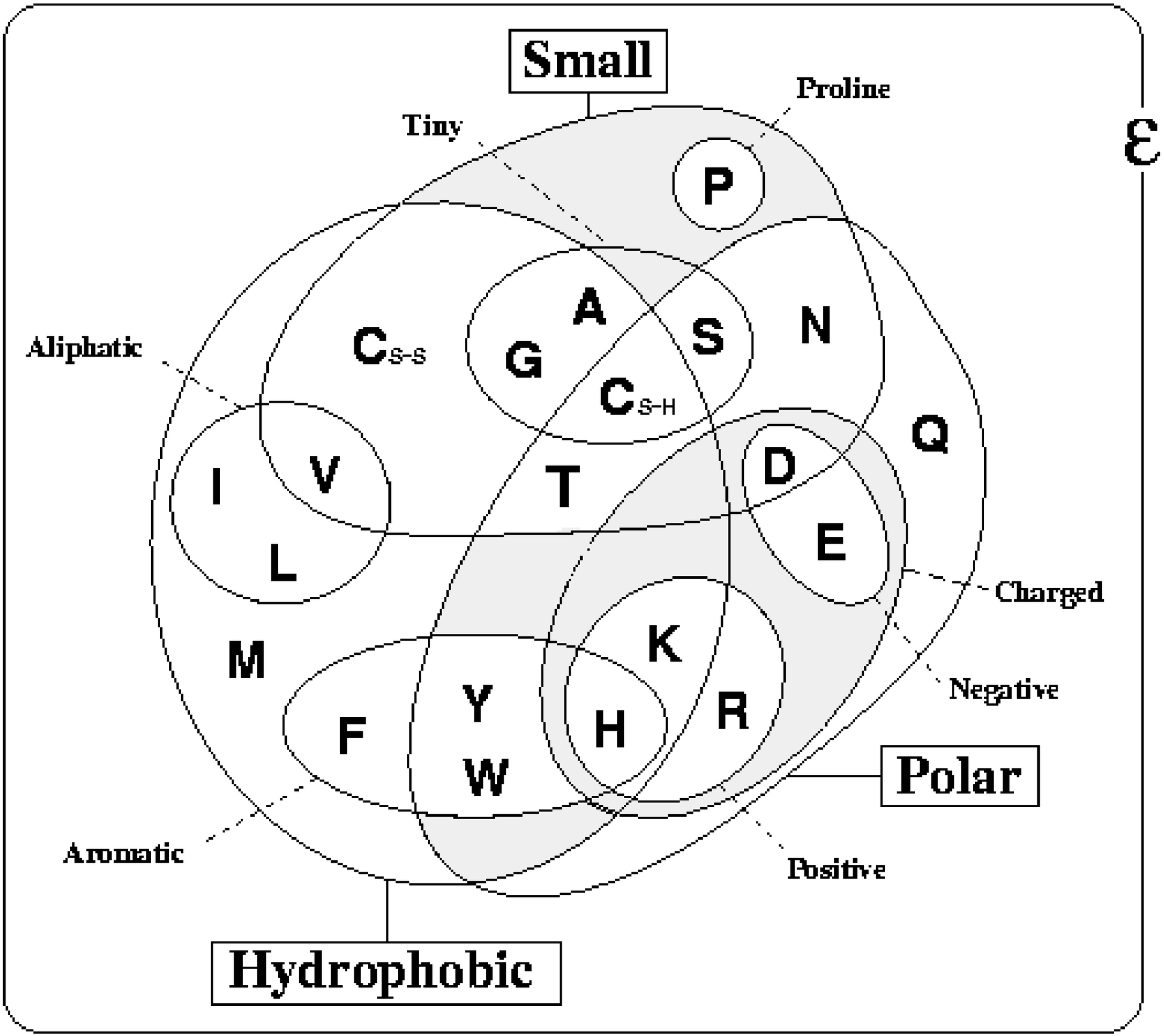
- The quality of each pair-wise alignment is represented as a score and the scores are ranked.

- 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)




Small
Proline
Tiny
ε
P
A
Aliphatic
C_{S-S}
G
S
N
C_{S-H}
Q
D
I
V
T
L
E
Charged
K
M
Y
Negative
F
H
R
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

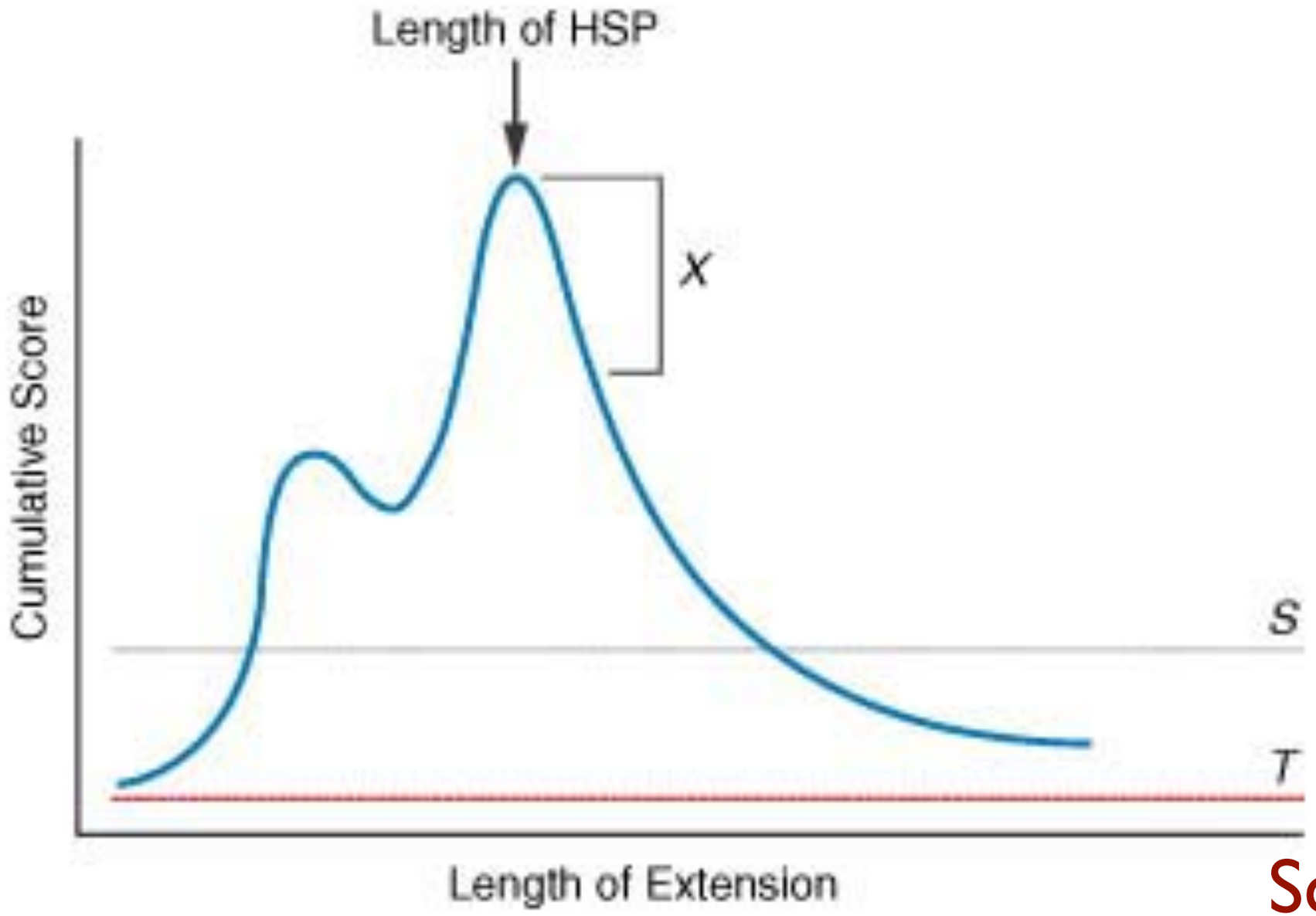
RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9

...

Extension using neighborhood words greater than neighborhood score threshold (T = 11)

```
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

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

LABORATORY BIOINFORMATICS

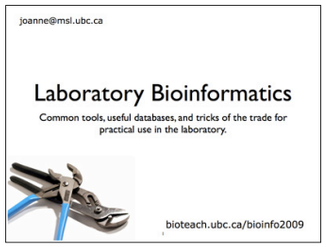
This workshop will focus on bioinformatics techniques for practical use in the laboratory. Hands-on exercises for retrieving data, primer design, BLAST searching, and genomics data navigation will be covered. Primarily aimed at researchers who are new to the area, or familiar but require a quick updating,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



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 9:30am

BLAST background, guided tour & practical exercises

joanne@msl.ubc.ca