joanne@msl.ubc.ca

Laboratory Bioinformatics

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

bioteach.ubc.ca/bioinfo2009

Computer Lab

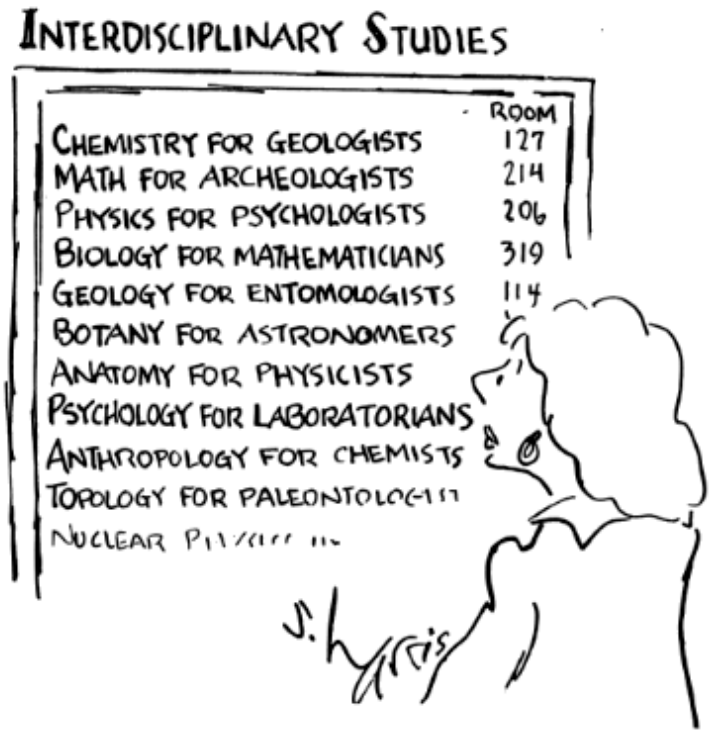
- Computers, available here for your use
- wireless login

DETAILS WILL BE AVAILABLE ONSITE

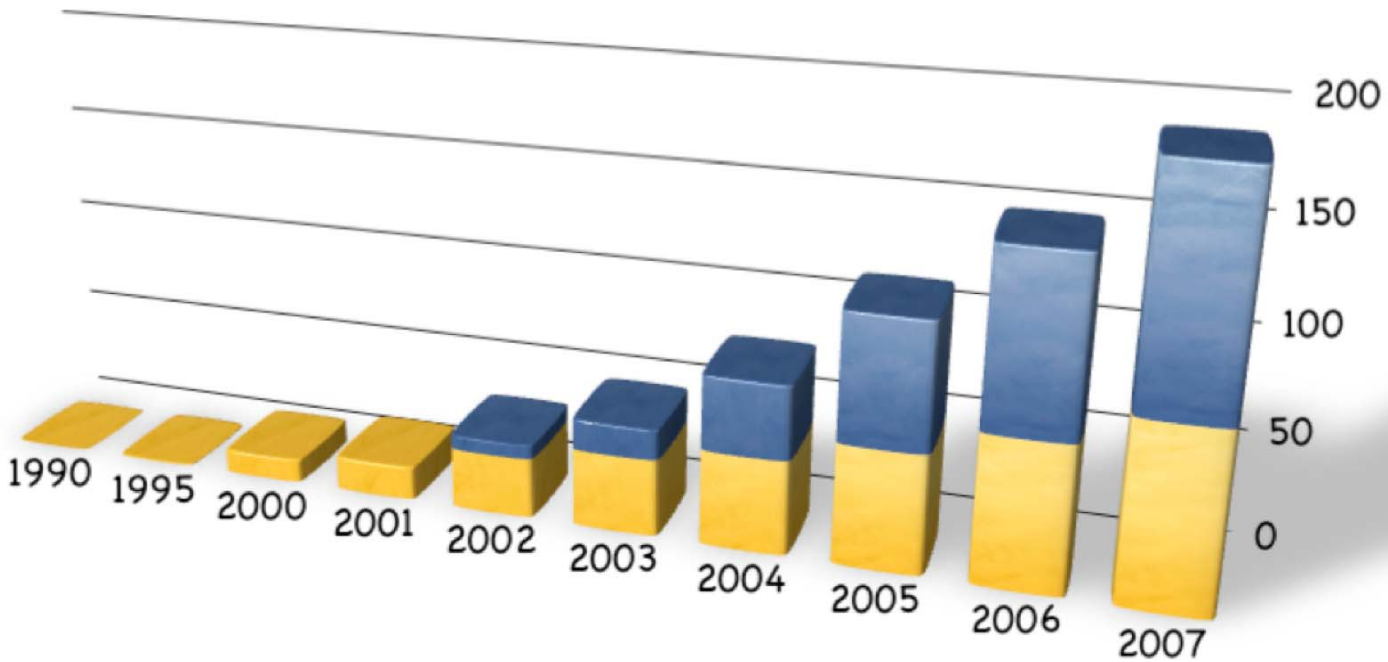
Module 1 Topics

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

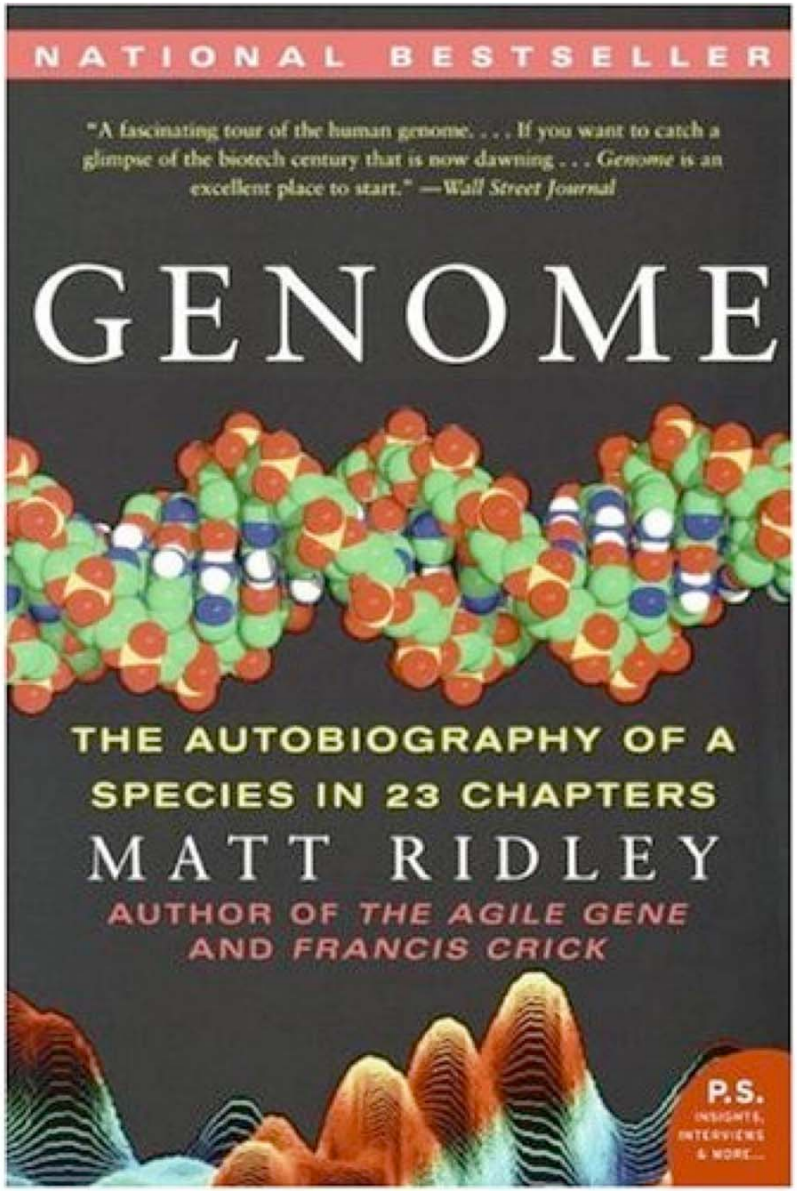
Bioinformatics for Biologists



Growth of GenBank

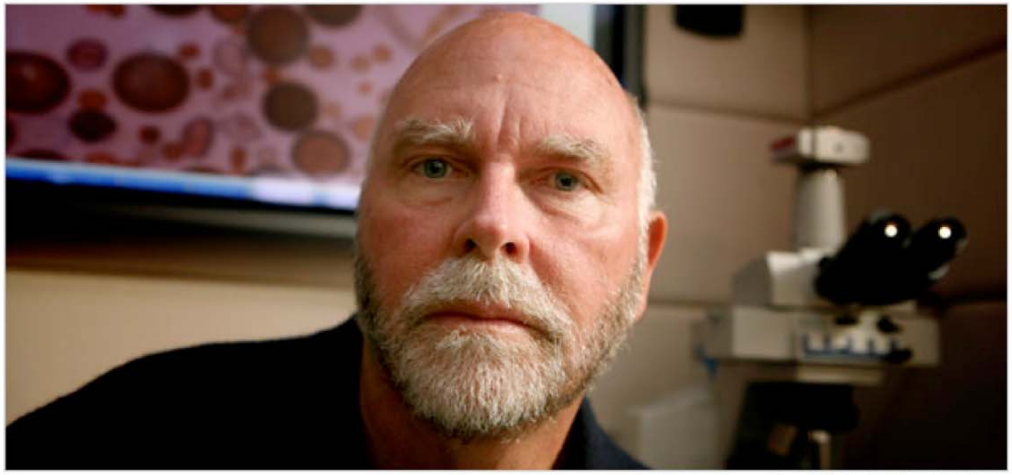


In 2005, International sequence databases exceed 100 gigabases


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.

Personalized Medicine?

In the Genome Race, the Sequel Is Personal

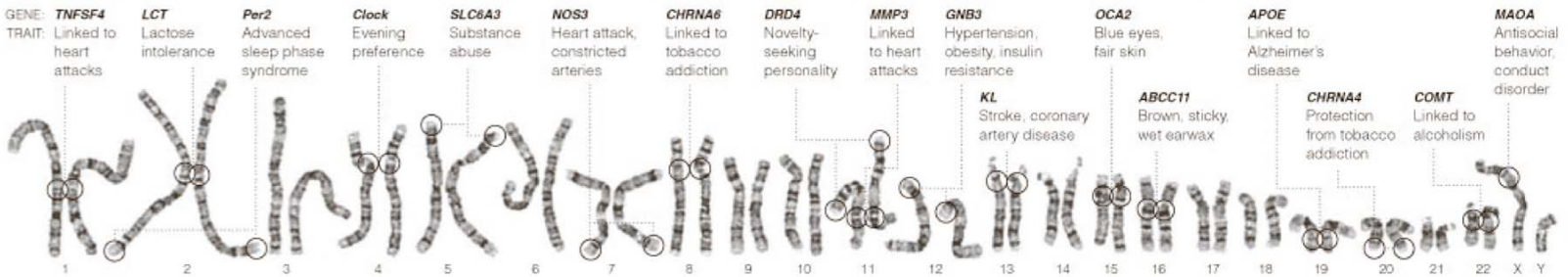


Thor Swift for The New York Times

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

The New York Times — September 3, 2007

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

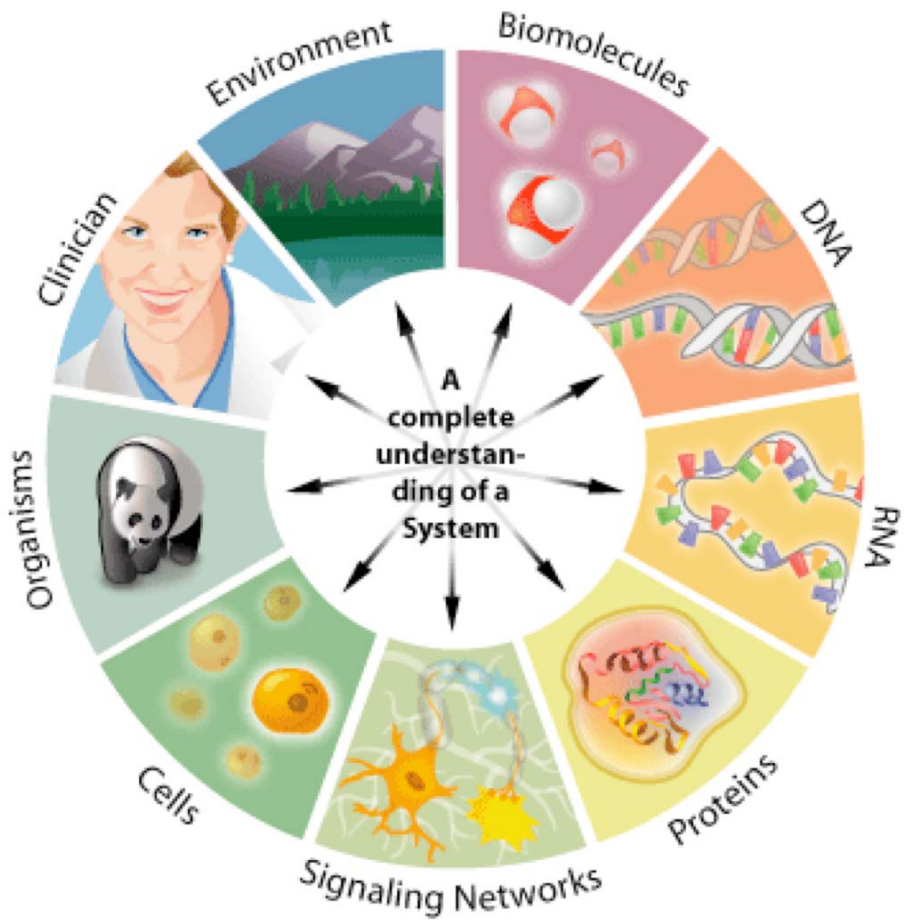


Sources: PLoS Biology; GeneCards

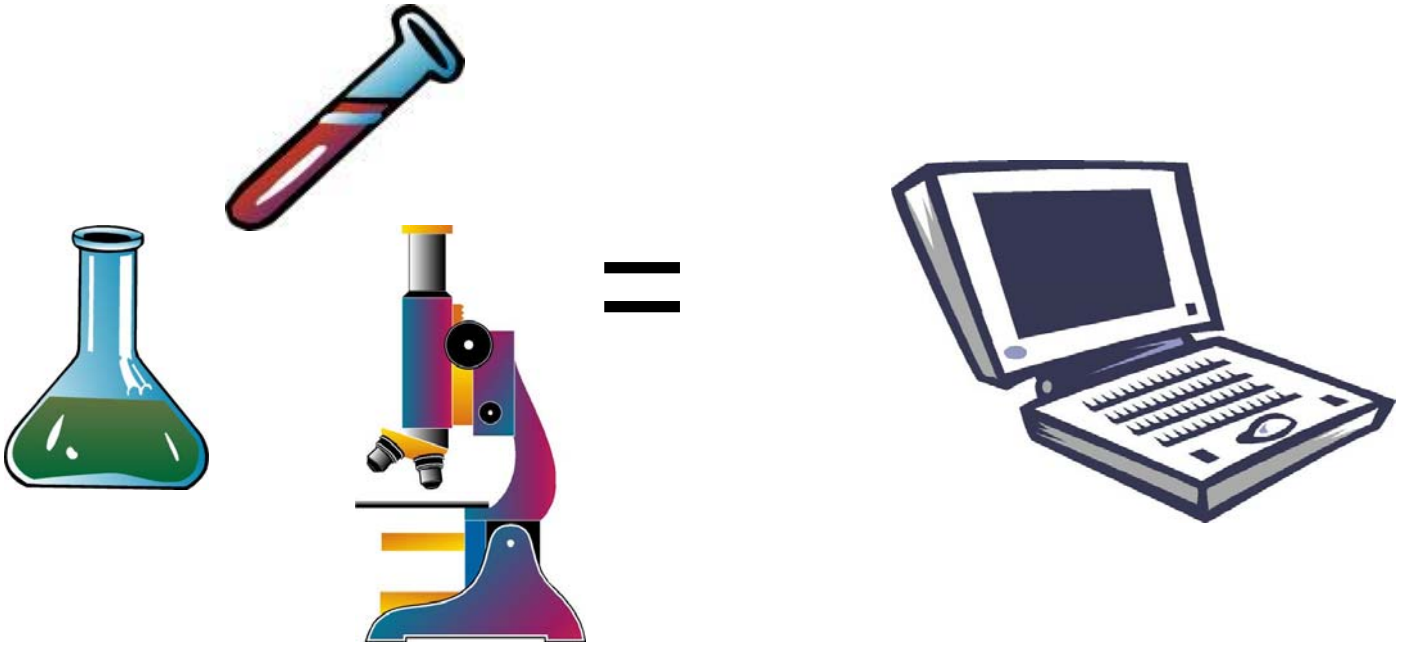
Genes highlighted above are linked to specific traits but do not mean the conditions are inevitable.

THE NEW YORK TIMES

What is Bioinformatics?

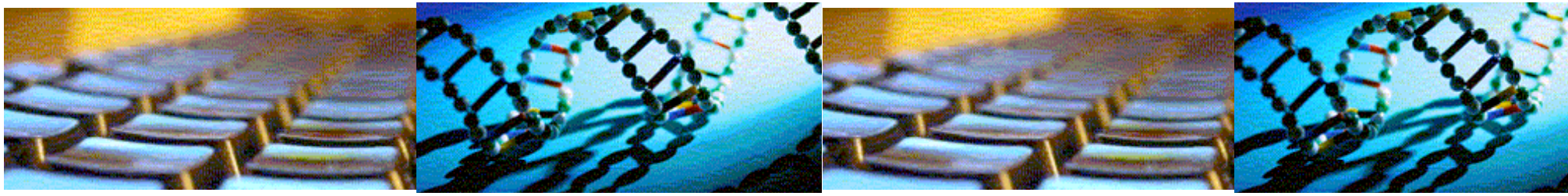


Laboratory Bioinformatics



What is Bioinformatics?

Goals & Priorities



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

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

Bioinformatics Questions

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

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

Overview of Topics*

✓ Module 1 - Public Database Resources NCBI

✓ Module 2 - BLAST, Primer Design, MSA

✓ Module 3 - Genome Browsers, Special Topics*

*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

National Library of Medicine National Institutes of Health

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

Search All Databases for Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

Genomic biology
The human genome, whole

What does NCBI do?

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

GenBank® Celebrating 25 Years

NCBI will hold a scientifc meeting to celebrate the 25th anniversary of GenBank.
April 7-8, 2008
Natcher Auditorium, NIH Campus, Bethesda MD

click here for more information

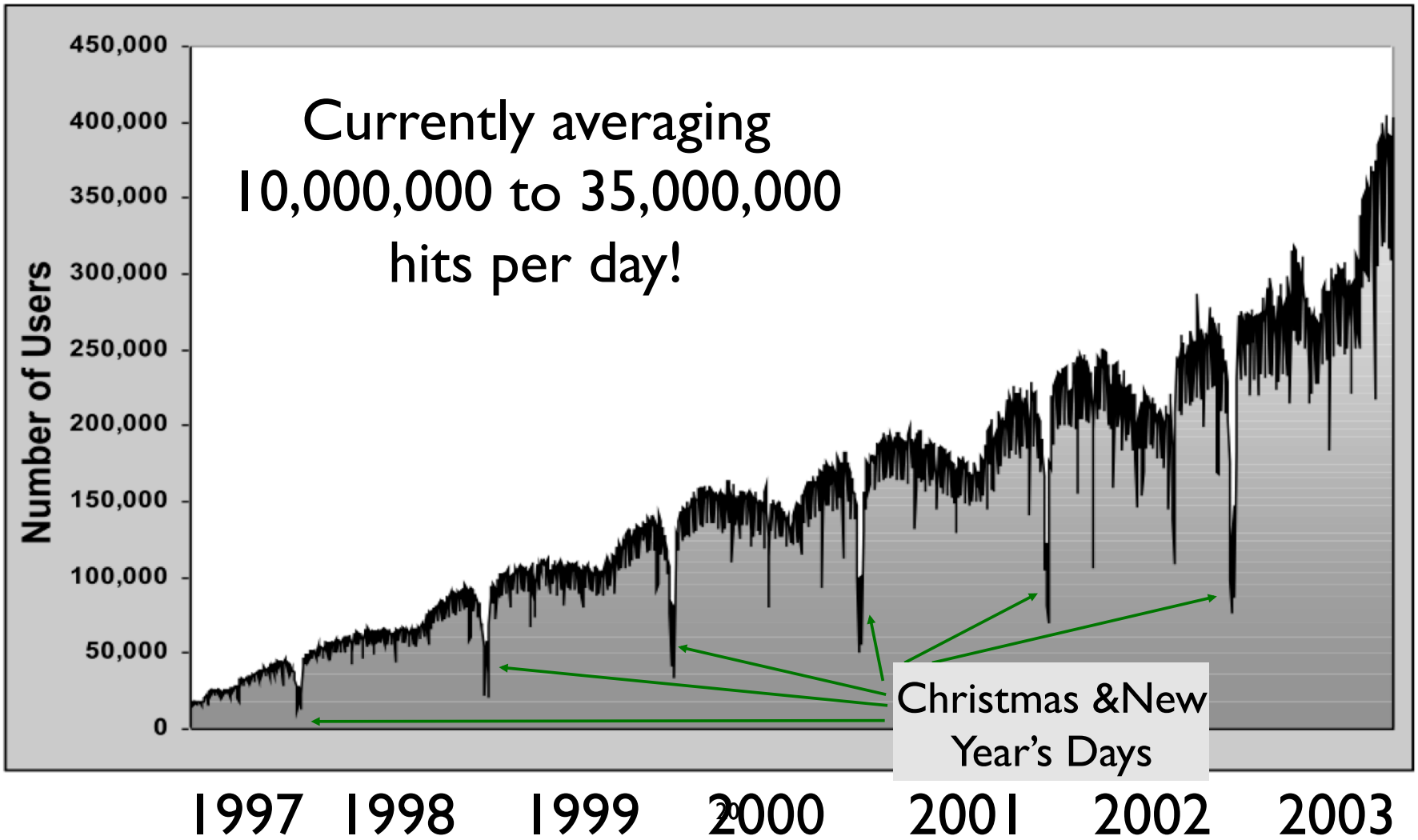
New **Protein Clusters**
Entrez Protein Clusters database

The new Entrez Protein Clusters database is a collection of Reference Sequence (RefSeq) proteins, from the complete

Hot Spots

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

Number of Users and Hits Per Day



NCBI Site Map

Search NCBI

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

Databases

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

Literature Databases

- PubMed
- PubMed Central
- OMIM
- Books

Molecular Databases

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

Genomes

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

Query all Entrez Databases

NCBI

About
Contact
Research
Software Engineering
Site Search

Tools

Entrez Tools

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

Data Analysis Tools

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

FTP

- Databases and Software

Programming Tools

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

Data Submissions

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

Nucleotide Sequences

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

Polymorphisms

- dbSNP (Single Nucleotide Polymorphisms)

Expression

- GEO (Gene Expression Omnibus)

Cytogenetic Data

- SKY/M-FISH & CGH Database

Trace Data

- Trace Archive

Education

News

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

Bookshelf

- Coffee Break
- Genes and Disease
- NCBI Handbook

Resources

- Tutorials, Courses, and more

The NCBI ftp site

NCBI **FTP site**

PubMed Entrez BLAST OMIM Books TaxBrowser Structure

Search All Databases for Go

NCBI

SITE MAP
Guide to NCBI resources

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

GenBank
sequence submission support and software

Molecular databases
sequences, structures and taxonomy

Literature databases
PubMed and OMIM

Genomic Biology

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

- BLAST Basic Local Alignment Search Tool

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

- CDD Data

Download data from the Conserved Domain Database.

- CD-Tree

Download the protein domain hierarchy viewer and editor.

- Cn3D

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

- Data Repository

Download collections of contributed molecular biology

- dbGaP

Download open access Genotype and Phenotype dat

- GenBank

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

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

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

NCBI Databases & Services

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

Types of Databases

Primary Databases

✓ Original submissions by experimentalists

✓ Content controlled by the submitter

✓ Examples: GenBank, SNP, GEO

Derivative Databases

✓ Built from primary data

✓ Content controlled by third party (NCBI)

✓ Examples: Refseq, TPA, RefSNP, UniGene, NCBI Protein, Structure, Conserved Domain

What is GenBank?

NCBI's Primary Sequence Database

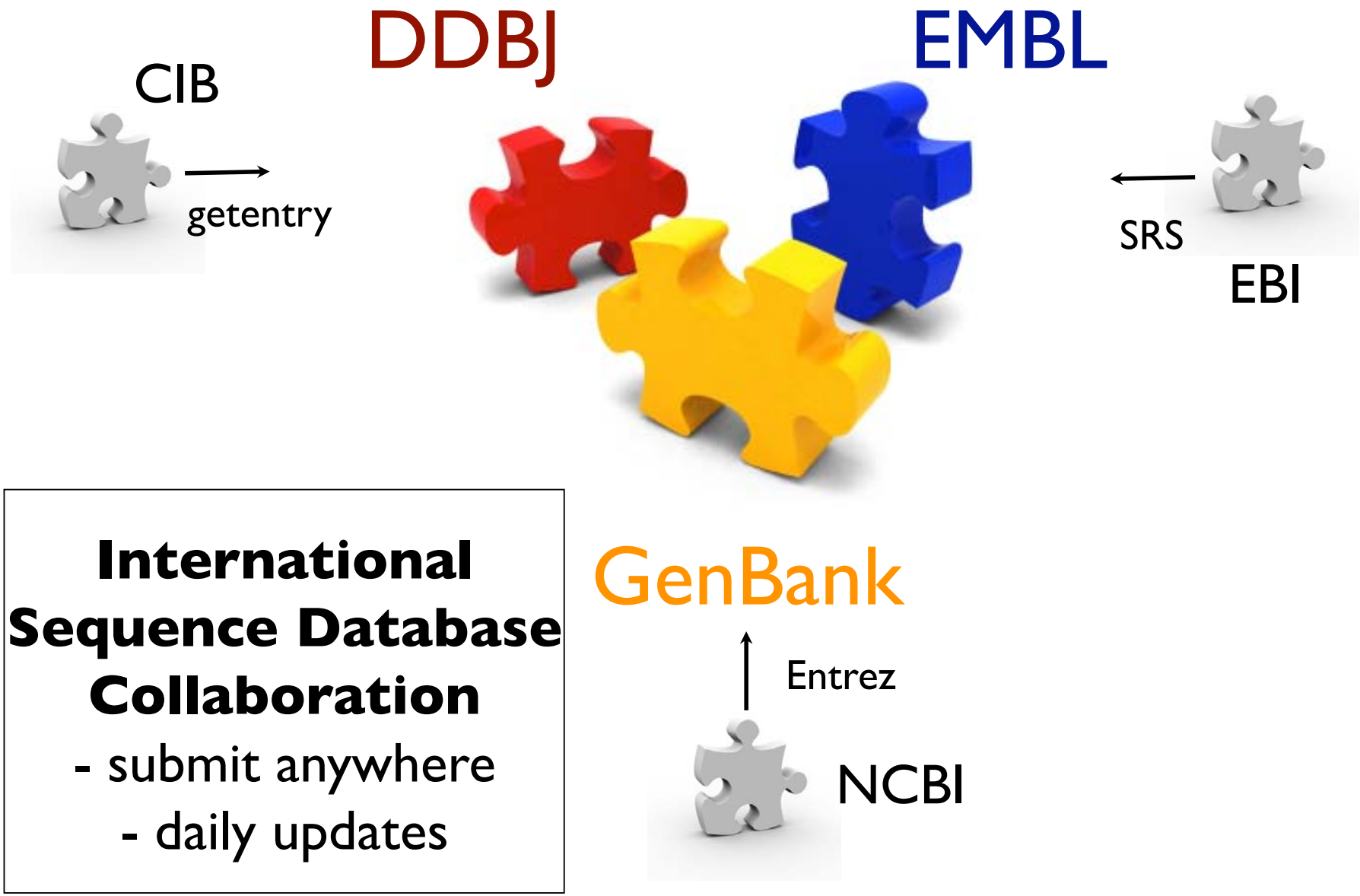
- Nucleotide only sequence database
- Archival in nature
- Historical
- Reflective of submitter point of view (subjective)
- Redundant

GenBank Data

✓ Direct submissions (traditional records)

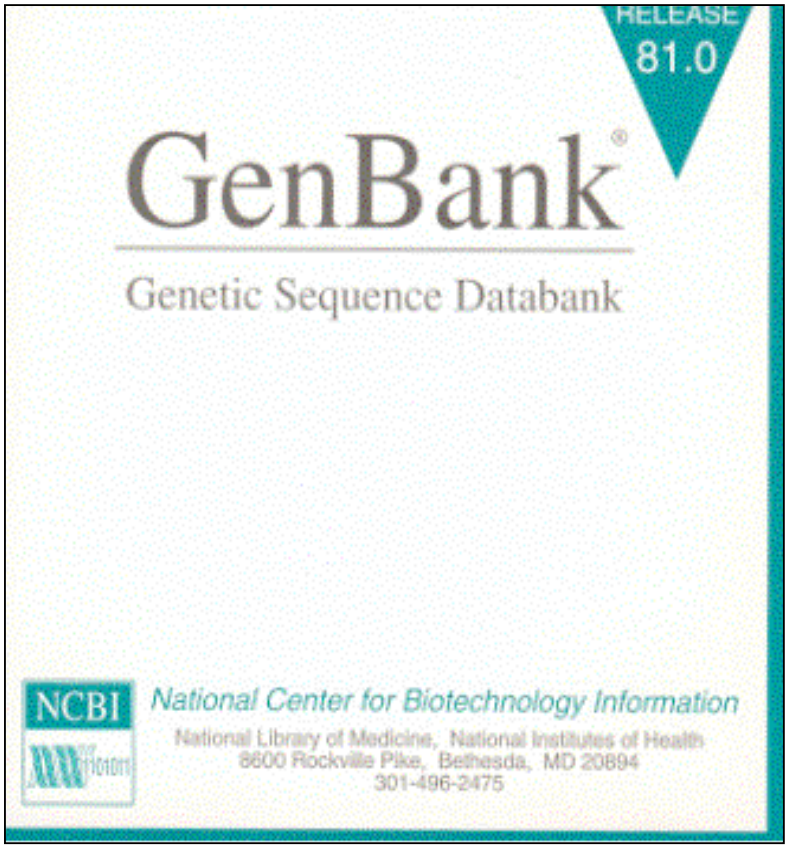
✓ Batch submissions (EST, GSS, STS)

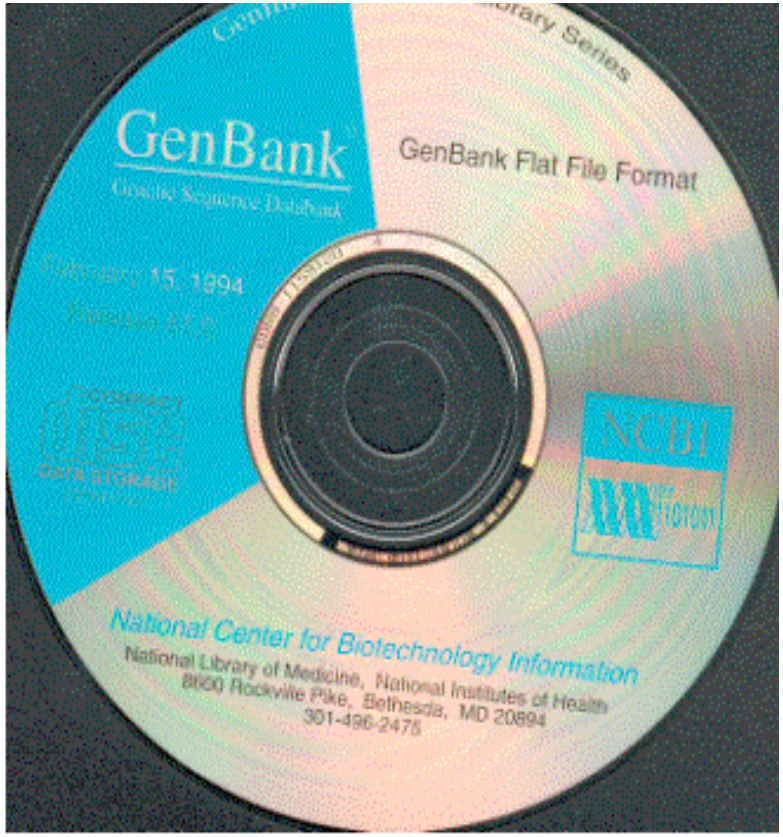
✓ ftp accounts (genome data)



International Sequence Database Collaboration

- submit anywhere
- daily updates





GenBank: NCBI's Primary Sequence Database

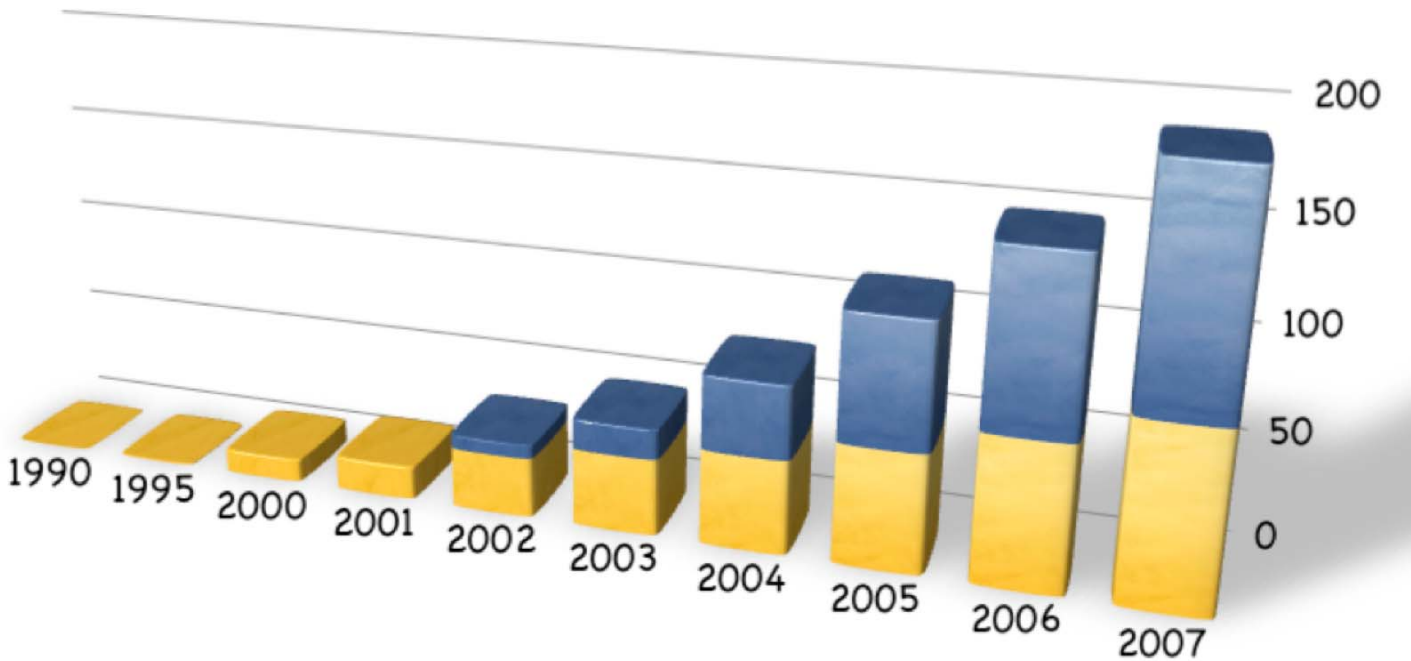
ftp://ftp.ncbi.nih.gov/genbank/

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

*includes WGS

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

Growth of GenBank



Current Release 169
Doubling time 12-14 months

GenBank WGS

Organization of GenBank

Records are divided into 18 Divisions.

☑ Traditional:

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

☑ BULK Divisions:

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

Entrez query: gbdiv_xxx[Properties]

A Traditional GenBank Record

Header

Feature Table

Sequence

Traditional GenBank Record

```
LOCUS       HSHMLHI                 2503 bp    mRNA    linear   PRI 31-MAR-1994
DEFINITION  Human DNA mismatch repair (hmlh1) mRNA, complete cds.
ACCESSION   U07418
VERSION     U07418.1  GI:466461
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vert          teleostomi;
            Mammalia; Eutheria; Primates; Catarrhini
REFERENCE   1  (bases 1 to 2503)
  AUTHORS   Papadopoulos,N., Nicolaides,N.C., Wei
            Carter,K.C., Rosen,C.A., Haseltine,W
            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 heredi
  JOURNAL   Science 263 (5153), 1625-1629 (1994)
```

Accession
- Stable
- Reportable
- Universal

```
ACCESSION    U07418

VERSION      U07418.1  GI:466461
```

Version
- Tracks changes in sequence

GI number
- NCBI internal use

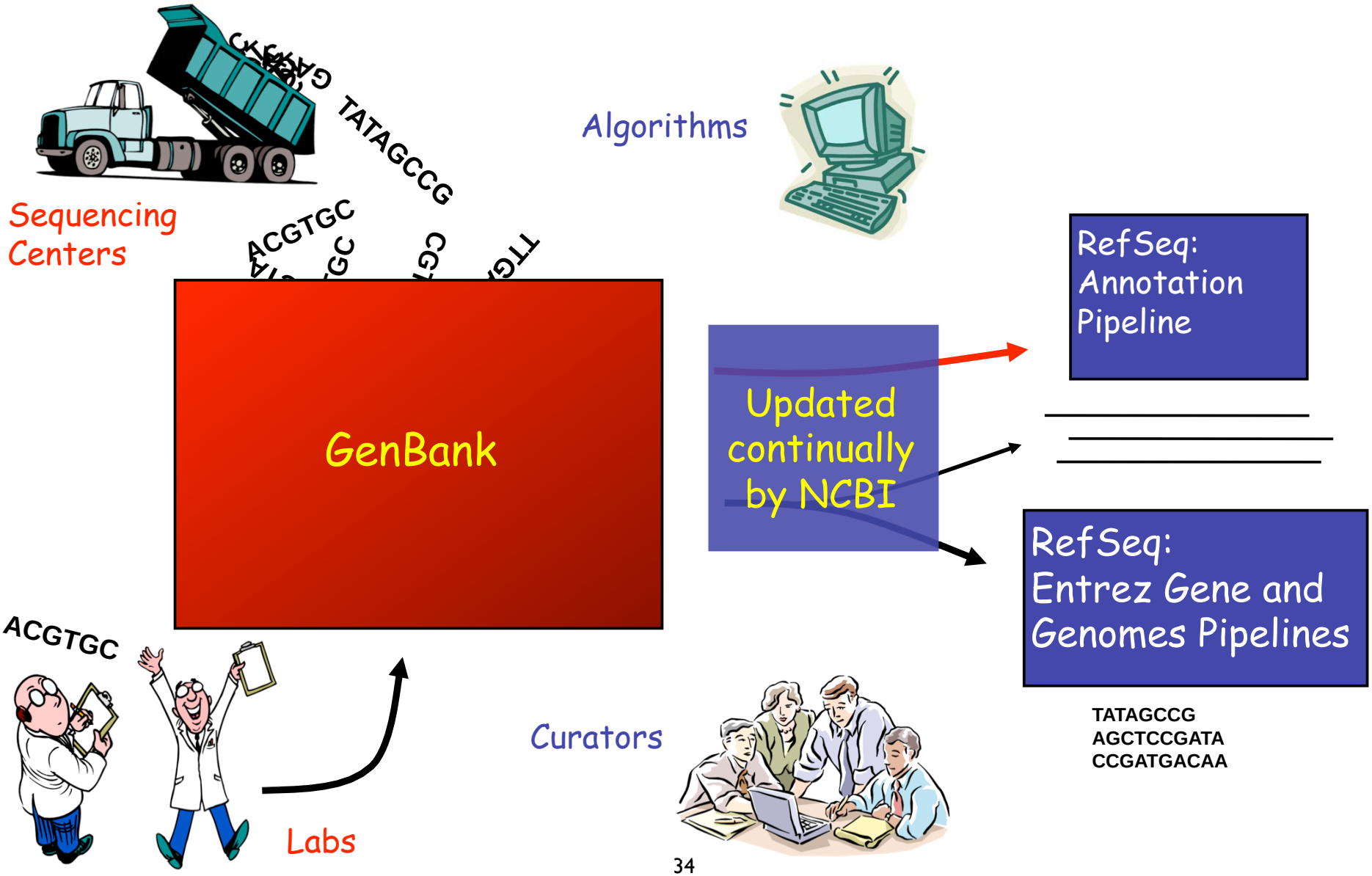
```
FEATURES             Location/Qualifiers
     source          1..2503
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="3"
                     /map="p21"
                     /tissue_type="gall bladder"
                   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

```
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
//
```

well annotated

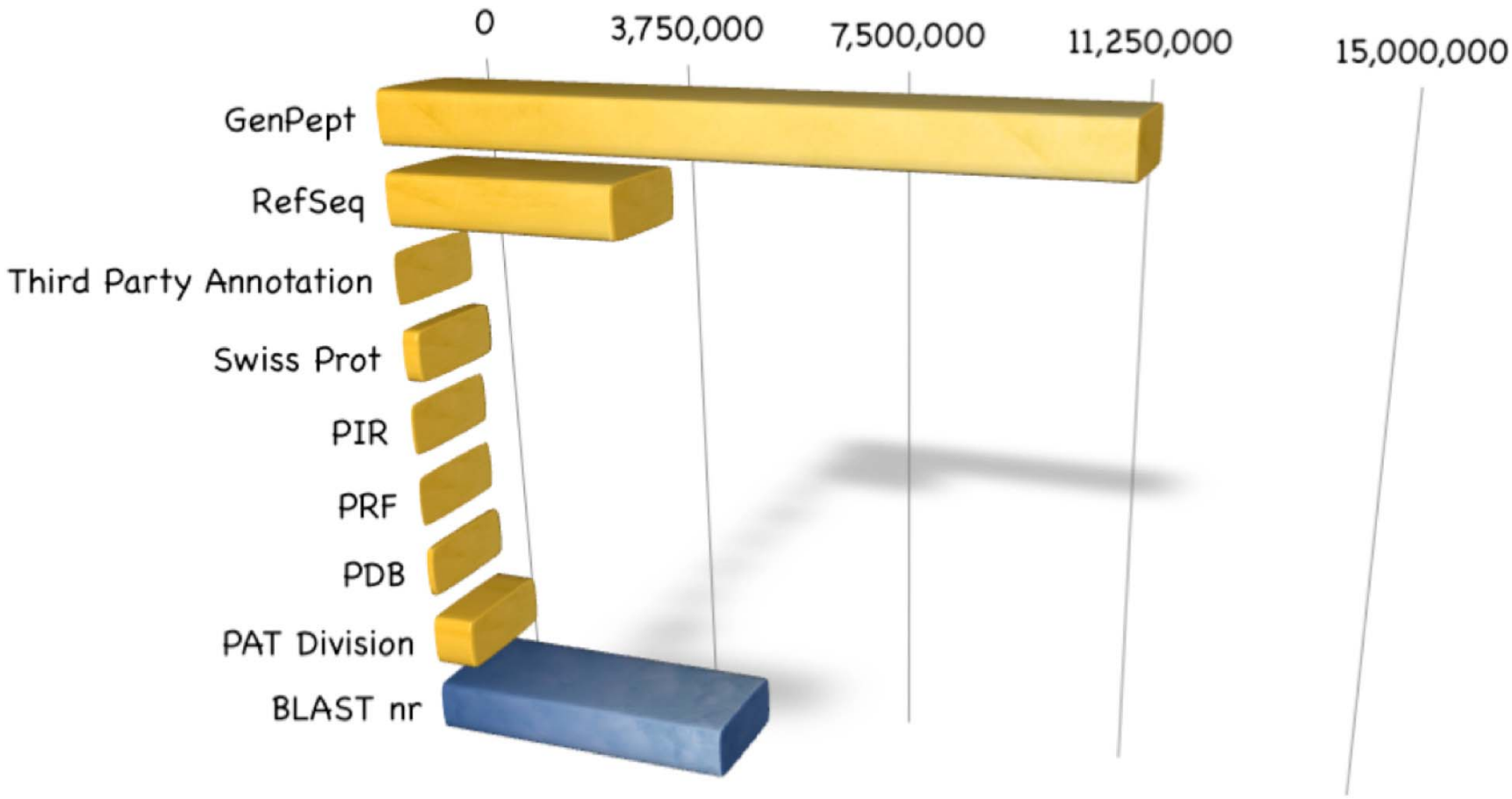
the sequence is the data

Primary vs. Derivative Databases



Derivative Databases

Entrez Protein



GenPept

- GenBank CDS translations

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

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

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.

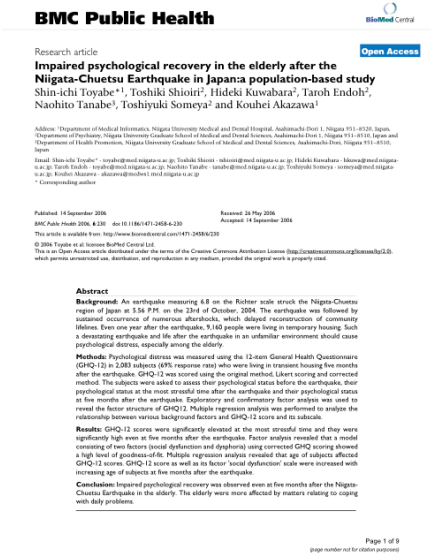
SCIENTIFIC AMERICAN

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@medw1.med.niigata-u.ac.jp

* Corresponding author

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

Received: 26 May 2006
Accepted: 14 September 2006

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

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

Abstract

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

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

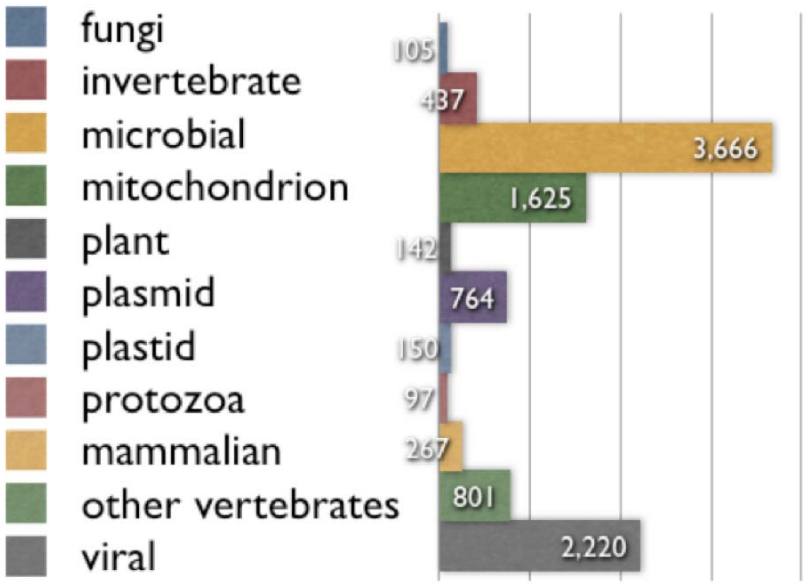
Results: GHQ-12 scores were significantly elevated at the most stressful time and they were significantly high even at five months after the earthquake. Factor analysis revealed that a model consisting of two factors (social dysfunction and dysphoria) using corrected GHQ scoring showed a high level of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

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

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

RefSeq

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



*refseq release 33

RefSeq Accession Numbers*

- prefix indicates the molecule type.

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

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

RefSeq Accession Numbers

- mRNAs and Proteins

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

- Genomic Records

NG_123456 Reference Genomic Sequence

- Chromosome

NC_123455 Microbial replicons, organelle, genomes, human chromosomes

- Assemblies

NT_123456	Contig
NW_123456	WGS Supercontig

Other NCBI Databases

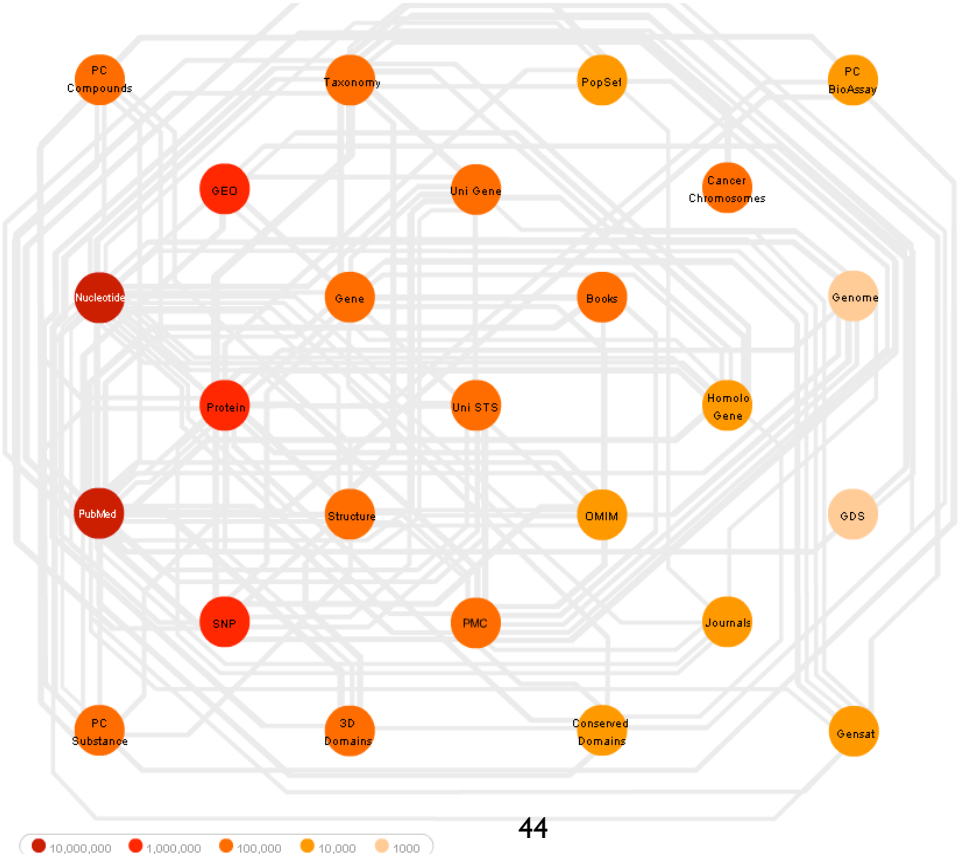
Structure:	imported structures (PDB)	Cn3D viewer, NCBI curation
CDD:	conserved domain database	Protein families (COGs and KOGs); Single domains (PFAM, SMART, CD)
dbSNP:	nucleotide polymorphism	variation data
Gene:	gene records	unified searchable database of genes, replaces locuslink
HomoloGene:	homologs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

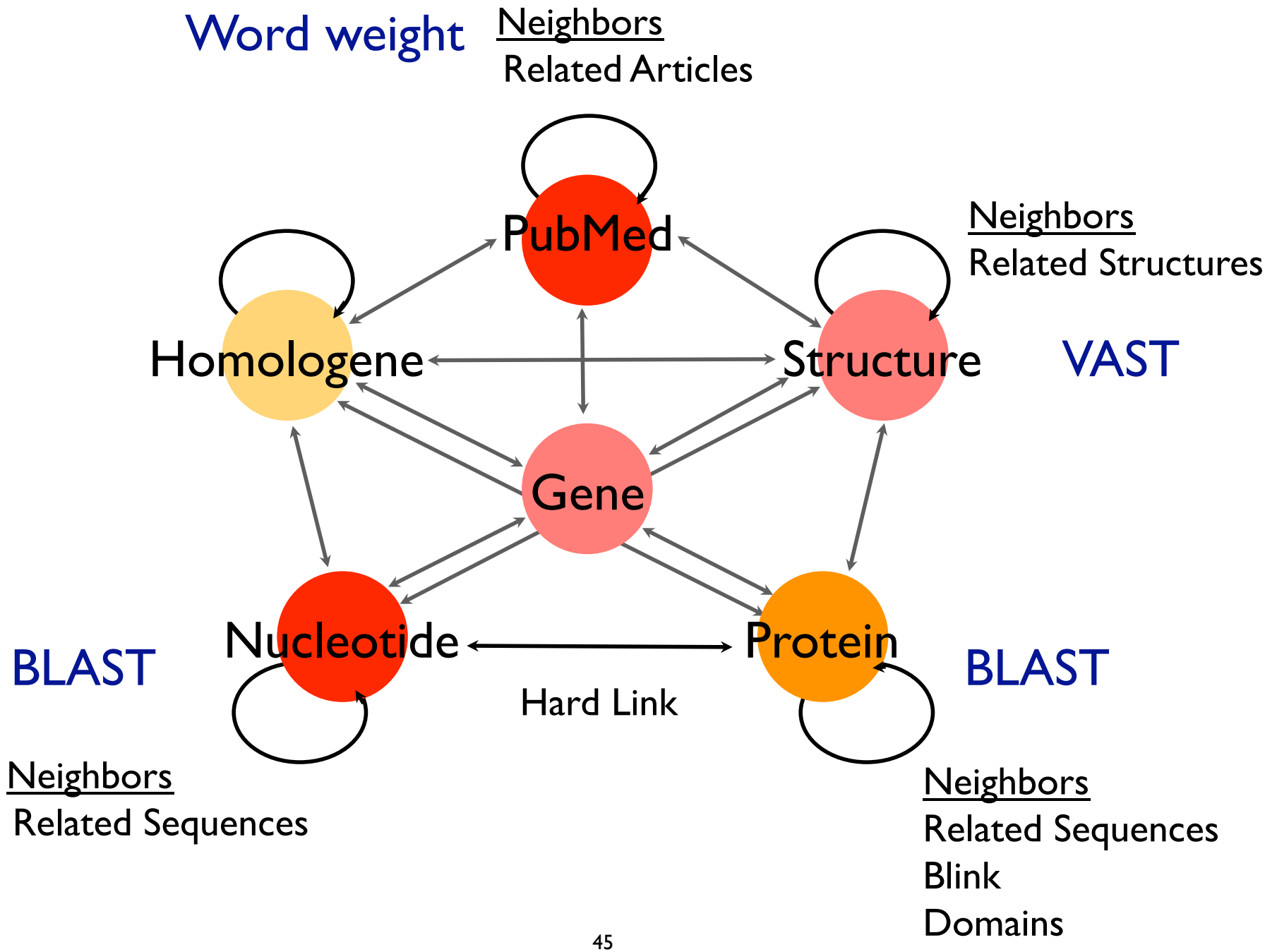
http://www.ncbi.nih.gov/Database/datamodel

NCBI Model of Entrez Databases Search Entrez for Go

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

This diagram requires Flash for viewing.




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

Neighbors in Entrez

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

Links

SNP

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

158 samples

Profile Neighbors, Sequence Neighbors, Links

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

Reporter: AF116645

Large-scale analysis of the human

GEO

1: MLH1

Order cDNA clone, Links

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

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

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

Location: 3p21.3

Gene

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

Related Articles, Links

PubMed

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

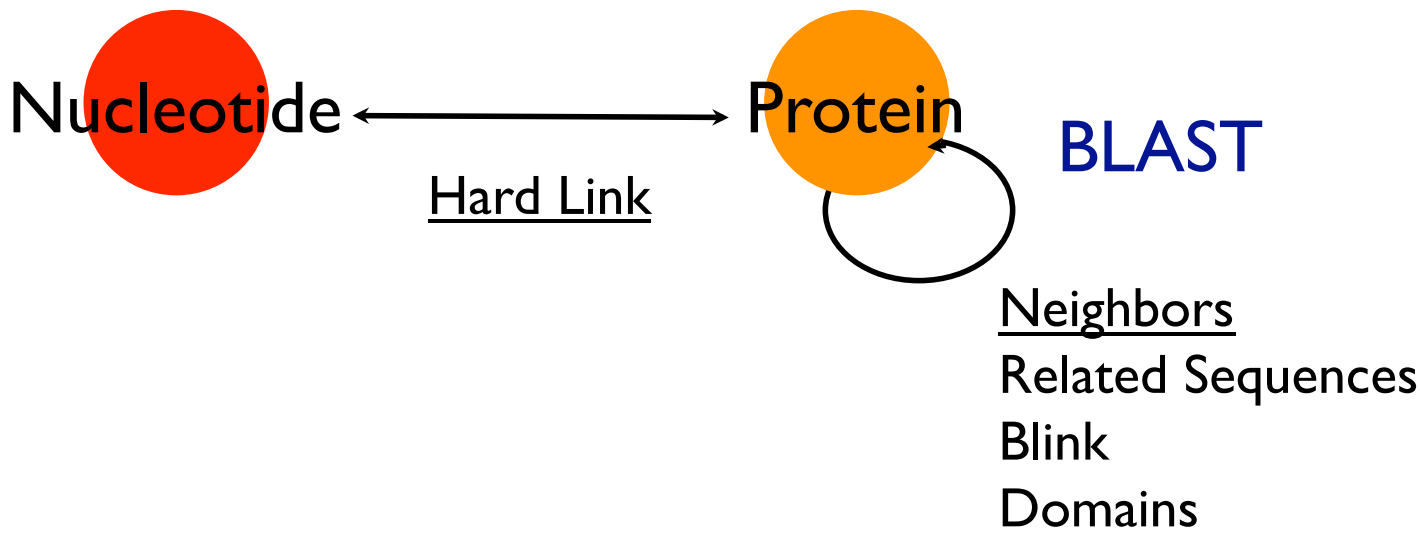
BLink, Conserved Domains, Links

Comment Features Sequence

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

Protein

Entrez - Linking Data



Blink & Domains

Neighbors: BLAST Link
pre-computed BLAST

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

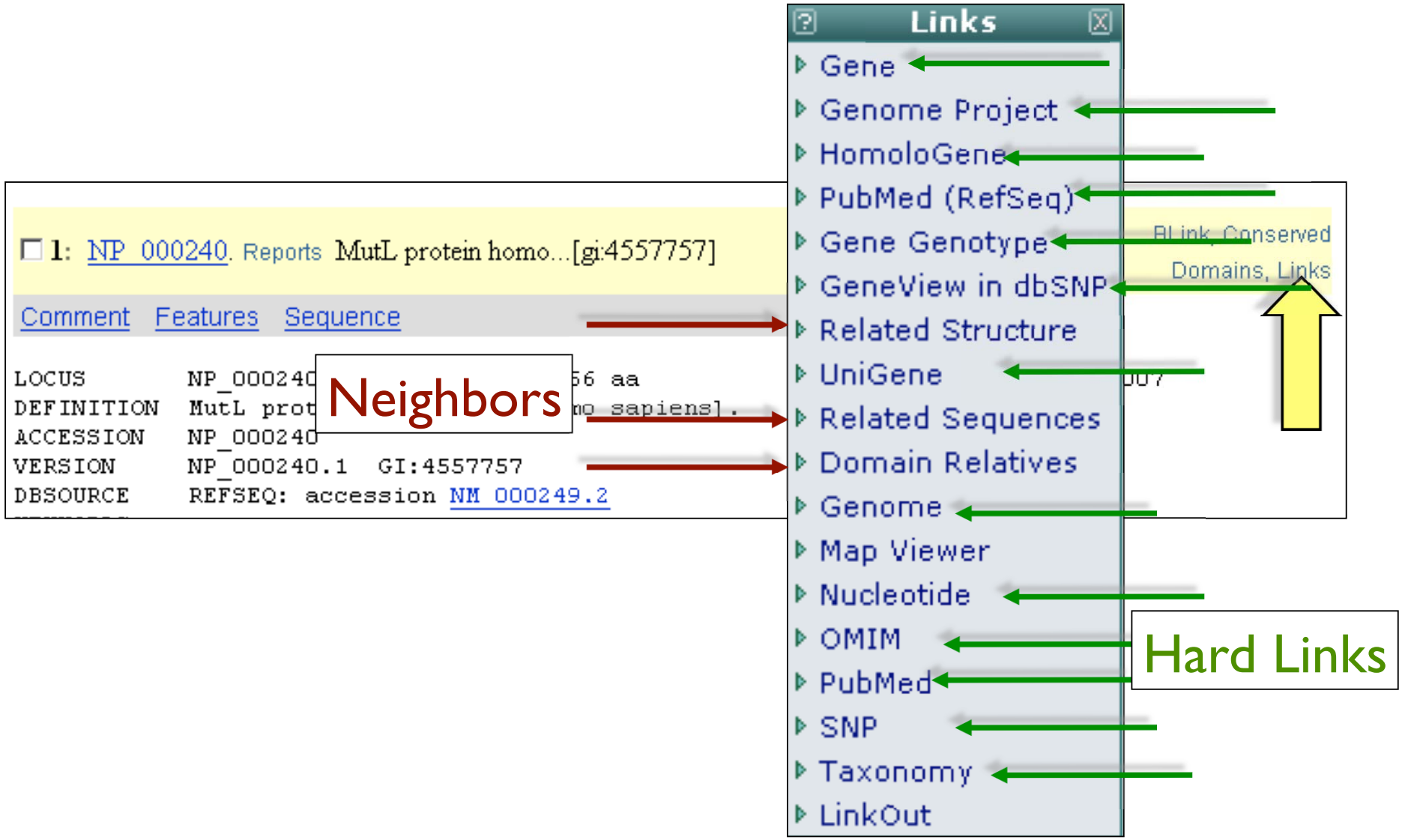
BLink, Conserved Domains, Links

Comment Features Sequence

Neighbors:
pre-computed CDD search

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

Links


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

Sequence Databases

GUIDED TOUR: Retrieving Data

Laboratory Bioinformatics Scenario:

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

transactions www.biochemsoctrans.org

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

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

Abstract

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

Mismatch repair

The mismatch repair system serves to correct errors that occur during DNA replication. These errors can take the form of misincorporated nucleotides that result in mispaired bases or insertion/deletion loops that can result from replication slippage at polynucleotide tracts [1,2]. The mismatch repair proteins are conserved from prokaryotes to humans. *Escherichia coli* uses homodimers of MutS and MutL proteins, while yeast and humans utilize multiple orthologues to each of MutS and MutL. The mismatch repair proteins function

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).

HNPCC (hereditary non-polyposis colorectal cancer)

HNPCC is an autosomal dominant disease that accounts

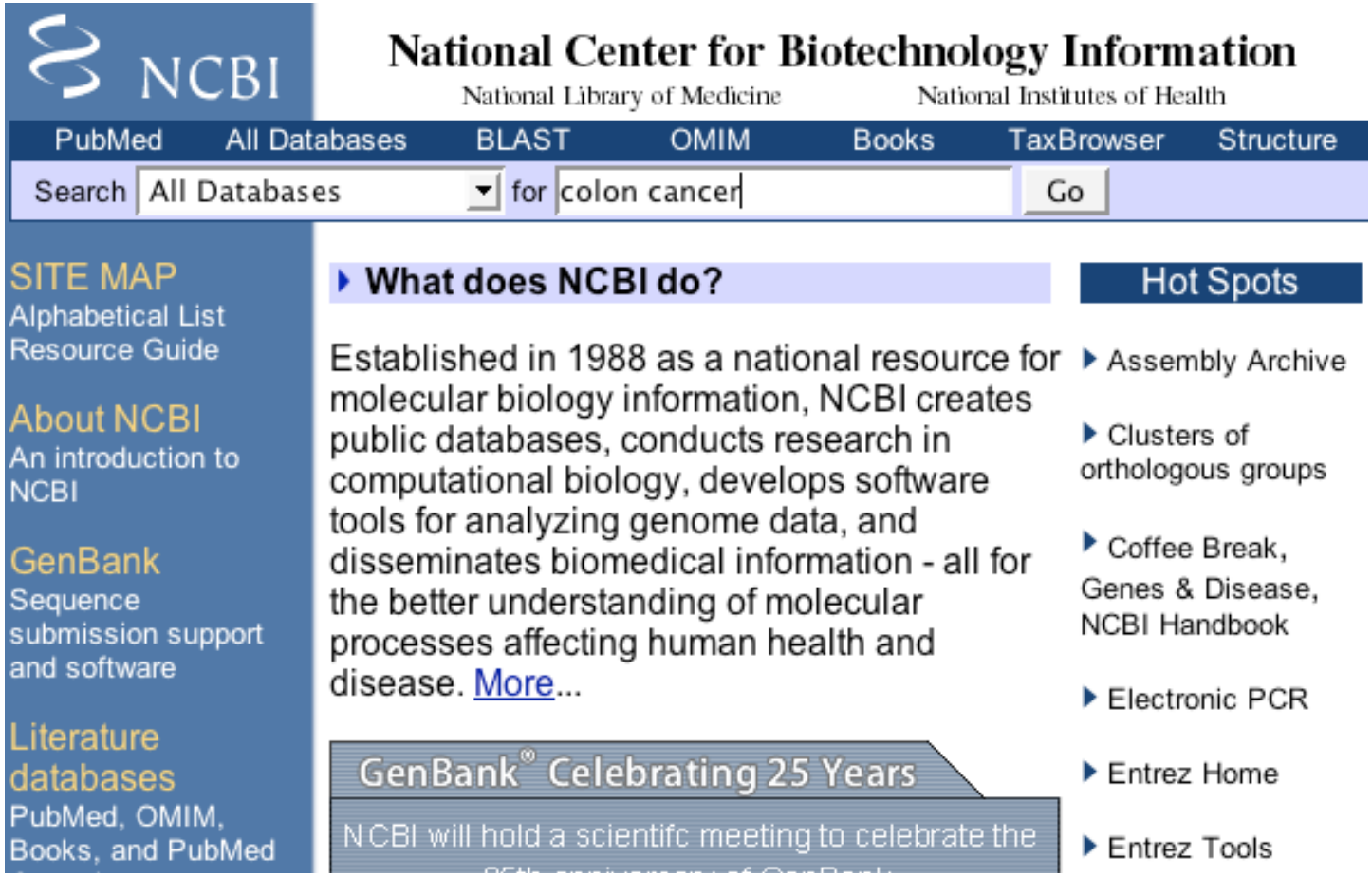
Database searching with Entrez

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

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

✓ Linking and neighboring with MLH1

Start with a search for "colon cancer"



NCBI

Entrez, The Life Sciences Search Engine

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

Search across databases colon cancer GO Clear Help

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

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

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

Human Disease Genes

NCBI

MIM *120436
Description
Cloning
Gene Function
Biochemical Features
Gene Structure
Mapping
Molecular Genetics
Animal Model
Allelic Variants
• View List
References
Contributors
Creation Date
Edit History

• Clinical Synopsis
• Gene map

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

LinkOut
HNPCC
HGVS
HGMD
GAD

OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University

My NCBI
[Sign In] [Reg

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

Search OMIM for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Detailed Show 20 Send to

*120436 GeneTests, Links

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

Gene map locus 3p21.3

TEXT

DESCRIPTION

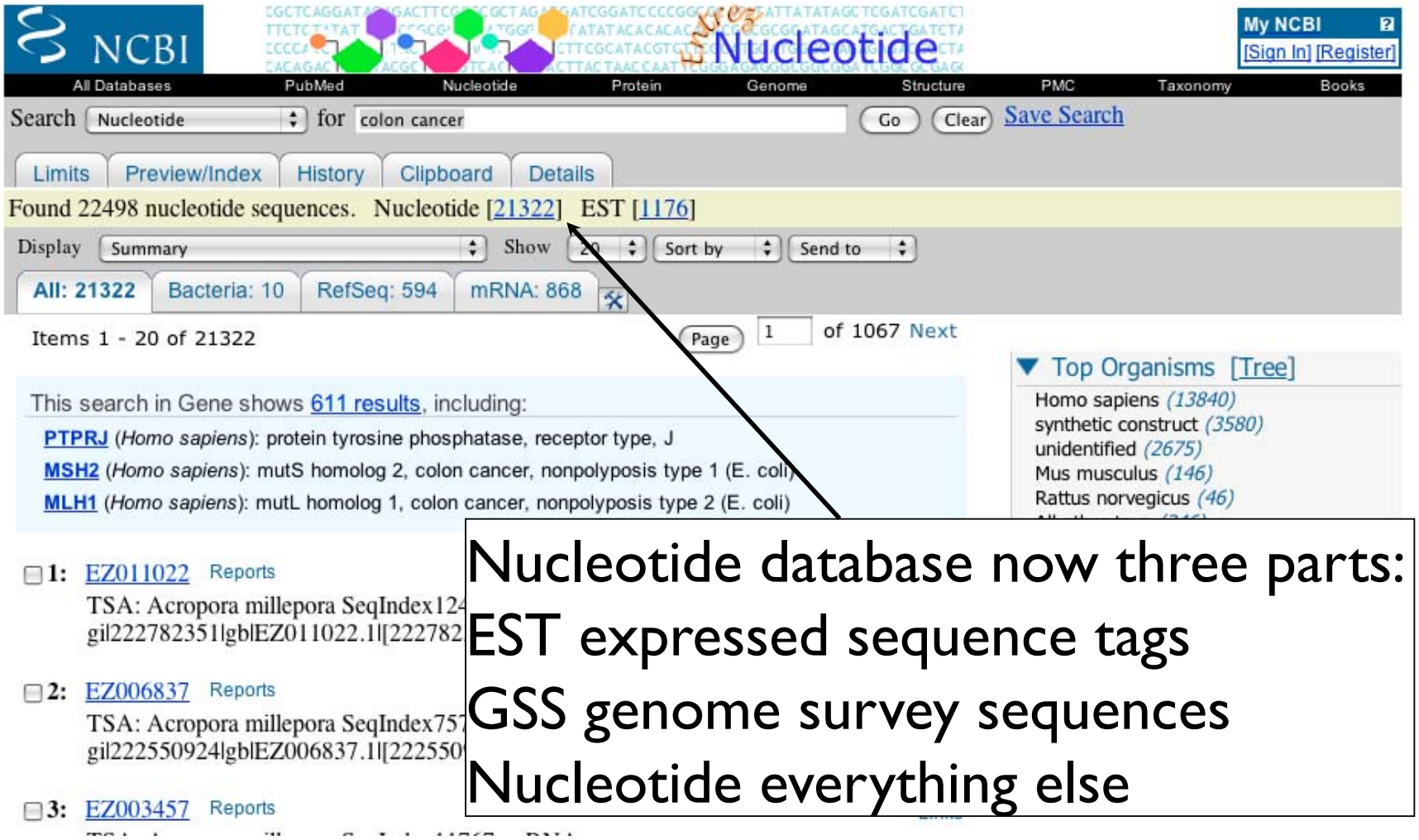
MLH is homologous to the E. coli MutL gene and is involved in DNA mismatch repair. Heterozygous mutations in the MLH1 gene result in hereditary nonpolyposis colorectal cancer-2 (HNPCC2; 609310) (Papadopoulos et al., 1994).

CLONING

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

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

Search Nucleotide



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

Advanced Search Options

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

Search Nucleotide for colon cancer Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Tabs

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

Display Summary Show 20 Sort by Send to

All 21322 | Bacteria: 10 | RefSeq: 594 | mRNA: 868

Items 1 - 20 of 21322 Page 1 of 1067 Next

This search in Gene shows 611 results, including:

- PTPRJ (*Homo sapiens*): protein tyrosine phosphatase, receptor type, J
- MSH2 (*Homo sapiens*): mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)
- MLH1 (*Homo sapiens*): mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)

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

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

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

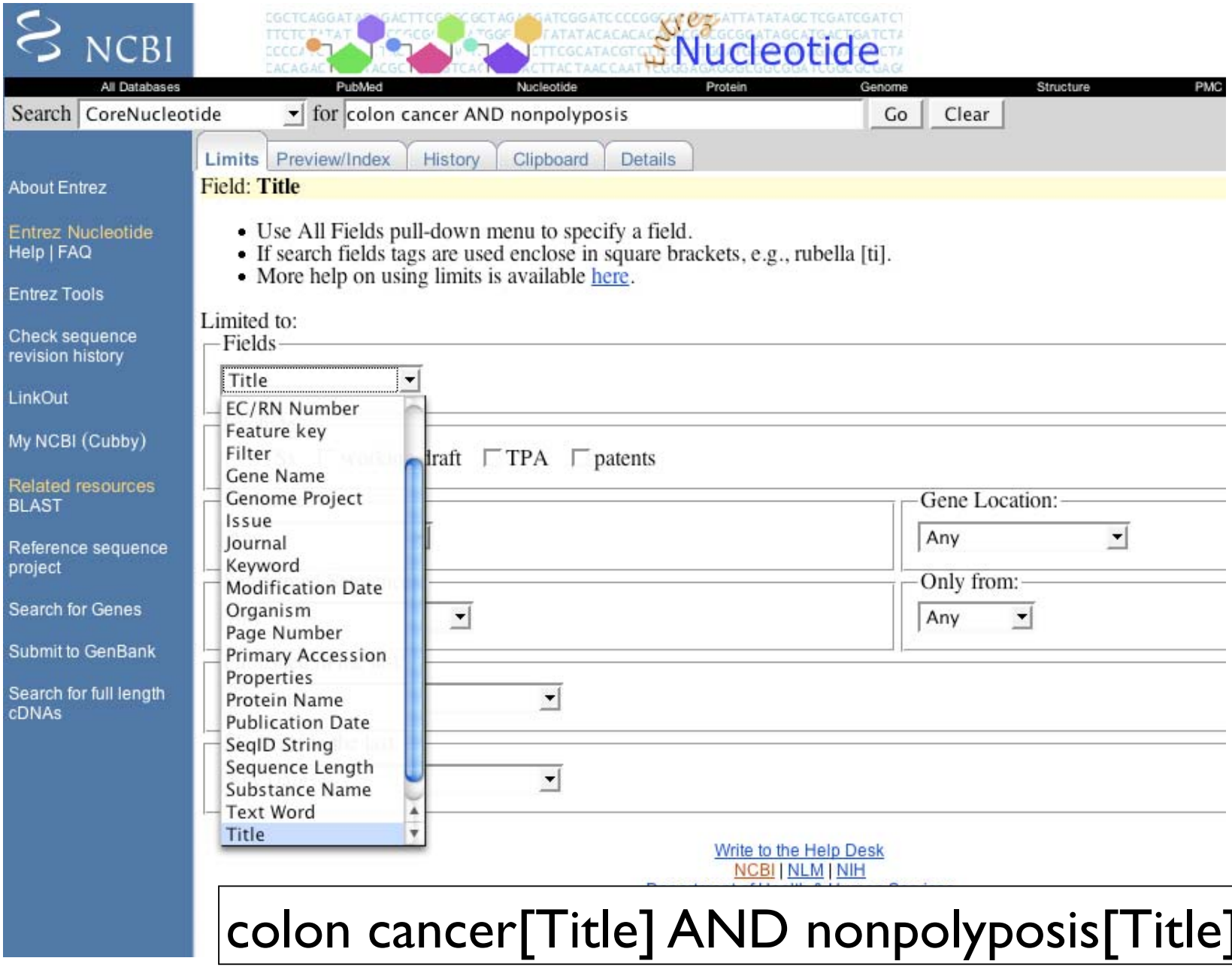
Top Organisms [Tree]
- Homo sapiens (13840)
- synthetic construct (3580)
- unidentified (2675)
- Mus musculus (146)
- Rattus norvegicus (46)
- All other taxa (246)

More...

Recent Activity

Tur

Your browsing activity is empty



colon cancer[Title] AND nonpolyposis[Title]

NCBI

Entrez Nucleotide

My NCBI
Welcome joannealisonfox. [Sign Out]

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

Search CoreNucleotide for colon cancer AND nonpolyposis Go Clear

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI (Cubby)

Related resources
BLAST

Reference sequence project

Search for Genes

Submit to GenBank

Search for full length cDNAs

Field: **Title**

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

Limited to:

Fields
Title

Exclude
☐ STSs ☐ working draft ☐ TPA ☐ patents

Molecule:
mRNA

Gene Location:
Any

Segmented Sequences:
Any

Only from:
RefSeq

Published in the last:
Any Date

Modified in the last:
Any Date

Write to the Help Desk

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

Advanced Search Options

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

Search Nucleotide for colon cancer Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Tabs

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

Display Summary Show 20 Sort by Send to

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

Items 1 - 20 of 22 Page 1 of 1067 Next

This search in Gene shows 611 results, including:

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

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

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

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

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

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

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

Recent Activity
Tur
Your browsing activity is empty

NCBI

Entrez Nucleotide

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

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

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI (Cubby)

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

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

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

Add Term(s) to Query or View Index:

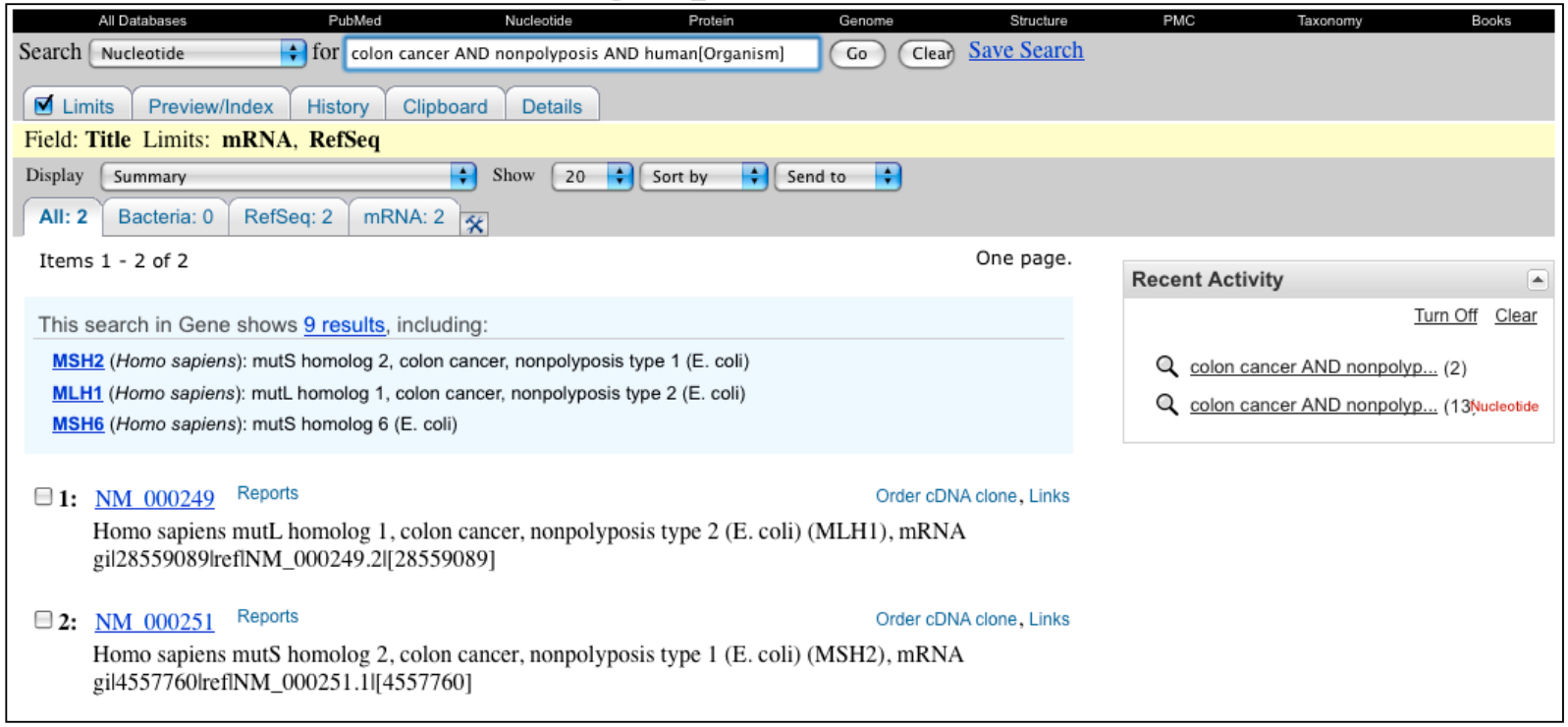
- Enter a term in the text box; use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within a field.

All Fields | Preview | Index

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

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

Refining your Search



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

Useful Field Restrictions

- **[Title]:** Definition line in GenBank / GenPept format shown in Summary format
  - `glyceraldehyde 3 phosphate dehydrogenase[Title]`
- **[Organism]:** NCBI's taxonomy. Organizing system for molecular databases
  - `mouse[organism]; green plants[organism]; Streptomyces coelicolor[organism]`
- **[Properties]:** molecule type, location, database source
  - `biomol_mrna[properties]; biomol_genomic[properties]; gene_in_mitochondrion[properties]; srcdb_pdb[properties]`
- **[Filter]:** subsets of data, Entrez links
  - `all[filter]; nucleotide mapview[filter]; nucleotide_omim[filter]`

NCBI

Entrez Nucleotide

My NCBI
Welcome joannealisonfox. [Sign Out]

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

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

Limits | Preview/Index | History | Clipboard | Details

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

Display Summary Show 20 Sort by Send to

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

Items 1 - 2 of 2 One page.

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

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

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

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

NCBI Entrez Nucleotide

My NCBI [Sign In] [Register]

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

Search CoreNucleotide for Go Clear

Limits Preview/Index History Clipboard Details

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

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

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

Comment Features Sequence

```
LOCUS       NM_000249                2524 bp    mRNA    linear   PRI 20-AUG-2007
DEFINITION  Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E.
            coli) (MLH1), mRNA.
ACCESSION   NM_000249
VERSION     NM_000249.2  GI:28559089
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2524)
  AUTHORS   Perri,F., Cotugno,R., Piepoli,A., Merla,A., Quitadamo,M.,
            Gentile,A., Pilotto,A., Annese,V. and Andriulli,A.
  TITLE     Aberrant DNA methylation in non-neoplastic gastric mucosa of H.
            Pylori infected patients and effect of eradication
  JOURNAL   Am. J. Gastroenterol. 102 (7), 1361-1371 (2007)
   PUBMED   17509026
  REMARK    GeneRIF: While CDH1 methylation seems to be an early event in Hp
            gastritis, MLH1 methylation occurs late along with IM.
REFERENCE   2  (bases 1 to 2524)
  AUTHORS   Bettstetter,M., Dechant,S., Ruemmele,P., Grabowski,M., Keller,G.,
            Holinski-Feder,E., Hartmann,A., Hofstaedter,F. and Dietmaier,W.
  TITLE     Distinction of hereditary nonpolyposis colorectal cancer and
            sporadic microsatellite-unstable colorectal cancer through
            quantification of MLH1 methylation by real-time PCR
  JOURNAL   Clin. Cancer Res. 13 (11), 3221-3228 (2007)
   PUBMED   17545526
  REMARK    GeneRIF: quantitative MLH1 methylation analysis in MSI-H CRC is a
            valuable molecular tool to distinguish between HNPCC and sporadic
            MSI-H CRC
REFERENCE   3  (bases 1 to 2524)
  AUTHORS   Takahashi,M., Shimodaira,H., Andreutti-Zaugg,C., Iggo,R.,
            Kolodner,R.D. and Ishioka,C.
  TITLE     Functional analysis of human MLH1 variants using yeast and in vitro
            mismatch repair assays
  JOURNAL   Cancer Res. 67 (10), 4595-4604 (2007)
   PUBMED   17510385
  REMARK    GeneRIF: The 101 MLH1 variants were examined for the dominant
```

Links

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

NCBI

Entrez Nucleotide

My NCBI [Sign In] [Register]

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

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

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Entrez Nucleotide
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI (Cubby)

Related resources
BLAST

Reference sequence project

Search for Genes

Submit to GenBank

Search for full length cDNAs

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

Found 2 nucleotide sequences. CoreNucleotide [2]

Display Summary Show 20 Sort by Send to

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

Items 1 - 2 of 2 One page.

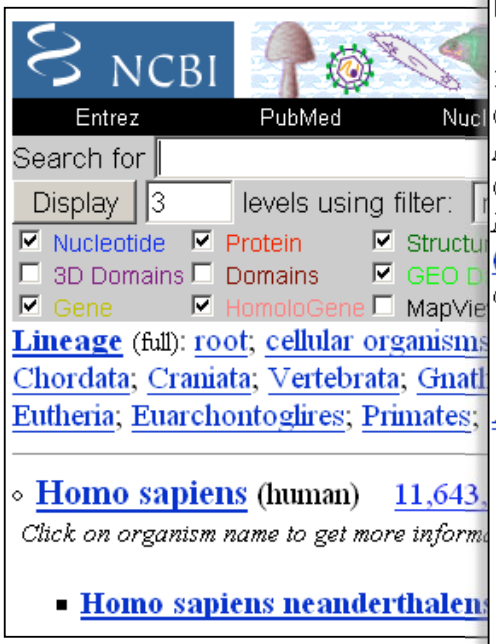
1: NM_000249 Reports
Homo sapiens mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1), mRNA
gi|28559089|ref|NM_000249.2|[28559089]

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

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

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

Taxonomy



Homo sapiens

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

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

Entrez records

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

Genome Information

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

Genome view: 24 chromosomes

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

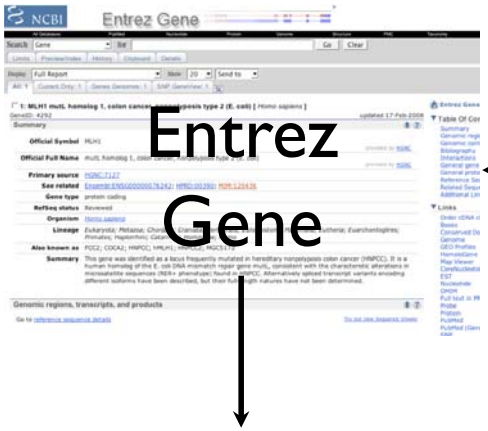
All molecular databases

Goal: Find MLH1 homologs

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

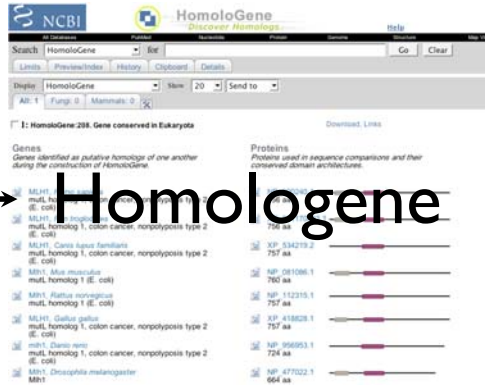
Protein

BLink



Entrez Gene

Other Entrez Databases



Homologene

Gene neighbors

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

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

Limits | Preview/Index | History | Clipboard | Details

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

Display Summary Show 20 Sort by Send to

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

Items 1 - 2 of 2 One page.

This search in Gene shows 9 results, including:

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

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

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

1: NM_000249 Reports

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

2: NM_000251 Reports

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

Recent Activity

Turn Off Clear

colon cancer AND nonpolyp... (2)

colon cancer AND nonpolyp... (13 Nucleotide

Links

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

MLH1 Gene Record

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

GeneID: 4292 updated 10-Apr-2007

Summary

Official Symbol	MLH1	provided by HGNC
Official Full Name	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	provided by HGNC
Primary source	HGNC:712	
See related	HPRD:003	
Gene type	protein co	
RefSeq status	Reviewed	
Organism	*Homo sap*	

Genomic regions, transcripts, and products

Go to reference sequence details

NC_000003.10

[37009983 ▶ [37067341 ▶

5' 3'

NM_000249.2 NP_000240.1 CCDS2663.1

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

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

See MLH1 in MapViewer

[37383246 ▶

A4

TCEA1P2

Interactions + GO

Interactions

Description .....

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

GeneOntology Provided by GOA

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

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

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

Sequences

NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. Explain

mRNA and Protein(s)

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

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

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

RefSeqs of Annotated Genomes: Build 36.2

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

Reference assembly

Genomic

1. **NC_000003.10 Reference assembly**

Range	37009983..37067341
Download	GenBank FASTA

2. **NT_022517.17**

Range	36974983..37032341
Download	GenBank FASTA

Alternate assembly (based on Celera assembly)

Genomic

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

Range	36977744..37035102
Download	GenBank FASTA

2. **NW_921651.1**

Range	36977744..37035102
Download	GenBank FASTA

Related Sequences

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

MLH1: Sequence Links

Genomic regions, transcripts, and products

Go to refere sequence details

NC_000003.10

NM_000249.2 NP_000240.1 CCDS2663.1

- coding region - untranslated region

Links

mRNA LINKS

- FASTA
- GENBANK

Links

PROTEIN LINKS

- FASTA
- GENPEPT
- Blink
- Conserved Domains

chromosome: 3; Location: 3p21.3

in MapViewer

[36992791 [37383246

LOC645571 LRRFIP2 GOLGA4

EPM2AIP1 TCEA1P2

MLH1

Links Explain

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

NCBI Entrez Gene

My NCBI
Welcome joannealisonfox. [Sign Out]

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

Search Gene for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Full Report Show 20 Send to

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

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

GeneID: 4292 updated 16-Sep-2007

Summary

Official Symbol	MLH1 provided by HGNC
Official Full Name	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) provided by HGNC
Primary source	HGNC:7127
See related	Ensembl:ENSG00000076242; HPRD:00390; MIM:120436
Gene type	protein coding
RefSeq status	Reviewed
Organism	*Homo sapiens*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Also known as	FCC2; COCA2; HNPCC; hMLH1; HNPCC2; MGC5172
Summary	This gene was identified as a locus frequently mutated in hereditary nonpolyposis colon cancer (HNPCC). It is a human homolog of the E. coli DNA mismatch repair gene mutL, consistent with the characteristic alterations in microsatellite sequences (RER+ phenotype) found in HNPCC. Alternatively spliced transcript variants encoding different isoforms have been described, but their full-length natures have not been determined.

Genomic regions, transcripts, and products

Go to reference sequence details

NC_000003.10

Entrez Gene Home

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

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

Finding Homologs:

1: HomoloGene:208. Gene conserved in Eukaryota

Download, Links

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

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

HomoloGene Downloader

Homologene:208. Gene conserved in Eukaryota

Download | Protein | sequences (in FASTA format)

Include 0 bp upstream of gene

Include 0 bp downstream of gene

Select which sequences should be included

Select All | Unselect All

Species	Gene		
H.sapiens	MLH1	NM_0	
P.troglodytes	MLH1	XM_0	
C.familiaris	LOC477019	XM_5	
M.musculus	Mlh1	NM_0	
R.norvegicus	Mlh1	NM_031053.1	NP_112315.1
G.gallus	MLH1	XM_418828.1	XP_418828.1
D.melanogaster	Mlh1	NM_057674.2	NP_477022.1
A.gambiae	AgaP_ENSANGG00000011527	XM_320342.2	XP_320342.2
A.gambiae	ENSANGG00000010995	XM_307435.2	XP_307435.2
S.pombe	SPBC1703.04	NM_001022118.1	NP_596199.1
S.cerevisiae	MLH1	MLH1_6323819	NP_013890.1
K.lactis	KLLA0D09955g	XM_453504.1	XP_453504.1
E.gossypii	GeneID:2757243	NM_210705.1	NP_985351.1
N.crassa	NCU08309.1	XM_329014.1	XP_329015.1
A.thaliana	ATMLH1	NM_116882.2	NP_567345.2

Protein
mRNA
Genomic

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	724 aa

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

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

mutL homolog 1 (E. coli)

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

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

D.melanogaster Mlh1
Mlh1

A.gambiae AgaP_ENSANGG000000
ENSANGP00000014016

A.gambiae ENSANGG00000010995
ENSANGP00000013484

S.pombe SPBC1703.04
hypothetical protein

S.cerevisiae MLH1
Mlh1p

K.lactis KLLA0D09955g
mRNA gene KLLA0D09955g

E.gossypii GeneID:2757243
Eremothecium gossypii AFL199C gen

N.crassa NCU08309.1
hypothetical protein

A.thaliana ATMLH1
ATMLH1

O.sativa Os01g0958900
mRNA gene Os01g0958900

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

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

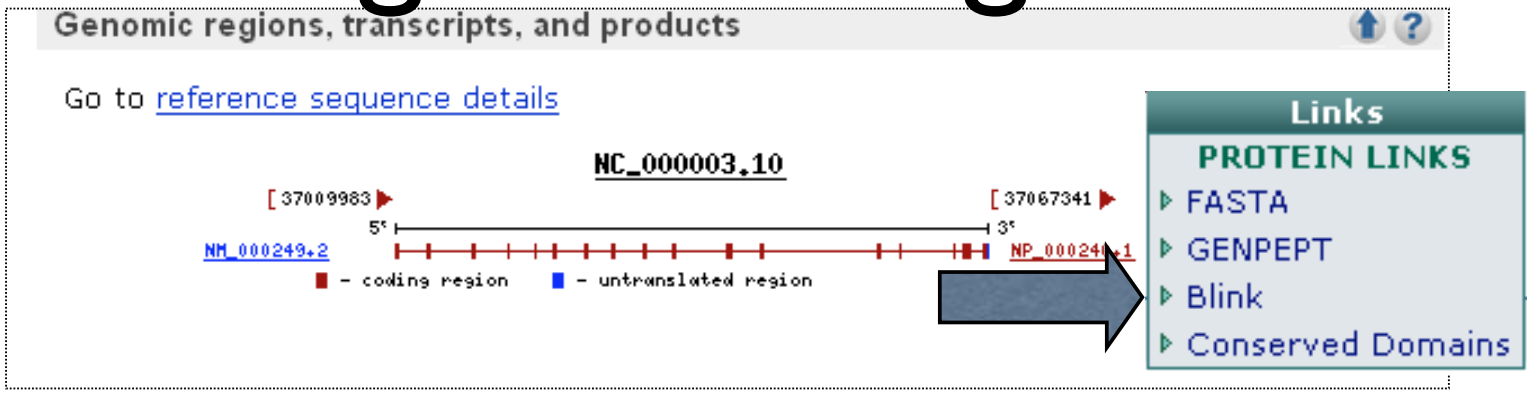
NP_985351.1
771 aa

NP_001045457.1
724 aa

Gene Links

Protein Links

Finding Homologs 2: BLink



BLink: BLAST Link

BLINK *precomputed BLAST*

Home | Taxonomy Report | Multiple Alignment | Blast | Help

My NCBI [Sign In] [Register]

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

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

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

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

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

Reset all filters

Choose Display Options

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

Results: 1 - 100 Next Page Last

% hits

reset selection

756 aa

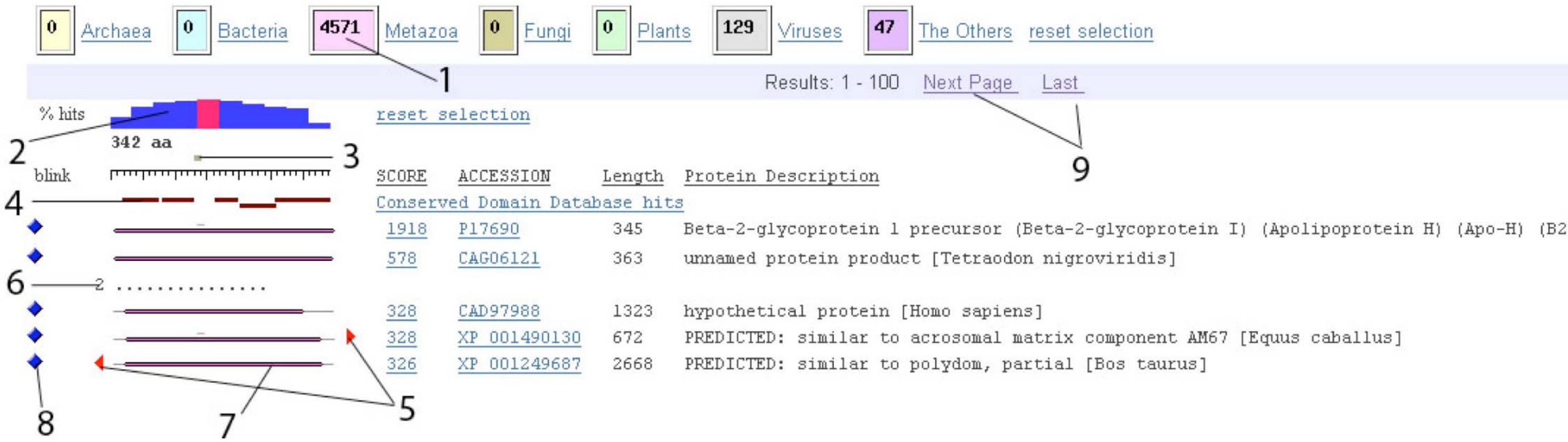
blink

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

BLINK

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

BLINK precomputed BLAST



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

Sample Questions that can be answered with BLink

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

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

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

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

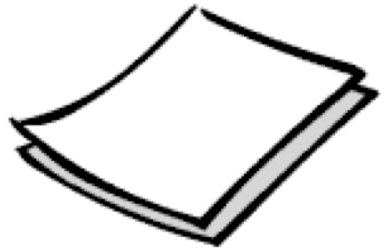
5. What domains are present in the query sequence?

Sequence Databases

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

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

navigate to:
bioteach.ubc.ca/bioinfo2009



Let's compare our results

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

Use the preview tab and feature keys

Strategy #1: search nt

Strategy #2: search entrez gene

Check your History

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

Advanced Tips & Tricks for Searching PubMed

My NCBI

Bookshelf

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

NCBI

PubMed www.pubmed.gov

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

My NCBI [Sign In] [Register]

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

Search PubMed for cancer AND carrots Preview Go Clear

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Text Version

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

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

Related Resources
Order Documents

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

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

Clear History

New PubMed display search: TPH1

All: 128 Review: 11

Gene Information

TPH1 tryptophan hydroxylase 1 [Homo sapiens]

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

Location: 11p15.3-p14

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

Gene

Items 1 - 20 of 128

Page 1 of 7 Next

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

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

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

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

Also try:

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

Titles with your search terms

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

» See all...

Recent Activity

Turn Off Clear

TPH1 (128)

The medical treatment of obsessive-compulsive disorder and anxiety.

clomipramine (3295) Pubmed

Maylandia zebra M...[gi:193992698]

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

The Abstract plus page

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

Links

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

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

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

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

PMID: 18923670 [PubMed - indexed for MEDLINE]

PMCID: PMC2565062

Related Articles

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

» See Reviews... | » See All...

Recent Activity

Turn Off Clear

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

NCBI

National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

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

Search All Databases for Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

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

Tools
Data mining

What does NCBI do?

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

New **Protein Clusters**
Entrez Protein Clusters database

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

1 Billion Live Traces

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

PubMed Central

Hot Spots

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

NCBI **Bookshelf**

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

Search Books for plasmodium Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Books Show 20 Send to

All: 328 | Figures: 18

6 items in **The Intolerable Burden of Malaria: II. What's New, What's Needed**

Breman, Joel G.; Alilio, Martin S.; Mills, A., editors
Northbrook (IL): The American Society of Tropical Medicine and Hygiene; c2004

3 items in **Molecular Biology of the Cell**

Alberts, Bruce; Johnson, Alexander; Lewis, Julian; Raff, Martin; Roberts, Keith; Walter, Peter
New York and London: Garland Science; c2002

3 items in **Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach**

Gruber, Arthur; Durham, Alan M.; Huynh, Chuong; del Portillo, Hernando A., editors
Bethesda (MD): National Library of Medicine (US), NCBI; 2008

2 items in **Medical Microbiology**

Baron, Samuel, editor.
Galveston (TX): University of Texas Medical Branch; c1996

Control of Gene Expression in *Plasmodium*

Mauro Ferreira de Azevedo[1] and Hernando A. del Portillo[2]

[1]*Departamento de Parasitologia, Instituto de Ciências Biomédicas, Universidade de São Paulo. Av. Lineu Prestes 1374, São Paulo, SP 05508-900, Brazil*

[2]*Present Address: Barcelona Centre for International Health Research (CRESIB), Hospital Clinic/IDIBAPS, Universitat de Barcelona, Roselló 132, 4a planta, 08036, Barcelona, Spain. Phone: 34 93 2275706; Fax: 34 93 4515272*

Created July 17, 2006.
Last update May 11, 2007.

Malaria parasites have more than 10 stages of cellular differentiation and invade at least four types of cells in two different hosts with a considerable variation in temperature between them. All of this complex biology depends on the efficient control of gene expression, about which our knowledge still has many shortcomings. Although this parasite has some general mechanisms in common with yeast and higher eukaryotes, many aspects of its genetic regulation seem to be specific to this genus: (i) during the asexual blood stages, the parasites seem to turn on a rigid, viral-like program of early, middle, and late genes expressed as a cascade of continuous events; (ii) it seems likely that malaria parasites have acquired unique and yet-to-be-described transcription factors; (iii) antisense transcription has been described in about 10% of the coding genome, clearly indicating as-yet-undefined, post-transcriptional control mechanisms; and (iv) control gene expression of the var subtelomeric multigene family involves a gene-specific cross-talk between intron and exon, as well as epigenetic mechanisms to control allelic exclusion. Here, we review our present knowledge on control of gene expression in malaria parasites and illustrate the importance of bioinformatics in advancing our knowledge in this area, with illustrative examples on promoters, transcription

In this page

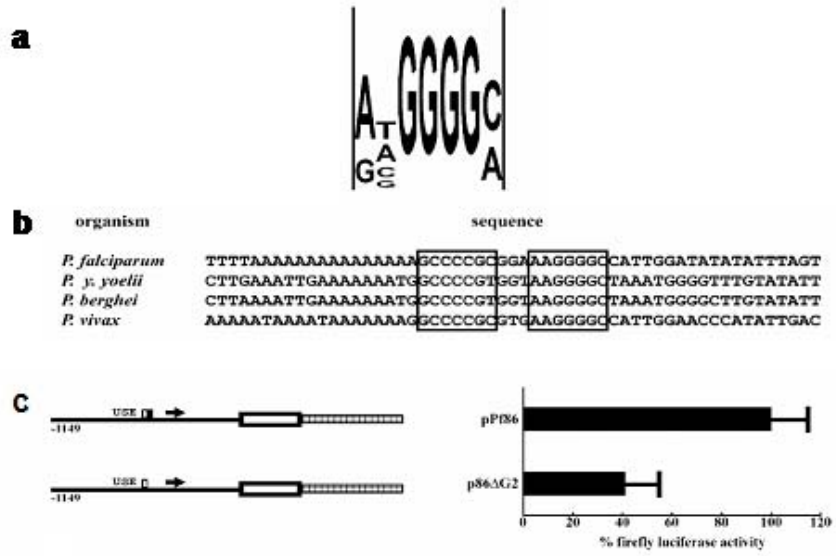


Figure 3. A functional G-rich palindromic element is conserved among the intergenic regions of *hsp* genes in *Plasmodium*. a. G-rich element. b. G-rich palindromic element in different malaria species. c. A reporter plasmid containing the intergenic region of the *hsp* gene of *P. falciparum* and driving the expression of the luciferase reporter gene is functional. Data and figures obtained with permission from Dyanne Wirth.

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

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

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

NCBI Home PubMed GenBank BLAST Sign In | My NCBI

My NCBI

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

Table of Contents

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

Welcome to My NCBI

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

Sign into My NCBI

Username

Password

Keep me signed in

Remember my username

Sign In

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

See more sign in options for My NCBI partner organizations.

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

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

My NCBI

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

Table of Contents

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

My NCBI Home » Saved Data

My Saved Data

Bibliographies

My Bibliography	6 Items
Other Citations	Not Created

Saved Searches (Manage)

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

Collections (Manage)

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

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

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

NCBI

PubMed www.pubmed.gov

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

My NCBI [Sign In] [Register]

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

Search PubMed for Go Clear Advanced Search

Limits | Preview/Index | History | Clipboard | Details

About Entrez
Text Version

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

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

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

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

Search terms may be topics, authors or journals.

The NIH Public Access Policy May Affect You

Does NIH fund your work?

Then your manuscript must be made available in PubMed Central

How?

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

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

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

search with a gene name of interest to you

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

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

Bibliography

Related Articles in PubMed

PubMed links

GeneRIFs: Gene References Into Function

What's a Gene

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

Submit: N

Follow link from PubMed to Entrez Gene

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

NCBI

PubMed www.pubmed.gov

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

My NCBI
Welcome joannealisonfox. [Sign Out]

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

Search PubMed for Go Clear Advanced Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort By Send to

Send to
Text
File
Printer
Clipboard
Collections
E-mail
Order

All: 493 Review: 20

Items 1 - 20 of 493

Page 1 of 25 Next

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

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

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

Recent Activity
Turn Off Clear

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

PTEN phosphatase and tensin homolog [Homo sapiens]

pten (3788) PubMed

mlh1 AND cmed6[book] (13) Books

Toward a confocal subcellular atlas of the human proteome.

Save your PubMed results to your MyNCBI collections

Credits

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

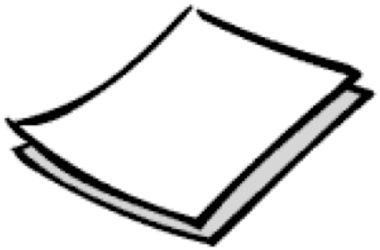
  NCBI HelpDesk - Field Guide Course Materials

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

- Questions? Please contact:

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


AMBL I The Educational Facilities of the Michael Smith Labs
AMBL
(don't be a fly on the wall - participate!)
LABORATORY BIOINFORMATICS
This workshop will focus on bioinformatics techniques for practical use in the laboratory. Hands-on exercises for retrieving data, primer design, BLAST searching, and genomics data navigation will be covered. Primarily aimed at researchers who are new to the area, or familiar but require a quick updating, where content covered can be tailored to laboratory needs.
Written by AMBL
RESOURCES
UNIVERSITY »
LABORATORY BIOINFORMATICS WORKSHOP, FEBRUARY 16-18TH, 2009
This workshop will focus on bioinformatics techniques for practical use in the laboratory. Hands-on exercises for retrieving data, primer design, BLAST searching, and genomics data navigation will be covered. Primarily aimed at researchers who are new to the area, or familiar but require a quick updating, where content covered can be tailored to laboratory needs.
joanne@msl.ubc.ca
Laboratory Bioinformatics
Common tools, useful databases, and tricks of the trade for practical use in the laboratory.
bioteach.ubc.ca/bioinfo2009
Inside
Pages
ABOUT
GENETICS PHENOTYPE
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 Module #2

BLAST background, guided tour & practical exercises

joanne@msl.ubc.ca

BLink: BLAST Link

BLINK *precomputed BLAST*

Home | Taxonomy Report | Multiple Alignment | Blast | Help

My NCBI [Sign In] [Register]

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

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

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

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

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

Reset all filters

Choose Display Options

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

Results: 1 - 100 Next Page Last

% hits

reset selection

756 aa

blink

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

BLAST

Finding Function By Sequence Similarity

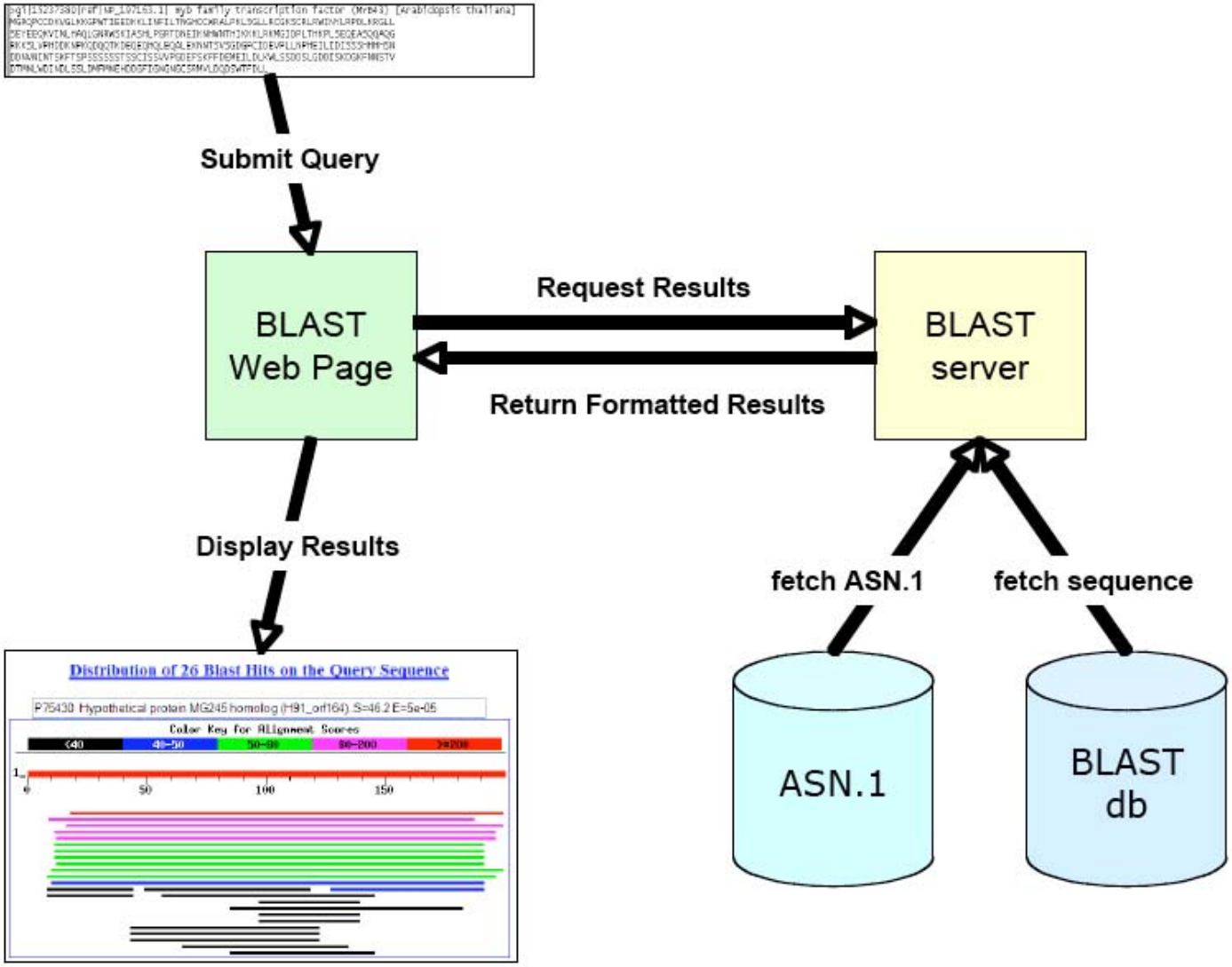
Concepts of Sequence Similarity Searching

- The premise:

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

The BLAST algorithm

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


Submit Query
BLAST
Web Page
Request Results
Return Formatted Results
BLAST
server
Display Results
fetch ASN.1
fetch sequence
Distribution of 26 Blast Hits on the Query Sequence
Color Key for Alignment Scores
ASN.1
BLAST
db

What BLAST tells you ...

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

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

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

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

BLAST programs

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

more BLAST programs

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

nucleotide only

protein only

BLAST Algorithm

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

Sequence Similarity Searching – The statistics are important

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

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

Where does the score (S) come from?

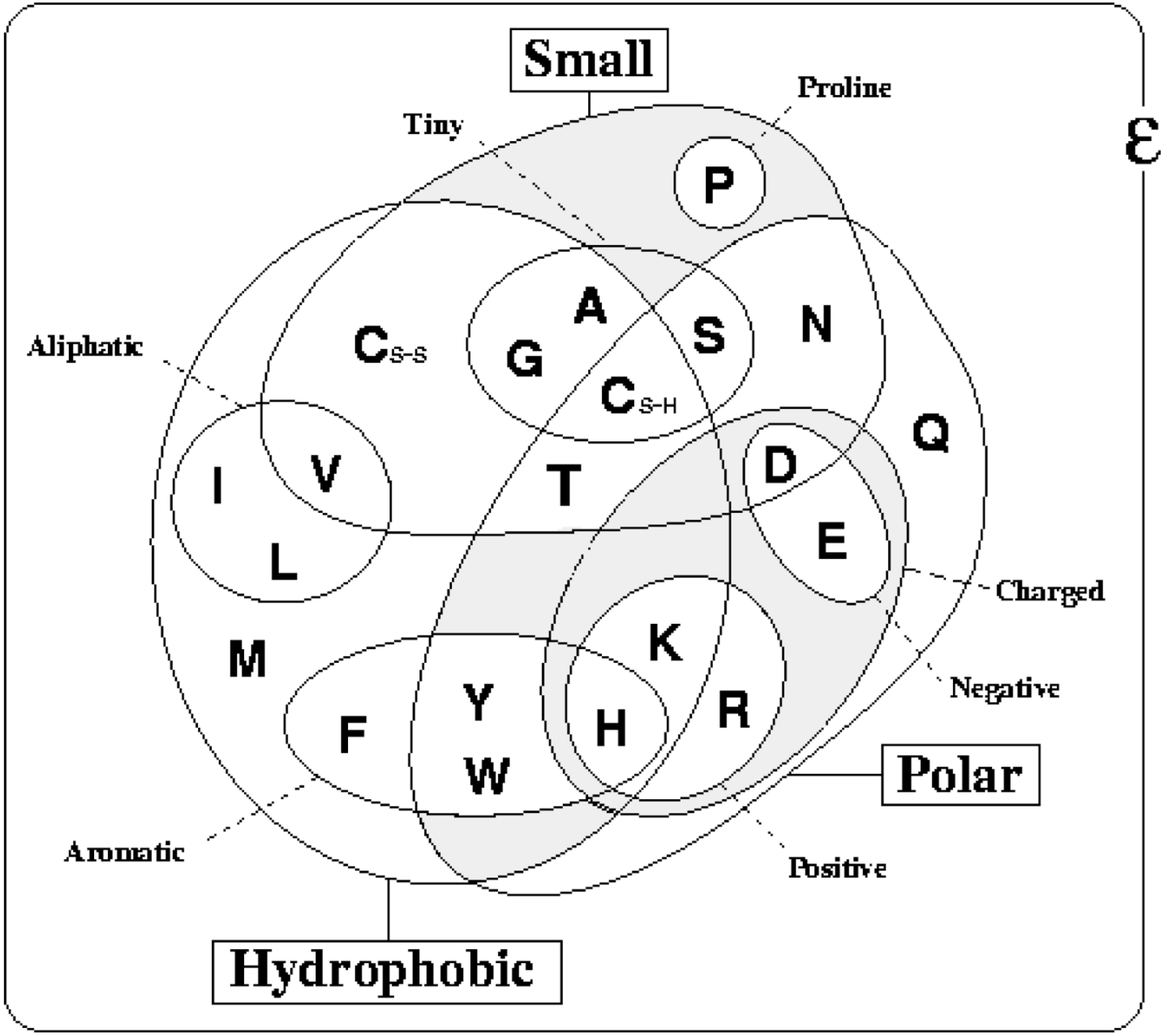
- The quality of each pair-wise alignment is represented as a score and the scores are ranked.
- Scoring matrices are used to calculate the score of the alignment base by base (DNA) or amino acid by amino acid (protein).
- The alignment score will be the sum of the scores for each position.

What's a scoring matrix?

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

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

BLOSUM 62


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

BLOSUM vs PAM

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

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

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

- The quality of the alignment is represented by the Score (S).
- The score of an alignment is calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (PAM, BLOSUM) whereas gap scores are assigned empirically .
- The significance of each alignment is computed as an E value (E).
- Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

Notes on E-values

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

Homology: Some Guidelines

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

Suggested BLAST Cutoffs

Take Home Message:
Always look at your alignments

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

BLAST Algorithm

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

How Does BLAST Really Work?

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

BLAST Algorithm

Query Word (W = 3)

TLSHAWRLSNETDKRPFIETAERL**RDQ**HKKDYPEYKYQPRRRKNGKPGSSSEADAHSE

Determine neighborhood

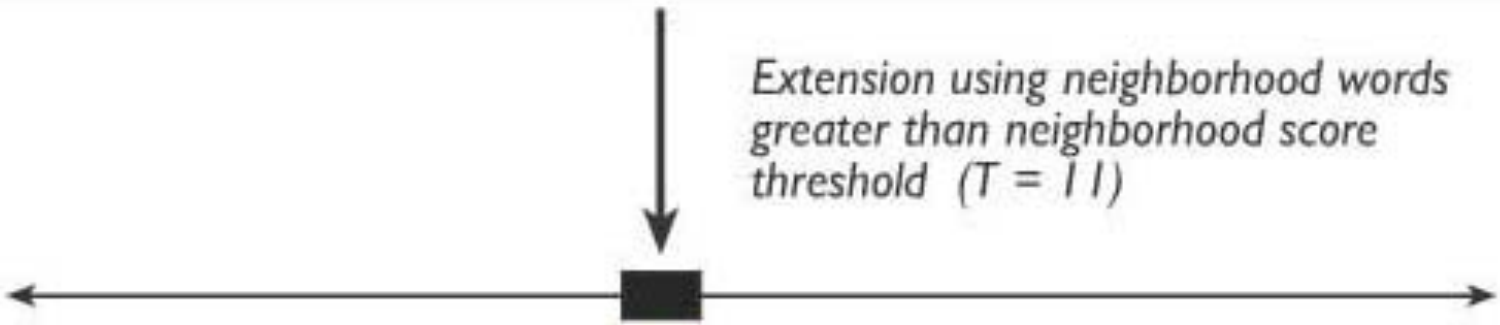
RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...

How Does BLAST Really Work?

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

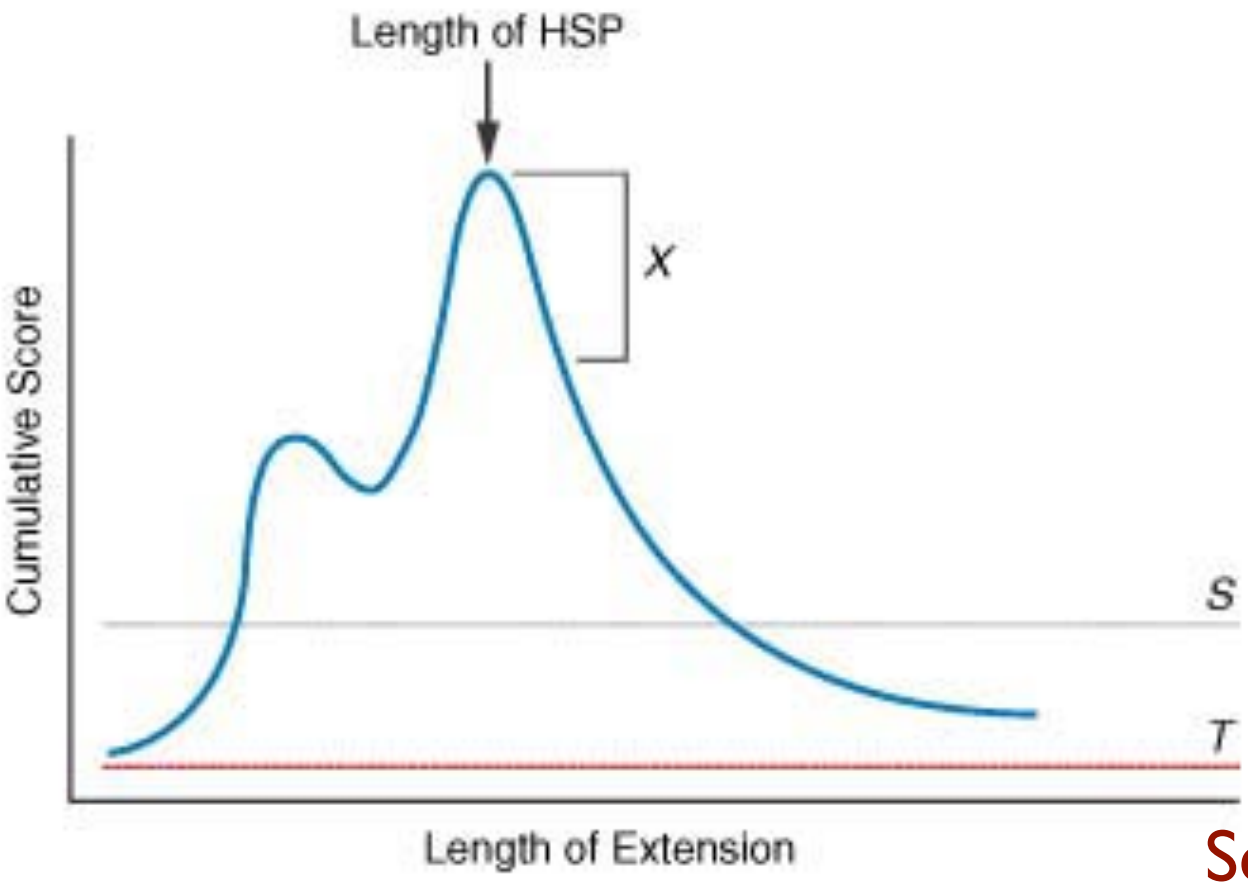
BLAST Algorithm

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




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

Extending the High Scoring Segment Pair (HSP)



Minimum Score (S)

Neighborhood Score Threshold (T)

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

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

Query  2    IVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  61
            +VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI
Sbjct  8    MVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKI  67

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


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

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

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


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

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

Query  3    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIY  62
            VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHK+HYKIY
Sbjct  9    VSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKDHYKIY  68

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

BLAST Algorithm

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

Credits

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

  NCBI HelpDesk - Field Guide Course Materials

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

- Questions? Please contact:

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

joanne@msl.ubc.ca

Laboratory Bioinformatics

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

bioteach.ubc.ca/bioinfo2009

Module 2 Topics

- **BLAST** - Finding Function by Sequence Similarity
- **GUIDED TOUR** - Advanced Tips & Tricks for Using BLAST
- **PRACTICAL EXERCISES** - The Plasmodium HSP86 Story
- **COMMON TASKS** - Basic Search; Searching Sets of Sequences (multiple inputs; small custom databases); Primer Design

BLAST

GUIDED TOUR: Advanced Tips & Tricks for Using BLAST

http://blast.ncbi.nlm.nih.gov/

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New Designing or Testing PCR Primers? Try your search in Primer-BLAST. Go

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

News

Align two sequences form.

The Align two sequences link on the BLAST home page now uses the standard BLAST submission form.
Tue, 03 Feb 2009 16:00:00 EST

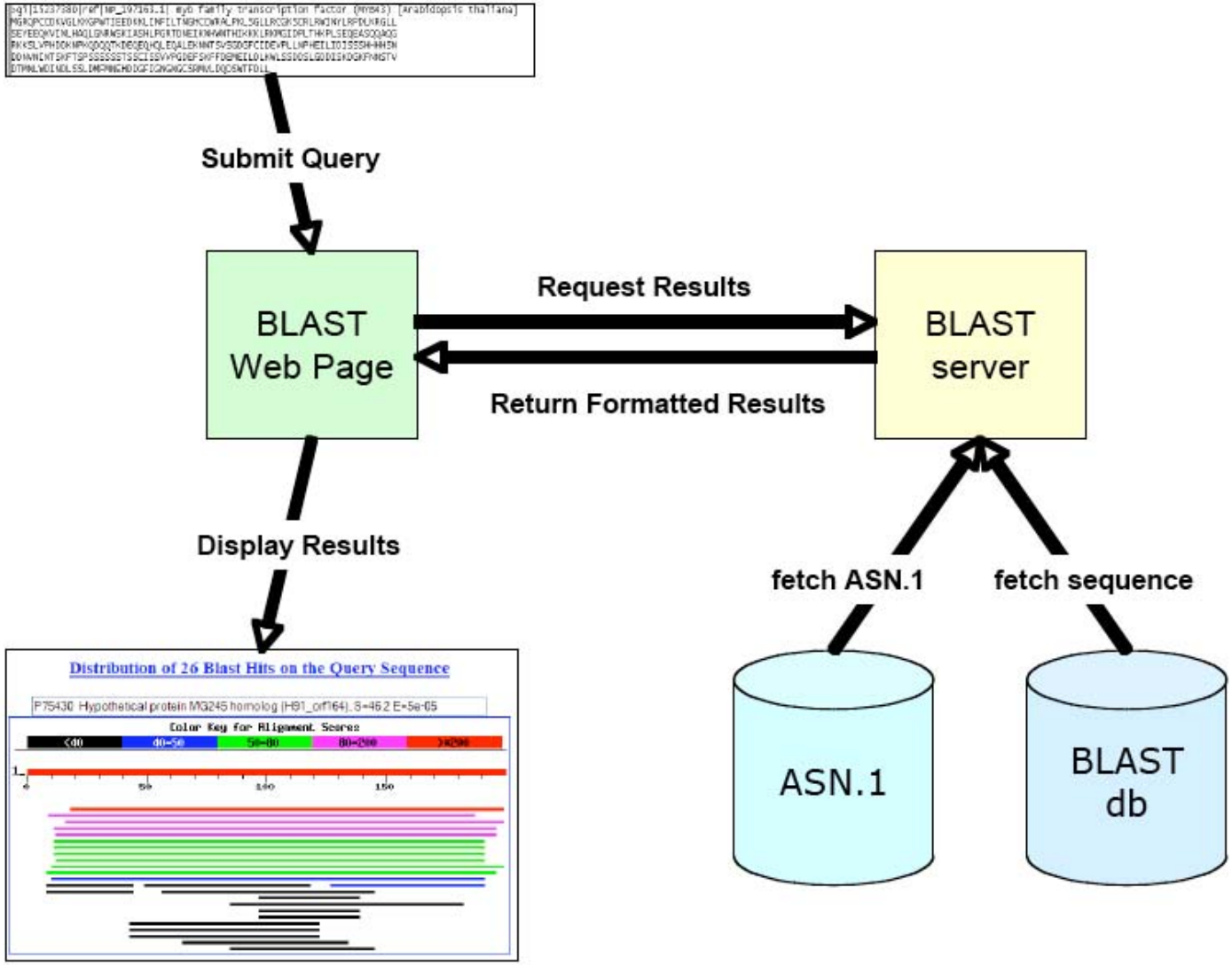
More BLAST news...

Tip of the Day

How to do Batch BLAST jobs.

BLAST makes it easy to examine a large group of potential gene candidates.

More tips...


Submit Query
BLAST
Web Page
Request Results
Return Formatted Results
BLAST
server
Display Results
fetch ASN.1
fetch sequence
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
ASN.1
BLAST
db

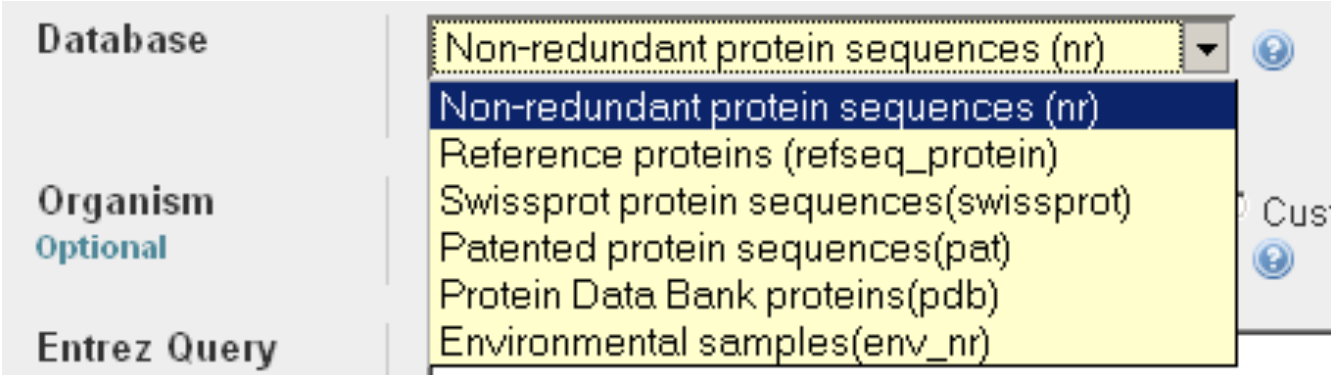
Consider your research question ...

- Are you looking for an particular gene in a particular species?
- Are you looking for additional members of a protein family across all species?
- Are you looking to annotate genes in your species of interest?

Know your reagents

- Changing your choice of database is changing your search space
- Database size affects the BLAST statistics
- Databases change rapidly and are updated frequently

Protein Databases: nr



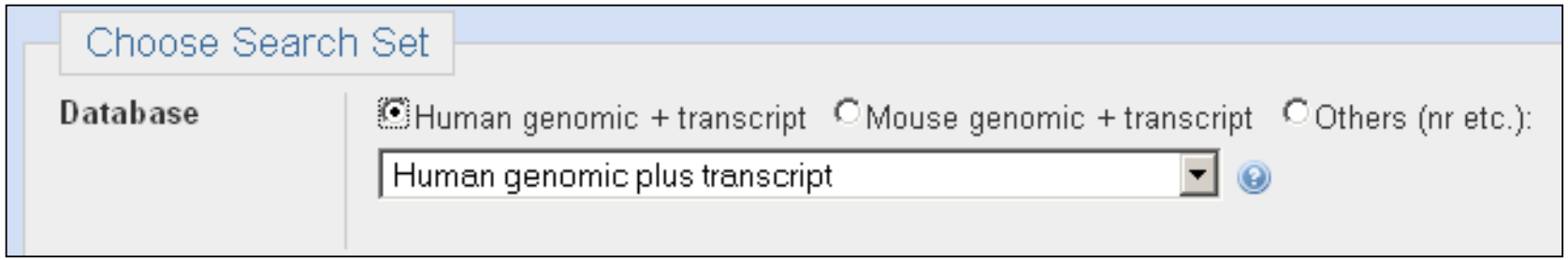
- nr (non-redundant protein sequences) default
  - GenBank CDS translations
  - NP_ RefSeqs
  - Outside Protein
    - PIR, Swiss-Prot, PRF
    - PDB (sequences from structures)
- pat protein patents
- env_nr environmental samples

Services

Blastp

blastx

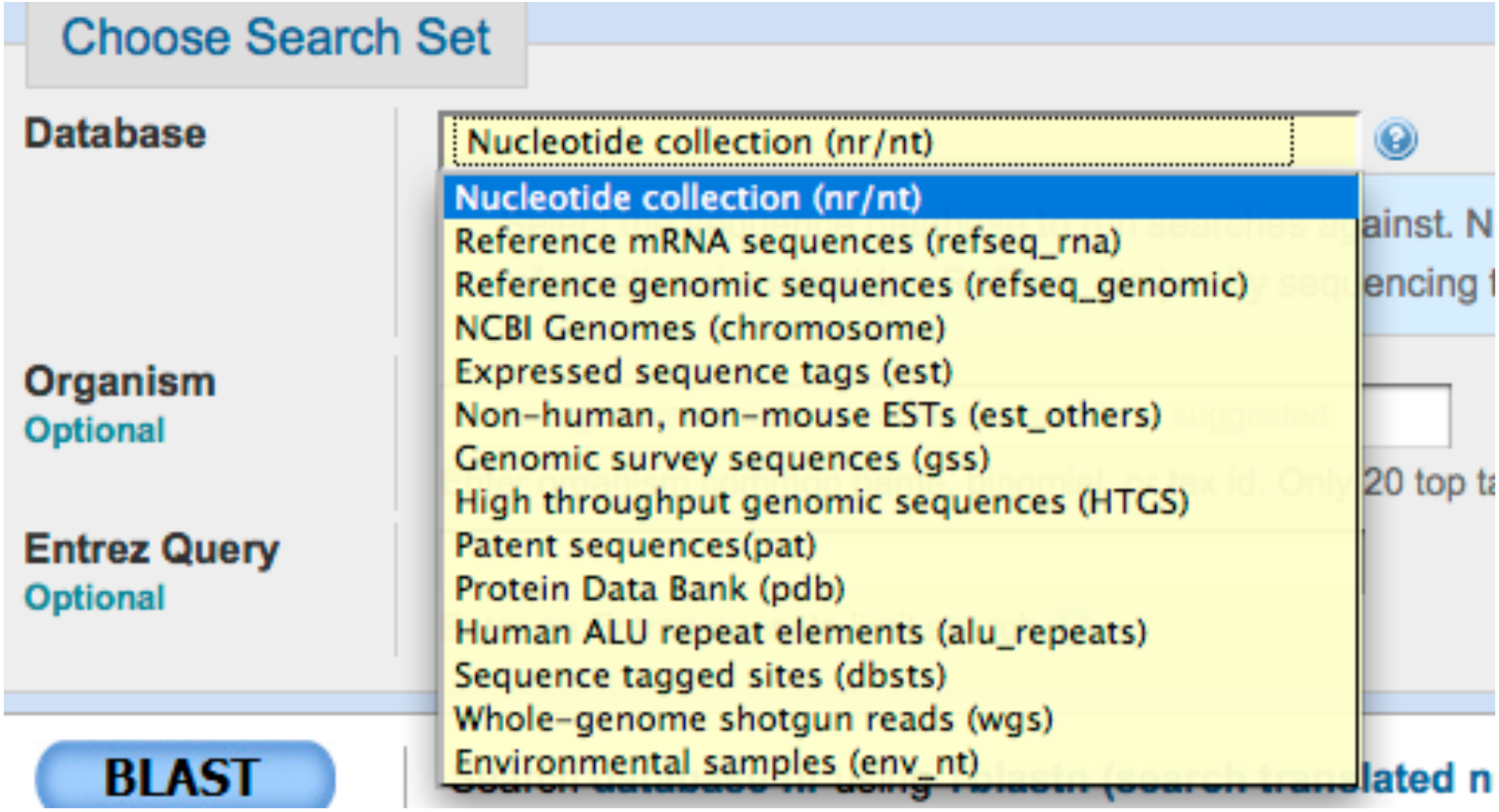
Nucleotide Databases: Human and Mouse



- Human and mouse genomic + transcript default
- Separate sections in output for mRNA and genomic
- Direct links to Map Viewer for genomic sequences

Megablast, blastn service

Nucleotide Databases: Traditional



Services
blastn
tblastn
tblastx

Nucleotide Databases:

- **nr (nt)** Traditional GenBank
  + RefSeq nucleotides
  + PDB sequences
- **refseq_rna**
- **refseq_genomic** NC_
- **NCBI genomes**
  complete genomes
  + chromosomes from RefSeq
- **est** expressed sequence tags
  - human + mouse, others
- **htgs** high throughput genomic
  - unfinished
- **gss** genome survey sequence
  - single-pass genomic data
- **pdb** protein data bank
  - derived from 3D structures
- **wgs**
  - whole genome shotgun
- **env_nt**
  - environmental samples

Databases are mostly non-overlapping

http://blast.ncbi.nlm.nih.gov/

BLAST

Home | Recent Results | Saved Strategies | Help

Alignment Search Tool

My NCBI

Welcome joannealisonfox. [Sign Out]

NCBI/ BLAST/ Help

Browse BLAST documentation.

Getting Started

- BLAST short course
- BLAST program selection guide

Program Selection Guide

Getting Help

- Email blast-help
- Mailing list

About BLAST

- Frequently Asked Questions
- NCBI Handbook: BLAST
- The Statistics of Sequence Similarity Scores
- NAR 2004 Web server issue
- NAR 2006 Web server issue
- BLAST glossary
- References

Other BLAST information

- Download BLAST Software and Databases
- Developer information

BLAST News

BLAST News directory

3. Program Selection Tables

The appropriate selection of a BLAST program for a given search is influenced by the following three factors **1)** the nature of the query, **2)** the purpose of the search, and **3)** the database intended as the target of the search and its availability. The following tables provide recommendations on how to make this selection.

Table 3.1 Program Selection for Nucleotide Queries

Length [1]	Database	Purpose	Program	Explanation
20 bp or longer 28 bp or above for megablast	Nucleotide	Identify the query sequence	discontiguous megablast, megablast, or blastn	Learn more ...
		Find sequences similar to query sequence	discontiguous megablast or blastn	Learn more ...
		Find similar sequence from the Trace archive	Trace megablast, or Trace discontiguous megablast	Learn more ...
		Find similar proteins to translated query in a translated database	Translated BLAST (tblastx)	Learn more ...
	Peptide	Find similar proteins to translated query in a protein database	Translated BLAST (blastx)	Learn more ...
7 - 20 bp	Nucleotide	Find primer binding sites or map short contiguous motifs	Search for short, nearly exact matches	Learn more ...

NOTE:
[1] The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in the Section 4 below. With default setting, the shortest unambiguous query one can use is 11 for blastn and 28 for MEGABLAST.

Table 3.2 Program Selection for Protein Queries				
Length [1]	Database	Purpose	Program	Explanation
15 residues or longer	Peptide	Identify the query sequence or find protein sequences similar to the query	Standard Protein BLAST (blastp)	Learn more ...
		Find members of a protein family or build a custom position-specific score matrix	PSI-BLAST	Learn more ...
		Find proteins similar to the query around a given pattern	PHI-BLAST	Learn more ...
		Find conserved domains in the query	CD-search (RPS-BLAST)	Learn more ...
		Find conserved domains in the query and identify other proteins with similar domain architectures	Conserved Domain Architecture Retrieval Tool (CDART)	Learn more ...
	Nucleotide	Find similar proteins in a translated nucleotide database	Translated BLAST (tblastn)	Learn more ...
5-15 residues	Peptide	Search for peptide motifs	Search for short, nearly exact matches	Learn more ...

Note:
[1] The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in Section 4 below.

As genomic and other specialized sequence information is made available to the public, NCBI creates specialized BLAST pages for those sequences. The table below provides a general guide on how to select and use those special BLAST databases.

Table 3.3 Search against Organism Specific or Genome Databases [1]

Query [2]	Database	Purpose	BLAST Pages to Use [3]	Explanation
Nucleotide: 20 or 28 bp and above Protein: 15 residues and above	Human Genome	Map the query sequence Determine the genomic structure Identify novel genes Find homologs Other data mining	Human	Learn more ...
	Mouse Genome		Mouse	Learn more ...
	Rat Genome		Rat	Learn more ...
	Chimp, Cow, Dog, or Chicken Genome		Chimp, or Cow, Dog, Chicken	Learn more ...
	Cat, Sheep, or Pig Genome		Cat, Sheep, or Pig	Learn more ...
	Zebrafish or Fugu (Pufferfish)		Zebrafish or Fugu rubripes	Learn more ...
	Insects (flies and honeybees)		Insects	Learn more ...
	Nematodes (worms)		Nematodes	Learn more ...
	Plants		Plants	Learn more ...
	Fungi Genomes (including yeasts)		Fungi	Learn more ...
	Protozoa		Protozoa	Learn more ...
	Environmental Samples		Environmental Samples	Learn more ...
	Other Lower Eukaryotic Genomes		Other eukaryotes genomes	Learn more ...
	Microbial Genomes		Microbial genomes	Learn more ...

NOTE:

[1] Those pages access the genome database consisting of contig assemblies and other sequences specific to the organisms. Not all organisms listed here have genome assemblies available.

[2] Sequence length is only a suggestion. For most of the pages, the search parameters can be modified to enable searches with a short query by pasting additional options in the "Advanced Options" text box. For protein comparisons, -F F -e 20000 -W 2 should be used. For nucleotide comparison, use -F F -e 1000 -W 7. This also requires the uncheck of the megablast checkbox.

[3] Available databases and their contents are described in Section 5.

BLAST pages for special purposes are listed under Special and Meta sections. Their functions are described in Table 3.4 below.

Table 3.4 Function of Special BLAST Pages under Special/Meta Sections				
Query [1]	Database	Purpose	BLAST Page to Use	Explanation
Nucleotide: 11 bp or above Protein: 15 or above	- [2]	Compare two sequences directly	Align two sequences	Learn more ...
	Immunoglobulin sequences	Find matches to curated immunoglobulin sequences	igBLAST	Learn more ...
Nucleotide: 20 or 28 bp and above	UniVec	Screen for vector contamination	VecScreen	Learn more ...
	GEO	Find matches to sequences with MicroArray information	GEO BLAST	Learn more ...
	SNP	Find matches to human reference SNPs	SNP BLAST	Learn more ...
-	- [3]	To retrieve results for a search with its RID	Retrieve result for an RID	Learn more ...

Note:
[1] The query sequence length is only a suggestion. For most of the pages, the search parameters can be modified to enable better handling of short query by pasting additional options in the "Advanced Options" text box. For protein comparisons, -F F -e 20000 -W 2 should be used. For nucleotide comparison, use -F F -e 2000 -W 7.
[2] "Align two sequences" treats the second sequence as the database.
[3] Requires valid RIDs that are assigned within the past 24 hours.

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI
Welcome joannealisonfox. [Sign Out]

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

News

Old BLAST Web Pages to be deleted June 11th 2007

As previously announced access to the old pages will be removed on June 11, 2007.

2007-06-01 12:15:00

More BLAST news...

Tip of the Day

How to use BLAST to find human sequences in a database that can be amplified with a particular primer pair.

A frequent use of nucleotide-nucleotide BLAST is to check the specificity of oligonucleotides for hybridization in PCR. The goal is usually to make sure that the primers will give a unique product from the target genome or cDNA

NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query. more... Reset page Bookmark

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

231571

231571

Query subrange

From

To

Or, upload file Browse...

Job Title Q02067:Achaete-scute homolog 1 (Mash-1)

Enter a descriptive title for your BLAST search

Choose Search Set

Database Swissprot protein sequences(swissprot)

Organism Optional Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

Let's look at some of the options!

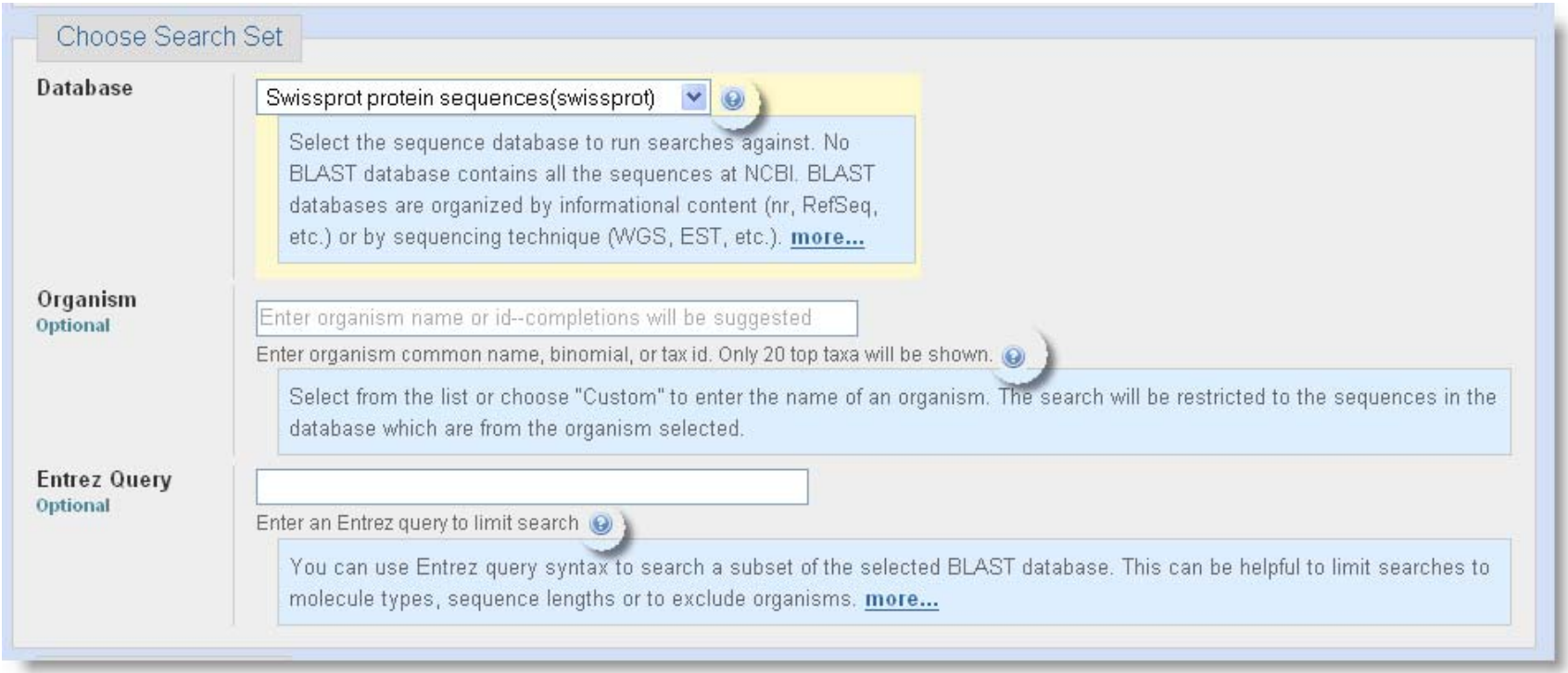
BLAST Search database swissprot using Blastp (protein-protein BLAST)

Show results in a new window

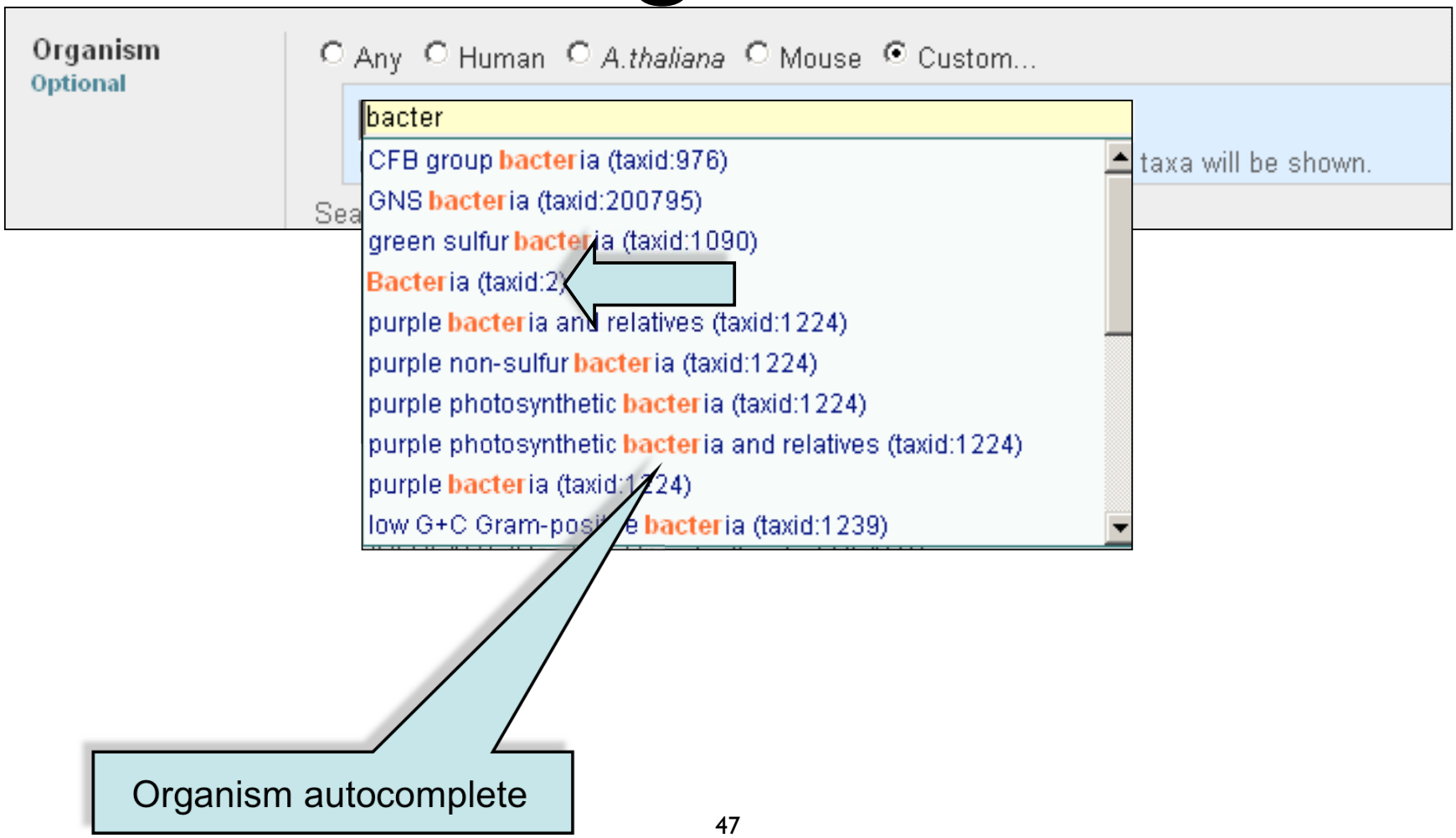
Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

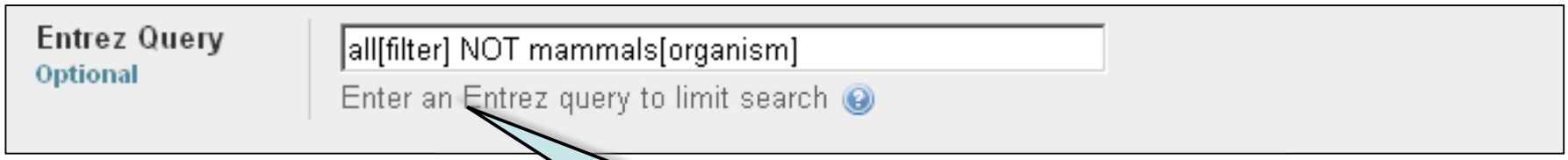
Context Specific Help



Limiting Database: Organism



Limiting Database: Entrez Query



all[filter] NOT mammals[organism]

gene_in_mitochondrion[Properties]
2006:2007 [Modification Date]

Nucleotide
biomol_mrna[Properties]
biomol_genomic[Properties]

BLAST

Search database swissprot using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the defau

General Parameters

Max target sequences: 100

Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

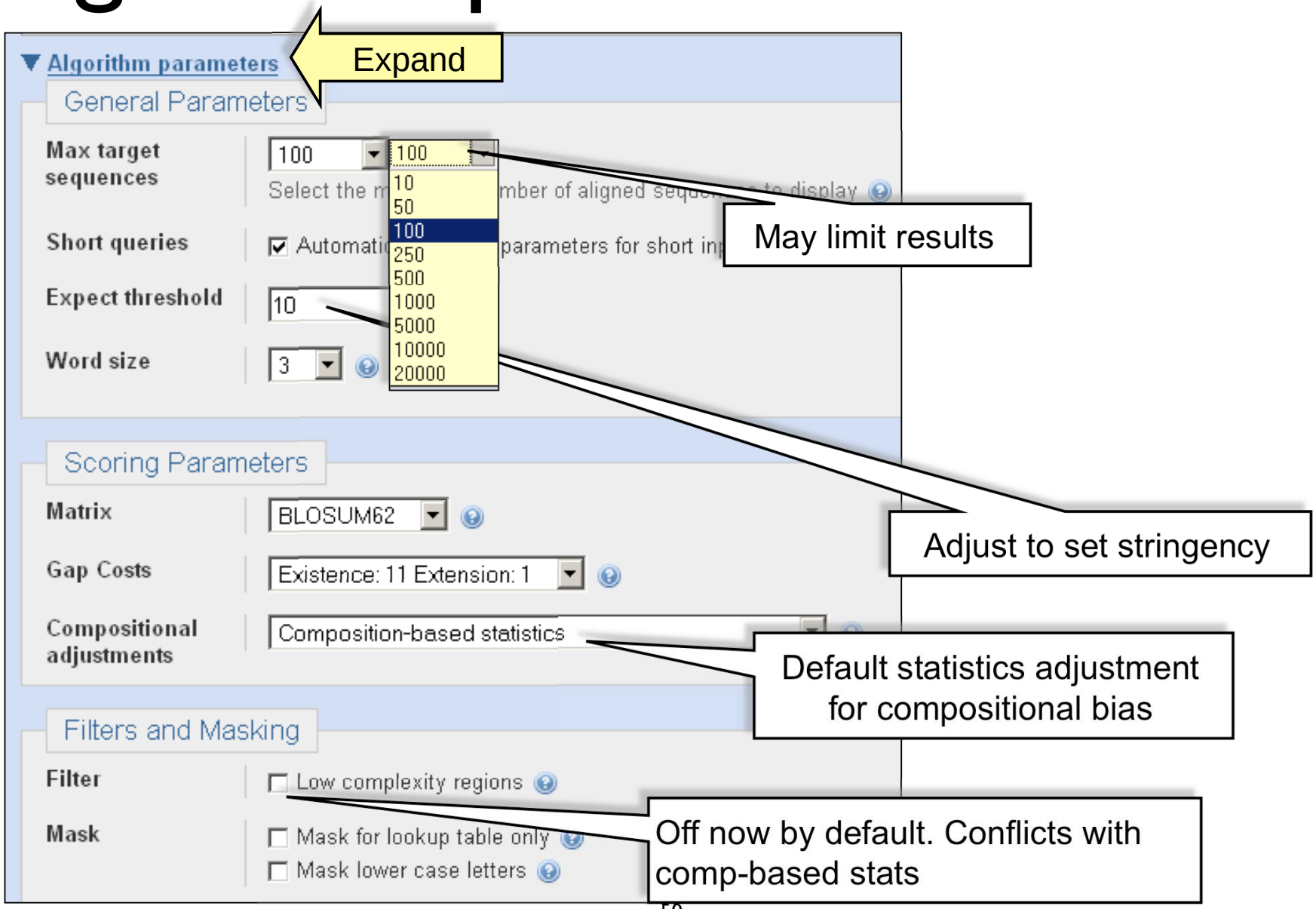
Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Algorithm parameters: Protein



Automatic Short Sequence Adjustment

Job Title: Elvis Lives!

No putative conserved domains have been detected

Your search parameters were adjusted to search for a short input sequence.

WAITING

Request ID	1WSB0FX012
Status	Searching

e-value	20000
Word Size	2
Matrix	PAM30
Comp Stats	Off
Low Comp Filter	Off

```
>ref|ZP_01712014.1|  conserved hypothetical protein [Pseudomonas putida GB-1
Length=245

 Score = 18.5 bits (36),  Expect = 15305
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query  1    ELVIS  5
            ELVIS
Sbjct  126  ELVIS  130


>ref|ZP_01712512.1|  Substrate-binding region of ABC-type glycine betaine tra
system [Pseudomonas putida GB-1]
Length=342

 Score = 18.5 bits (36),  Expect = 15305
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query  1    ELVIS  5
            ELVIS
Sbjct  172  ELVIS  176


>ref|XP_001366374.1|  PREDICTED: similar to R7 binding protein [Monodelphi
Length=257

 Score = 18.5 bits (36),  Expect = 15305
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query  1   ELVIS  5
           ELVIS
Sbjct  69  ELVIS  73


>ref|ZP_01711731.1|  GCN5-related N-acetyltransferase [Caldivirga maquilinge
Length=166

 Score = 18.5 bits (36),  Expect = 15305
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query  1   ELVIS  5
           ELVIS
Sbjct  20  ELVIS  24
```

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

```
>gi|231571|sp|Q02067|ASCL1_MOUSE Achaete-scute homolog 1
(Mash-1)
MESSGKMESGAGQQPQPPQPFLPPAACFFATAAAAAAAAAAAAQSAQQQQPQAPPQQAPQLS
GGGHKSAAKQDKRQRSSSPELMRCKRRLNFSGFGYSLPQQQPAAVARRNERERNRVKLVNLG
PNGAANKKMSKVETLRSAVQYIRALQQLLDEHDAVSAAFQAGVLSPTISPNYSNDLNSMAGS
```

Query subrange

From

To

Or, upload file Browse...

Job Title MASH1 BLAST for CBW

Enter a descriptive title for your BLAST search

Choose Search Set

Database Swissprot protein sequences(swissprot)

Organism Optional

Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

- (●) blastp (protein-protein BLAST)
- ( ) PSI-BLAST (Position-Specific Iterated BLAST)
- ( ) PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST

Search database **swissprot** using **Blastp (protein-protein BLAST)**

☐ **Show results in a new window**

BLAST

Basic Local Alignment Search Tool

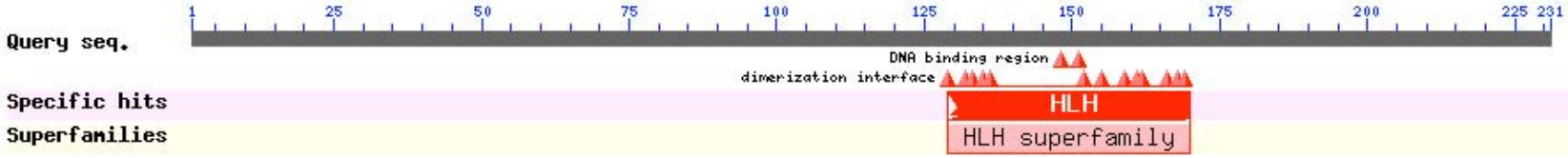
Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - T9U0ZFN4011 [Formatting options]

Job Title: Q02067:RecName: Full=Achaete-scute homolog...

Putative conserved domains have been detected, click on the image below for detailed results.



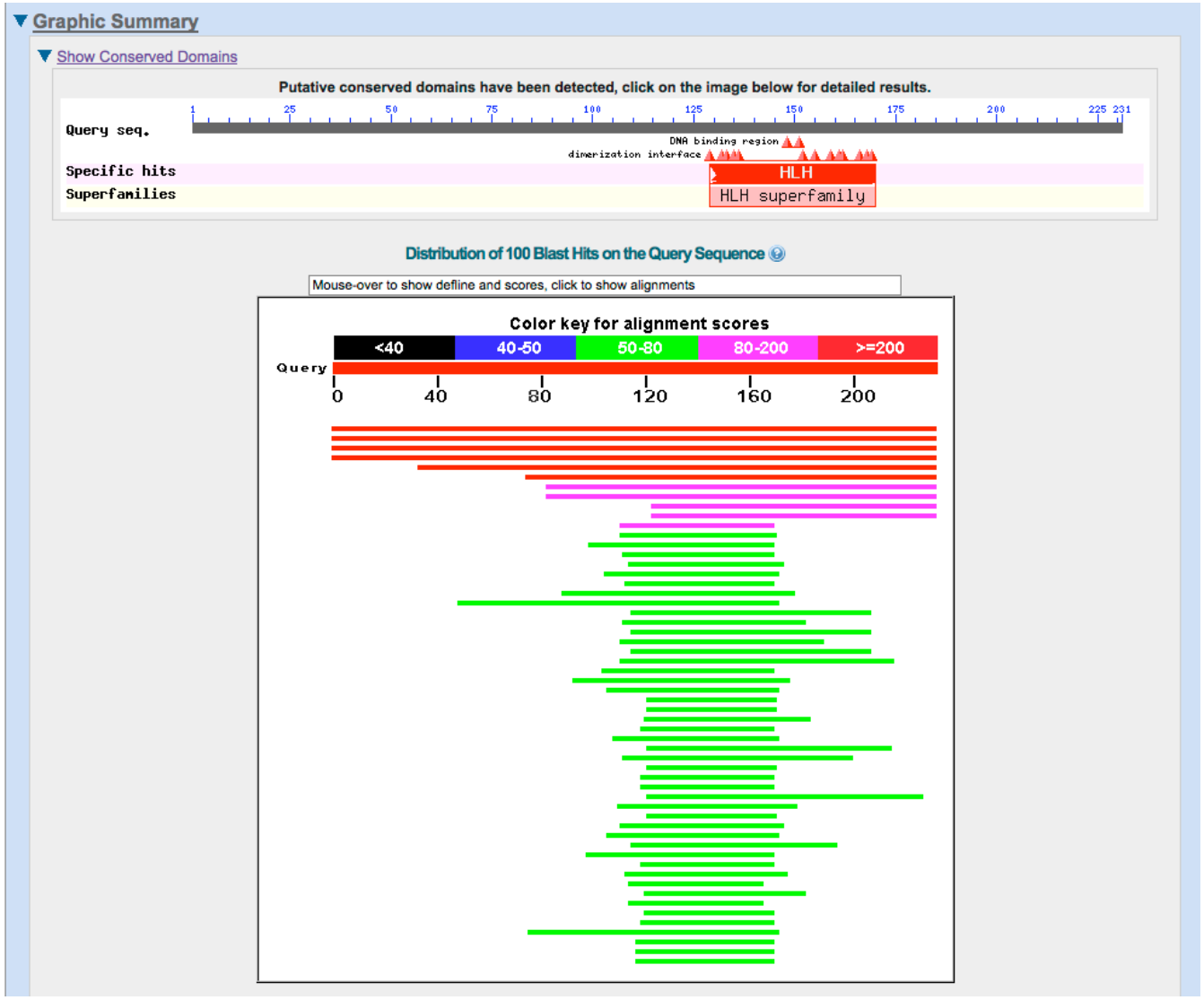
Request ID	**T9U0ZFN4011**
Status	Searching
Submitted at	Thu Feb 12 22:25:19 2009
Current time	Thu Feb 12 22:25:26 2009
Time since submission	00:00:06

This page will be automatically updated in **78** seconds

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback

NCBI | NLM | NIH | DHHS

A graphical view



The BLAST hit list

BLAST Home

My NCBI [Sign In] [Register]

NCBI/ BLAST

Edit and

Q02067:

Des

Full=Mash-1

Program BLASTP 2.2.19+ ▸Citation

Molecule type amino acid
Query Length 231

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results]

▸ Graphic Summary

▾ Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value	
sp|Q02067.1|ASCL1_MOUSE RecName: Full=Achaete-scute homolog 1...	466	4e-131	G
sp|P19359.1|ASCL1_RAT RecName: Full=Achaete-scute homolog 1	347	4e-95	G
sp|P50553.2|ASCL1_HUMAN RecName: Full=Achaete-scute homolog 1...	332	1e-90	G
sp|Q90259.1|ASL1A_DANRE RecName: Full=Achaete-scute homolog 1...	298	1e-80	G
sp|Q06234.1|ASCL1_XENLA RecName: Full=Achaete-scute homolog 1	289	9e-78	G
sp|Q90260.1|ASL1B_DANRE RecName: Full=Achaete-scute homolog 1...	217	3e-56	G
sp|Q2EGB9.1|ASCL2_BOVIN RecName: Full=Achaete-scute homolog 2...	135	1e-31	G
sp|Q99929.2|ASCL2_HUMAN RecName: Full=Achaete-scute homolog 2...	124	3e-28	G
sp|P19360.1|ASCL2_RAT RecName: Full=Achaete-scute homolog 2; ...	106	8e-23	G
sp|O35885.2|ASCL2_MOUSE RecName: Full=Achaete-scute homolog 2...	103	1e-21	G
sp|Q7RTU5.2|ASCL5_HUMAN RecName: Full=Achaete-scute homolog 5	80.5	6e-15	G
sp|Q6XD76.1|ASCL4_HUMAN RecName: Full=Achaete-scute homolog 4...	78.2	4e-14	G
sp|Q9NQ33.2|ASCL3_HUMAN RecName: Full=Achaete-scute homolog 3...	75.9	2e-13	G
sp|Q9JJR7.1|ASCL3_MOUSE RecName: Full=Achaete-scute homolog 3...	75.1	3e-13	G
sp|P10083.1|AST5_DROME RecName: Full=Achaete-scute complex pr...	74.7	3e-13	
sp|P10084.2|AST4_DROME RecName: Full=Achaete-scute complex pr...	71.6	3e-12	
sp|Q10007.1|HLH6_CAEEL RecName: Full=Helix-loop-helix protein 6	64.3	5e-10	G

BLAST Alignments

```
>sp|P20389|MYC2_MARMO  N-myc 2 proto-oncogene protein
Length=454

 Score = 35.8 bits (81),  Expect = 0.14, Method: Composition-based stats.
 Identities = 22/52 (42%), Positives = 30/52 (57%), Gaps = 4/52 (7%)

Query  133  FATLREHVPNGAANKKMSKVETLRSAVQYIRALQ----QLLDEHDAVSAAFQ  180
            F TLR+HVP    N+K +KV  L+ A +Y+  LQ    QLL E + + A  Q
Sbjct  391  FTTLRDHVPELVKNEKAAKVVILKKACEYVHYLQAKEHQLLMEKEKLQARQQ  442
```

Identical match

positive score (conservative)

Negative or zero

gap

BLAST Alignments

```
>sp|P04198|MYCN_HUMAN N-myc proto-oncogene protein
Length=464

 Score = 35.4 bits (80),  Expect = 0.025, Method: Composition-based stats.
 Identities = 22/52 (42%), Positives = 31/52 (59%), Gaps = 4/52 (7%)

Query  133  FATLREHVPNGAANKKMSKVETLRSAVQYIRALQ----QLLDEHDAVSAAFQ  180
            F TLR+HVP    N+K +KV  L+ A +Y+ +LQ    QLL E + + A  Q
Sbjct  401  FLTLRDHVPELVKNEKAAKVVILKKATEYVHSLQAEEHQLLLEKEKLQARQQ  452


>sp|Q02363|ID2_HUMAN DNA-binding protein inhibitor ID-2 (Inhibitor of DNA binding
2)
Length=134

 Score = 35.4 bits (80),  Expect = 0.025, Method: Composition-based stats.
 Identities = 19/47 (40%), Positives = 29/47 (61%), Gaps = 0/47 (0%)

Query  129  VNLGFATLREHVPNGAANKKMSKVETLRSAVQYIRALQQLLDEHDAV  175
            +N  ++ L+E VP+   NKK+SK+E L+  + YI  LQ  LD H  +
Sbjct  39   MNDCYSKLKELVPSIPQNKKVSKMEILQHVIDYILDLQIALDSHPTI  85


>sp|P12980|LYL1_HUMAN Protein lyl-1 (Lymphoblastic leukemia-derived sequence 1)
Length=267

 Score = 35.4 bits (80),  Expect = 0.025, Method: Composition-based stats.
 Identities = 22/50 (44%), Positives = 31/50 (62%), Gaps = 0/50 (0%)

Query  129  VNLGFATLREHVPNGAANKKMSKVETLRSAVQYIRALQQLLDEHDAVSAA  178
            VN  FA LR+ +P    ++K+SK E LR A++YI  L +LL +  A  AA
Sbjct  153  VNGAFAELRKLLPTHPPDRKLSKNEVLRLAMKYIGFLVRLLRDQAAALAA  202
```

- ## Similarity

The extent to which nucleotide or protein sequences are related. The extent of similarity between two sequences can be based on percent sequence identity and/or conservation. In BLAST similarity refers to a positive matrix score.

- ## Identity

The extent to which two (nucleotide or amino acid) sequences are invariant.

- ## Homology

Similarity attributed to descent from a common ancestor.

It is your responsibility as an informed bioinformatician to use these terms correctly: A sequence is either homologous or not. Don't use % with this term!

Re-Format and/or Download your BLAST results

Edit and Resubmit | Save Search Strategies | Formatting options | Download

Formatting options — Reformat

Show Alignment as HTML — Advanced View — Use old BLAST report format — Reset form to defaults

Alignment View Pairwise

Display Graphical Overview — Linkout — Sequence Retrieval — NCBI-gi

Masking Character: Lower Case — Masking Color: Grey

Limit results Descriptions: 100 — Graphical overview: 100 — Alignments: 100

Organism — Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.

Entrez query:

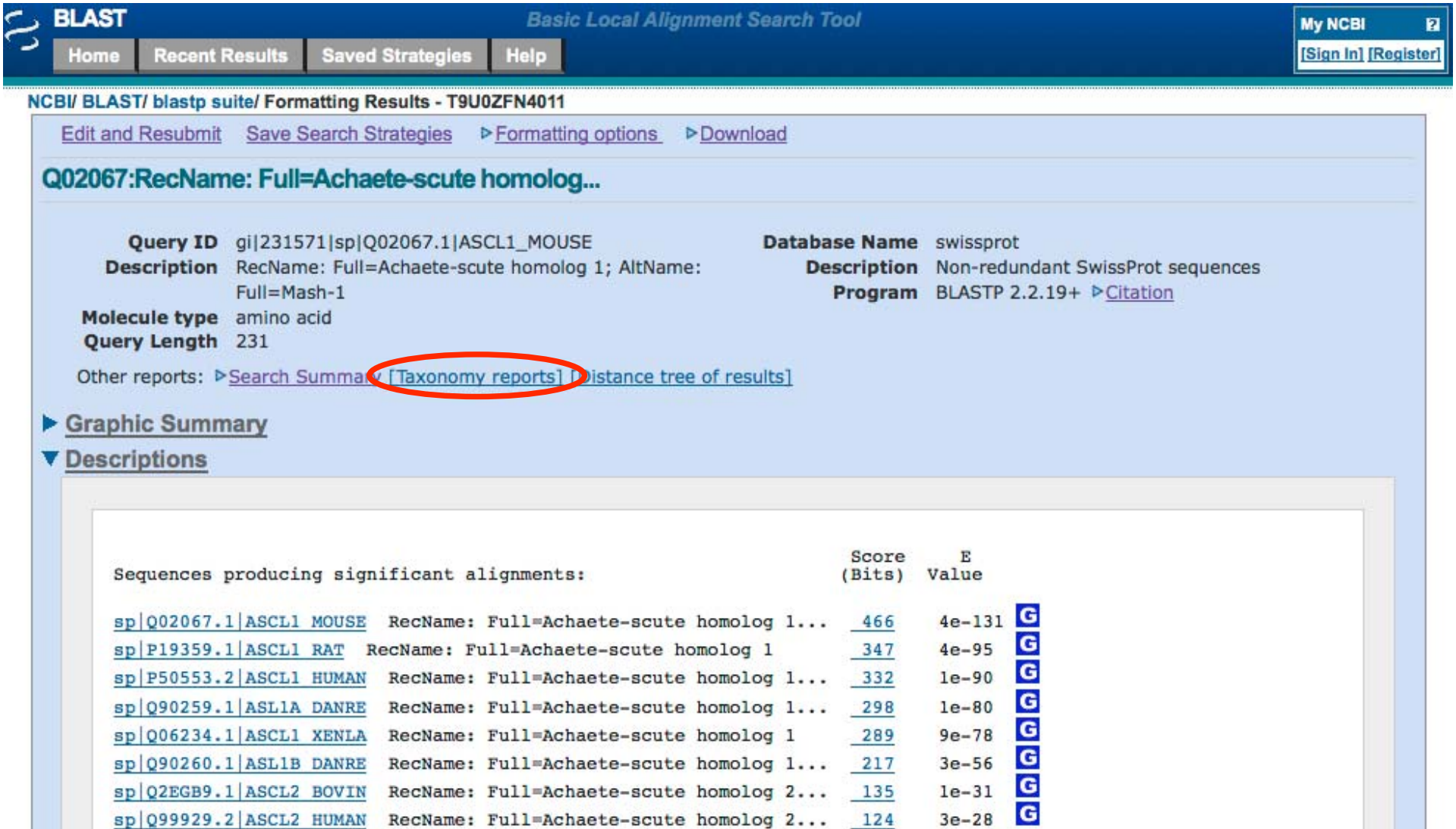
Expect Min: — Expect Max:

Format for PSI-BLAST — with inclusion threshold:

Download

Alignment	Search Strategies	Bioseq
Text, XML, ASN.1, Hit Table(text), Hit Table(csv)	ASN.1	ASN.1

Sorting BLAST by Taxonomy



BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp/ Formatting Results - VVH6PBD3011 [Reformat these Results] [Edit and Resubmit] [Sign in above to save your search strategy]

Job Title: gi|231571 (231 letters)

Show Conserved Domains

Tax BLAST Report

Index

- Lineage Report
- Organism Report
- Taxonomy Report
- Help

Lineage Report

```
Bilateria          [animals]
. Coelomata          [animals]
. . Euteleostomi       [vertebrates]
. . . Tetrapoda          [vertebrates]
. . . . Amniota            [vertebrates]
. . . . . Eutheria           [placentals]
. . . . . . Euarchontoglires [placentals]
. . . . . . . Glires           [placentals]
. . . . . . . . Muroidea         [rodents]
. . . . . . . . Murinae            [rodents]
. . . . . . . . . Mus musculus (mouse) ----------------  466 22 hits [rodents]              Achaete-scute homolog 1 (Mash-1)
. . . . . . . . . Rattus norvegicus (brown rat) .......  347 10 hits [rodents]              Achaete-scute homolog 1
. . . . . . . . . Mesocricetus auratus (Syrian hamster) -   50  2 hits [rodents]              Neurogenic differentiation factor 1 (NeuroD1)
. . . . . . . . . Oryctolagus cuniculus (domestic rabbit) -  49  1 hit  [rabbits & hares]      Heart- and neural crest derivatives-expressed
. . . . . . . . . Homo sapiens (man) ------------------------  332 25 hits [primates]          Achaete-scute homolog 1 (HASH1)
. . . . . . . . . Macaca fascicularis (cynomolgus monkey) ...   48  1 hit  [primates]          Neurogenic differentiation factor 6 (NeuroD6)
. . . . . . Bos taurus (cow) ----------------------------  135  4 hits [even-toed ungulates]  Achaete-scute homolog 2 (Mash2)
. . . . . . Ovis aries (domestic sheep) ................   50  1 hit  [even-toed ungulates]  Heart- and neural crest derivatives-expressed
. . . . . Gallus gallus (bantam) ------------------------   60  8 hits [birds]               Heart- and neural crest derivatives-expressed
. . . . . Coturnix japonica ............................   50  1 hit  [birds]               Myogenic factor 5 (Myf-5) (Myogenic factor 3)
. . . . Xenopus laevis (common platanna) ----------------  289 10 hits [frogs & toads]        Achaete-scute homolog 1
. . . . Notophthalmus viridescens (red-spotted newt) ....   49  1 hit  [salamanders]           Myogenic factor 5 (Myf-5)
. . . Danio rerio (leopard danio) ------------------------  298  8 hits [bony fishes]          Achaete-scute homolog 1a (Zash-1a) (Pituitary
. . Drosophila melanogaster ------------------------------   74  5 hits [flies]               Achaete-scute complex protein T5 (Achaete)
. Caenorhabditis elegans (nematode) ----------------------   64  4 hits [nematodes]           Helix-loop-helix protein 6
```

Organism Report

```
 Mus musculus (mouse) [rodents] taxid 10090
sp|Q02067|ASCL1_MOUSE Achaete-scute homolog 1 (Mash-1)          466  4e-131
sp|O35885|ASCL2_MOUSE Achaete-scute homolog 2 (Mash-2)          103  9e-22
sp|Q9JJR7|ASCL3_MOUSE Achaete-scute homolog 3 (bHLH transc...    75  2e-13
sp|Q61039|HAND2_MOUSE Heart- and neural crest derivatives-...    60  7e-09
sp|P27792|LYL1_MOUSE Protein lyl-1 (Lymphoblastic leukemia...    53  8e-07
```

Distance Tree of Results

Tree view for rid: T9U0ZFN4011, query ID: sp|Q02067.1, database: swissprot

This tree was produced using BLAST pairwise alignments. more...

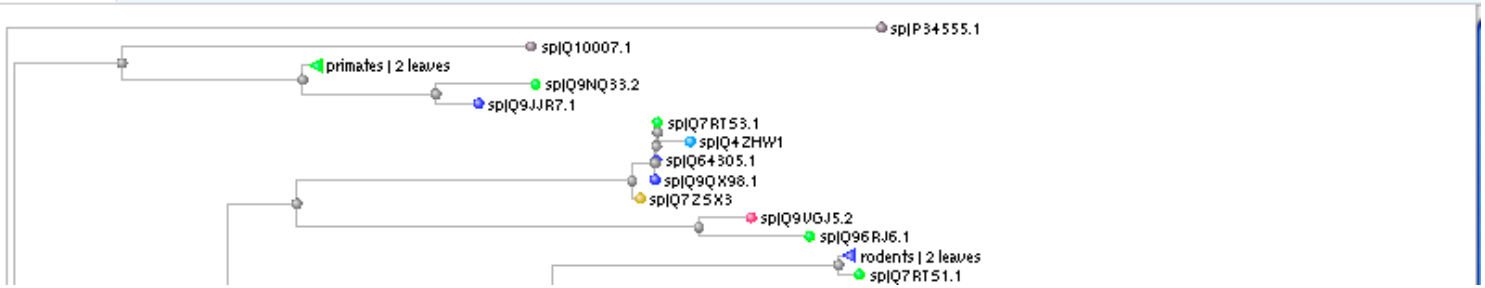
BLAST computes a pairwise alignment between a query and the database sequences searched. It does not explicitly compute an alignment between the different database sequences (i.e., does not perform a multiple alignment). For purposes of this sequence tree presentation an implicit alignment between the database sequences is constructed, based upon the alignment of those (database) sequences to the query. It may often occur that two database sequences align to different parts of the query, so that they barely overlap each other or do not overlap at all. In that case it is not possible to calculate a distance between these two sequences and only the higher scoring sequence is included in the tree.

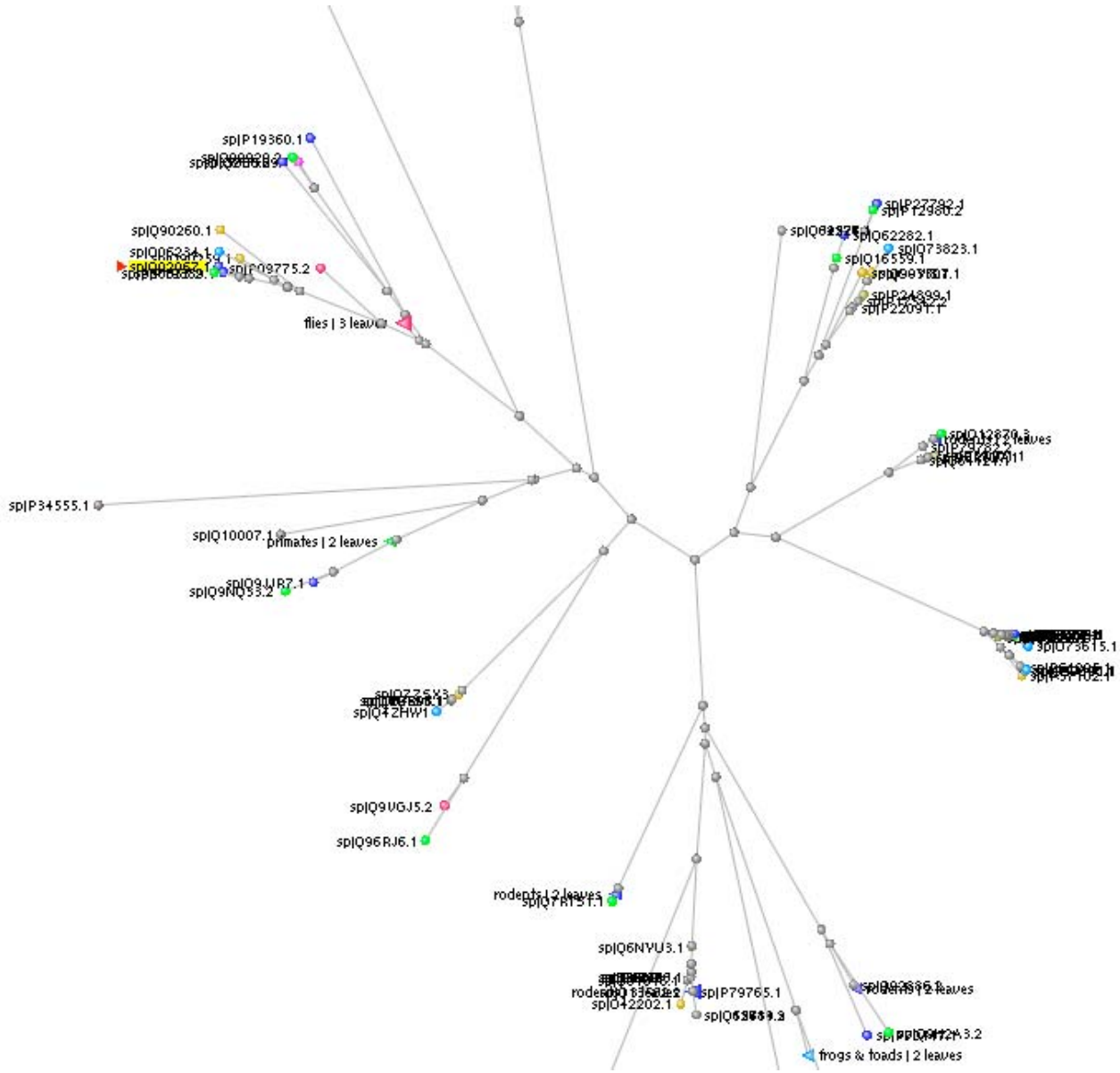
Tree method Fast Minimum Evolution | **Max Seq Difference** 0.85 | **Distance** Grishin (protein) | Reset | Download

Tree Method:
Algorithm used to produce a tree from given distances (or dissimilarities) between sequences. Available options:
1) Fast Minimum Evolution (*Desper R and Gascuel O, Mol Biol Evol 21:587-98, 2004*)
2) Neighbor Joining (*Saitou N and Nei M, Mol Biol Evol, 4:406-25, 2004*)
Note: Both algorithms produce un-rooted tree such as ones shown as *radial* or *force* in the tabs below. The rooted trees are created by placing a root in the middle of the longest edge.

read more in context specific help menus

rectangle | slanted | radial | force | Show distance | Mouse over an internal node for a subtree or alignment





- ✓ Rectangle: rectangular shaped rooted tree, where root is places in the longest edge
- ✓ Slanted: similar to rectangle, but with triangular tree shape
- ✓ Radial: un-rooted tree
- ✓ Force: similar to radial, where nodes are pushed away from one another for better presentation.

Nucleotide BLAST

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI

Welcome joannealisonfox. [Sign Out]

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

News

New Human and Mouse pre-indexed databases

Human and mouse genomic + transcript megablast searches now use a faster, indexed algorithm that typically reduces run time by two thirds, as compared with standard megablast.

2007-09-04 10:55:00

More BLAST news...

Tip of the Day

Using Genomic BLAST

Genomic BLAST pages are helpful because they allow the genomic context of a BLAST search to be displayed in the Map Viewer. For example, discontiguous (cross-species) MegaBLAST against the human RefSeq transcript for albumin (NM_000477) can be used to identify the homolog in the rat genome.

More tips...

nt BLAST: New Output

NCBI/ BLAST/ blastn suite: BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page Bookmark

Enter Query Sequence

Enter accession number, gi, or FAST AB168636 Clear

```
>Crab eating macaque CDC20 mRNA
AGCGGAGAGTTTAAGAGGCGTAAGCGAGGCGTGTTAAACCCGGTCGGAACTGCAACTTGCTC
ACGGGCTCCGCAGGCACCAACTGCAAGGACCCCTCCCGCTGCGGGCGTTCCCATGGCACAAT
GAGAGTGACCTGCACTCGCTGCTTCAGCTGGATGCACCCATCCCCAATGCACCCCCTGCGCG
GCAAAGCCAAGGAAGCCTCAGGCCCGGCCCCCTCACCCATGCGGGCCGCCAACCGATCCCAC
```

Query subrange

From

To

Or, upload file Browse...

Job Title Crab eating macaque CDC20 mRNA

Enter a descriptive title for your BLAST search

Choose Search Set

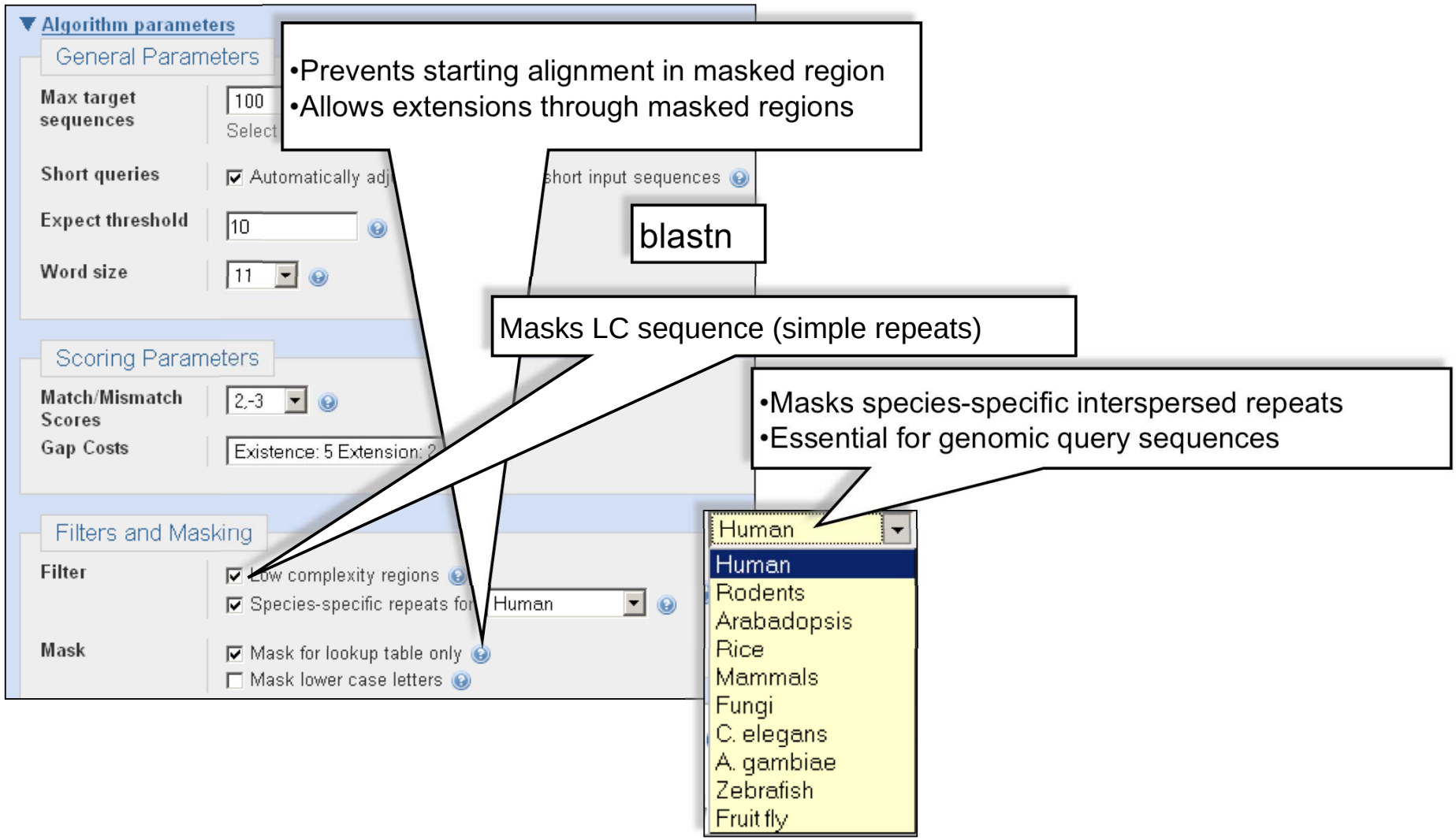
Database ◉ Human genomic + transcript ○ Mouse genomic + transcript ○ Others (nr etc.):

Human genomic plus transcript

Entrez Query
Optional

Enter an Entrez query to limit search

Algorithm parameters: Nucleotide



Sortable Results

Separate Sections for Transcript and Genome

Legend for links to other resources: U UniGene E GEO G ene S Structure

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Tot score	Query coverage	E value	Max ident	Links
Transcripts							
NM_001255.1	Homo sapiens CDC20 cell division cycle 20 ho	2876	2876	95%	0.0	97%	U E G M
Genomic sequences [show first]							
NT_023935.17	Homo sapiens chromosome 9 genomic contig	2629	2629	94%	0.0	95%	
NW_924484.1	Homo sapiens chromosome 9 genomic contig	2601	2601	94%	0.0	95%	
NT_032977.8	Homo sapiens chromosome 1 genomic contig	428	3002	95%	9e-117	100%	
NW_921351.1	Homo sapiens chromosome 1 genomic contig	428	3010	95%	9e-117	100%	

Pseudogene on Chromosome 9

Functional Gene on Chromosome 1

Total Score: All Segments

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Tot score	Query coverage	E value	Max ident	Links
Transcripts							
NM_001255.1	Homo sapiens CDC20 cell division cycle 20 hc	2876	2876	95%	0.0	97%	U E G M
Genomic sequences [show first]							
NW_921351.1	Homo sapiens chromosome 1 genomic contig	428	3010	95%	9e-117	100%	
NT_032977.8	Homo sapiens chromosome 1 genomic contig	428	3002	95%	9e-117	100%	
NT_023935.17	Homo sapiens chromosome 9 genomic contig	2629	2629	94%	0.0	95%	
NW_924484.1	Homo sapiens chromosome 9 genomic contig	2601	2601	94%	0.0	95%	

Functional Gene
Now First

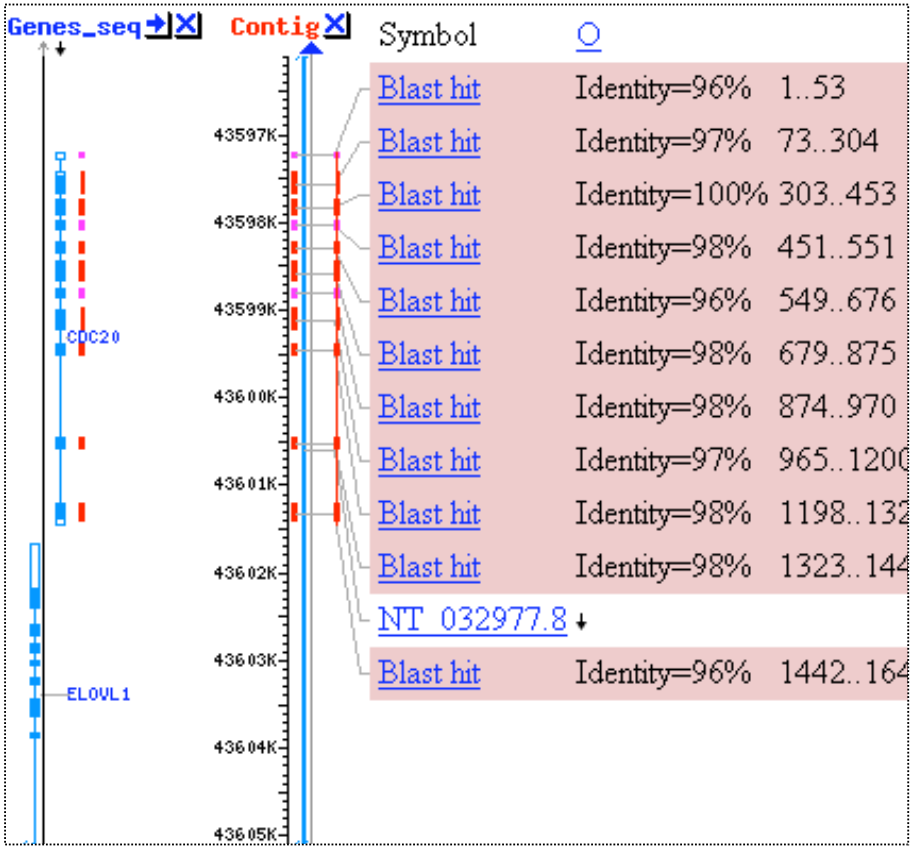
Sorting in Exon Order

```
> ref|NT_032977.8|Hs1_33153  Homo sapiens chromosome 1 genomic contig, reference assembly
Length=73835825

                                                    Sort alignments for this subject sequence by:
                                                      E value  Score  Percent identity
                                                      Query start position  Subject start position

 Features in
   cell div

 Score =  42
 Identities
 Strand=Plus

Query  965
Sbjct  13796

Query  1025
Sbjct  13796

Features flanking this part of subject sequence:
  6169 bp at 5' side: myeloproliferative leukemia virus oncogene
  223 bp at 3' side: cell division cycle 20

Score = 89.7 bits (45),  Expect = 1e-14
Identities = 51/53 (96%), Gaps = 0/53 (0%)
Strand=Plus/Plus

Query  1         AGCGGAGAGTTTAAGAGGCGTAAGCGAGGCGTGTTAAACCCGGTCGGAACTGC  53
                 ||||||||||||||||||||||||| ||||||||||| ||||||||||||||
Sbjct  13796530  AGCGGAGAGTTTAAGAGGCGTAAGCCAGGCGTGTTAAAGCCGGTCGGAACTGC  13796582

Features in this part of subject sequence:
  cell division cycle 20

Score =  412 bits (208),  Expect = 5e-112
Identities = 226/232 (97%), Gaps = 0/232 (0%)
Strand=Plus/Plus

Query  73        GGGCTCCGCAGGCACCAACTGCAAGGACCCCTCCCGCTGCGGGCGTTCCCATGGCACAAT  132
                 |||||||| |||||||||||||||||||||||| ||||||||| ||||||||||||| |
Sbjct  13796755  GGGCTCCGTAGGCACCAACTGCAAGGACCCCTCCCCCTGCGGGCGCTCCCATGGCACAGT  13796814

Query  133       TCGCGTTCGAGAGTGACCTGCACTCGCTGCTTCAGCTGGATGCACCCATCCCCAATGCAC  192
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  13796815  TCGCGTTCGAGAGTGACCTGCACTCGCTGCTTCAGCTGGATGCACCCATCCCCAATGCAC  13796874
```

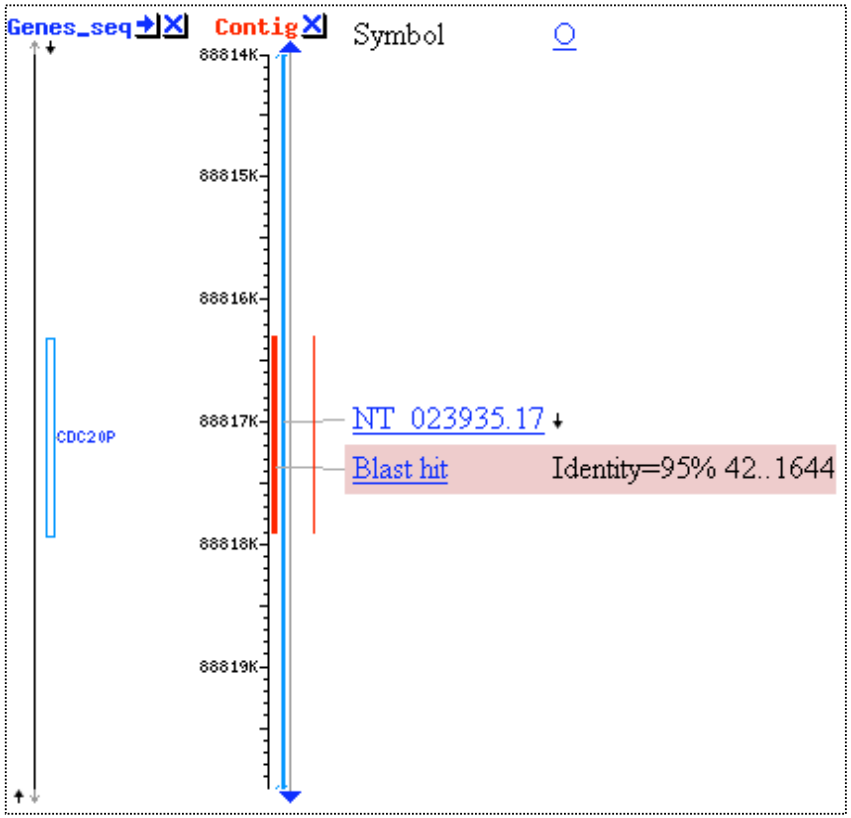
Defau
Lon

Query start position
Exon order

Links to Map Viewer



Chromosome 1



Chromosome 9

Recent and Saved Strategies

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI

Welcome joannealisonfox. [Sign Out]

NCBI/ BLAST/ Recent Results

Links to your unexpired BLAST jobs appear below. more...

Login to My NCBI to save search strategies

Lookup BLAST Job

Request ID: Go

Your Recent Results

(Click headers to sort columns)

Submitted at	Request ID	Status	Program	Title	Qlength	Database	Expires at		
09-26 18:40	FNRZKDEZ012	Done	blastp	Q02067:Achaete-scute homolog 1 (Mash-1)	231	swissprot	09-28 06:40	save	✖
09-26 18:20	FNPT3VP9015	Done	blastp	unknown protein - predict two seperate HSPs	169	nr	09-28 06:20	save	✖
09-26 15:09	FNBKFCA3014	Done	blastx	DinoDNA from THE LOST WORLD p. 135	1435	nr	09-28 03:09	save	✖
09-26 14:57	FNAXJ9F4015	Done	blastn	DinoDNA from JURASSIC PARK p. 103 nt 1-1200	1200	nr	09-28 02:57	save	✖
09-26 12:43	FN31TZK015	Done	megablast	dbj|AB168636| (1696 letters)	1696	Human G+T	09-28 00:43	save	✖

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface

NCBI | NLM | NIH | DHHS

Genomic and Specialized BLAST pages

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with **Primer-BLAST**
- Search **trace archives**
- Find **conserved domains** in your sequence (cds)
- Find sequences with similar **conserved domain architecture** (cdart)
- Search sequences that have **gene expression profiles** (GEO)
- Search **immunoglobulins** (IgBLAST)
- Search for **SNPs** (snp)
- Screen sequence for **vector contamination** (vecscreen)
- **Align** two sequences using BLAST (bl2seq)
- Search **protein** or **nucleotide** targets in PubChem BioAssay

Service Addresses

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

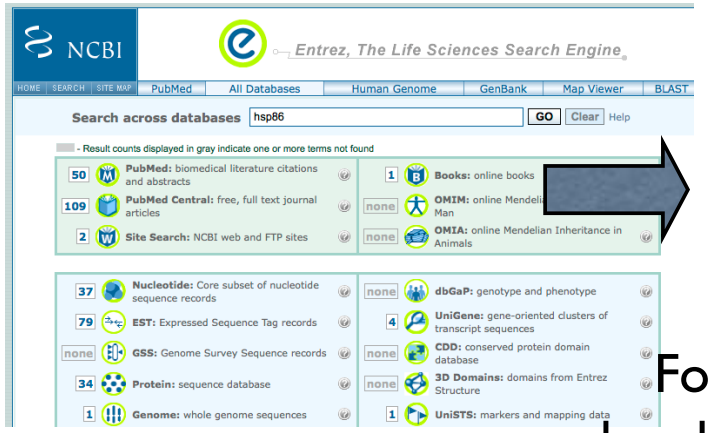
Telephone support: 301- 496- 2475

BLAST

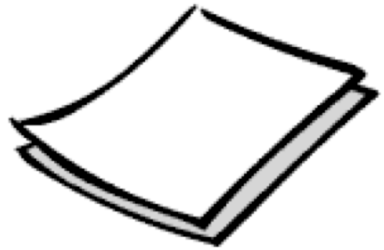
PRACTICAL EXERCISE: The Plasmodium Hsp86 Story

I am studying the control of gene expression in P. *falciparum* and would like to use BLAST to determine whether the coding and intergenic regions of hsp86 are conserved in P. *berghei*, P. *yoelli*, and P. *vivax*.

navigate to:
ncbi.nlm.nih.gov



Step #2: search using Entrez



Follow step-by-step instructions in handout and carry out BLAST searches to complete the practical exercise

Let's compare our results

Use BLAST to find hsp86 orthologues

Step #3: search using tblastn

Step #4: search using blastn

Try some BLAST searches with your own sequence of interest…

Explore what happens when you change advanced parameters…

Step #2 – Search using Entrez

- Use hsp86 to search Entrez
  - ✓ Download protein sequence
  - ✓ Assumes gene name is annotated already

NCBI Entrez Protein

My NCBI [Sign In] [Register]

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

Search Protein for hsp86 Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort By Send to

All: 34 | Bacteria: 0 | RefSeq: 13 | Related Structures: 34

This search in Gene shows 12 results, including:

- hsp90aa1.1 (*Xenopus (Silurana) tropicalis*): heat shock protein 90kDa alpha (cytosolic), class
- Hsp90aa1 (*Rattus norvegicus*): heat shock protein 90, alpha (cytosolic), class A member 1
- HSP90AA1 (*Homo sapiens*): heat shock protein 90kDa alpha (cytosolic), class A member 1

Items 1 - 20 of 34 Page 1 of 2 Next

1: P07901 Reports Conserved Domains, BLink, Links
RecName: Full=Heat shock protein HSP 90-alpha; AltName: Full=HSP 86; AltName: Full=Tumor-specific transplantation 86 kDa antigen; Short=TSTA
gi|1170384|sp|P07901.4|HS90A_MOUSE[1170384]

2: CAA34748 Reports Conserved Domains, BLink, Links
heat shock-like protein [Mus musculus]
gi|51457|emb|CAA34748.1|[51457]

3: AAA37867 Reports Conserved Domains, BLink, Links
heat shock protein

Top Organisms [Tree]
- Mus musculus (12)
- Homo sapiens (11)
- Rattus norvegicus (7)
- Xenopus (Silurana) tropicalis (2)
- Plasmodium falciparum (2)
- All other taxa (1)

More...

Recent Activity
Turn Off Clear
Your browsing activity is empty.

NCBI Entrez Protein

My NCBI [Sign In] [Register]

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

Search Protein for (hsp86) AND "Plasmodium falciparum"[porgn:__txid5833] Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort By Send to

All: 2 | Bacteria: 0 | RefSeq: 1 | Related Structures: 2

This search in Gene shows 1 result.

PF07_0029 (*Plasmodium falciparum 3D7*): heat shock protein 86

Chromosome 7, NC_004328.1 (286894 .. 289916)

Gene ID: 2655065; Other Aliases: PF07_0029

Items 1 - 2 of 2 One page.

1: AAC47837 Reports Conserved Domains, BLink, Links

heat shock protein 86 [Plasmodium falciparum]
gi|2642495|gb|AAC47837.1|[2642495]

2: XP_001348998 Reports Conserved Domains, BLink, Links

heat shock protein 86 [Plasmodium falciparum 3D7]
gi|124511730|ref|XP_001348998.1|[124511730]

Top Organisms [Tree]

Plasmodium falciparum *(2)*
Plasmodium falciparum 3D7 *(1)*

Recent Activity

Turn Off Clear

(hsp86) AND "Plasmodium f... (2) Protein

NCBI Entrez Protein

My NCBI [Sign In] [Register]

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

Search Protein for Go Clear

Limits | Preview/Index | History | Clipboard | Details

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

- GenPept
- GenPept(Full)
- FASTA
- ASN.1
- XML
- INSDSeq XML
- TinySeq XML
- Feature Table

NCBI Reference Sequence: XP_001348998.1

heat shock protein 86 [Plasmodium falciparum 3D7]

Comment Features Sequence

```
LOCUS       XP_001348998             745 aa            linear   INV 0
2008
DEFINITION  heat shock protein 86 [Plasmodium falciparum 3D7].
ACCESSION   XP_001348998
VERSION     XP_001348998.1  GI:124511730
DBSOURCE    REFSEQ: accession XM_001348962.1
KEYWORDS    .
SOURCE      Plasmodium falciparum 3D7
  ORGANISM  Plasmodium falciparum 3D7
            Eukaryota; Alveolata; Apicomplexa; Aconoidasida; Haemosporida;
            Plasmodium; Plasmodium (Laverania).
REFERENCE   1  (residues 1 to 745)
  AUTHORS   Seeger,K., Murphy,L., Harris,D., Berriman,M., Pain,A., Hall,N.,
            Quail,M. and Barrell,B.
  JOURNAL   Unpublished
REFERENCE   2  (residues 1 to 745)
  AUTHORS   Seeger,K., Murphy,L., Harris,D., Berriman,M., Pain,A., Hall,N.,
            Quail,M. and Barrell,B.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-SEP-2002) P.falciparum Genome Sequencing
Consortium,
```

on Shown

ew

alysis Tools

ence

Domains

the PF07_0029 gene

uence of the human malaria parasite Plasm [Nature. 2002]

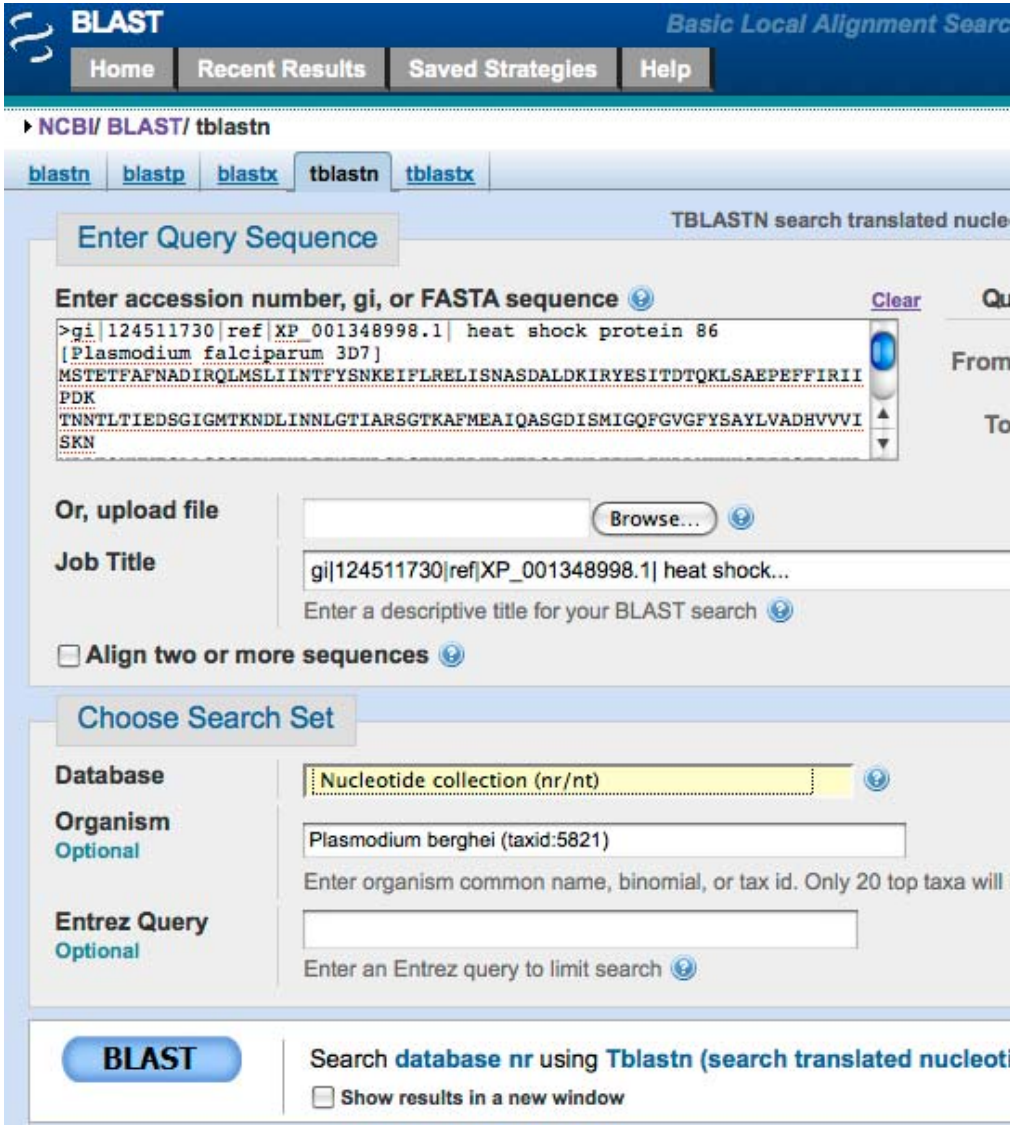
» See all...

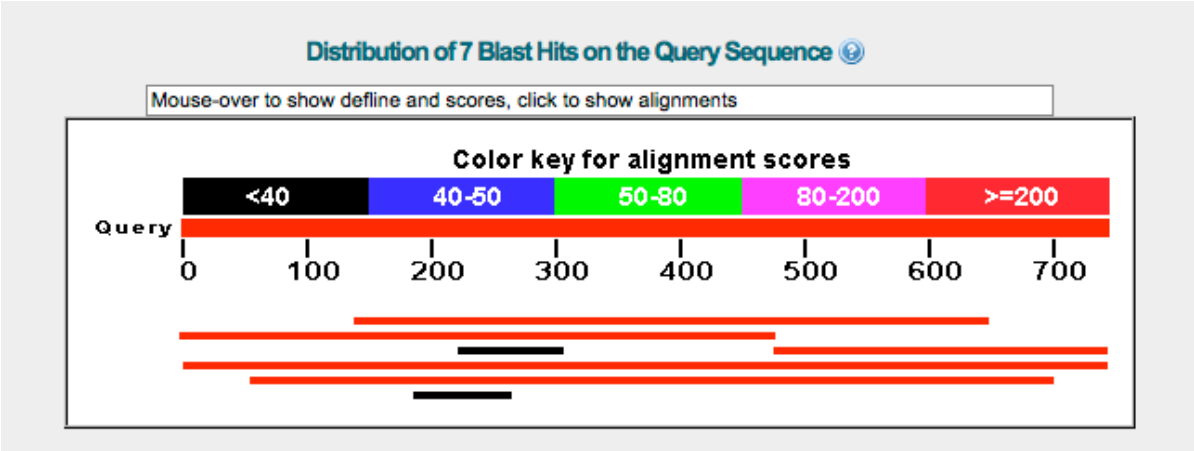
Identical Proteins for XP_001348998.1

- heat shock protein 86 [P [CAD50836]
- heat shock protein 86 [Pl [AAC47837]
- heat shock protein 86 [Pl [AAA66178]

Step #3- tblastn against nr

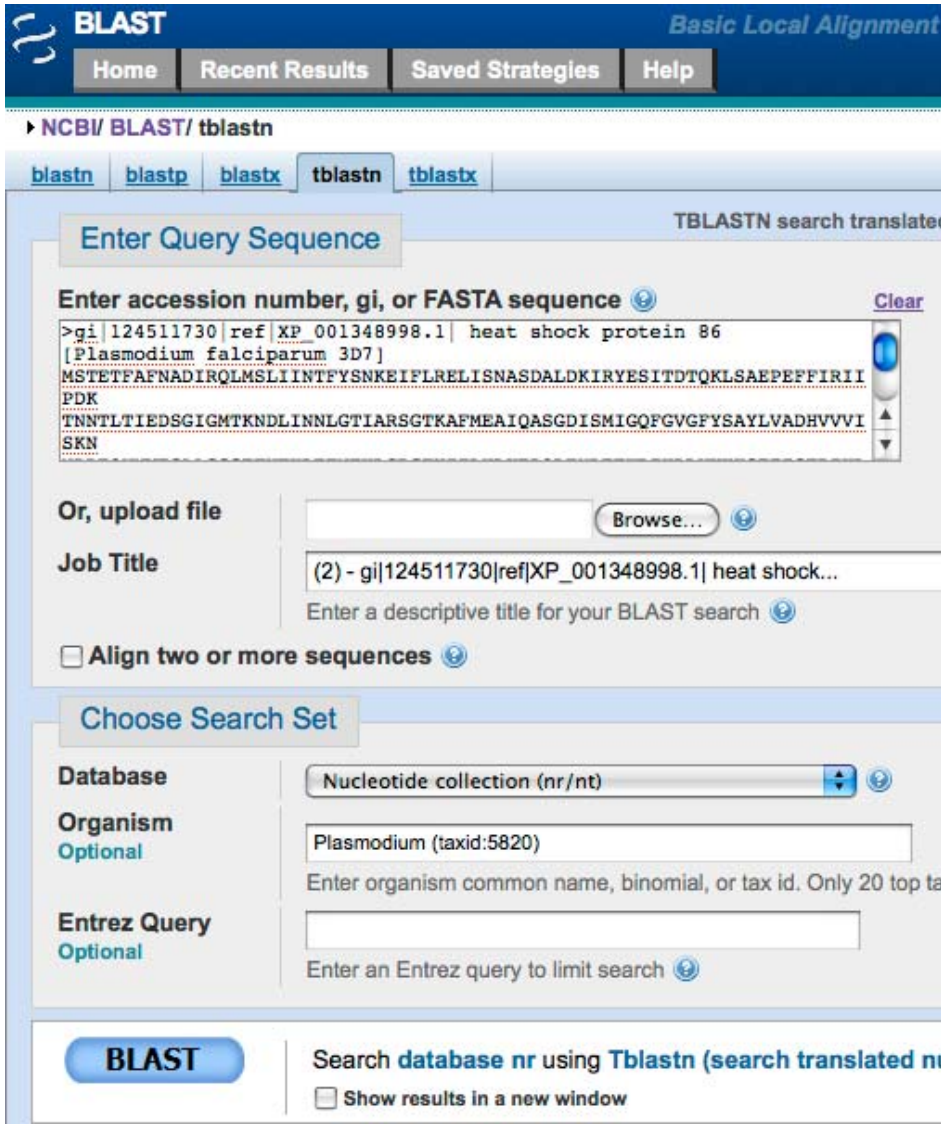
- Translating BLAST programs (blastx, tblastn, tblastx)
  - ✓ Look for similar proteins
  - ✓ Identify potential homologs in other species






```
>ref|XM_669671.1| Plasmodium berghei strain ANKA heat shock protein 86 (PB300823.00.0)
partial mRNA
Length=1455

 GENE ID: 3423212 PB300823.00.0 | heat shock protein 86
[Plasmodium berghei strain ANKA] (10 or fewer PubMed links)

 Score =  870 bits (2249),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 453/510 (88%), Positives = 468/510 (91%), Gaps = 25/510 (4%)
 Frame = +1

Query  141   NDDEQYVWESAAGGSFTVTKDETNEKLGRGTKIILHLKEDQLEYLEEKRIKDLVKKHSEF  200
             NDDEQYVWESAAGGSFTVTKDETNEK+GRGTKIILHLKEDQLEYLEEKRIKDLVKKHSEF
Sbjct  1     NDDEQYVWESAAGGSFTVTKDETNEKIGRGTKIILHLKEDQLEYLEEKRIKDLVKKHSEF  180

Query  201   ISFPIKLYCERQNEKEITASEEEGEGEGEREGEEEEKKKKTGEDKNADESKEENEDEE  260
             ISFPIKLYCERQNEKEIT SEEE  +GE                        K+E ED E
Sbjct  181   ISFPIKLYCERQNEKEITESEEEAQDGE------------------------KKEGEDAE  288

Query  261   KKEDNEEDDNKTDHPKVEDVTEELENAEKKKKEKRKKKIHTVEHEWEELNKQKPLWMRKP  320
             KKED+ E  +  + PKVEDVTEE     +KKKEKRKKKIHTVEHEWEELNKQKPLWMRKP
Sbjct  289   KKEDDGEQKDGEERPKVEDVTEE-LENAEKKKEKRKKKIHTVEHEWEELNKQKPLWMRKP  465

Query  321   EEVTNEEYASFYKSLTNDWEDHLAVKHFSVEGQLEFKALLFIPKRAPFDMFENRKKRNNI  380
             EEVTNEEYASFYKSLTNDWEDHLAVKHFSVEGQLEFKALLFIPKRAPFDMFENRKKRNNI
Sbjct  466   EEVTNEEYASFYKSLTNDWEDHLAVKHFSVEGQLEFKALLFIPKRAPFDMFENRKKRNNI  645

Query  381   KLYVRRVFIMDDCEEIIPEWLNFVKGVVDSEDLPLNISRESLQQNKILKVIKKNLIKKCL  440
             KLYVRRVFIMDDCEEIIPEWLNFVKGVVDSEDLPLNISRESLQQNKILKVIKKNLIKKCL
Sbjct  646   KLYVRRVFIMDDCEEIIPEWLNFVKGVVDSEDLPLNISRESLQQNKILKVIKKNLIKKCL  825

Query  441   DMFSELAENKENYKKFYEQFSKNLKLGIHEDNANRTKITELLRFQTSKSGDEMIGLKEYV  500
             DMF+ELAENK+NYKKFYEQFSKNLKLGIHEDNANR KITELLRFQTSKSGDEMIGLK+YV
Sbjct  826   DMFAELAENKDNYKKFYEQFSKNLKLGIHEDNANRAKITELLRFQTSKSGDEMIGLKDYV  1005

Query  501   DRMKENQKDIYYITGESINAVSNSPFLEALTKKGFEVIYMVDPIDEYAVQQLKDFDGKKL  560
             DRMK+NQKDIYYITGESINAVSNSPFLEALTK+G+EVIYMVDPIDEYAVQQLKDFDGKKL
Sbjct  1006  DRMKDNQKDIYYITGESINAVSNSPFLEALTKRGYEVIYMVDPIDEYAVQQLKDFDGKKL  1185
```

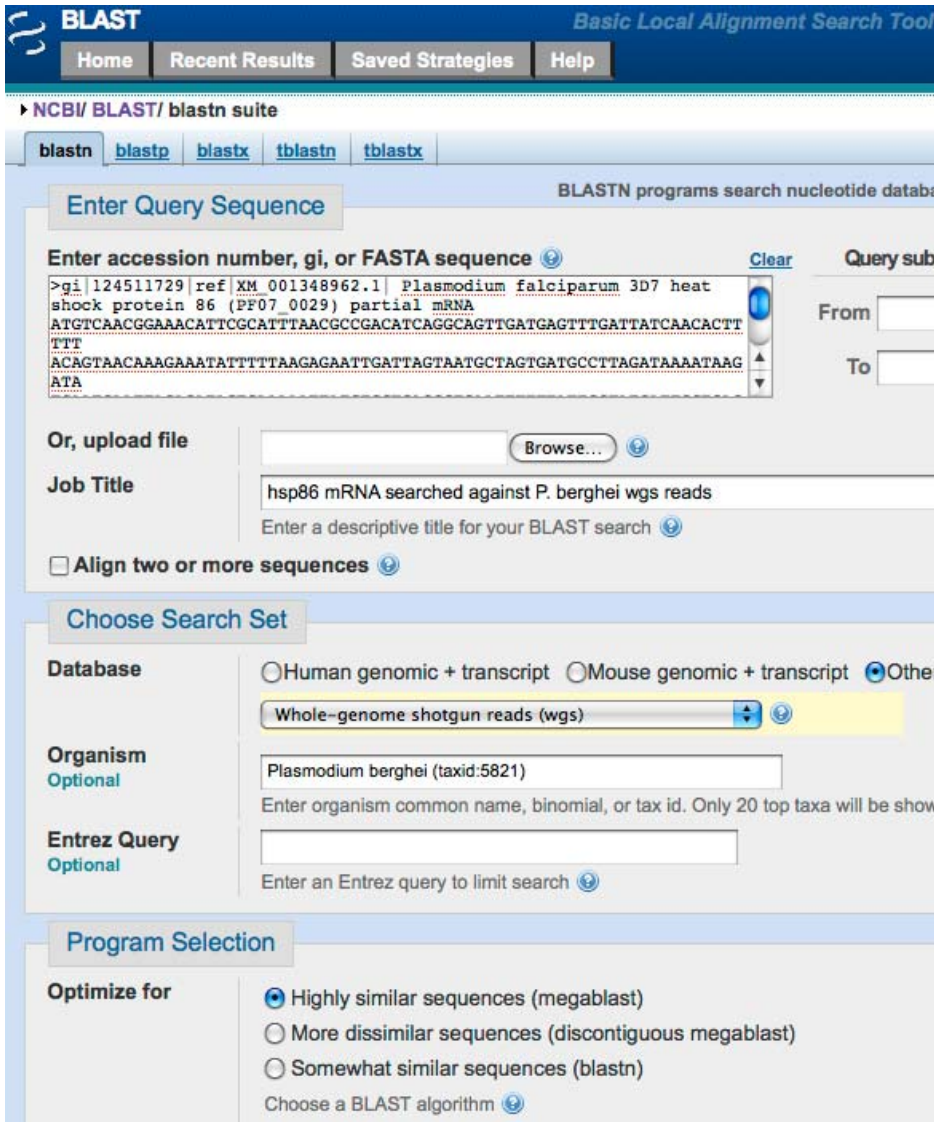
Step #4 – tblastn for all Plasmodium



- Modify tblastn search strategy
  - ✓ Organism limits
  - ✓ Plasmodium (taxid:5820)

Sequences producing significant alignments:	Score (Bits)	E Value	
ref|XM_001348962.1| Plasmodium falciparum 3D7 heat shock prot...	1516	0.0	G
emb|Z29667.1| P.falciparum (7) mRNA for heat-shock protein	1511	0.0	
ref|XM_001613401.1| Plasmodium vivax SaI-1 heat shock protein...	1366	0.0	G
gb|AF030694.2| Plasmodium falciparum strain Dd2 heat shock pr...	969	0.0	
gb|L34027.1|PFAHSP86A Plasmodium falciparum (clone Dd2) heat ...	969	0.0	
emb|AL844506.2| Plasmodium falciparum 3D7 chromosome 7	969	0.0	
ref|XM_669671.1| Plasmodium berghei strain ANKA heat shock pr...	870	0.0	G
ref|XM_724064.1| Plasmodium yoelii yoelii str. 17XNL heat sho...	842	0.0	G
gb|L34028.1|PFAHSP86B Plasmodium falciparum (clone HB3) heat ...	827	0.0	
ref|XM_675364.1| Plasmodium berghei strain ANKA hypothetical ...	808	0.0	G
emb|AM910983.1| Plasmodium knowlesi strain H chromosome 1, co...	551	0.0	
ref|XM_736288.1| Plasmodium chabaudi chabaudi heat shock prot...	706	0.0	G
ref|XM_671644.1| Plasmodium berghei strain ANKA hypothetical ...	528	2e-149	G
ref|XM_724063.1| Plasmodium yoelii yoelii str. 17XNL heat sho...	527	4e-149	G
ref|XM_671514.1| Plasmodium berghei strain ANKA endoplasmin p...	523	1e-147	G
ref|XM_720375.1| Plasmodium yoelii yoelii str. 17XNL heat sho...	520	8e-147	G
ref|XM_001617261.1| Plasmodium vivax SaI-1 endoplasmin precur...	518	2e-146	G
gb|AE014188.2| Plasmodium falciparum 3D7 chromosome 12, compl...	517	4e-146	
emb|AM910996.2| Plasmodium knowlesi strain H chromosome 14, c...	516	6e-146	
ref|XM_001350584.1| Plasmodium falciparum 3D7 endoplasmin hom...	516	6e-146	G
ref|XM_002262256.1| Plasmodium knowlesi strain H endoplasmin ...	516	7e-146	G
emb|X13014.1| Plasmodium falciparum mRNA for HSP90 like protein	385	2e-106	
ref|XM_002259147.1| Plasmodium knowlesi strain H Heat shock p...	336	1e-91	G
emb|AM910991.1| Plasmodium knowlesi strain H chromosome 9, co...	336	1e-91	

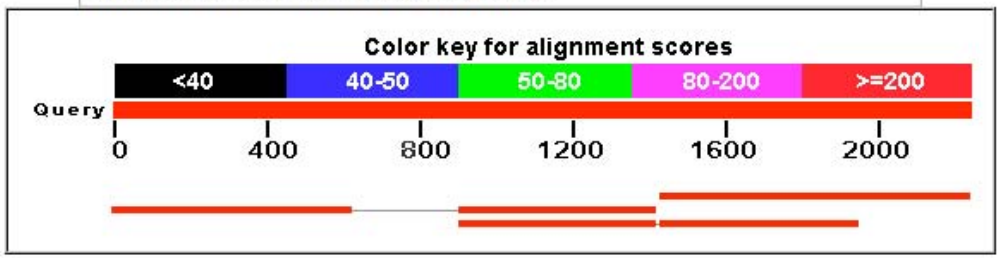
Step #5 – blastn against wgs



- Most common use of blastn
  - ✓ Sequence identification
  - ✓ Establish whether an exact match for a sequence is already present in the database, may need to search additional datasets i.e., wgs
  - ✓ For highly similar sequences use megablast

Distribution of 5 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Distance tree of results NEW

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
gi|221251489|NZ_CAAI01002816.1	Plasmodium berghei whole genome shotgun assembly, contig PB_RP3	870	870	36%	0.0	85%	
gi|221250961|NZ_CAAI01002550.1	Plasmodium berghei whole genome shotgun assembly, contig PB_RP2	660	1287	51%	0.0	87%	
gi|221259295|NZ_CAAI01005066.1	UNBERG.2.12326, whole genome shotgun sequence >gi|56493543|en	627	1211	46%	6e-179	87%	

>gi|221250961|ref|NZ_CAAI01002550.1| D Plasmodium berghei whole genome shotgun assembly, contig PB_RP2909, whole genome shotgun sequence
gi|56501377|emb|CAAI01002550.1| D Plasmodium berghei whole genome shotgun assembly, contig PB_RP2909, whole genome shotgun sequence
Length=30031

Sort alignments for this subject sequence by:
E value Score Percent identity
Query start position Subject start position

```
 Score =  660 bits (357),  Expect = 0.0
 Identities = 545/635 (85%), Gaps = 16/635 (2%)
 Strand=Plus/Plus

Query  1      ATGTCAACGGAAACATTCGCATTTAACGCCGACATCAGGCAGTTGATGAGTTTGATTATC  60
              |||||||  |||||||| |||||||| ||||| || ||||| ||||||||||| || |||
Sbjct  28471  ATGTCAAAAGAAACATTTGCATTTAATGCCGATATTAGGCAATTGATGAGTTTAATCATC  28530

Query  61     AACACTTTTTACAGTAACAAAGAAATATTT-TTAAGAGAA-TTGATTAGTAATGCTAGTG  118
              |||||||| ||||| ||||||||||| ||| ||||||||| || ||||| ||||||||||
Sbjct  28531  AACACTTTCTACAGCAACAAAGAAAT-TTTCTTAAGAGAACTT-ATTAGCAATGCTAGTG  28588

Query  119    ATGC-CTTAGATAAAATAAGATATGAATCAATTACAGATACTCAAAAATTATCT-GCTGA  176
              |||| ||| || |||||||||||||||||||||||||||||| || ||| | ||  || ||
Sbjct  28589  ATGCTCTT-GACAAAATAAGATATGAATCAATTACAGATACCCAGAAACT-TCAAGCCGA  28646

Query  177    GCCTGAATTTTTTATTCGTATCATTCCTGACAAAACCAACAATACATTAACTATTGAAGA  236
               ||||||||||| ||| | || ||||| || |||||||| || || |||||||||||||
Sbjct  28647  AGCTGAATTTTTCATTACAATTATTCGCATAAAACGAATAAGAGTTTAACTATTGAAGA  28706
```

NCBI 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▼

Showing 1.50kb region from base 27392 to 28891.

NCBI Reference Sequence: NZ_CAAI01002550.1

Plasmodium berghei whole genome shotgun assembly, contig PB_RP2909, whole genome shotgun sequence

Comment Features Sequence

```
LOCUS       NZ_CAAI01002550         1500 bp    DNA     linear   INV 05-FEB-2009
DEFINITION  Plasmodium berghei whole genome shotgun assembly, contig PB_RP2909,
            whole genome shotgun sequence.
ACCESSION   NZ_CAAI01002550 REGION: 27392..28891
VERSION     NZ_CAAI01002550.1  GI:221250961
DBLINK      Project:15588
KEYWORDS    WGS.
SOURCE      Plasmodium berghei
  ORGANISM  Plasmodium berghei
            Eukaryota; Alveolata; Apicomplexa; Aconoidasida; Haemosporida;
            Plasmodium; Plasmodium (Vinckeia).
REFERENCE   1
  AUTHORS   Hall,N., Karras,M., Raine,J.D., Carlton,J.M., Kooij,T.W.,
            Berriman,M., Florens,L., Janssen,C.S., Pain,A., Christophides,G.K.,
            James,K., Rutherford,K., Harris,B., Harris,D., Churcher,C.,
            Quail,M.A., Ormond,D., Doggett,J., Trueman,H.E., Mendoza,J.,
            Bidwell,S.L., Rajandream,M.A., Carucci,D.J., Yates,J.R. III,
            Kafatos,F.C., Janse,C.J., Barrell,B., Turner,C.M., Waters,A.P. and
            Sinden,R.E.
  TITLE     A comprehensive survey of the Plasmodium life cycle by genomic,
            transcriptomic, and proteomic analyses
  JOURNAL   Science 307 (5706), 82-86 (2005)
   PUBMED   15637271
REFERENCE   2  (bases 1 to 1500)
  AUTHORS   Hall,N.
  TITLE     Direct Submission
```

Change Region Shown

- Whole sequence
- Selected Region

from: 27392 to: 28891

Update View

Customize View

Sequence Analysis Tools

- BLAST Sequence
- Pick Primers

Recent Activity

Turn Off Clear

- Plasmodium berghei whole genome shotgun assembly, contig PB_RP2909,
- hsp86 mRNA searched again... BLAST
- UNBERG.2.12326, whole genome shotgun sequence
- Plasmodium berghei strain ANKA, whole genome shotgun sequence
- PB300823.00.0 heat shock protein 86

BLAST

COMMON TASKS - Basic Search; Searching Sets of Sequences (multiple inputs; small custom databases); Primer Design

BMC Genomics

BioMed Central

Research article

Open Access

A salmonid EST genomic study: genes, duplications, phylogeny and microarrays

Ben F Koop*[1,6], Kristian R von Schalburg[1], Jong Leong[1], Neil Walker[1], Ryan Lieph[1], Glenn A Cooper[1], Adrienne Robb[1], Marianne Beetz-Sargent[1], Robert A Holt[2], Richard Moore[2], Sonal Brahmbhatt[3], Jamie Rosner[3], Caird E Rexroad III[4], Colin R McGowan[5] and William S Davidson[5]

Address: [1]Centre for Biomedical Research, University of Victoria, Victoria, British Columbia, V8W 3N5, Canada, [2]Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, V5Z 4S6, Canada, [3]Prostate Centre, Vancouver, British Columbia, V6H 3Z6, Canada, [4]ARS, USDA, Natl Ctr Cool & Cold Water Aquaculture, Kearneysville, WV 25430, USA, [5]Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada and [6]Department of Biology, University of Victoria, P.O. Box 3020, Victoria, British Columbia, V8W 3N5, Canada

Email: Ben F Koop* - bkoop@uvic.ca; Kristian R von Schalburg - krvs@uvic.ca; Jong Leong - jong@uvic.ca; Neil Walker - nwalker@uvic.ca; Ryan Lieph - handsomryan@gmail.com; Glenn A Cooper - gac@uvic.ca; Adrienne Robb - arobb@uvic.ca; Marianne Beetz-Sargent - marianbs@uvic.ca; Robert A Holt - rholt@bcgsc.ca; Richard Moore - rmoore@bcgsc.ca; Sonal Brahmbhatt - Sonal.Brahmbhatt@vch.ca; Jamie Rosner - Jamie.Rosner@vch.ca; Caird E Rexroad - caird.rexroadIII@ARS.USDA.GOV; Colin R McGowan - cmcgowan@icywaters.com; William S Davidson - wdavidso@sfu.ca

* Corresponding author

Published: 17 November 2008

BMC Genomics 2008, **9**:545 doi:10.1186/1471-2164-9-545

Received: 13 June 2008
Accepted: 17 November 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/545

Abstract

Background: Salmonids are of interest because of their relatively recent genome duplication, and their extensive use in wild fisheries and aquaculture. A comprehensive gene list and a comparison of genes in some of the different species provide valuable genomic information for one of the most widely studied groups of fish.

Results: 298,304 expressed sequence tags (ESTs) from Atlantic salmon (69% of the total), 11,664 chinook, 10,813 sockeye, 10,051 brook trout, 10,975 grayling, 8,630 lake whitefish, and 3,624 northern pike ESTs were obtained in this study and have been deposited into the public databases. Contigs were built and putative full-length Atlantic salmon clones have been identified. A database containing ESTs, assemblies, consensus sequences, open reading frames, gene predictions and putative annotation is available. The overall similarity between Atlantic salmon ESTs and those of rainbow trout, chinook, sockeye, brook trout, grayling, lake whitefish, northern pike and rainbow smelt is 93.4, 94.2, 94.6, 94.4, 92.5, 91.7, 89.6, and 86.2% respectively. An analysis of 78 transcript sets show *Salmo* as a sister group to *Oncorhynchus* and *Salvelinus* within Salmoninae, and Thymallinae as a sister group to Salmoninae and Coregoninae within Salmonidae. Extensive gene duplication is consistent with a genome duplication in the common ancestor of salmonids. Using all of the available EST data, a new expanded salmonid cDNA microarray of 32,000 features was created. Cross-species hybridizations to this cDNA microarray indicate that this resource will be useful for studies of all 68 salmonid species.

Conclusion: An extensive collection and analysis of salmonid RNA putative transcripts indicate that Pacific salmon, Atlantic salmon and charr are 94–96% similar while the more distant whitefish, grayling, pike and smelt are 93, 92, 89 and 86% similar to salmon. The salmonid transcriptome reveals a complex history of gene duplication that is consistent with an ancestral salmonid genome duplication hypothesis. Genome resources, including a new 32 K microarray, provide valuable new tools to study salmonids.

Open the Koop et al paper:

BMC Genomics

Research article

Open Access

A salmonid EST genomic study: genes, duplications, phylogeny and microarrays

Ben F Koop[*1,6], Kristian R von Schalburg[1], Jong Leong[1], Neil Walker[1], Ryan Lieph[1], Glenn A Cooper[1], Adrienne Robb[1], Marianne Beetz-Sargent[1], Robert A Holt[2], Richard Moore[2], Sonal Brahmbhatt[3], Jamie Rosner[3], Caird E Rexroad III[4], Colin R McGowan[5] and William S Davidson[5]

Published: 17 November 2008

Received: 13 June 2008

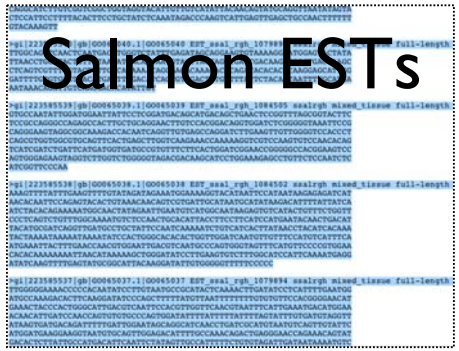
Accepted: 17 November 2008

Abstract

Background: Salmonids are of interest because of their relatively recent genome duplication, and their extensive use in wild fisheries and aquaculture. A comprehensive gene list and a comparison of genes in some of the different species provide valuable genomic information for one of the most widely studied groups of fish.

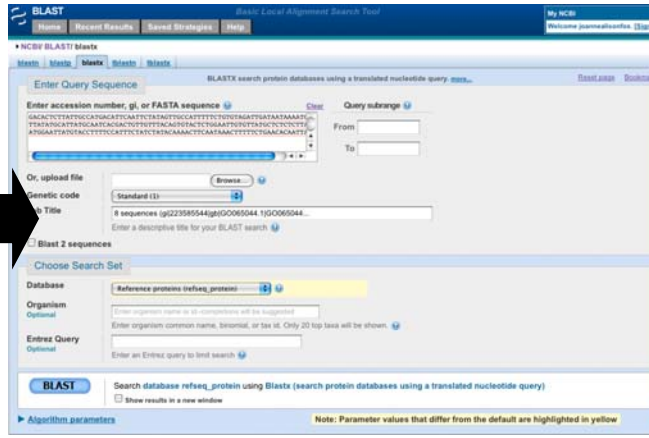
Results: 298,304 expressed sequence tags (ESTs) from Atlantic salmon (69% of the total), 11,664 chinook, 10,813 sockeye, 10,051 brook trout, 10,975 grayling, 8,630 lake whitefish, and 3,624 northern pike ESTs were obtained in this study and have been deposited into the public databases. Contigs were built and putative full-length Atlantic salmon clones have been identified. A database containing ESTs, assemblies, consensus sequences, open reading frames, gene predictions and putative annotation is available. The overall similarity between Atlantic salmon ESTs and those of rainbow trout, chinook, sockeye, brook trout, grayling, lake whitefish, northern pike and rainbow smelt is 93.4, 94.2, 94.6, 94.4, 92.5, 91.7, 89.6, and 86.2% respectively. An analysis of 78 transcript sets show Salmo as a sister group to Oncorhynchus and Salvelinus within Salmoninae, and Thymallinae as a sister group to Salmoninae and Coregoninae within Salmonidae. Extensive gene duplication is consistent with a genome duplication in the common ancestor of salmonids. Using all of the available EST data, a new expanded salmonid cDNA microarray of 32,000 features was created. Cross-species hybridizations to this cDNA microarray indicate that this resource will be useful for studies of all 68 salmonid species.

Conclusion: An extensive collection and analysis of salmonid RNA putative transcripts indicate that Pacific salmon, Atlantic salmon and charr are 94–96% similar while the more distant whitefish, grayling, pike and smelt are 93, 92, 89 and 86% similar to salmon. The salmonid transcriptome reveals a complex history of gene duplication that is consistent with an ancestral salmonid genome duplication hypothesis. Genome resources, including a new 32 K microarray, provide valuable new tools to study salmonids.



Get the Salmon sequences and carry out the BLAST searches

We'll walk through this example together



Can you identify the ESTs?

Search #1: Use multiple EST sequences as input query

use blastx

Is the hbaa1 gene present?

Search #2: Use the hbaa1 sequence as input, search against Salmon EST custom database

use blast2seq option with tblastn

Searching with Multiple Sequences as Input

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI — Welcome joannealisonfox. [Sign Out]

NCBI/ BLAST/ blastx

blastn | blastp | blastx | tblastn | tblastx

BLASTX search protein databases using a translated nucleotide query. more...

Reset page | Bookmark

Enter Query Sequence

Enter accession number, gi, or FASTA sequence — Clear

Query subrange

From

To

Or, upload file — Browse...

Genetic code — Standard (1)

Job Title — 8 sequences (gi|223585544|gb|GO065044.1|GO065044...

Enter a descriptive title for your BLAST search

Blast 2 sequences

Choose Search Set

Database — Reference proteins (refseq_protein)

Organism (Optional) — Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query (Optional)

Enter an Entrez query to limit search

BLAST — Search database refseq_protein using Blastx (search protein databases using a translated nucleotide query)

Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow

11 sequences (gi|223585544|gb|GO065044.1|GO065044...

Results for: 7:lcl|5773 gi|223585538|gb|GO065038.1|GO065038 EST_ssal_rgh_1084502 ssalrgh mixed_tissue full-length Salmo sal...(614bp)

What's this?

Query ID
Description

1:lcl|5767 gi|223585544|gb|GO065044.1|GO065044 EST_ssal_rgh_1084509 ssalrgh mixed_tissue full-length Salmo sal...(725bp)
2:lcl|5768 gi|223585543|gb|GO065043.1|GO065043 EST_ssal_rgh_1079901 ssalrgh mixed_tissue full-length Salmo sal...(897bp)
3:lcl|5769 gi|223585542|gb|GO065042.1|GO065042 EST_ssal_rgh_1079900 ssalrgh mixed_tissue full-length Salmo sal...(266bp)
*4:lcl|5770 gi|223585541|gb|GO065041.1|GO065041 EST_ssal_rgh_1084506 ssalrgh mixed_tissue full-length Salmo sal...(290bp)
*5:lcl|5771 gi|223585540|gb|GO065040.1|GO065040 EST_ssal_rgh_1079898 ssalrgh mixed_tissue full-length Salmo sal...(310bp)
6:lcl|5772 gi|223585539|gb|GO065039.1|GO065039 EST_ssal_rgh_1084505 ssalrgh mixed_tissue full-length Salmo sal...(432bp)
7:lcl|5773 gi|223585538|gb|GO065038.1|GO065038 EST_ssal_rgh_1084502 ssalrgh mixed_tissue full-length Salmo sal...(614bp)
8:lcl|5774 gi|223585537|gb|GO065037.1|GO065037 EST_ssal_rgh_1079894 ssalrgh mixed_tissue full-length Salmo sal...(629bp)
9:lcl|5775 gi|223585536|gb|GO065036.1|GO065036 EST_ssal_rgh_1079893 ssalrgh mixed_tissue full-length Salmo sal...(884bp)
10:lcl|5776 gi|223585535|gb|GO065035.1|GO065035 EST_ssal_rgh_1084500 ssalrgh mixed_tissue full-length Salmo sal...(821bp)
11:lcl|5777 gi|223585534|gb|GO065034.1|GO065034 EST_ssal_rgh_1079892 ssalrgh mixed_tissue full-length Salmo sal...(791bp)

Molecule type
Query Length

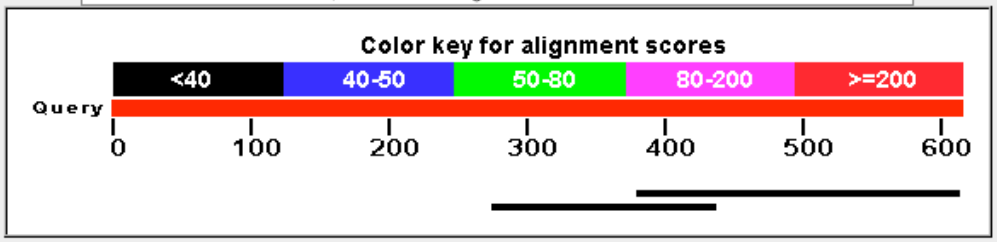
Other reports: ▷Search Summary | Taxonomy reports

Results for: pull down list

Graphic Summary

Distribution of 2 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Descriptions

Alignments

Select All Get selected sequences

```
>ref|YP_934206.1|  hypothetical protein azo2703 [Azoarcus sp. BH72]
Length=774

 GENE ID: 4607585 azo2703 | hypothetical protein [Azoarcus sp. BH72]
(10 or fewer PubMed links)

 Score = 35.0 bits (79),  Expect = 4.3
 Identities = 20/80 (25%), Positives = 36/80 (45%), Gaps = 2/80 (2%)
 Frame = -2

Query  613  GEKPPQYPCNAAYSKL--DILILNGCQRHFKDIPAFYVNFFCVFHGEHETTHWALTSIPR  440
            G++PP  P + A + L  D L+L    +H+K   A +  +  +  G  +   W L   P
Sbjct  557  GQRPPVTPLSRAEAGLPDDALVLAAFHQHYKITRASFALWMRLLRGLPDALLWLLEGAPS  616

Query  439  WFKVISLK*HGNNIDPTSVC  380
                +S +   + +DP  +C
Sbjct  617  AMARLSQEARAHGVDPARLC  636
```

BLAST

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ tblastn

blastn | blastp | blastx | tblastn | tblastx

TBLASTN search translated nucleotide subjects using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA

paste hbaa1 sequence

```
>gi|47271417|ref|NP_571332.2| hemoglo
MSLSDTDKAVVKAIWAKISPKADEIGAEALARMLTVYPQTKTYFSHWADLSPGSGPVKKHGKTIMGAVG
E
AVSKIDDLVGGLAALSELHAFKLRVDPANFKILSHNVIVVIAMLFPADFTPEVHVSVDKFFNNLALALS
E
KYR
```

From

To

Or, upload file

Browse...

Job Title

Enter a descriptive title for your BLAST sea

Use BLAST 2 Sequences for Searching against small custom databases

Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA s

paste Salmon ESTs

```
>gi|223585544|gb|GO065044.1|GO065044 ES
ssalrgh mixed_tissue full-length Salmo salar cDNA Salmo salar
cDNA clone ssal_rgh_520_381 3', mRNA sequence
AACTTGCAGCAAATACAAAAACAATAAATGATCAAACGAAACGTGACAACAGTGACATGCAAAC
AGGCAC
CTACACAAAAACAAGATCCCACAAACCAGTGGGGAAATGGCTGCCGAAATATGATCCCCAATCA
```

To

Or, upload file

Browse...

BLAST

Search nucleotide sequence using Tblastn (search translated nucleotide subjects using a protein qu

Show results in a new window

Search against small custom database

Blast 2 sequences

gi|47271417|ref|NP_571332.2| hemoglobin alpha...

Query ID lcl|20148
Description gi|47271417|ref|NP_571332.2| hemoglobin alpha adult-1 [Danio rerio]
Molecule type amino acid
Query Length 143

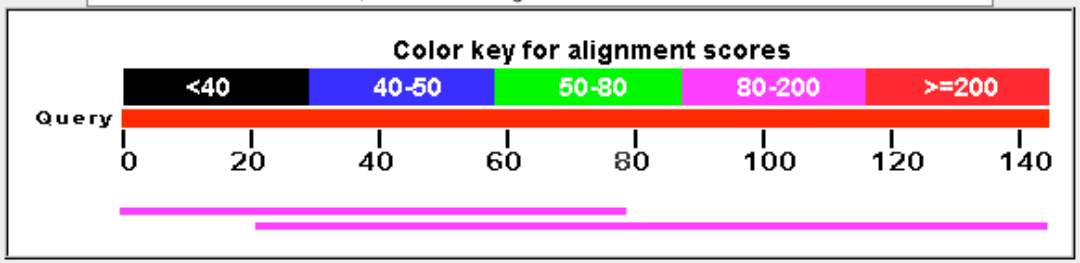
Subject ID 8 subjects
Description See details
Molecule type nucleic acid
Subject Length n/a
Program TBLASTN 2.2.19+ Citation

Other reports: Search Summary [Taxonomy reports]

Graphic Summary

Distribution of 2 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Descriptions

```
                                                          Score     E
Sequences producing significant alignments:              (Bits)  Value

lcl|20152  gi|223585542|gb|GO065042.1|GO065042 EST_ssal_rgh_10...   116     3e-31
lcl|20155  gi|223585539|gb|GO065039.1|GO065039 EST_ssal_rgh_10...   178     4e-50
```

BLAST tasks

- Basic BLAST
  - ✓ Hsp86 examples
- Batch BLAST searching
  - ✓ Use Salmon ESTs as input
- Search against a small custom database
  - ✓ Use BLAST 2 Sequences utility

Primer-BLAST

- NCBI's Primer Designer and Specificity Checker

http://www.ncbi.nlm.nih.gov/tools/primer-blast/

Primer-BLAST — A tool for finding specific primers

NCBI/ Primer-BLAST: Finding primers specific to your PCR template (using Primer3 and BLAST). more... Tips for finding specific primers

Reset page | Save search parameters

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred) Clear

Range — From / To
Forward primer
Reverse primer
Clear

Or, upload FASTA file — Choose File no file selected

Primer Parameters

Use my own forward primer (5'->3' on plus strand) Clear
Use my own reverse primer (5'->3' on minus strand)

PCR product size — Min 200, Max 1000
# of primers to return — 10

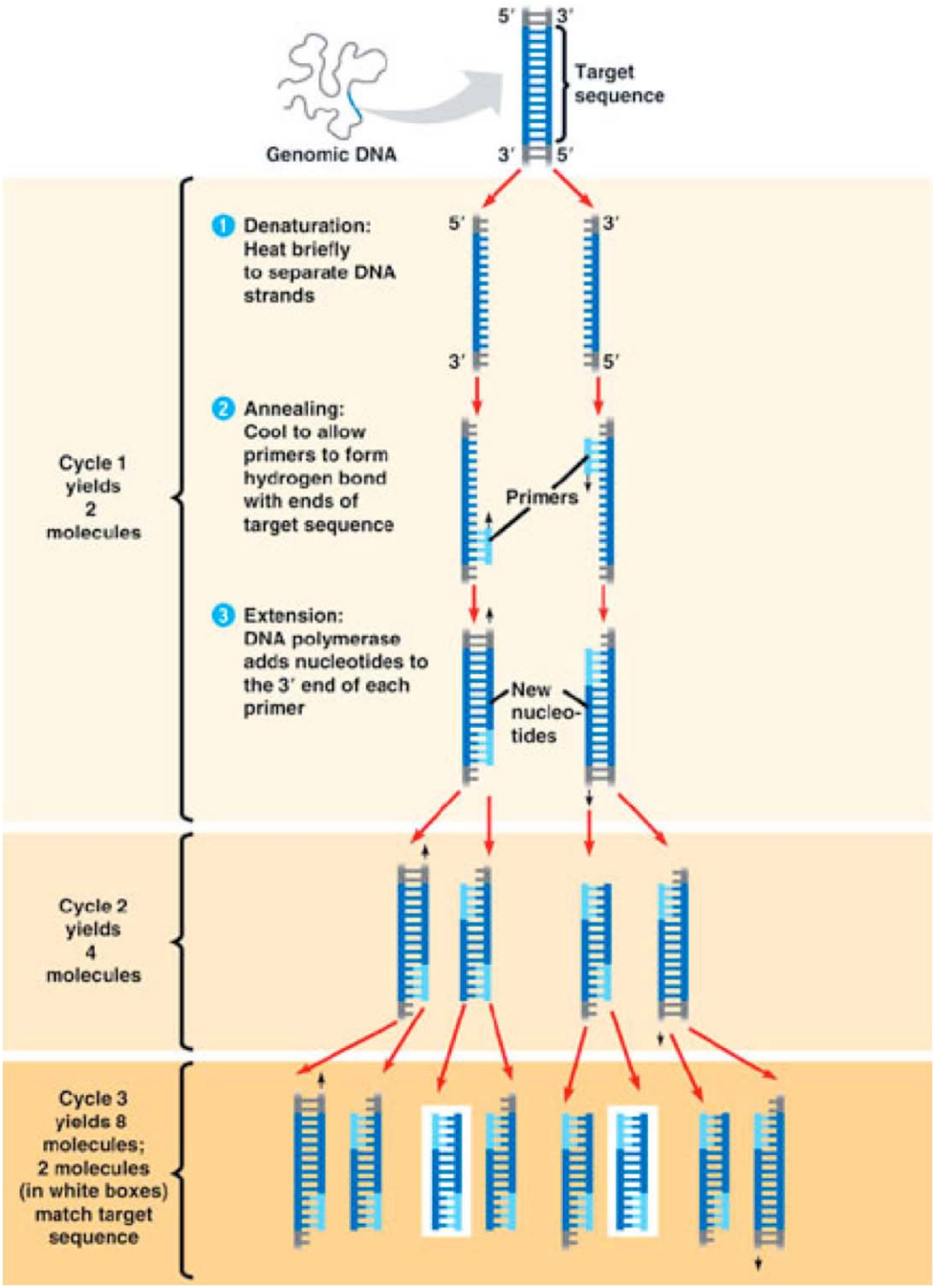
Primer melting temperatures (T_m) — Min 57.0, Opt 60.0

offers integrated primer design with Primer3 & specificity check with custom BLAST

Primer Pair Specificity Checking Parameters

Specificity check — ☑ Enable search for primer pairs specific to the intended PCR template

Organism — Homo sapiens


Genomic DNA
5′
3′
Target sequence
3′
5′
1 Denaturation: Heat briefly to separate DNA strands
5′
3′
3′
5′
2 Annealing: Cool to allow primers to form hydrogen bond with ends of target sequence
Primers
3 Extension: DNA polymerase adds nucleotides to the 3′ end of each primer
New nucleo-tides
Cycle 1 yields 2 molecules
Cycle 2 yields 4 molecules
Cycle 3 yields 8 molecules; 2 molecules (in white boxes) match target sequence

Primer Design

Balance:

✓ Specificity - frequency of mispriming

✓ Efficiency of Amplification - 2X increase

Consider:

- primer length (18-24nt)
- primer Tm (>54°C)
- 3' end (G or C)
- GC content (45-55%)
- primer dimers
- for cDNA - coding region; across intron/exon boundary

General Concepts for PCR Primer Design.
Dieffenback CW, Lowe TMJ, Dveksler GS Genome Research 3 (1993) S30-37 [PMID:8118394]

Primer-BLAST input

designs primers specific to target template and unique in the target database

NCBI/ Primer-BLAST: Finding primers

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred) Clear

Range

From To

Forward primer

Reverse primer

Clear

Or, upload FASTA file Browse...

Primer Parameters

Use my own forward primer (5'->3' on plus strand) Clear

Use my own reverse primer (5'->3' on minus strand) Clear

Min Max

PCR product size 200 1000

# of primers to return 10

Min Opt Max Max Tm difference

Primer melting temperatures (Tm) 57.0 60.0 63.0 3

can specify primer sequence(s), desired product size, Tm ranges, Tm difference (can be used with or without template)

Primer-BLAST Specificity

By default human sequences are searched in specificity check

Primer Pair Specificity Checking Parameters

Specificity check

☑ Enable search for primer pairs specific to the intended PCR template

With this option on, the program will search the primers against the selected database and determine whether a primer pair can generate a PCR product on any targets in the database based on their matches to the targets and their orientations. The program will return, if possible, only primer pairs that do not generate a valid PCR product on unintended sequences and are therefore specific to the intended template. Note that the specificity is checked not only for the forward-reverse primer pair, but also for forward-forward as well as reverse-reverse primer pairs.

Organism

Homo sapiens

Enter an organism name, taxonomy id or select from the suggestion list as you type.

Database

Refseq mRNA (refseq_rna)

Primer specificity stringency

At least 2 total mismatches to unintended targets, including

at least 2 mismatches within the last 5 bps at the 3' end

The larger the mismatches (especially those toward 3' end) are between primers and the unintended targets, the more specific the primer pair is to your template (i.e., it will be difficult to anneal to and amplify unintended targets). However, specifying a larger mismatch value may make it more difficult to find such specific primers. Try to lower the mismatch value in such case.

custom BLAST; focus on 3' end to avoid mispriming

Misprimed product size deviation

1000

Splice variant handling

☐ Allow primer to amplify mRNA splice varia

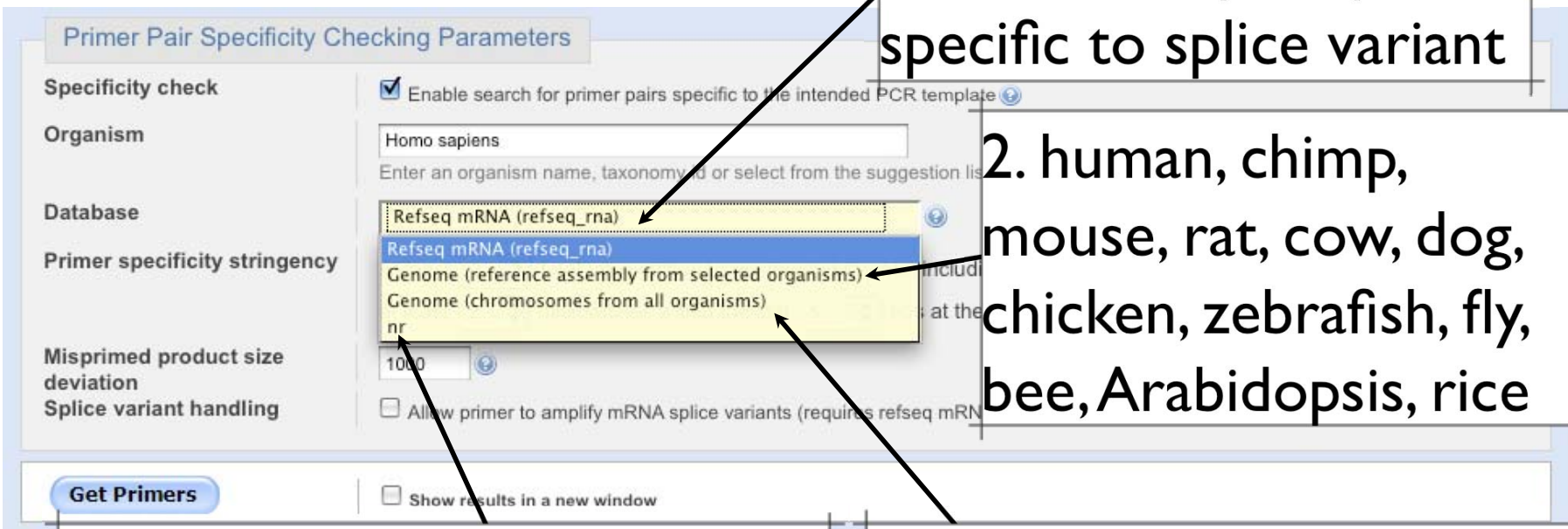
Get Primers

☐ Show results in a new window

Advanced parameters

Primer-BLAST Specificity

Four BLAST nucleotide databases available for searching



1. with refseq template, specific to splice variant

2. human, chimp, mouse, rat, cow, dog, chicken, zebrafish, fly, bee, Arabidopsis, rice

3. all NC_ (includes above) + other organisms, microbes

4. nr = database with widest coverage of organisms

Primer-BLAST Advanced

Adjustable settings from Primer3 see Primer 3 Input Help: http://fokker.wi.mit.edu/primer3/input-help-040.htm

Useful options specific to Primer-BLAST:
1. avoid regions that contain SNPs
2. avoid repetitive regions

Advanced parameters

Primer Pair Specificity Checking Parameters

Blast max number of hit sequences	250 (default)
Blast expect (E) value	1000 (default)
Max primer pairs to screen	3000 (default)

Primer Parameters

	Min	Opt	Max
PCR Product Tm			
Primer Size	15	20	27
Primer GC content (%)	20.0		80.0

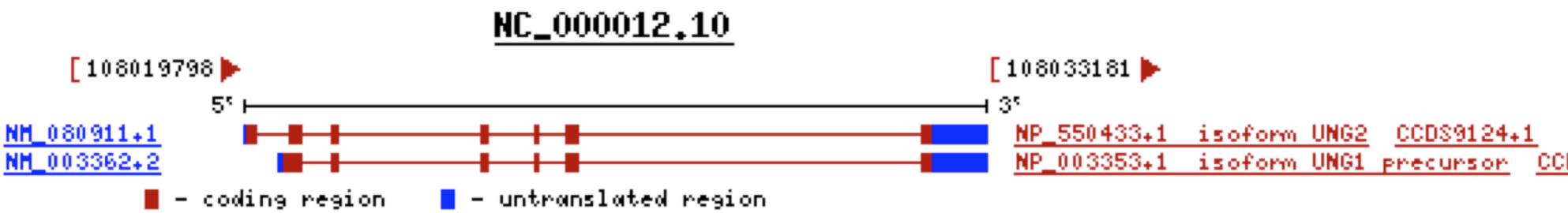
GC clamp	0
Max self complementarity:	8.00
Max 3' end complementarity:	3.00
SNP handling	☐ Primer binding site may not contain known SNP
Repeat filter	Automatic Avoid repeat region for primer selection by filtering with repeat database
Low complexity filter	☑ Avoid low complexity region for primer selection
Concentration of monovalent cations	50.0
Concentration of divalent cations	0.0
Concentration of dNTPs	0.0
Salt correction formula:	Schildkraut and Lifson 1965
Annealing Oligo Concentration	50.0

Internal hybridization oligo parameters

Hybridization oligo	☐ Pick internal hybridization oligo

Min Opt Max

Primer-BLAST example



Task #1: Use Primer BLAST to design primers specific to the UNG2 splice variant, NM_080911.

Task #2: Use Primer BLAST to design primers that will identify both splice variants.

Task #3: Carry out a specificity check for one of your primer pairs. Will this primer pair (designed against the human UNG transcripts) also amplify transcripts from other primate species?

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

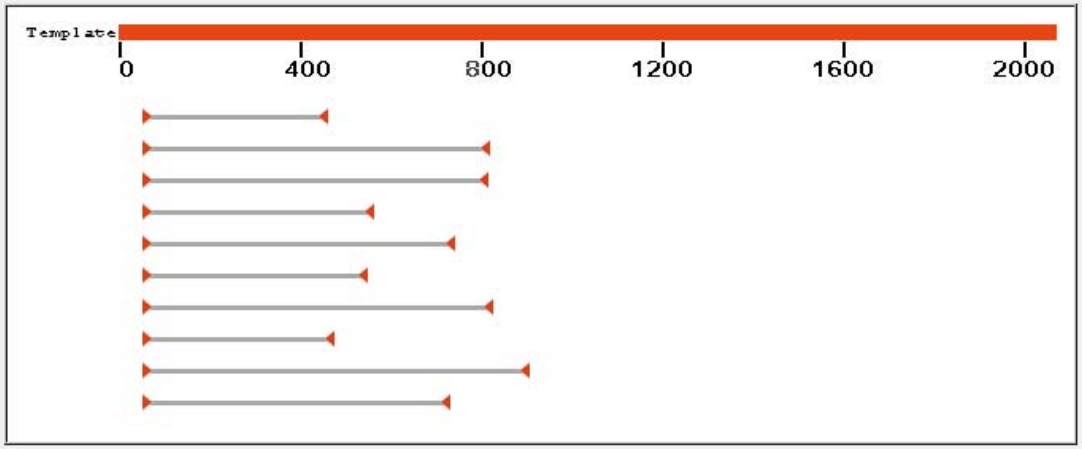
- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search for SNPs (snp)
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA transcript libraries
- Constraint Based Protein Multiple Alignment Tool

Task #1: Use Primer BLAST to design primers specific to the UNG2 splice variant, NM_080911.

enter NM_080911 as template

use all default settings

▼ Summary of primer pairs



▼ Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	CTCCTCAGCTCCAGGATGAT	Plus	20	56	75	59.36	55.00%
Reverse primer	AGGTGAAGACTTGGTGTGGG	Minus	20	479	460	60.00	55.00%
Product length	424						

Products on intended target

>NM_080911.1 Homo sapiens uracil-DNA glycosylase (UNG), transcript variant 2, mRNA

```
product length = 424
Forward primer  1    CTCCTCAGCTCCAGGATGAT  20
Template        56   ....................  75

Reverse primer  1     AGGTGAAGACTTGGTGTGGG  20
Template        479  ....................  460
```

Task #2: Use Primer BLAST to design primers that will identify both splice variants.

▼ Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	CCCACACCAAGTCTTCACCT	Plus	20	460	479	60.00	55.00%
Reverse primer	CACCCCAACATCTGTCACTG	Minus	20	1407	1388	60.00	55.00%
Product length	948						

Products on intended target

>NM_080911.1 Homo sapiens uracil-DNA glycosylase (UNG), transcript variant 2, mRNA

```
product length = 948
Forward primer  1     CCCACACCAAGTCTTCACCT  20
Template        460   ....................  479

Reverse primer  1      CACCCCAACATCTGTCACTG  20
Template        1407   ....................  1388
```

Products on allowed transcript variants

>NM_003362.2 Homo sapiens uracil-DNA glycosylase (UNG), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA

```
product length = 948
Forward primer  1     CCCACACCAAGTCTTCACCT  20
Template        488   ....................  507

Reverse primer  1      CACCCCAACATCTGTCACTG  20
Template        1435   ....................  1416
```

enter NM_080911 as template

- Allow primer to amplify mRNA splice variants

Note: Parameter values that differ from the default are highlighted in yellow

Primer Pair Specificity Checking Parameters

Specificity check

☑ Enable search for primer pairs specific to the intended PCR template

Organism

Homo sapiens

Enter an organism name, taxonomy id or select from the suggestion list as you type.

Add more organisms

Database

Refseq RNA (refseq_rna)

Primer specificity stringency

At least 2 total mismatches to unintended targets, including

at least 2 mismatches within the last 5 bps at the 3' end

Misprimed product size deviation

1000

Splice variant handling

☑ Allow primer to amplify mRNA splice variants (requires refseq mRNA sequence as PCR template input)

If enabled, this program will NOT exclude the primer pairs that can amplify the mRNA splice variants of the same gene as your PCR template, thus making primers gene specific rather than transcript specific. This option requires you to enter a refseq mRNA accession or gi or fasta sequence as PCR template input because other type of input may not allow the program to properly interpret the result.

Get Primers

☐ Show results in a new window

▶ Advanced parameters

Note: Parameter values that differ from the default are highlighted in yellow

Task #3: Carry out a specificity check for one of your primer pairs. Will this primer pair (designed against the human UNG transcripts) also amplify transcripts from other primate species?

no template

use my own:

- forward primer
- reverse primer

- organism; specify primate

- database; specify nr

Primer pair 1

	Sequence (5'->3')	Length	Tm	GC%
Forward primer	GCCTTGTTTTCTTGCTCTGG	20	59.99	50.00%
Reverse primer	CACCCCAACATCTGTCACTG	20	60.00	55.00%

Products on target templates

>AK291341.1 Homo sapiens cDNA FLJ76845 complete cds, highly similar to Homo sapiens uracil-DNA glycosylase (UNG), transcript variant 1, mRNA

```
product length = 595
Forward primer  1    GCCTTGTTTTCTTGCTCTGG  20
Template        849  ....................  868

Reverse primer  1     CACCCCAACATCTGTCACTG  20
Template        1443  ....................  1424
```

>XM_001136198.1 PREDICTED: Pan troglodytes uracil-DNA glycosylase, transcript variant 1 (UNG), mRNA

```
product length = 595
Forward primer  1    GCCTTGTTTTCTTGCTCTGG  20
Template        925  ....................  944

Reverse primer  1     CACCCCAACATCTGTCACTG  20
Template        1519  ....................  1500
```

>XM_509349.2 PREDICTED: Pan troglodytes uracil-DNA glycosylase, transcript variant 2 (UNG), mRNA

```
product length = 595
Forward primer  1    GCCTTGTTTTCTTGCTCTGG  20
Template        848  ....................  867

Reverse primer  1     CACCCCAACATCTGTCACTG  20
Template        1442  ....................  1423
```

>XM_001104421.1 PREDICTED: Macaca mulatta similar to uracil-DNA glycosylase isoform UNG1 precursor, transcript variant 2 (LOC706816), mRNA

```
product length = 603
Forward primer  1    GCCTTGTTTTCTTGCTCTGG  20
Template        868  ....................  887

Reverse primer  1     CACCCCAACATCTGTCACTG  20
```

Things you can do to maximize the chance of finding primers specific for your template.

- Use refseq accession or GI (rather than the raw DNA sequence) as template whenever possible. Even if you are only interested in part of the sequence, you can still use the accession or GI but you do need to specify the range (use forward primer "From" field for your sequence start position and reverse primer "To" field for your sequence stop position). The reason is that an accession or GI carries accurate information about its identity which allows primer-blast to better distinguish between intended template and off-targets.

- Choose a non-redundant database (such as refseq_rna or genome database). The nr database contains redundant entries which can interfere with the process of finding specific primers.

- Specify an organism for database search if you are only amplifying DNA from a specific organism. Searching all organisms will be much slower and off-target priming from other organisms are irrelevant.

Credits

- Materials for this presentation have been adapted with permission from the following NCBI HelpDesk course materials:

  Field Guide Course Materials

  Advanced Workshop for Bioinformatics Information Specialists

  NCBI News

- NCBI BLAST

  http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

MSA

MSA = Multiple Sequence Alignments

Examples


globin.aln
CLUSTAL 2.0.9 multiple sequence alignment
HBB_HUMAN --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
HBB_HORSE --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
HBA_HUMAN ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
HBA_HORSE ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
MYG_PHYCA ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
LGB2_LUPLU --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
HBB_HUMAN PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
HBB_HORSE PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
HBA_HUMAN ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
HBA_HORSE ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
GLB5_PETMA ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
MYG_PHYCA EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
LGB2_LUPLU VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV
HBB_HUMAN LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
HBB_HORSE LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
HBA_HUMAN LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
HBA_HORSE LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
GLB5_PETMA LAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
MYG_PHYCA ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LGB2_LUPLU VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---


ClustalX 2.0.9
Mode:
Multiple Alignment Mode
Font:
10
HBB_HUMAN
HBB_HORSE
HBA_HUMAN
HBA_HORSE
GLB5_PETMA
MYG_PHYCA
LGB2_LUPLU

Multiple Sequence Alignment

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPSD--IAVEWESNG--
```

The sole purpose of multiple sequence alignments is to place *homologous positions* of *homologous sequences* into the *same column.*

Clustal

- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994)
  - CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice.
    - Nucleic Acids Research, 22:4673-4680.

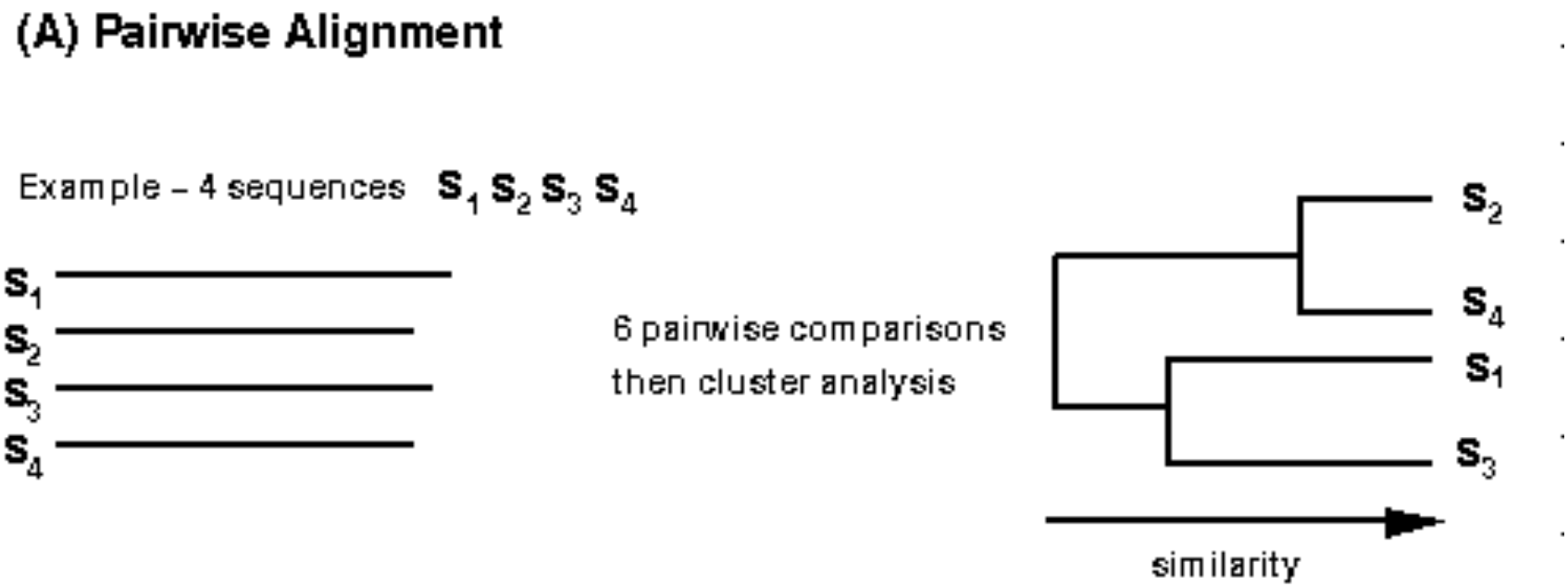
Differences between CLUSTAL and BLAST?

CLUSTAL	BLAST
• global alignment method	• local alignment method
- Align complete sequence	- Search for HSP
• Assumes homology	• Test for homology
• Complex gap penalties	• Simple gap penalties
• Slower	• Fast
• Align protein-protein or nucleotide-nucleotide only	• Translated searches

CLUSTAL Algorithm Steps

1. Pairwise alignment of each sequence pair
   - Number of comparisons depends on how many sequences
2. Compute distance matrix
   - Percent non-identity between each alignment pair
   - Lower distance means more similar
3. Construct a sequence similarity tree
   - Cluster sequences according to distance (similarity)
4. Progressive alignment of sequences according to a tree

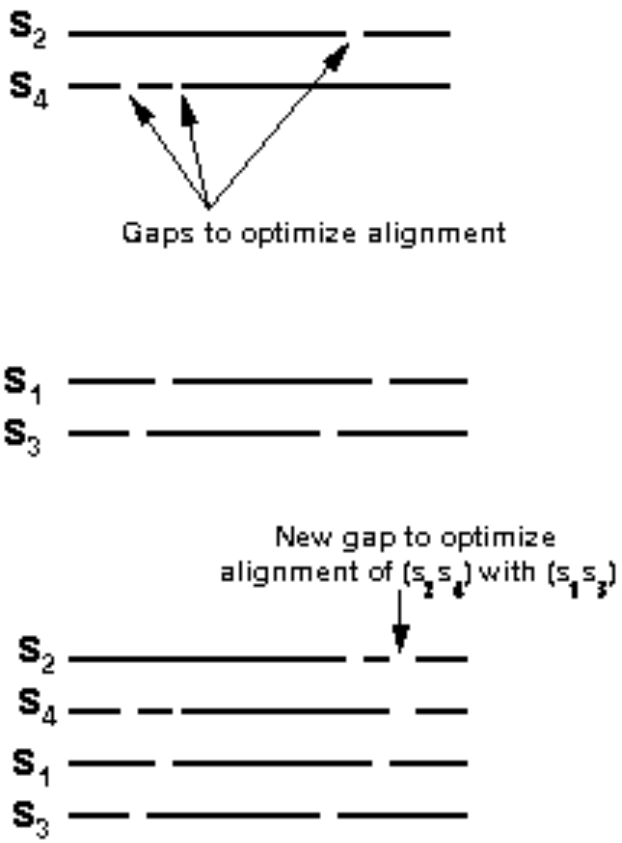
How does the Clustal algorithm actually work?



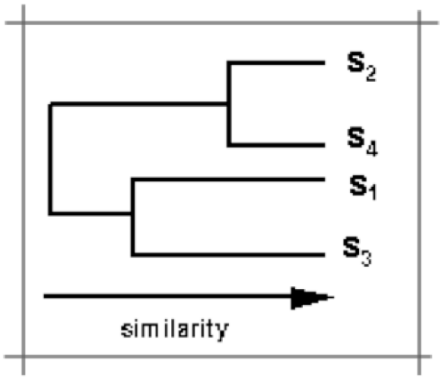
Which sequences would be aligned first?

Steps in a Multiple Sequence Alignment continued …

(B) Multiple alignment following the tree from A



align most similar pair



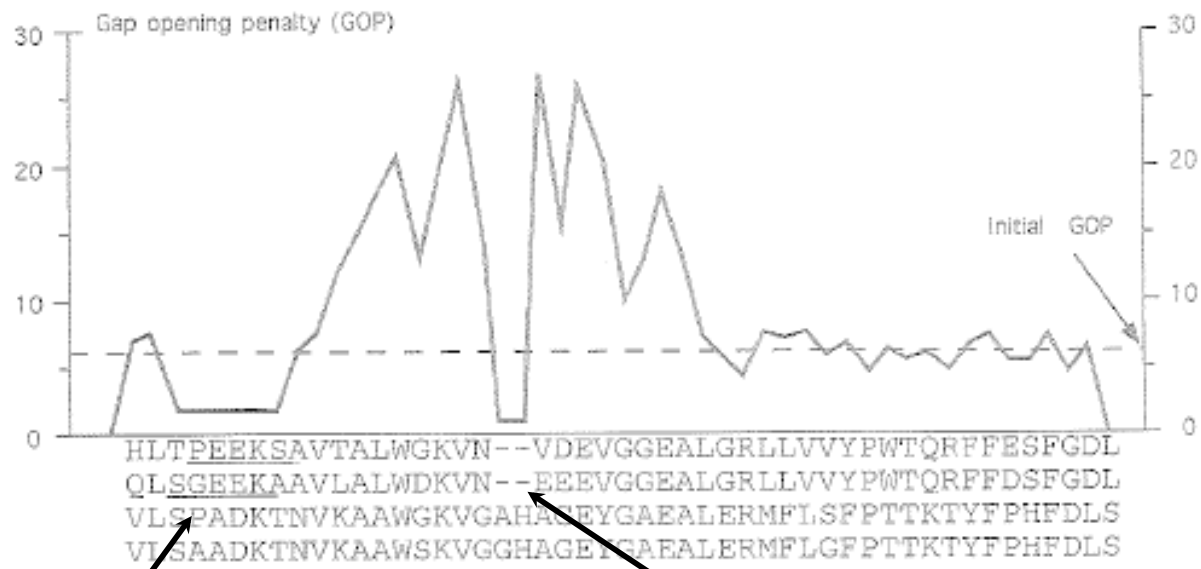
align next most similar pair

align alignments – preserve gaps

Position Specific Gap Penalities

- There are two type of gap opening penalities: gap opening and gap extension
  - Determined empirically by user
- Decrease penalties where gaps already occurs
- Increase penalties in adjacent positions to where gap already occurs
  - Encourage extension of gaps in loop regions vs. introduction of new gaps
- Increase or decrease gap penalties according to amino acid type
  - Increase penalties in stretches of hydrophobic residues
  - Discourage the disruption of secondary structure elements

Gap Penalties Example



hydrophilic stretch

existing gap

Figure from Higgens et al, Methods in Enzymology 266: 383

Standard Multiple Sequence Alignment Approach

- Be as sure as possible that the sequences included are homologous
- Know as much as possible about the gene/ protein in question before trying to create an alignment (secondary structure, domains etc..)
- Start with an automated alignment: preferably one that utilizes some evolutionary theory such as CLUSTAL

http://www.ebi.ac.uk/Tools/clustalw2/index.html

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- ClustalW2 Help
- ClustalW2 FAQ
- Jalview Help
- Scores Table
- Alignment
- Guide Tree
- Colours

- Similar Applications
  - Align
  - Kalign
  - MAFFT
  - MUSCLE
  - T-Coffee

- ClustalW Programmatic Access

- www.clustal.org

Clustal Related Literature

Search for Clustal related literature in Medline...

EBI > Tools > Sequence Analysis > ClustalW2

ClustalW2

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

New users, please read the FAQ.

>> Download Software

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	
	Sequence	interactive	full	
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	NO END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	yes	def	def
	ITERATION		NUMITER	
	none		1	

OUTPUT		PHYLOGENETIC TREE			
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS	CLUSTERING
aln w/numbers	aligned	none	off	off	NJ

Enter or paste a set of sequences in any supported format: Help

http://www.ebi.ac.uk/Tools/muscle/index.html

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- Muscle Help
- Jalview Help

- Similar Applications
  - Align
  - ClustalW2
  - Kalign
  - MAFFT
  - T-Coffee

- Muscle Programmatic Access

EBI > Tools > Sequence Analysis

MUSCLE

MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

Download Software

MUSCLE multiple sequence alignment method with reduced time and space complexity

RESULTS	SEARCH TITLE	YOUR EMAIL
interactive	Sequence	
OUTPUT FORMAT	**OUTPUT TREE**	**OUTPUT ORDER**
FASTA	none	aligned

Enter or Paste a set of Sequences in any supported format: Help

Upload a file: Choose File no file selected

Run | Reset

If you plan to use these services during a course please contact us.

http://www.ebi.ac.uk/Tools/t-coffee/index.html

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- TCoffee Help
- Jalview Help
- Alignment
- Guide Tree
- Colours

- Similar Applications
  - Align
  - ClustalW2
  - Kalign
  - MAFFT
  - MUSCLE

- T-Coffee Programmatic Access

T-Coffee Related Literature

Search for T-Coffee related literature in Medline...
more

EBI > Tools > Sequence Analysis

T-Coffee

T-Coffee is a multiple sequence alignment program. Multiple sequence alignment programs are meant to align a set of sequences previously gathered using other programs such as blast, fast, sw ...
The main characteristic of T-Coffee is that it will allow you to combine results obtained with several alignment methods. For instance if you have an alignment coming from ClustalW2, an other alignment coming from Dialign, and a structural alignment of some of your sequences, T-Coffee will combine all that information and produce a new multiple sequence having the best agreement whith all these methods.
By default, T-Coffee will compare all you sequences two by two, producing a global alignment and a series of local alignments (using lalign). The program will then combine all these alignments into a multiple alignment.

TCoffee Multiple Sequence Alignment

Download Software

EMAIL	RESULTS	RUN NAME	MATRIX	ORDER
	interactive	Sequence	none	aligned

Enter or Paste a set of Sequences in any supported format: Help

Upload a file: Choose File no file selected Run Reset

Standard Multiple Sequence Alignment Approach

Examine alignment:

- Are you confident that aligned residues/bases evolved from a common ancestor?
- Are domains of the proteins/predicted secondary structures, etc. aligning correctly?
- Are most indels outside of known motifs or secondary structure?
- → No? May need to edit sequences and redo…

The Take Home Message

Why perform an MSA?

- Visualize trends between homologous sequences
  - Shared regions of homology
  - Regions unique to a sequence within a family
  - Consensus sequence
- As the first step in a phylogenetic analysis

The Take Home Message

How does one perform an MSA?

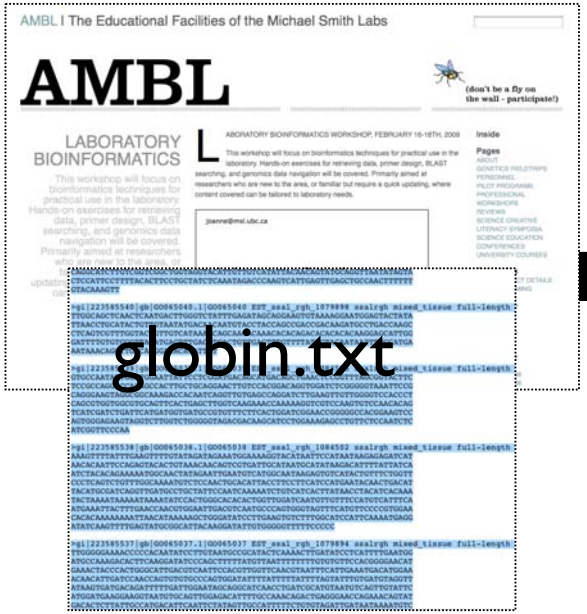
- By hand: too hard!
- Automated alignment: Fast, but doesn't necessarily produce the "correct" alignment

Best approach = Automated alignment with manual editing

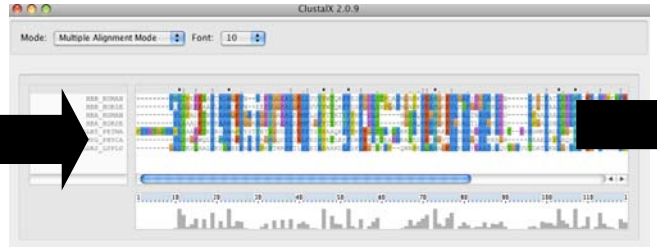
MSA

PRACTICAL EXERCISE: Comparing Sets of Protein Sequences

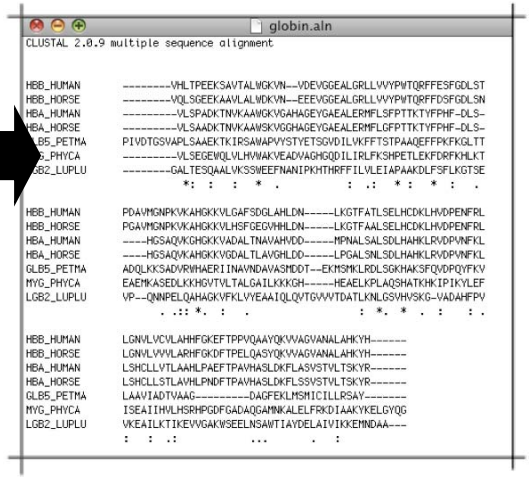
navigate to:
bioteach.ubc.ca/bioinfo2009



Clustal



We'll walk through install + do MSA #1 together



Install ClustalX on laptop

download program and install

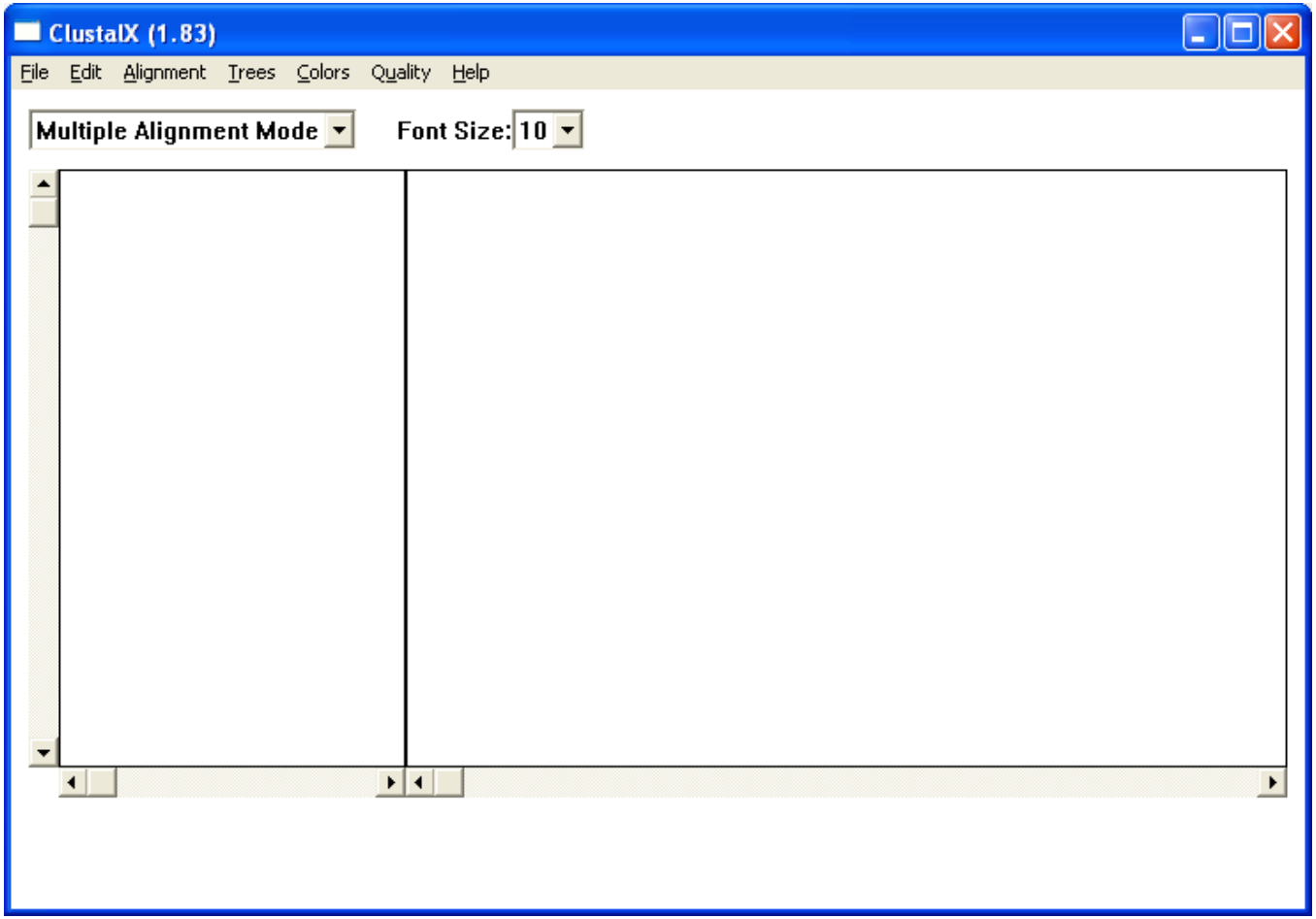
Use ClustalX to generate MSA

MSA #1: Use example sequences to generate alignment

MSA #2: Use your own sequences

Clustalx.exe

Open ClustalX



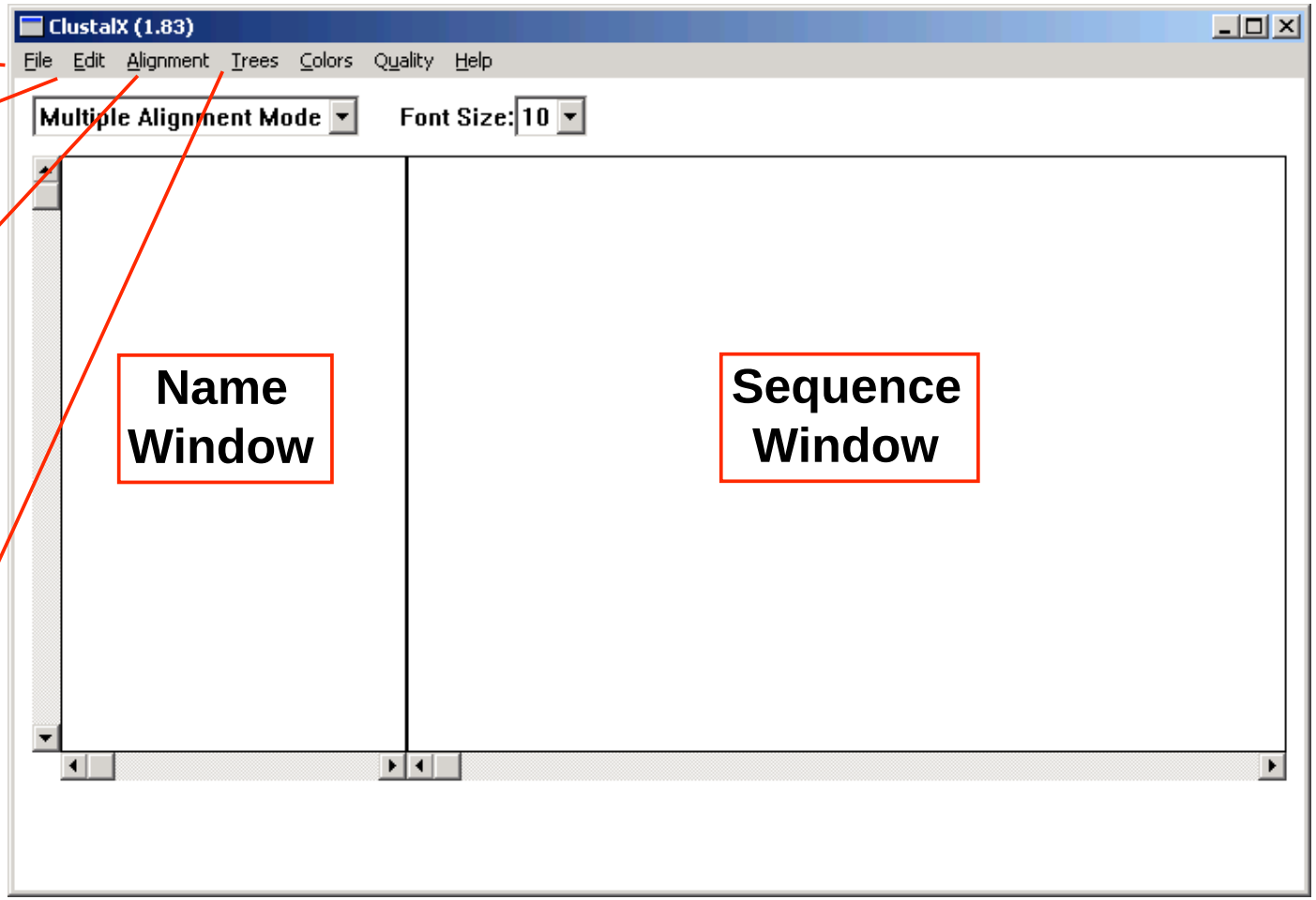
Starting up ClustalX

File:
-Load sequences

Edit:
-Remove all gaps

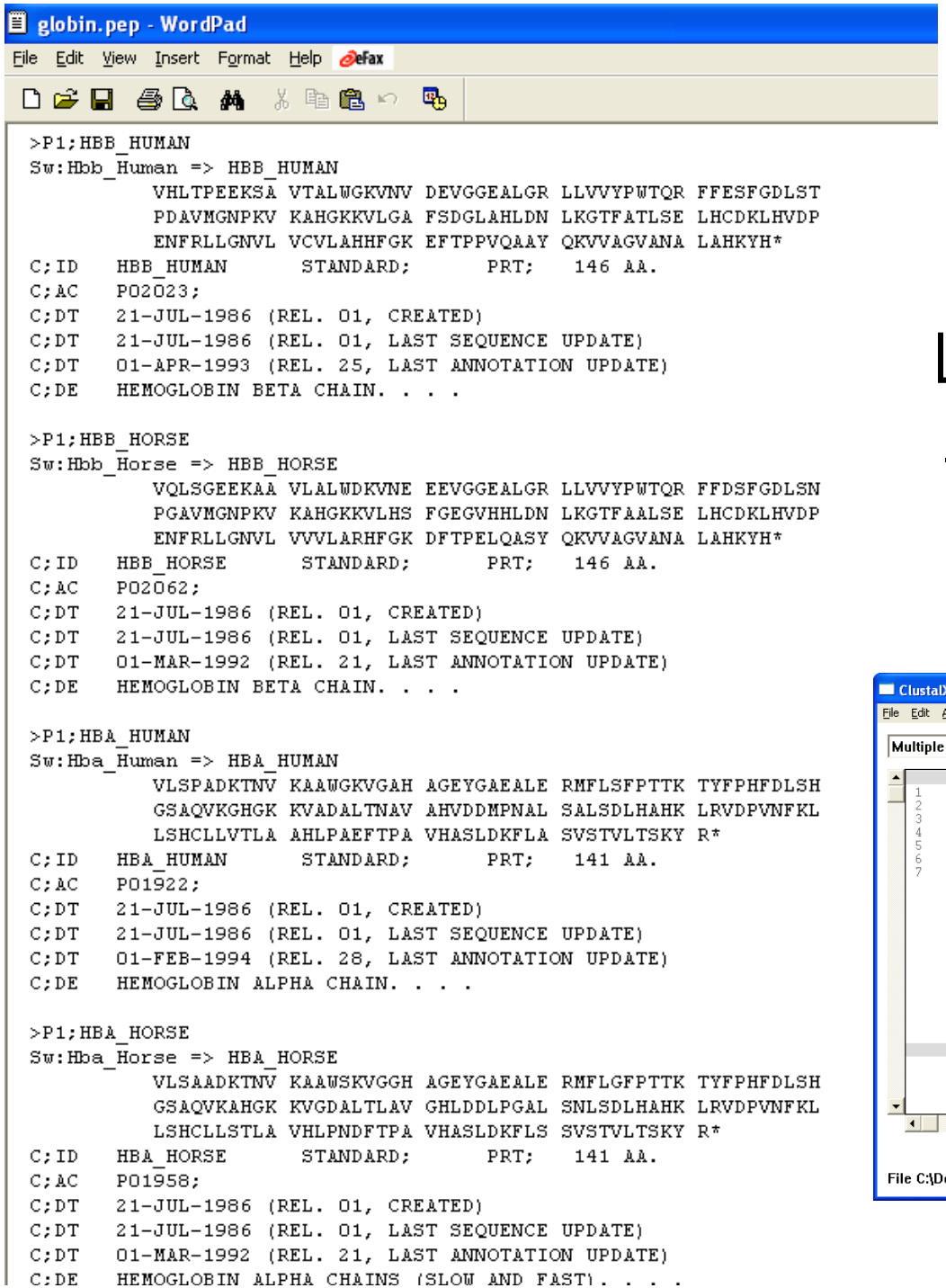
Alignment:
-Do complete alignment
-Alignment parameters

Trees:
-Bootstrapped NJ
-Output format options



Load the sequences

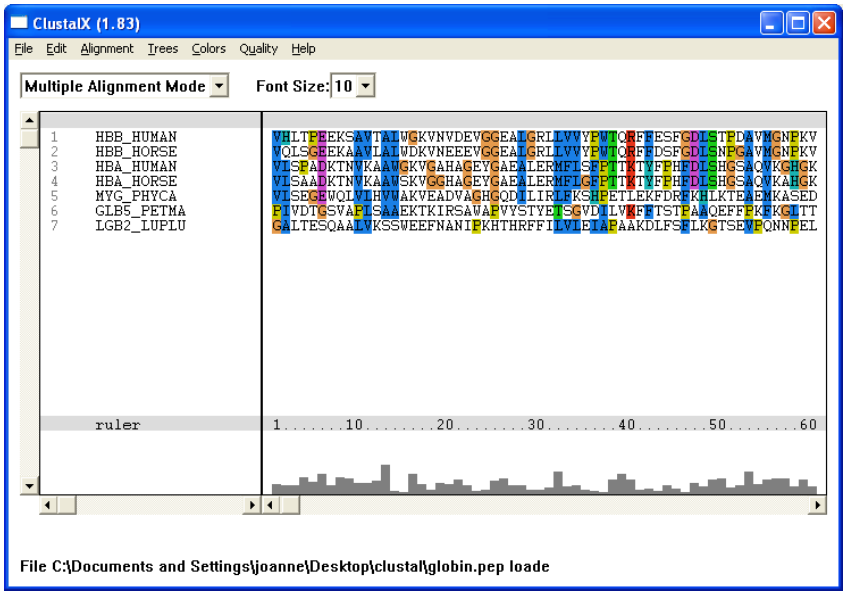
–globin.pep



globin.pep - WordPad

File Edit View Insert Format Help eFax

```
>P1;HBB_HUMAN
Sw:Hbb_Human => HBB_HUMAN
          VHLTPEEKSA VTALWGKVNV DEVGGEALGR LLVVYPWTQR FFESFGDLST
          PDAVMGNPKV KAHGKKVLGA FSDGLAHLDN LKGTFATLSE LHCDKLHVDP
          ENFRLLGNVL VCVLAHHFGK EFTPPVQAAY QKVVAGVANA LAHKYH*
C;ID   HBB_HUMAN      STANDARD;      PRT;   146 AA.
C;AC   P02023;
C;DT   21-JUL-1986 (REL. 01, CREATED)
C;DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
C;DT   01-APR-1993 (REL. 25, LAST ANNOTATION UPDATE)
C;DE   HEMOGLOBIN BETA CHAIN. . . .

>P1;HBB_HORSE
Sw:Hbb_Horse => HBB_HORSE
          VQLSGEEKAA VLALWDKVNE EEVGGEALGR LLVVYPWTQR FFDSFGDLSN
          PGAVMGNPKV KAHGKKVLHS FGEGVHHLDN LKGTFAALSE LHCDKLHVDP
          ENFRLLGNVL VVVLARHFGK DFTPELQASY QKVVAGVANA LAHKYH*
C;ID   HBB_HORSE      STANDARD;      PRT;   146 AA.
C;AC   P02062;
C;DT   21-JUL-1986 (REL. 01, CREATED)
C;DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
C;DT   01-MAR-1992 (REL. 21, LAST ANNOTATION UPDATE)
C;DE   HEMOGLOBIN BETA CHAIN. . . .

>P1;HBA_HUMAN
Sw:Hba_Human => HBA_HUMAN
          VLSPADKTNV KAAWGKVGAH AGEYGAEALE RMFLSFPTTK TYFPHFDLSH
          GSAQVKGHGK KVADALTNAV AHVDDMPNAL SALSDLHAHK LRVDPVNFKL
          LSHCLLVTLA AHLPAEFTPA VHASLDKFLA SVSTVLTSKY R*
C;ID   HBA_HUMAN      STANDARD;      PRT;   141 AA.
C;AC   P01922;
C;DT   21-JUL-1986 (REL. 01, CREATED)
C;DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
C;DT   01-FEB-1994 (REL. 28, LAST ANNOTATION UPDATE)
C;DE   HEMOGLOBIN ALPHA CHAIN. . . .

>P1;HBA_HORSE
Sw:Hba_Horse => HBA_HORSE
          VLSAADKTNV KAAWSKVGGH AGEYGAEALE RMFLGFPTTK TYFPHFDLSH
          GSAQVKAHGK KVGDALTLAV GHLDDLPGAL SNLSDLHAHK LRVDPVNFKL
          LSHCLLSTLA VHLPNDFTPA VHASLDKFLS SVSTVLTSKY R*
C;ID   HBA_HORSE      STANDARD;      PRT;   141 AA.
C;AC   P01958;
C;DT   21-JUL-1986 (REL. 01, CREATED)
C;DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
C;DT   01-MAR-1992 (REL. 21, LAST ANNOTATION UPDATE)
C;DE   HEMOGLOBIN ALPHA CHAINS (SLOW AND FAST). . . .
```



ClustalX (1.83)

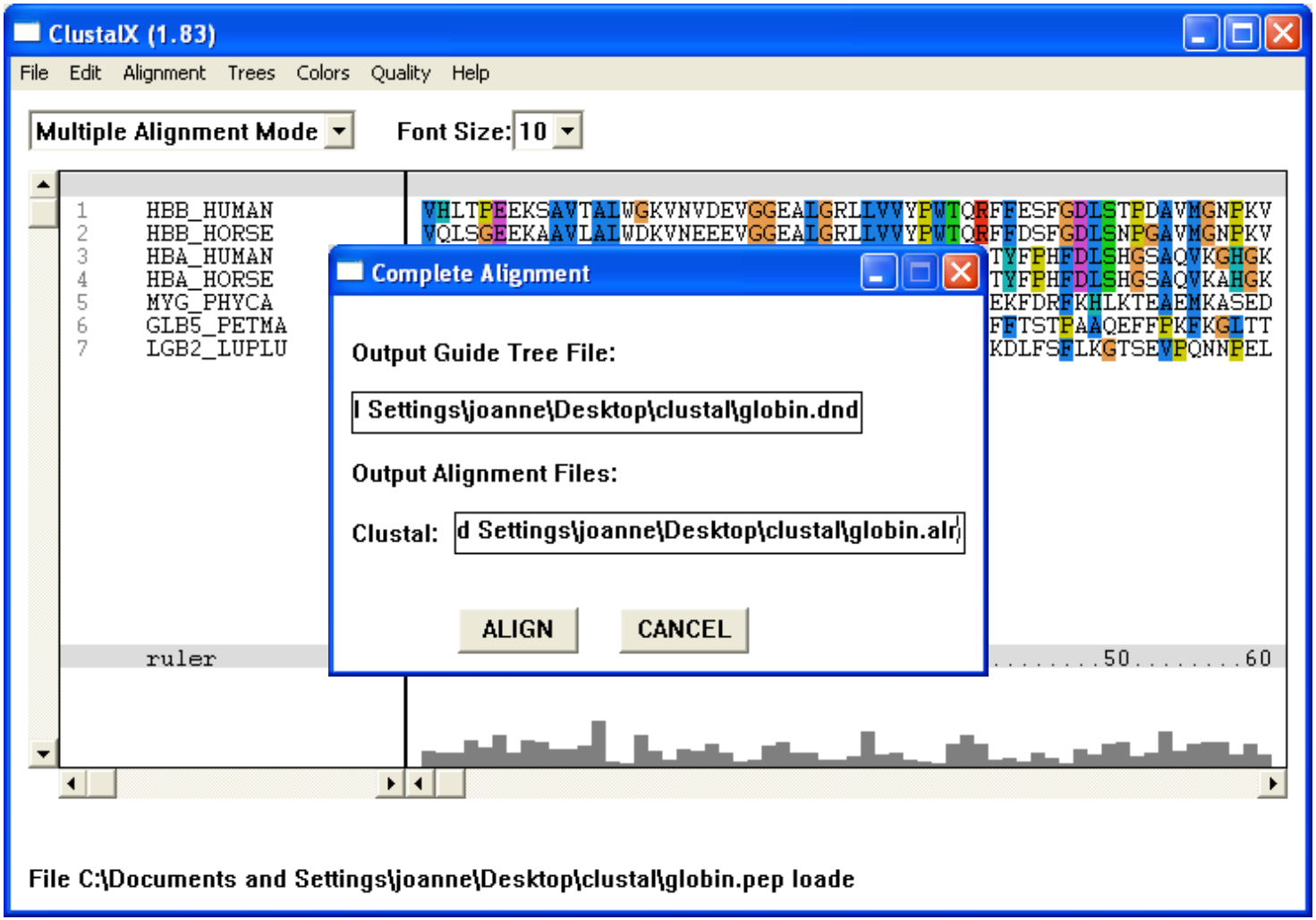
File Edit Alignment Trees Colors Quality Help

Multiple Alignment Mode Font Size: 10

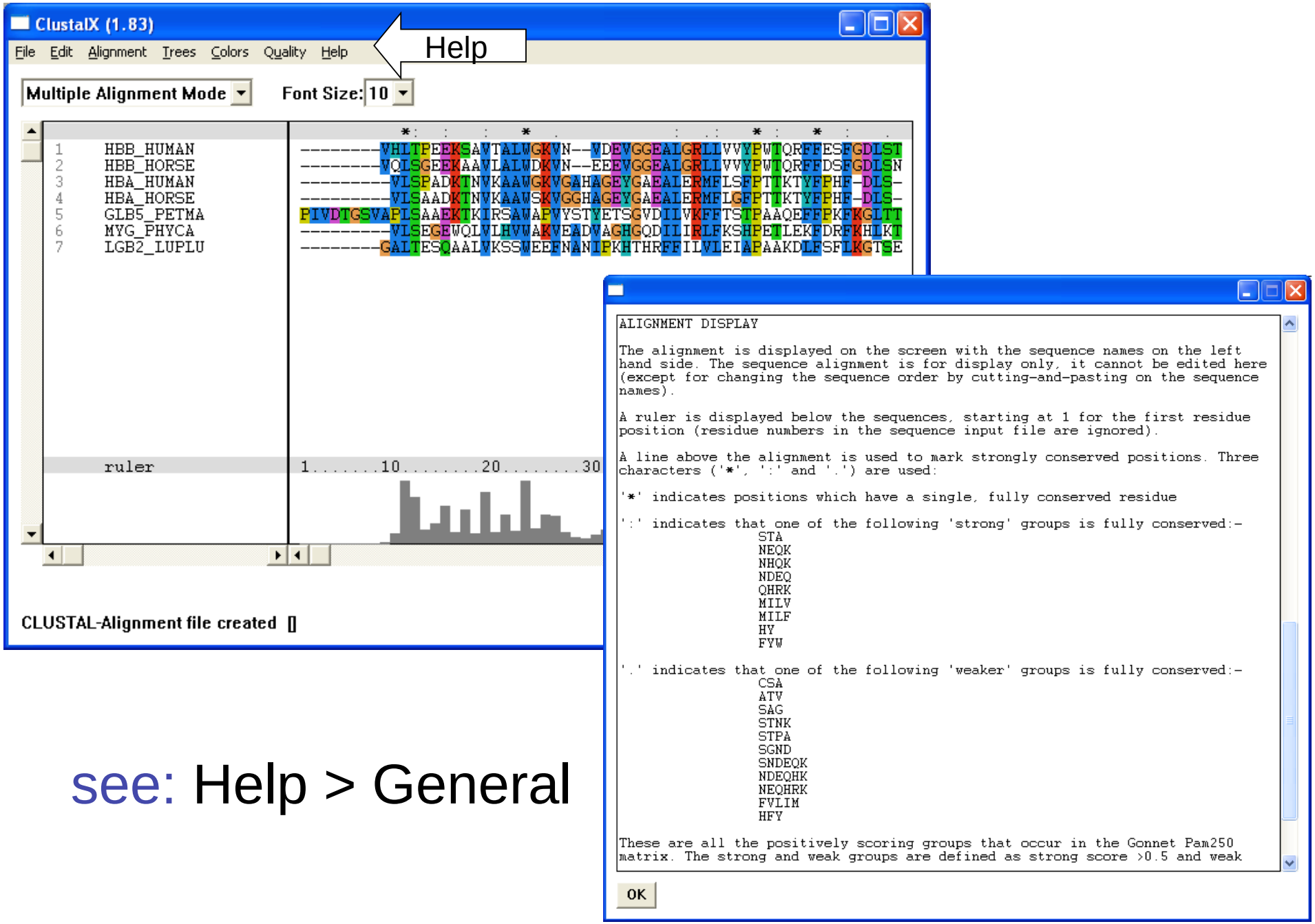
```
1   HBB_HUMAN   VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
2   HBB_HORSE   VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPKV
3   HBA_HUMAN   VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
4   HBA_HORSE   VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHGK
5   MYG_PHYCA   VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED
6   GLB5_PETMA  PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGITT
7   LGB2_LUPLU  GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPEL

    ruler       1.......10........20........30........40........50........60
```

File C:\Documents and Settings\joanne\Desktop\clustal\globin.pep loade

Alignment > Do Complete Alignment

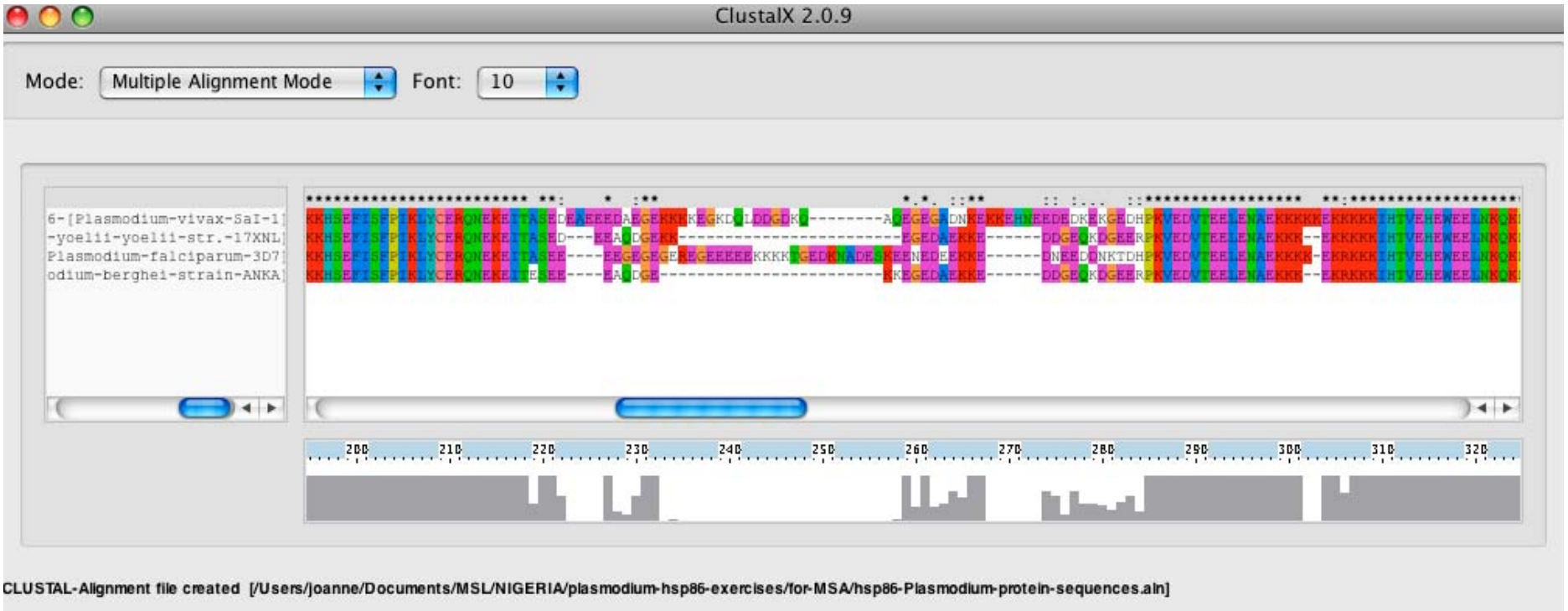


also see: Alignment > Alignment Parameters



see: Help > General

Can you create a MSA for the Plasmodium hsp86 protein sequences?



joanne@msl.ubc.ca

Bioinformatics

Common Tools & Tricks of the Trade

bioteach.ubc.ca/bioinfo2009

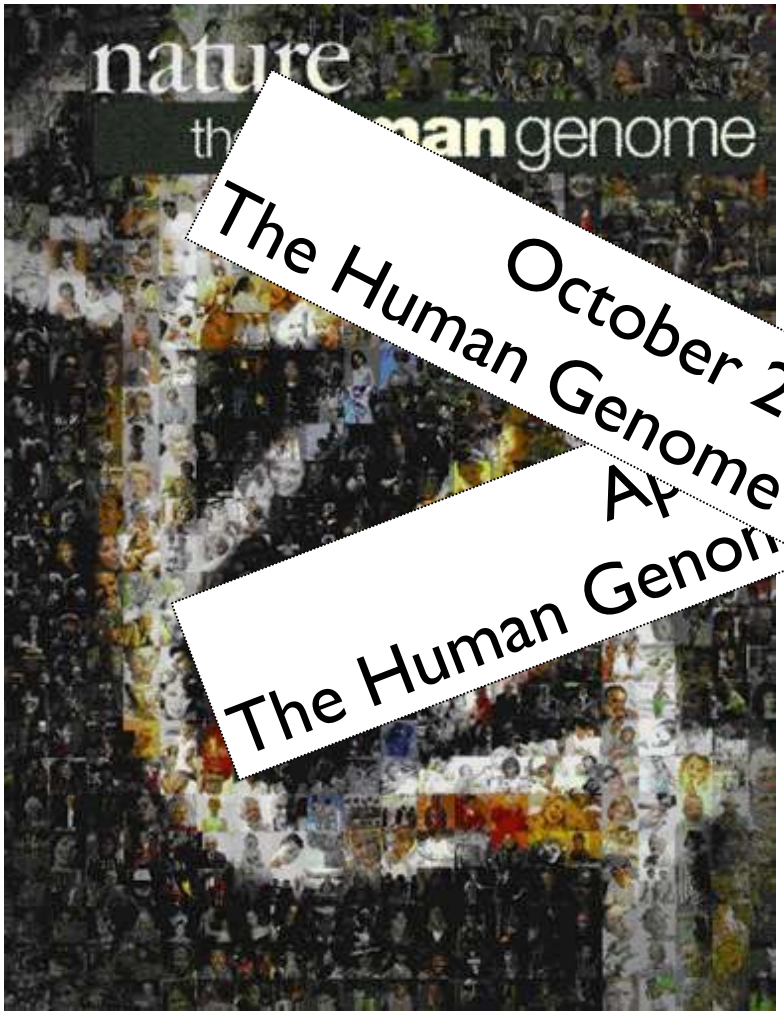
Module 3 Topics

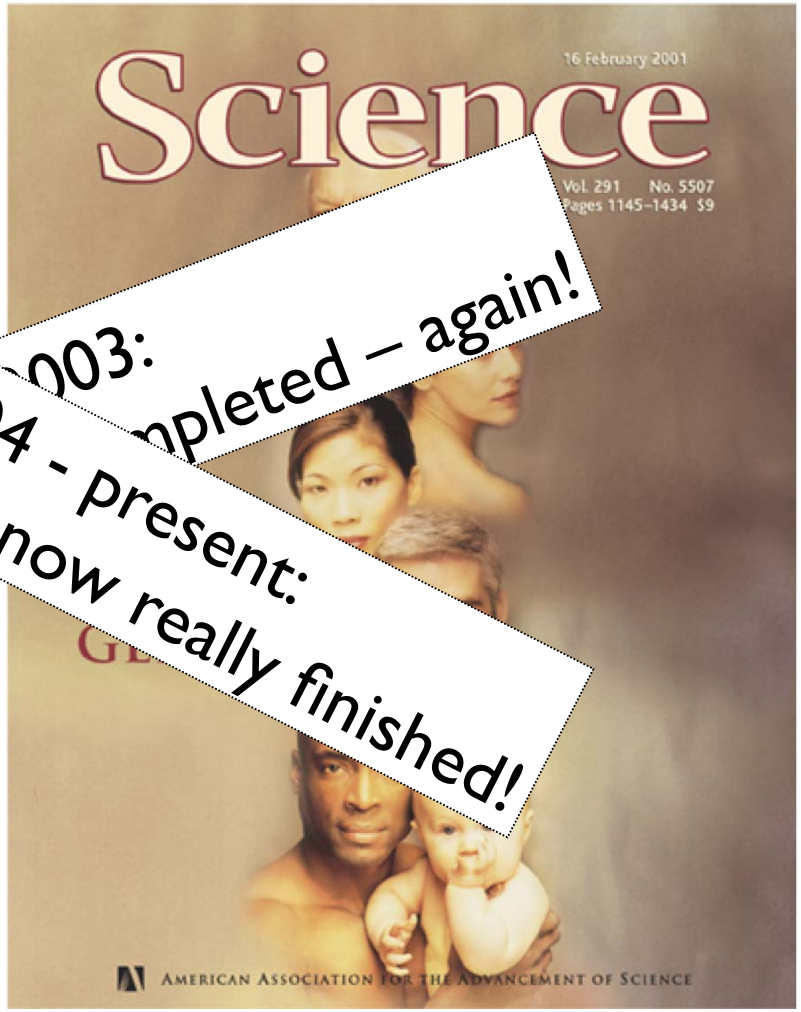
- **Genome Browsers**, Accessing Genome Annotations.
- **PRACTICAL EXERCISES**, three different views of the BRCA1 gene
- **Discovering GEO**, the Gene Expression Omnibus.
- **Pathway Resources** for Systems Biology
- **Bioinformatics Links Directory**, Conducting Research on the Web

Genome Browsers

Accessing Genome Annotations &
PRACTICAL EXERCISE: Three Different
Views of the BRCA1 Gene

The Human Genome Project





October 2004 - present:
The Human Genome is now really finished!

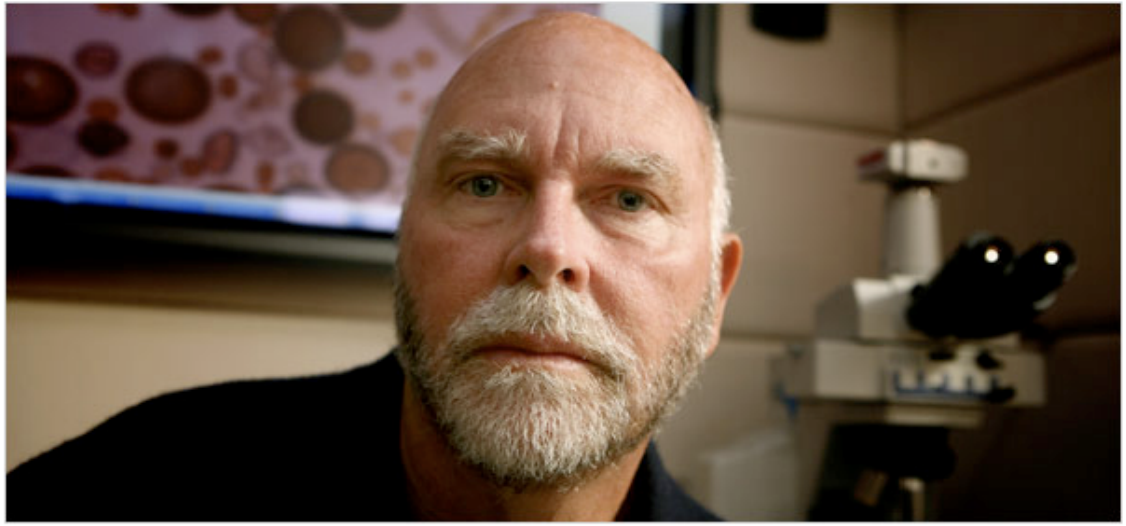
2003:
The Human Genome completed – again!

Public HGP

Celera Genomics

February 2001: Completion of the Draft Human Genome

In the Genome Race, the Sequel Is Personal



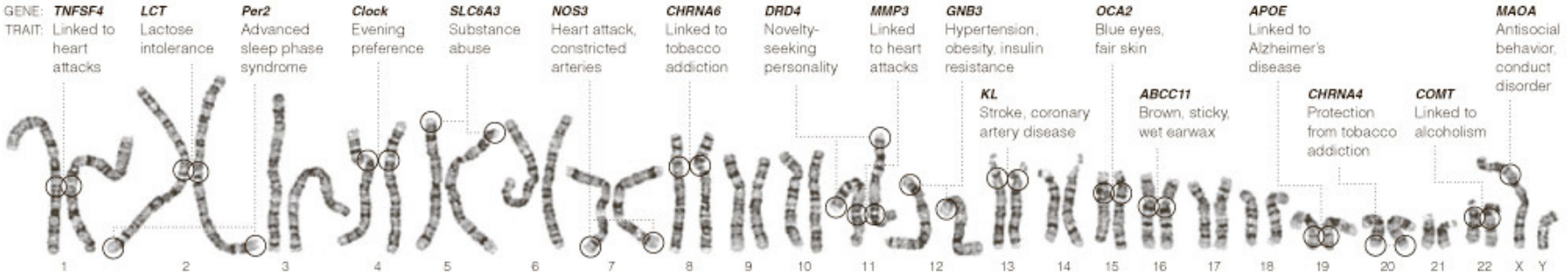
Thor Swift for The New York Times

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

The New York Times

September 3, 2007

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

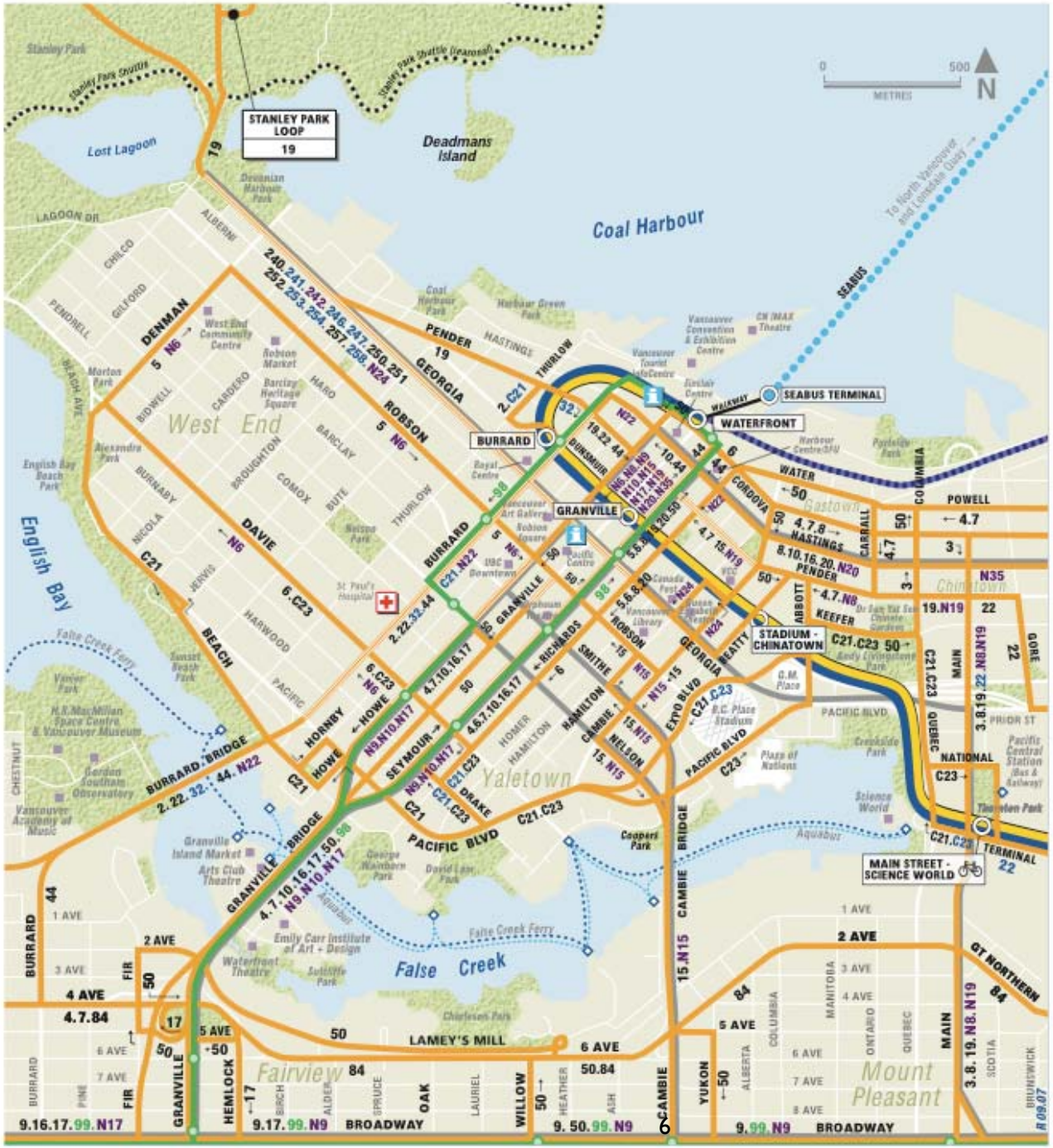


Sources: PLoS Biology; GeneCards

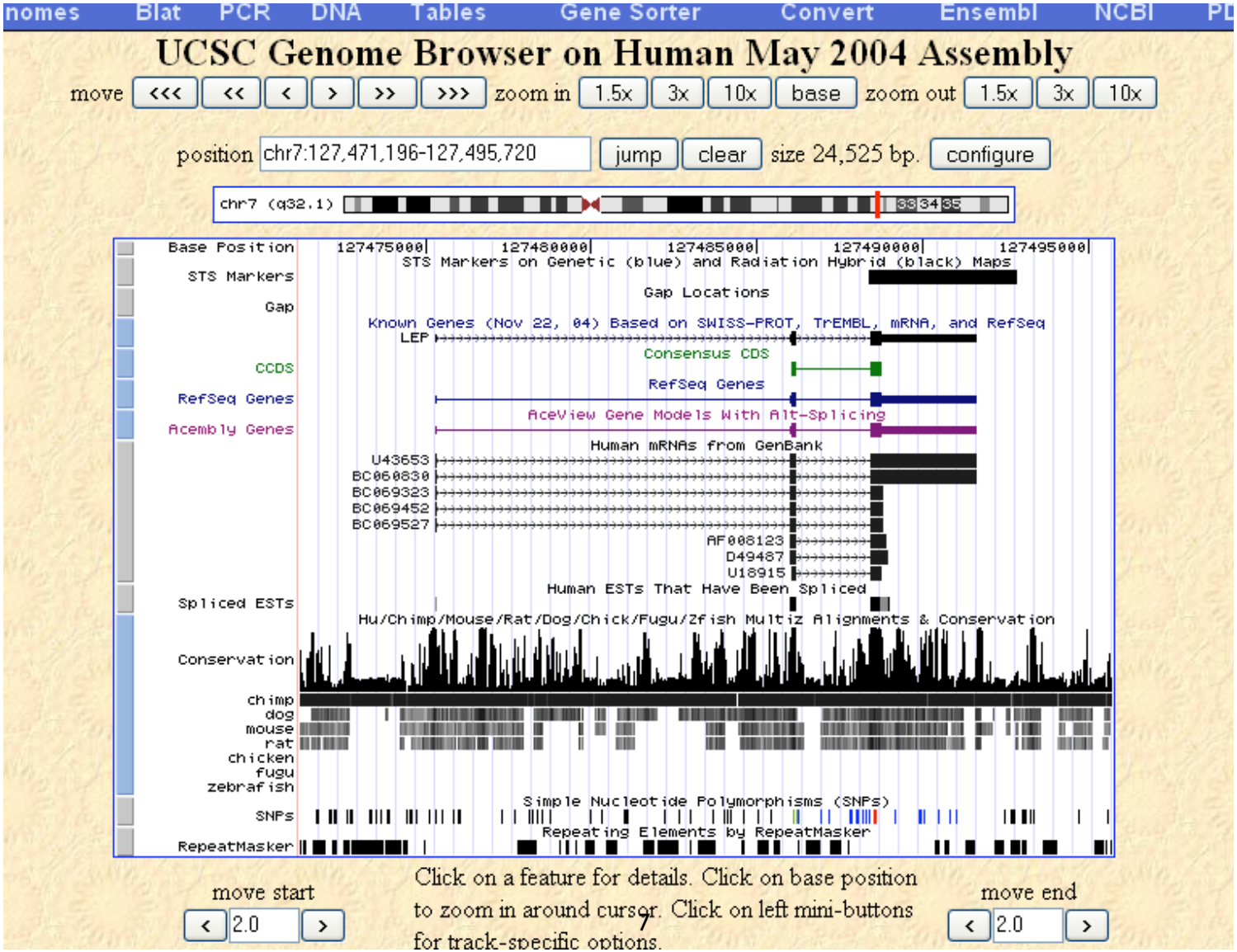
CHROMOSOMES

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

THE NEW YORK TIMES


STANLEY PARK LOOP
19
Lost Lagoon
Deadmans Island
Coal Harbour
West End
English Bay
False Creek
Yaletown
Fairview
Mount Pleasant
Gastown
Chinatown
SEABUS TERMINAL
WATERFRONT
BURRARD
GRANVILLE
STADIUM - CHINATOWN
MAIN STREET - SCIENCE WORLD
BROADWAY
LAMEY'S MILL
PACIFIC BLVD
BURRARD BRIDGE
GRANVILLE BRIDGE
CAMBIE BRIDGE
GEORGIA
DENMAN
ROBSON
PENDER
HASTINGS
DAVIE
BEACH
GRANVILLE
SEYMOUR
HOWE
HORNBY
BURRARD
MAIN
QUEBEC
CAMBIE
YUKON
HEMLOCK
OAK
WILLOW
GT NORTHERN
NATIONAL
POWELL
COLUMBIA
CARRALL
GORE
SEABUS
METRES
0
500
N

Let's Look at the Human Genome...



Objectives

- By the end of this module:

✓ You will be able to describe the following concepts: genome annotation, genome builds, and genome browsers.

✓ You will view the genomic location that contains the BRCA1 gene in the human genome using three different genome browsers.

✓ You will be able to compare and contrast the UCSC, Ensembl and MapViewer systems for visualizing genome information.

Genome Browsers

- What is a Genome Browser?
  - System for displaying, viewing, and accessing genome annotation data
- Genome annotations = knowledge attached to raw genome sequence.
  - Annotation information comes from many different sources
    - ✓ Computational pipelines
    - ✓ Research groups
    - ✓ Databases

Three different flavors of Genome Browsers:

- UCSC Genome Browser

  http://genome.cse.ucsc.edu/

- Ensembl

  http://www.ensembl.org/

- NCBI Map Viewer

  http://www.ncbi.nlm.nih.gov/mapview/

The underlying data is common for all three "flavors" of Genome Browsers.

- NCBI, UCSC and Ensembl use the same human genome assembly that is generated by NCBI
  - release timing is different between sites.
- Note the version of genome assembly to which you are referring
  - available precomputed info and locations of features will be different between different assemblies.

Let's compare the view of the BRCA1 gene in all three genome browsers.

Viewing the genomic region containing BRCA1

- Common features:

✓ Coordinate system is based on the build

✓ Zoom in and out

✓ Annotations displayed – ie. Gene features

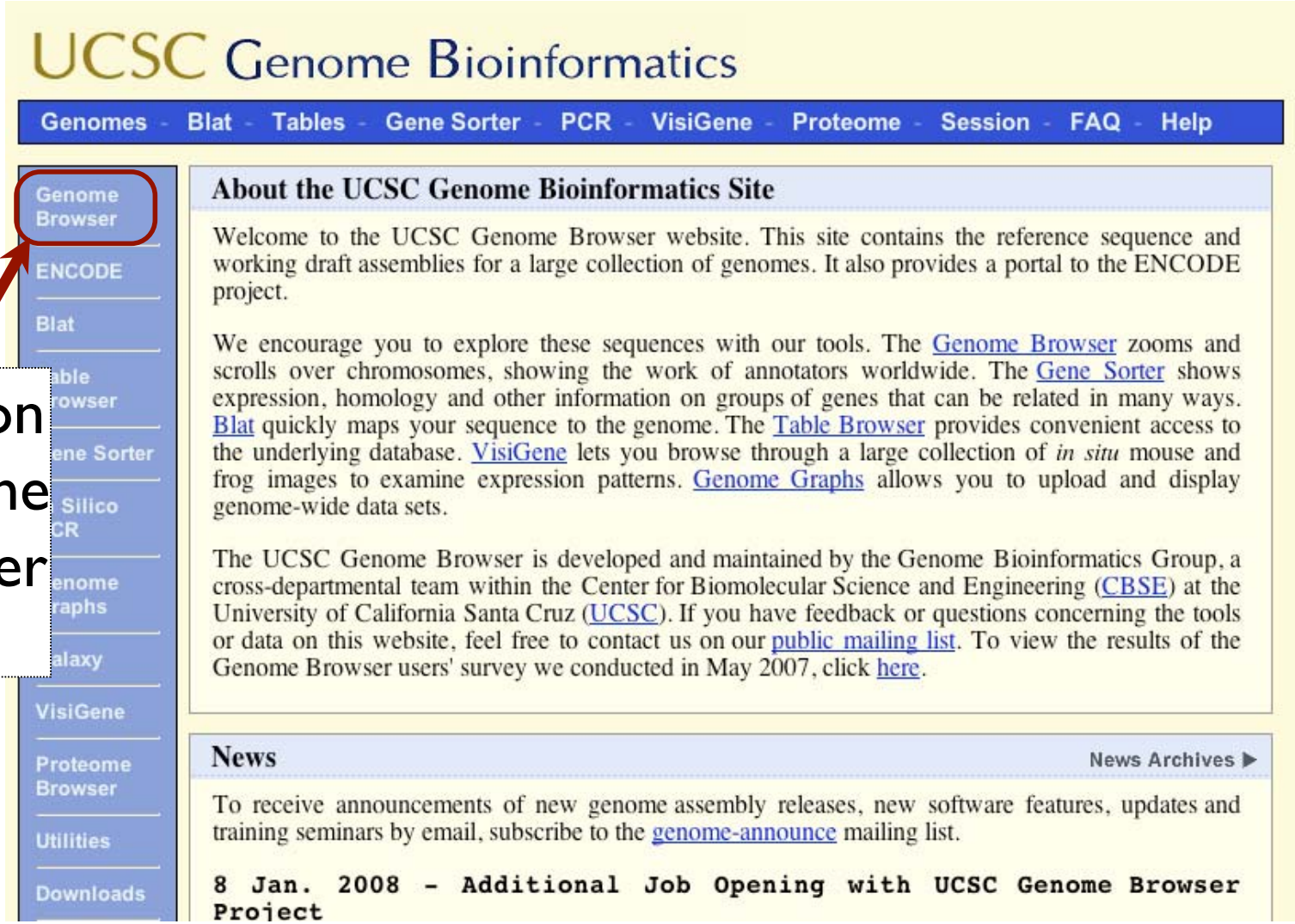
- Major Differences:

✓ Each Browser has a very different look and feel

✓ Annotation information displayed differently

✓ Different ways to navigate through the information

http://genome.cse.ucsc.edu/



Click on Genome Browser link

Home Genomes Blat Tables Gene Sorter PCR FAQ Help

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position or search term	image width	
Vertebrate	Human	May 2004	BRCA1	620	submit

Click here to reset the browser user interface settings to their defaults.

add your own custom tracks | configure tracks and display | clear position

About the Human May 2004 (hg17) assembly (sequences)

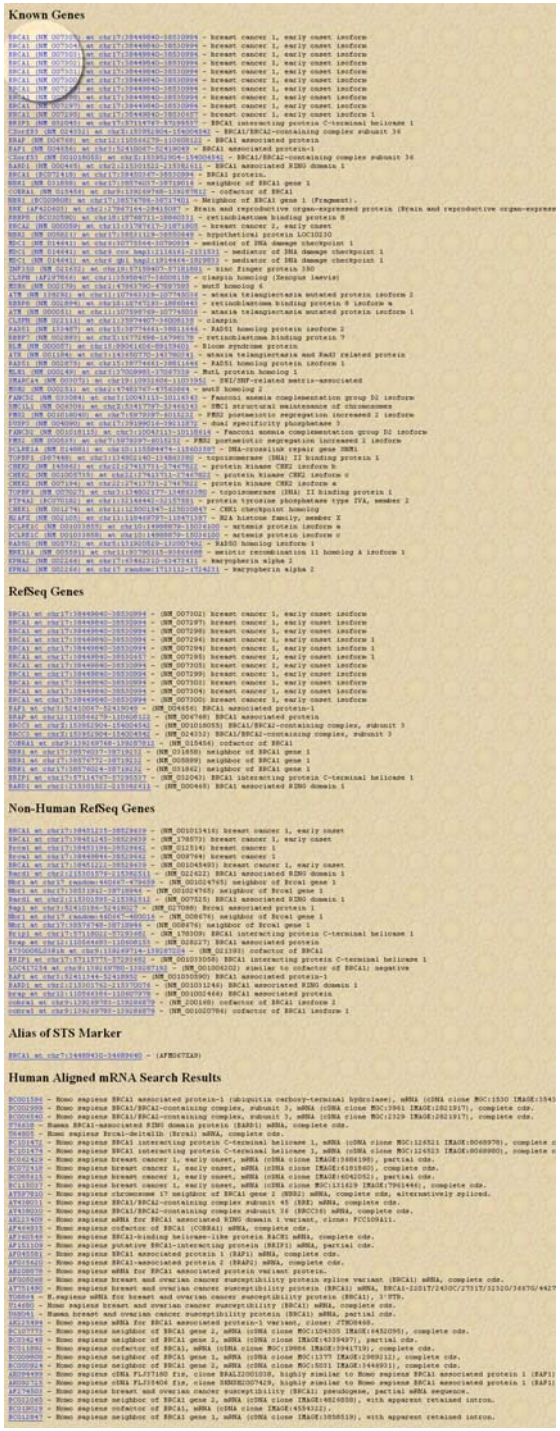
The May 2004 human reference sequence is based on NCBI Build 35 and was produced by the International Human Genome Se

Sample position queries

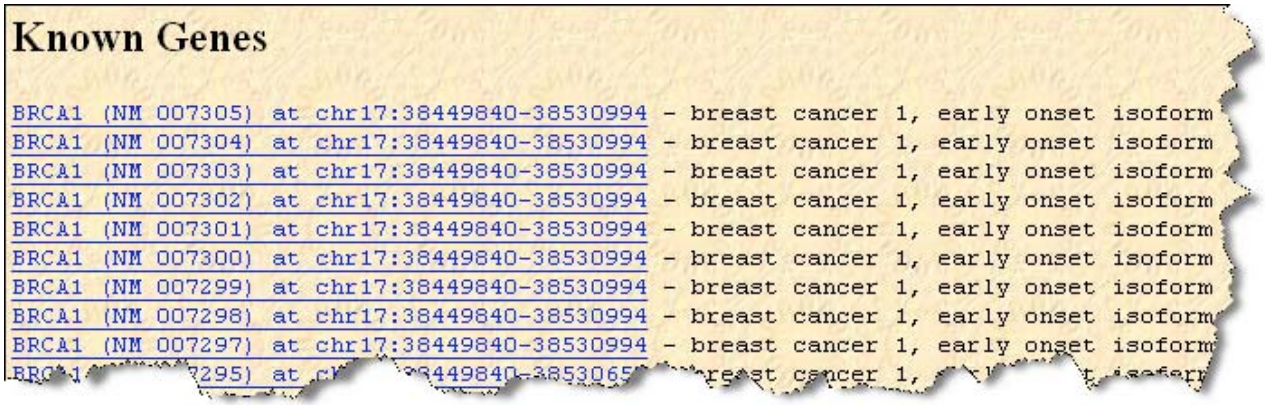
A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS m chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p arm telomere
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps. Includes 100,000 bases on each side as well.
RH18061;RH80175	Displays region between STS markers RH18061;RH80175. Includes 100,000 bases on each side as well.
AA205474	Displays region of EST with GenBank accession AA205474 in BRCA1 cancer gene on chr 17
AC008101	Displays region of clone with GenBank accession AC008101
AF083811	Displays region of mRNA with GenBank accession number AF083811
PRNP	Displays region of genome with HUGO Gene Nomenclature Committee identifier PRNP
NM_017414	Displays the region of genome with RefSeq identifier NM_017414
NP_059110	Displays the region of genome with protein accession number NP_059110
pseudogene mRNA	Lists transcribed pseudogenes, but not cDNAs

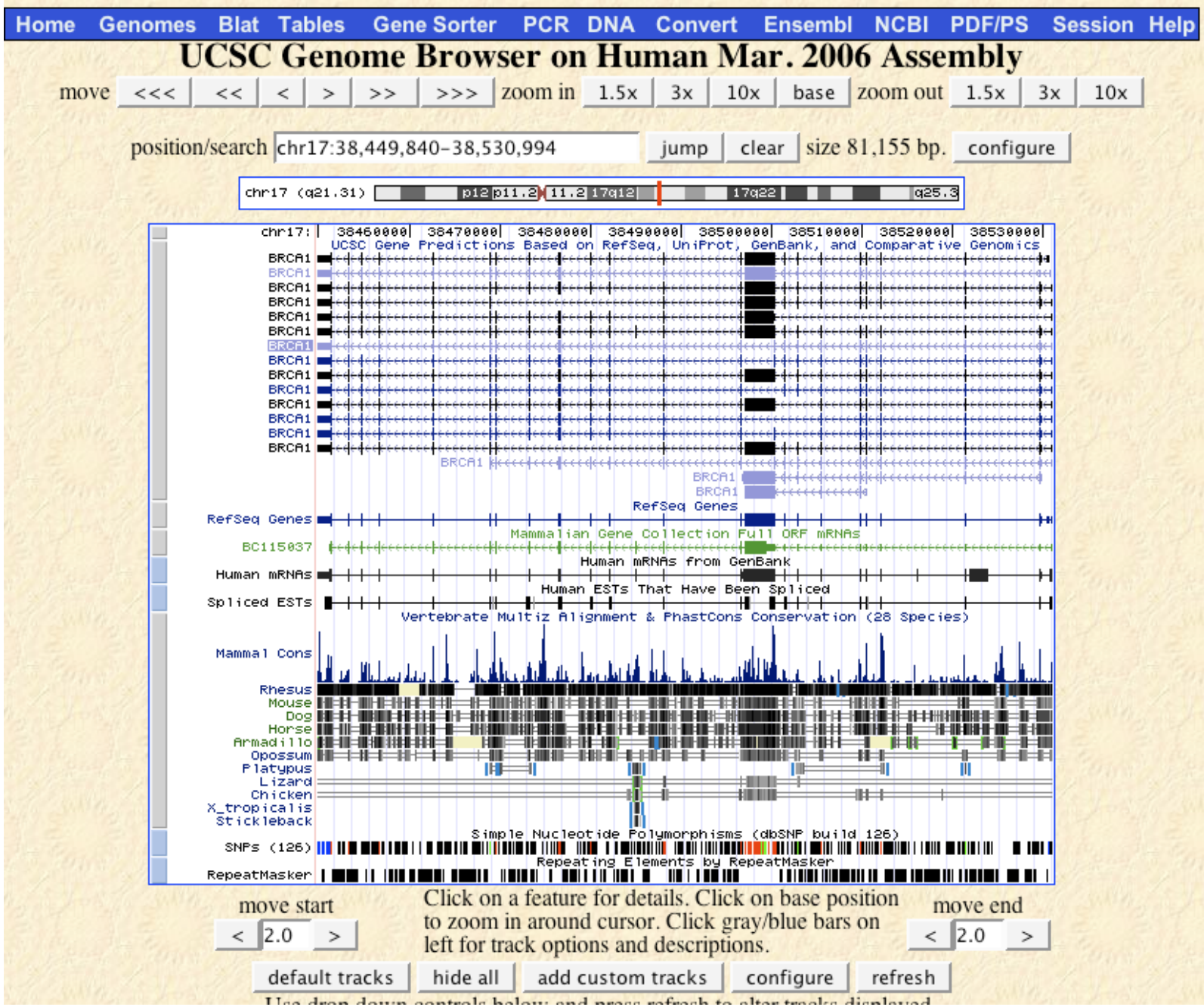
Search for BRCA1; Note sample queries



The Search Results



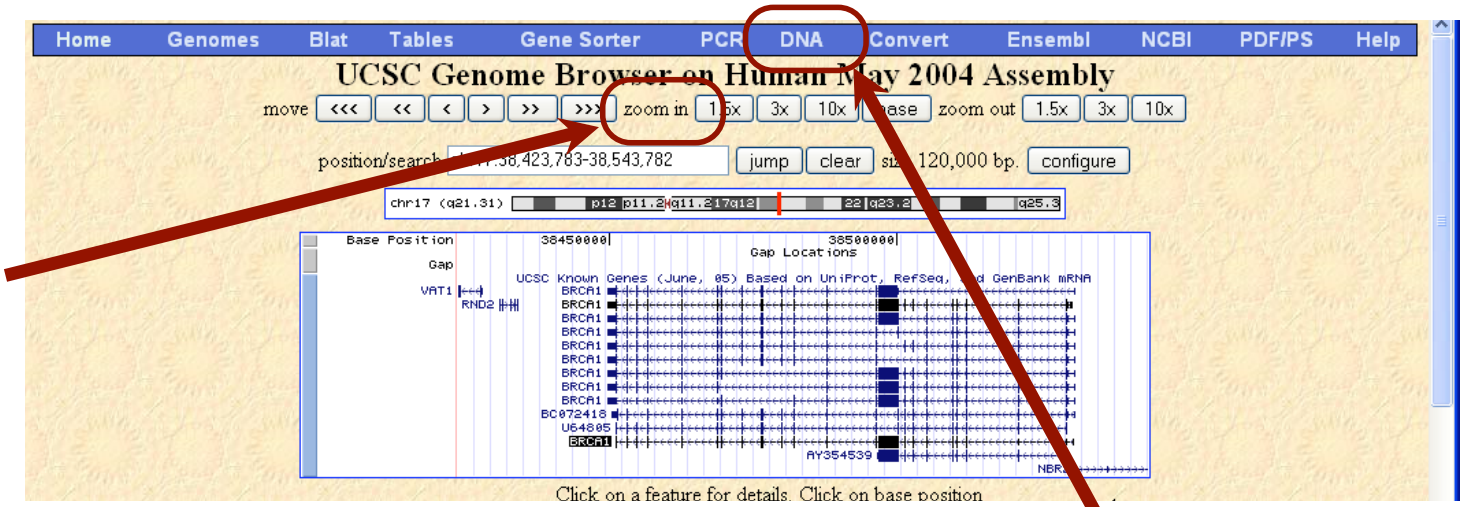
- Many BRCA1 isoforms
  - ✓ All located on chr 17
  - ✓ same chr coordinates
  - ✓ different gene structures


Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help
UCSC Genome Browser on Human Mar. 2006 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr17:38,449,840-38,530,994 jump clear size 81,155 bp. configure
chr17 (q21.31) p12 p11.2 11.2 17q12 17q22 q25.3
chr17: 38460000 38470000 38480000 38490000 38500000 38510000 38520000 38530000
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
BRCA1
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
BC115037
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Rhesus
Mouse
Dog
Horse
Armadillo
Opossum
Platypus
Lizard
Chicken
X_tropicalis
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 126)
SNPs (126)
Repeating Elements by RepeatMasker
RepeatMasker
move start
< 2.0 >
Click on a feature for details. Click on base position to zoom in around cursor. Click gray/blue bars on left for track options and descriptions.
move end
< 2.0 >
default tracks hide all add custom tracks configure refresh

Tasks

- What genes are on either side of BRCA1 on chr 17?
- Can you figure out how to download the genomic sequence for the BRCA1 region?
- Can you figure the display to add/remove tracks that are (or are not) of interest to you?

Zoom in
Zoom out



DNA link
Download
Sequence

collapse all

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

expand all

Mapping and Sequencing Tracks refresh

Base Position: dense	Chromosome Band: hide	STS Markers: hide	FISH Clones: hide	Recomb Rate: hide	Map Contigs: hide
Assembly: hide	Gap: hide	Coverage: hide	BAC End Pairs: hide	Fosmid End Pairs: hide	GC Percent: hide
Short Match: hide	Restr Enzymes: hide				

Phenotype and Disease Associations refresh

Genes and Gene Prediction Tracks refresh

UCSC Genes: pack	Old UCSC Genes: hide	Alt Events: hide	CCDS: hide	RefSeq Genes: dense	Other RefSeq: hide
MGC Genes: pack	ORFeome Clones: hide	TransMap...: hide	Vega Genes: hide	Ensembl Genes: hide	AceView Genes: hide
SIB Genes: hide	N-SCAN: hide	CONTRAST: hide	SGP Genes: hide		
Exoniphy: hide	Augustus: hide	RNA Genes: hide	ACEScan: hide		
Pos Sel Genes: hide					

mRNA and EST Tracks

Expression refresh

Regulation refresh

Comparative Genomics refresh

Variation and Repeats refresh

Pilot ENCODE Regions and Genes refresh

Drop down controls configure the data shown in the image above

http://www.ensembl.org/

Ensembl

Home

Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Search Ensembl

Search: All species for

Go

e.g. **human gene BRCA2** or **rat X:100000..200000** or **insulin**

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes (Log in to customize this list)

Human
NCBI36

Mouse
NCBIM37

Zebrafish
ZFISH7

All genomes

-- Select a species --

New to Ensembl?

Did you know you can:

- **Add custom tracks** using our new Control Panel
- **Upload your own data** and save it to your Ensembl account
- **Search for a DNA or protein sequence** using BLAST or BLAT
- **Fetch only the data you want** from our pu
- **Download o** in FASTA, MySQL and other formats
- **Mine Ensembl with BioMart** and export sequences or tables in text, html, or Excel format

Still got questions? Try our FAQs

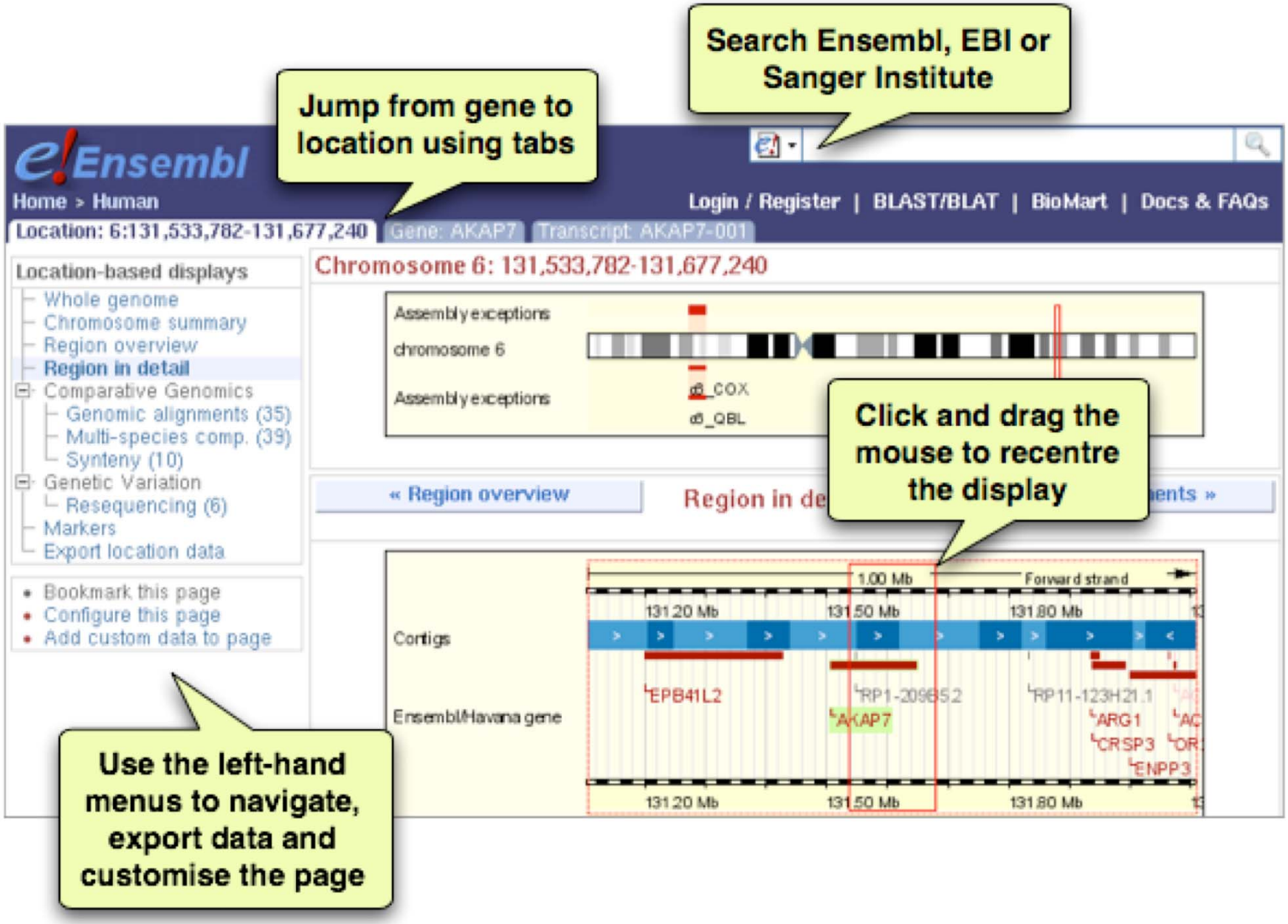
NEW!

The new Ensembl website
We've made some changes to our site, to make it faster and easier to use.
Find out more about what we've changed and why!

Click on Human

What's New in Release 52 (9 December 2008)

- **Homo sapiens core database** (Human)
- **Gorilla 2x assembly and genebuild** (Gorilla)


Search Ensembl, EBI or Sanger Institute
Jump from gene to location using tabs
Ensembl
Home > Human
Login / Register | BLAST/BLAT | BioMart | Docs & FAQs
Location: 6:131,533,782-131,677,240
Gene: AKAP7
Transcript: AKAP7-001
Location-based displays
Whole genome
Chromosome summary
Region overview
Region in detail
Comparative Genomics
Genomic alignments (35)
Multi-species comp. (39)
Synteny (10)
Genetic Variation
Resequencing (6)
Markers
Export location data
Bookmark this page
Configure this page
Add custom data to page
Chromosome 6: 131,533,782-131,677,240
Assembly exceptions
chromosome 6
Assembly exceptions
« Region overview
Click and drag the mouse to recentre the display
1.00 Mb
Forward strand
131.20 Mb
131.50 Mb
131.80 Mb
Contigs
Ensembl/Havana gene
EPB41L2
AKAP7
ARG1
CRSP3
ENPP3
Use the left-hand menus to navigate, export data and customise the page

About this species
- Description
- Genome Statistics
  - Assembly and Genebuil
  - Top 40 InterPro hits
  - Top 500 InterPro hits
- What's New
- Sample entry points
  - Karyotype
  - Location (AL032821.2
  - Gene (BRCA2)
  - Transcript (FOXP2-203

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

Search Ensembl Human

Search for: Go

e.g. **gene BRCA2** or **AL032821.2.1.143563** or **muscular dystrophy**

Description

Assembly and Genebuild »

Assembly

This release is based on the NCBI 36 assembly of the human genome [November 2005]. The data consists of a reference assembly of the complete genome plus the Celera WGS and a number of alternative assemblies of individual haplotypic chromosomes or regions. Full list of assemblies →

The International Human Genome Sequencing Consortium have published their scientific analysis of the finished human genome.

- Nature 431, 931 - 945 (21 October 2004)
- WT Sanger Institute Press Release

Annotation

Since release 38 (April 2006) the gene annotation presented has been a combined Ensembl-Havana, geneset which incorporates more than 18,000 full-length protein-coding transcripts annotated by the Havana team with the Ensembl automatic gene build. The human genome sequence is now considered sufficiently stable that since 2004 the major genome browsers have come together to produce a common set of identifiers where CDS annotations of transcripts can be agreed and these identifiers are also shown.

- More information about the CCDS project.

The ENCODE (ENCyclopedia Of DNA Elements) project aims to find functional elements in the human genome.

- More information about the ENCODE resources at Ensembl.

Vega Additional manual annotation of this genome can be found in Vega

Ensembl

Home > Human

Genome

Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Search Ensembl

- Feature type (22)
  - Domain (3)
    - Homo sapiens (3)
  - Gene (18)
    - Homo sapiens (18)
  - Marker (1)
    - Homo sapiens (1)
- Species (1)
  - Homo sapiens (22)
    - Domain (3)
    - Gene (18)
    - Marker (1)

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

Ensembl text search

brca1 corporate/tree:"Top/Species/Homo sapiens" corp [Search]

Your query matched 22 entries in the search database. Viewing hits 1-10

1 2 3

Ensembl Marker: BRCA1

A marker with 2 synonyms (262743 **BRCA1**)

Source: e52; **Feature type:** Marker; Homo sapiens; **Species:** Homo sapiens; Marker;

Ensembl protein_coding Gene: ENSG00000012048 (HGNC (automatic): BRCA1) [Region in detail]

Ensembl protein_coding gene ENSG00000012048 has 10 transcripts: ENST00000309486, ENST00000346315, ENST00000351666, ENST00000352993, ENST00000353540, ENST00000354071, ENST00000357654, ENST00000393680, ENST00000393683, ENST00000393691, associated peptides: ENSP00000013772, ENSP00000246907, ENSP00000310938, ENSP00000312236, ENSP00000326002, ENSP00000338097, ENSP00000350283, ENSP00000377285, ENSP00000377288, ENSP00000377294 and 35 exons: ENSE00000371140, ENSE00000729436, ENSE00000865492, ENSE00000865496, ENSE00000865503, ENSE00000865520, ENSE00000865521, ENSE00000865524, ENSE00000865528, ENSE00000865546, ENSE00000865551, ENSE00000865553, ENSE00000865557, ENSE00000865565, ENSE00001297284, ENSE00001312675, ENSE00001360157, ENSE00001360198, ENSE00001360203, ENSE00001368002, ENSE00001383775, ENSE00001383927, ENSE00001473234, ENSE00001473237, ENSE00001473240, ENSE00001473241, ENSE00001473245, ENSE00001516235, ENSE00001516237, ENSE00001516259, ENSE00001577499,

eptibility protein (RING finger protein 53) [Source:UniProtKB/Swiss-Prot;Acc:P38398]

g external identifiers mapped to it:

Affymx Microarray Focus: 204531_s_at

Affymx Microarray HCG110: 1993_s_at, 604_at

Affymx Microarray HuGeneFL: L78833_cds1_at, U64805_s_at

Affymx Microarray Human Exon 1.0 ST v2: 3722383, 3482826, 3800710, 2324530, 3722373, 3722386, 3722372, 3722385, 3722425, 3679671, 3282866

Affymx Microarray U133: 211851_x_at, g6552300_3p_a_at, g2218153_3p_a_at, 204531_s_at

Affymx Microarray U95: 1993_s_at, 604_at, 33724_at

Agilent CGH: A_14_P133777, A_14_P135846, A_14_P139703

Agilent Probe: A_32_P180603, A_32_P405851, A_23_P207400

CCDS: CCDS11458, CCDS11454, CCDS11457.1, CCDS11455.1, CCDS11459.1, CCDS11453, CCDS11458.1,

Click on ENSG00000012048

Ensembl

Home > Human

Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Location: 17:38,449,840-38,530,994 | Gene: BRCA1

Gene: BRCA1

- Gene summary
- Splice variants (10)
- Supporting evidence
- Sequence
- External references (15)
- Regulation
- Comparative Genomics
  - Genomic alignments (3
  - Gene Tree
    - Gene Tree (text)
    - Gene Tree (alignmen
- page
- Export data
- Bookmark this page

Gene: BRCA1 (ENSG00000012048)

Breast cancer type 1 susceptibility protein (RING finger protein 53) Source: UniProtKB/Swiss-Prot P38398

Location Chromosome 17: 38,449,840-38,530,994 reverse strand.

Transcripts There are 10 transcripts in this gene: hide transcripts

BRCA1-201	ENST00000309486	ENSP00000310938	protein_coding
BRCA1-202	ENST00000346315	ENSP00000246907	protein_coding
BRCA1-203	ENST00000351666	ENSP00000338007	protein_coding
BRCA1-204	ENST00000352993	ENSP00000312236	protein_coding
BRCA1-205	ENST00000353540	ENSP00000313772	protein_coding
BRCA1-206	ENST00000354071	ENSP00000326002	protein_coding
BRCA1-207	ENST00000357654	ENSP00000310283	protein_coding
BRCA1-208	ENST00000393680	ENSP000003	
BRCA1-209	ENST00000393685	ENSP000003	
BRCA1-210	ENST00000393691	ENSP000003	

Gene summary help

BRCA1 (HGNC (automatic))

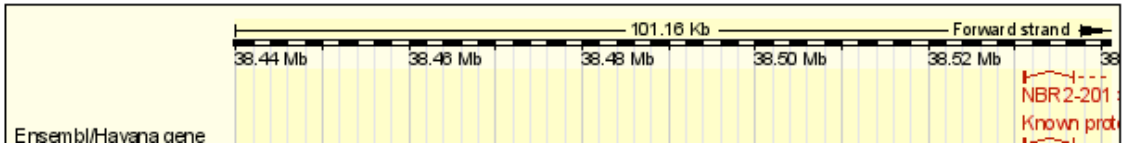
BRCC1, RNF53 [To view all Ensembl genes linked to the name click here.]

This gene is a member of the Human CCDS set: CCDS11453, CCDS11454, CCDS11455, CCDS11456, CCDS11457, CCDS11458, CCDS11459

Gene type Known protein coding

Prediction Method Transcripts were annotated by the Ensembl genebuild.

Transcripts



Gene Summary Shows you information about the gene

click here to view genomic location

Tasks

- Explore the information presented in the Gene Summary views.
  - Can you figure out how to visualize the alternatively spliced isoforms for BRCA1?
  - What can you find out about known variations in this gene?
- Using the Location Based Displays, can you figure out how to download the genomic sequence for the BRCA1 region?

The Splice Variants page shows you information about the transcripts

Gene: BRCA1

- Gene summary
- Splice variants (10)
- Supporting evidence
- Sequence
- External references (15)
- Regulation
- Comparative Genomics
  - Genomic alignments (3
  - Gene Tree
    - Gene Tree (text)
    - Gene Tree (alignment
  - Orthologues (28)
  - Paralogues (0)
  - Protein families (1)
- Genetic Variation
  - Variation Table
  - Variation Image
- External Data
- ID History
  - Gene history

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

Gene: BRCA1 (ENSG00000012048)

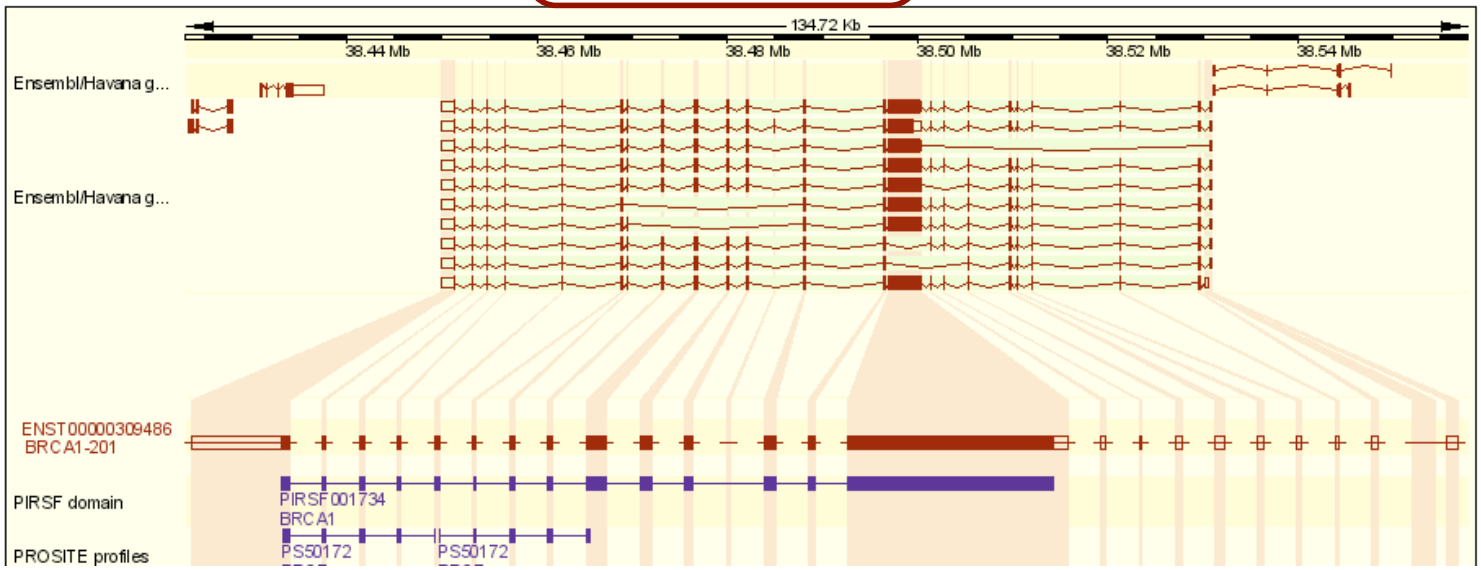
Breast cancer type 1 susceptibility protein (RING finger protein 53) Source: UniProtKB/Swiss-Prot P

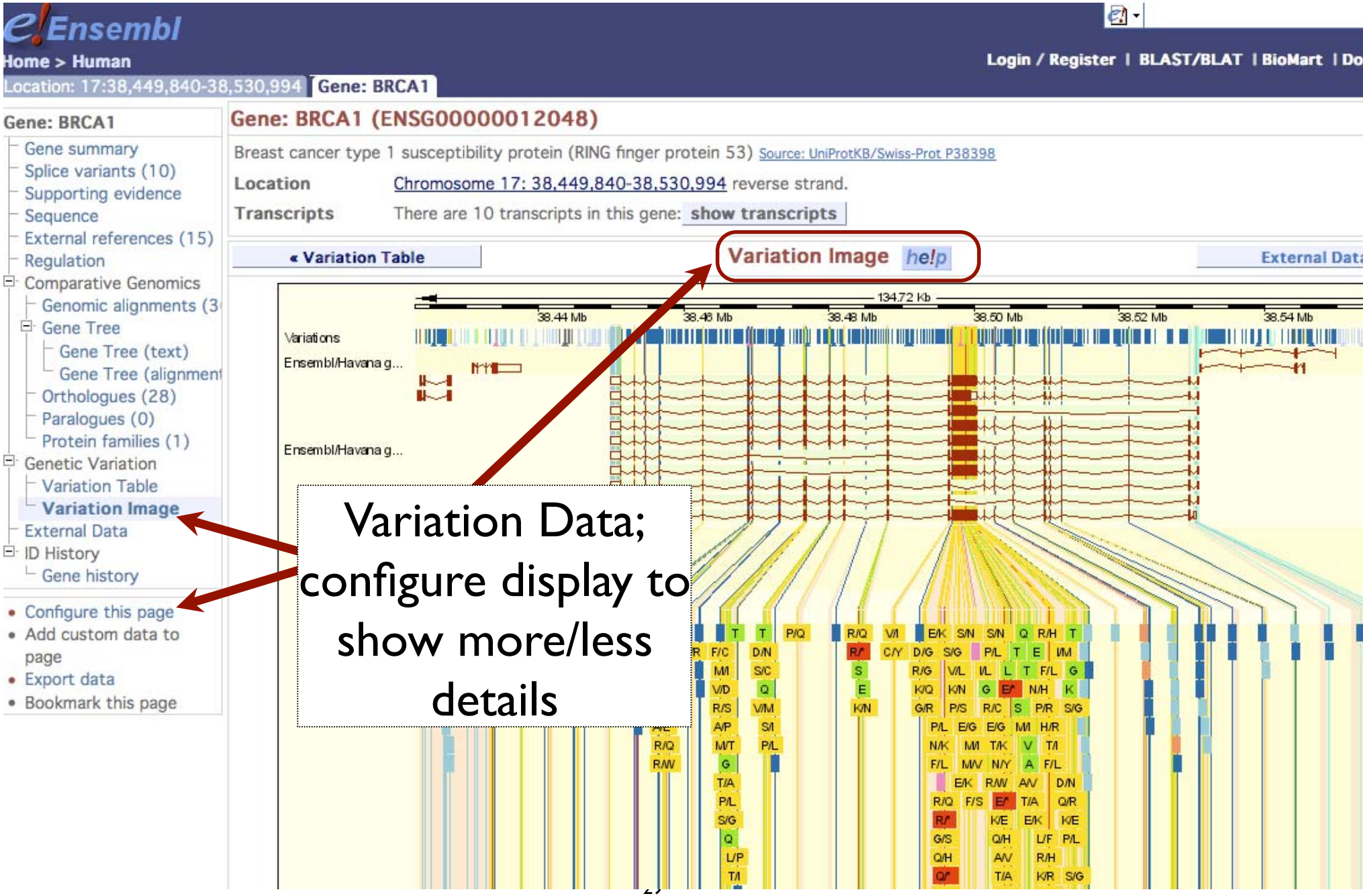
Location Chromosome 17: 38,449,840-38,530,994 reverse strand.

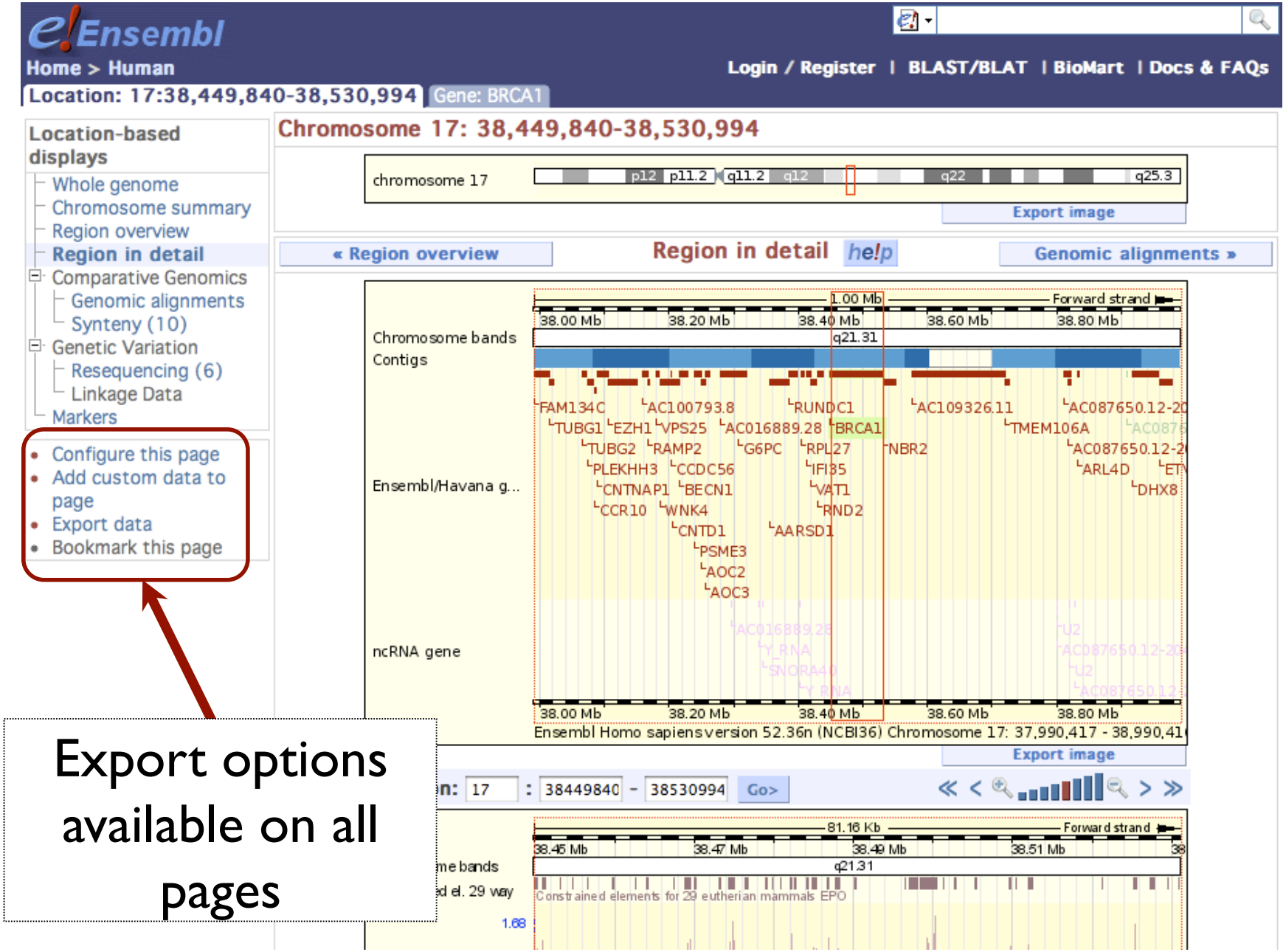
Transcripts There are 10 transcripts in this gene: hide transcripts

BRCA1-201	ENST00000309486	ENSP00000310938	
BRCA1-202	ENST00000346315	ENSP00000246907	
BRCA1-203	ENST00000351666	ENSP00000338007	
BRCA1-204	ENST00000352993	ENSP00000312236	
BRCA1-205	ENST00000353540	ENSP00000013772	
BRCA1-206	ENST00000354071	ENSP00000326002	
BRCA1-207	ENST00000357654	ENSP00000350283	protein_coding
BRCA1-208	ENST00000393680	ENSP00000377289	protein_coding
BRCA1-209	ENST00000393683	ENSP00000377288	protein_coding
BRCA1-210	ENST00000393691	ENSP00000377294	protein_coding

« Gene summary | Splice variants help | Supporting evidence »




Ensembl
Home > Human
Login / Register | BLAST/BLAT | BioMart | Do
Location: 17:38,449,840-38,530,994
Gene: BRCA1
Gene: BRCA1
Gene summary
Splice variants (10)
Supporting evidence
Sequence
External references (15)
Regulation
Comparative Genomics
Genomic alignments (3
Gene Tree
Gene Tree (text)
Gene Tree (alignment
Orthologues (28)
Paralogues (0)
Protein families (1)
Genetic Variation
Variation Table
Variation Image
External Data
ID History
Gene history
Configure this page
Add custom data to page
Export data
Bookmark this page
Gene: BRCA1 (ENSG00000012048)
Breast cancer type 1 susceptibility protein (RING finger protein 53) Source: UniProtKB/Swiss-Prot P38398
Location
Chromosome 17: 38,449,840-38,530,994 reverse strand.
Transcripts
There are 10 transcripts in this gene: show transcripts
« Variation Table
Variation Image help
External Data
134.72 Kb
38.44 Mb
38.46 Mb
38.48 Mb
38.50 Mb
38.52 Mb
38.54 Mb
Variations
Ensembl/Havana g...
Ensembl/Havana g...
Variation Data; configure display to show more/less details



Export options available on all pages

http://www.ncbi.nlm.nih.gov/mapview/

NCBI Home GenBank BLAST

Map Viewer Home | Help

The Map Viewer provides a wide variety of genome mapping and sequencing data. More..

Search

Search: Select Group or Organism

for:

Go

Tools Legend

- Search or Browse the Genome
- BLAST
- Clone Finder
- Genome Resources page

News

Annotation update released for human genome build 36 Mar 23, 2008

An annotation update for the human genome (NCBI Build 36.3) ... more

Show all

Related Resources

NCBI Home
NCBI Web Search
NCBI Site map
Genome Biology
Taxonomy
Entrez (Global Query)
BLAST
Map Viewer FTP

Small Genomes

Bacteria

Vertebrates (16)

Mammals (14)

Primates (3)

Scientific name	Common name	Build	Tools
Homo sapiens	human	Build 36.3	
		Build 35.1	
Macaca mulatta	rhesus macaque	Build 1.1	
Pan troglodytes	chimpanzee	Build 2.1	

Rodents (2)

Scientific name	Common name	Build	Tools
Mus musculus	laboratory mouse	Build 37.1	
		Build 36.1	
Rattus norvegicus	rat	RGSC v3.4	

Monotremes (1)

Marsupials (1)

Other Mammals (7)

Other Vertebrates (2)

Invertebrates (12)

Protozoa (18)

Plants (46)

Fungi (17)

Scientific name	Build
Aspergillus clavatus	
Aspergillus fumigatus	
Aspergillus niger	
Candida glabrata	Build 1.1
Cryptococcus neoformans	Build 2.1
Debaryomyces hansenii	Build 1.1
Encephalitozoon cuniculi	Build 1.1
Eremothecium gossypii	Build 3.1

Two builds of human; Note many genomes available

NCBI

NCBI Map Viewer

PubMed Nucleotide Protein Genome Gene Structure PopSet Taxonomy Help

Search for BRCA1 on chromosome(s) 17 assembly All Find Advanced Search

Map Viewer
Map Viewer Home
Map Viewer Help
Human Maps Help
Release Notes

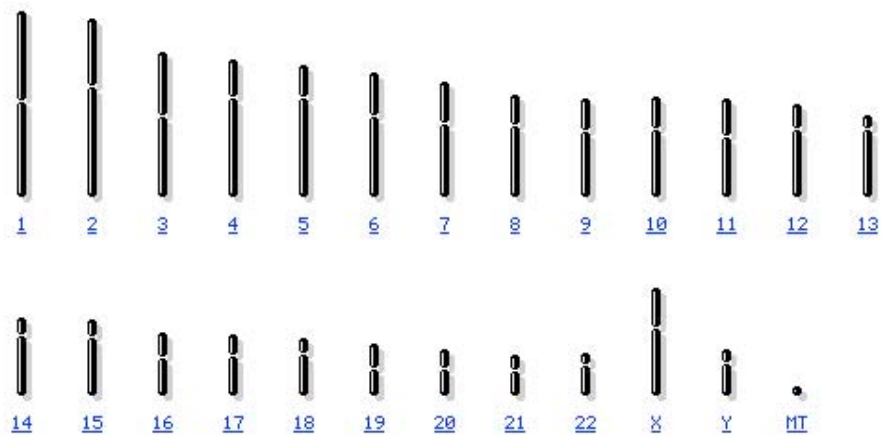
NCBI Resources
Genome Project
TaxPlot
Consensus CoDing Sequence (CCDS)
Human Genome Resources
NCBI Handbook
RefSeq
Whole Genome Association (WGA)

Organism Data in GenBank
EST
Genomic
mRNA
Protein

Homo sapiens (human) genome view

BLAST search the human genome

Build 36.2 statistics Switch to previous build



Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo; Homo sapiens

September 2006: NCBI released an annotation update for the human genome (NCBI Build 36.2); this update does not change the genome assembly. The previous version of the genome assembly, NCBI Build 35.1, can still be accessed for Map Viewer display and for BLAST. For additional information about changes, statistics, and the status of the CCDS project please refer to:

- Release Notes
- Statistics
- CCDS Project

The NCBI Map Viewer provides graphical displays of features on the human genome sequence assembly as well as

NCBI

NCBI Map Viewer

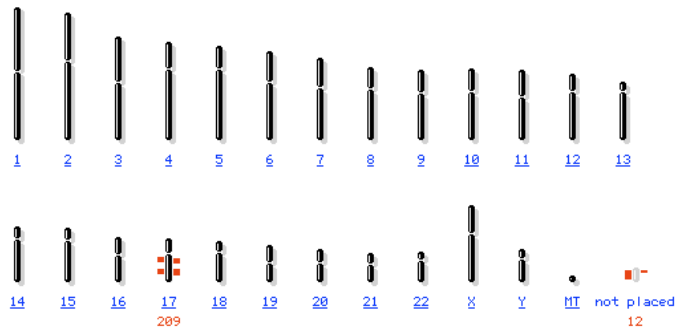
PubMed | Nucleotide | Protein | Genome | Gene | Structure | PopSet | Taxonomy | Help

Search for BRCA1 on chromosome(s) 17 assembly All Find Advanced Search

Homo sapiens (human) genome view

Build 36.2 statistics Switch to previous build

BLAST search the human genome



Search results for query "BRCA1": 221 hits

Hits shown: 1 - 100 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT not placed

Chr	Assembly	Match	Map element	Type	Maps
17	reference	all matches			
		Neighbor of **Brca1** gene 1	Rn.94975	Rn_EST_Cl	Rn UniG
		BRCA1 interacting protein C-terminal helicase 1	Mm.186143	Mm_EST_Cl	Mm UniG
		Neighbor of **Brca1** gene 1	Mm.784	Mm_EST_Cl	Mm UniG
		Neighbor of **BRCA1** gene 2 (9 hits)	Hs.559259	Hs_EST_Cl	Hs UniG
		Neighbor of **BRCA1** gene 1 (2 hits)	Hs.546264	Hs_EST_Cl	Hs UniG
		BRCA1 interacting protein C-terminal helicase 1	Hs.532799	Hs_EST_Cl	Hs UniG
		Neighbor of **BRCA1** gene 1	Hs.373818	Hs_EST_Cl	Hs UniG
		Neighbor of **BRCA1** gene 1 (2 hits)	Hs.277721	Hs_EST_Cl	Hs UniG
		BRCA1 interacting protein C-terminal helicase 1	Gga.17801	Gga_EST_Cl	Gga UniG

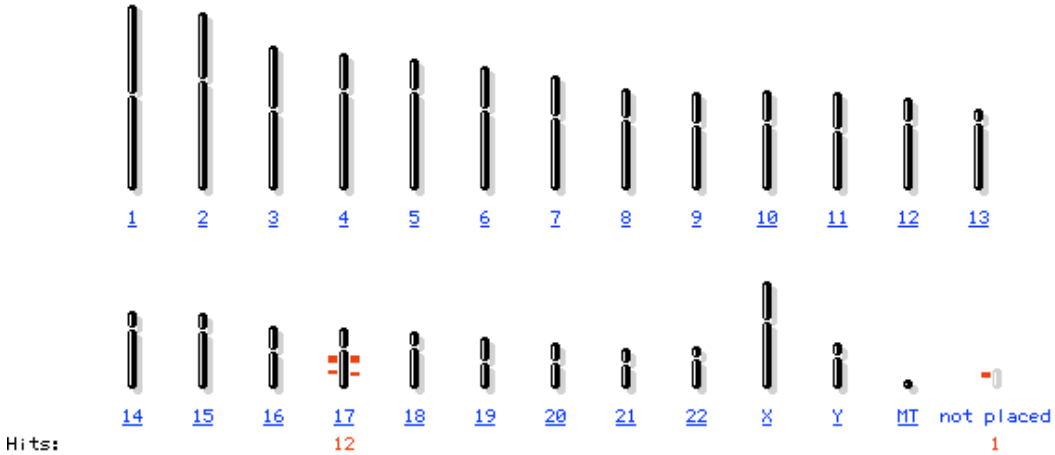
Quick Filter
- ☑ Gene
- Transcript :
  - ☐ all
  - ☐ RefSeq
- ☐ STS
- ☐ Unigene

Filter

Quick Filter
✓ Gene

Homo sapiens (human) genome view

Build 36.2 statistics Switch to previous build



Search results for query "BRCA1 AND gene[obj_type]": 13 hits

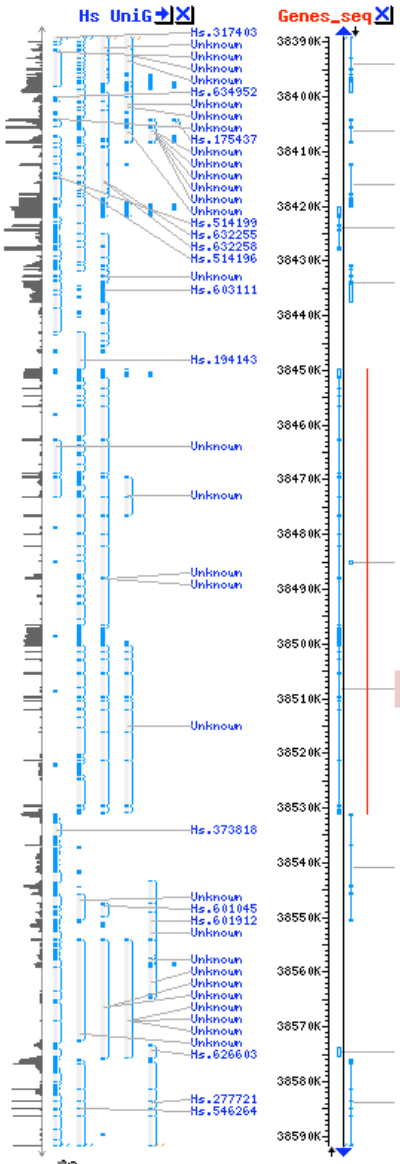
Chr	Assembly	Match	Map element	Type	Maps
17	reference	all matches			
		similar to neighbor of **BRCA1** gene 1	LOC728560	Gene	Genes_cyto | Genes_seq
		BRCA1P1 : like BRCA1	BRCA1P1	Gene	Genes_cyto | Genes_seq
		BRCA1-interacting protein 1	BRIP1	Gene	Genes_cyto | Genes_seq
		neighbor of **BRCA1** gene 2	NBR2	Gene	Genes_cyto | Genes_seq
		neighbor of **BRCA1** gene 1	NBR1	Gene	Genes_cyto | Genes_seq
		BRCA1 : breast cancer 1, early onset	BRCA1	Gene	Genes_cyto | Genes_seq
		BRCA1 : ENSG00000012048	BRCA1	GENE	ensGenes
17	Celera	all matches			
		BRCA1P1 : like BRCA1	BRCA1P1	GENE	Genes_seq
		BRCA1-interacting protein 1	BRIP1	GENE	Genes_seq
		neighbor of **BRCA1** gene 2	NBR2	GENE	Genes_seq
		neighbor of **BRCA1** gene 1	NBR1	GENE	Genes_seq
		BRCA1 : breast cancer 1, early onset	BRCA1	GENE	Genes_seq
17:not placed	reference	similar to neighbor of **BRCA1** gene 1	LOC727732	GENE	Genes_seq

Human genome overview page (Build 36.2)
Human genome overview page (Build 35.1)
Map Viewer Home
Map Viewer Help
Human Maps Help
FTP
Data As Table View
Maps & Options
Compress Map
Region Shown:
38,389K
38,592K
Go
out
zoom
in
You are here:
Ideogram
default
master

Master Map: Genes On Sequence | Summary of Maps | Maps & Optio

Region Displayed: **38,389K-38,592K bp**

Download/View Sequence/Evider



Symbol	O	Links	E	Cyto	Description
RUNDC1	↓	HGNC sv pr dl ev mm hm sts	best RefSeq	17q21.31	RUN domain containing 1
RPL27	↓	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq	17q21.1-q21.2	ribosomal protein L27
IFI35	↓	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq	17q21	interferon-induced protein 35
VAT1	↑	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq	17q21	vesicle amine transport protein 1 homolog (T cell
RND2	↓	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq	17q21	Rho family GTPase 2
RPL21P4	↓	HGNC sv dl ev mm	best RefSeq	17q21	ribosomal protein L21 pseudogene 4
BRCA1	↑	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq	17q21	breast cancer 1, early onset
NBR2	↓	HGNC sv pr dl ev mm sts	best RefSeq	17q21	neighbor of BRCA1 gene 2
BRCA1P1	↑	HGNC sv dl ev mm	best RefSeq	17q21	BRCA1 pseudogene 1
NBR1	↓	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq	17q21.31	neighbor of BRCA1 gene 1

Two tasks

- Can you figure out how to LinkOut to the OMIM and/or Homologene entries for BRCA1?
- Can you figure out how to download the genomic sequence for the BRCA1 region?

Human genome overview page (Build 36.2)
Human genome overview page (Build 35.1)
Map Viewer Home
Map Viewer Help
Human Maps Help
FTP
Data As Table View
Maps & Options
Compress Map
Region Shown:
38,389K
38,592K
Go

out
zoom
in

You are here:

Ideogram

17p13
17p12
17p11.2
17p11.1
17q11.1
17q11.2
17q12
17q21
17q22
17q23
17q24
17q25

default
master

Master Map: Genes On Sequence

Summary of Maps

Maps & Optio

Region Displayed: **38,389K-38,592K bp**

Download/View Sequence/Evider

Hs UniG

Genes_seq

Symbol	O	Links	E	Cyto	Description
RUNDC1	↓	HGNC sv pr dl ev mm hm sts	best RefSeq	17q21.31	RUN domain containing 1
RPL27	↓	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq	17q21.1-q21.2	ribosomal protein L27
IFI35	↓	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq	17q21	interferon-induced protein 35
VAT1	↑	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq	17q21	vesicle amine transport protein 1 homolog (T cell
RND2	↓	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq	17q21	Rho family GTPase 2
RPL21P4	↓	HGNC sv dl ev mm	best RefSeq	17q21	ribosomal protein L21 pseudogene 4
BRCA1	↑	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq	17q21	breast cancer 1, early onset
NBR2	↓	HGNC sv pr dl ev mm	best RefSeq		
BRCA1P1	↑	HGNC sv dl ev mm	best RefSeq		
NBR1	↓	OMIM HGNC sv pr dl ev mm hm sts	best RefSeq		

LinkOut

OMIM = disease

sv = sequence view

pr = protein record

dl = download

hm = Homologene

Bioinformatics

Session 3.1 - Discovering GEO, the Gene Expression Omnibus.

Functional Genomics

- What kinds of questions can you ask with microarray data?

✓ basic research

✓ drug target discovery

✓ biomarker discovery

✓ pharmacology & toxico-genomics

✓ clinical diagnosis - prognosis, diagnosis, & disease classification

✓ gene regulatory networks

✓ protein-DNA binding

✓ + more

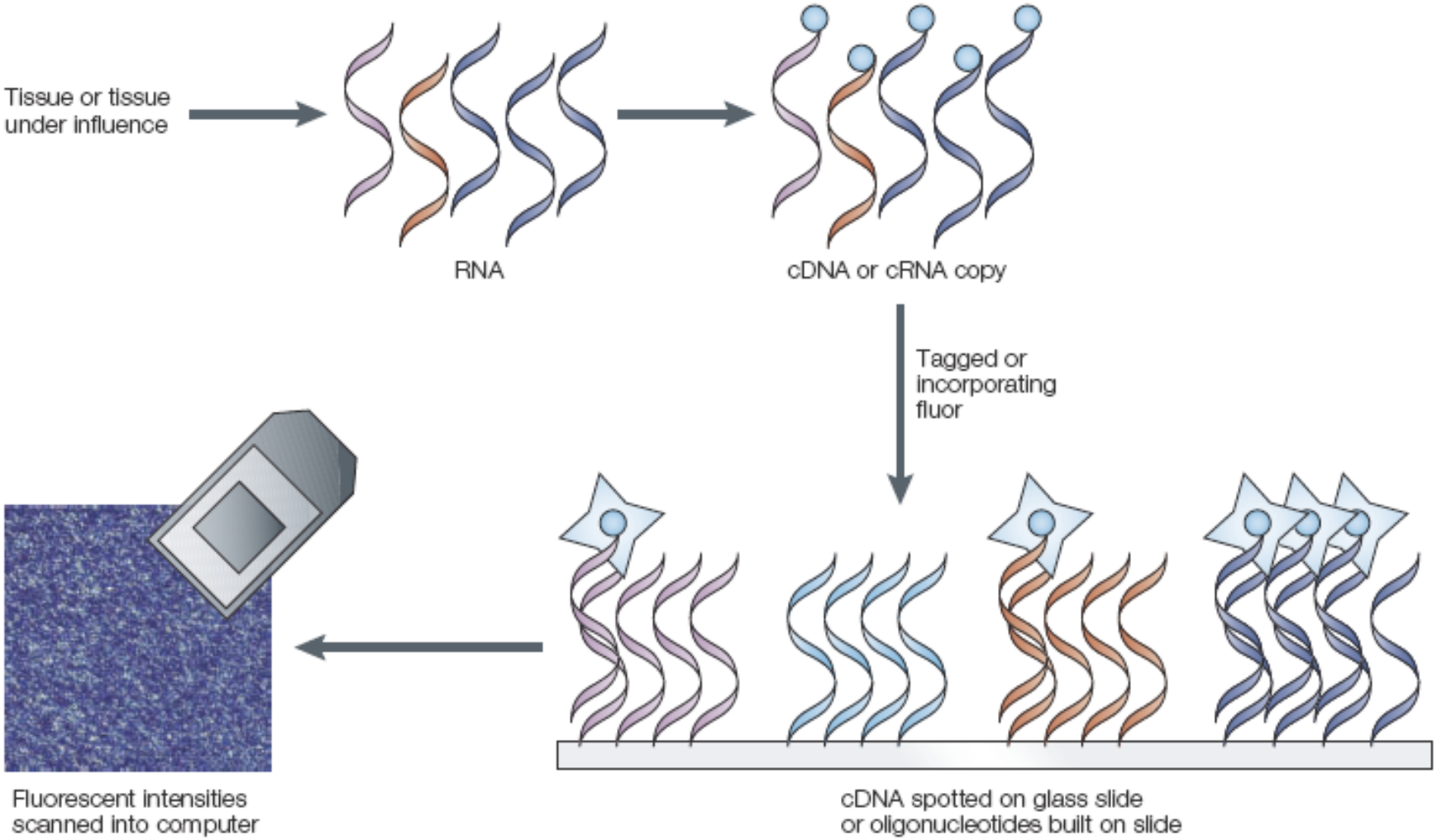


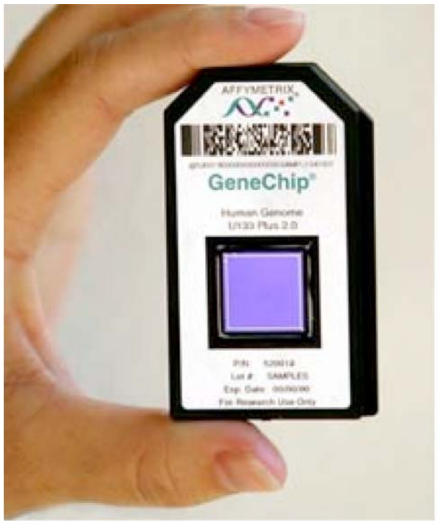
Figure 1 | **Schematized experimental process using a microarray.** Although the specific protocols differ, the microarray approach first involves isolating RNA or messenger RNA from appropriate biological samples, making the RNA (or a copy of it) fluorescent, hybridizing it to the microarray, washing off the excess and scanning the microarray under laser light.

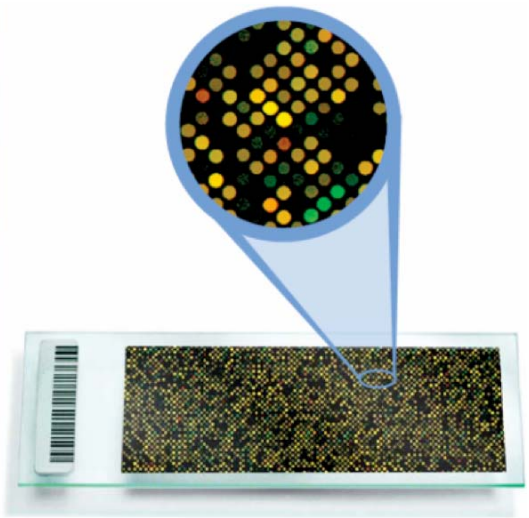
source: Butte, A Nature Reviews 2002

Different Platforms

in situ oligonucleotide
single sample, absolute levels

spotted DNA/cDNA
two samples, relative levels





Microarray Experiment

- Design
- Collect
- Pre-Process
- Analyze
- Interpret
- Submit
- Publish

Example = Normalization

Examples = Distance measures, data classification, clustering, + more

Rate Limiting Step = What do these results actually mean?

Public Microarray Data

- The Gene Expression Omnibus (GEO)
  - repository/archive gene expression data
- data submitted by the research community in fulfillment of journal requirements
- this public data represents an untapped resource; potential discovery from existing data sets is at your fingertips

GEO Database

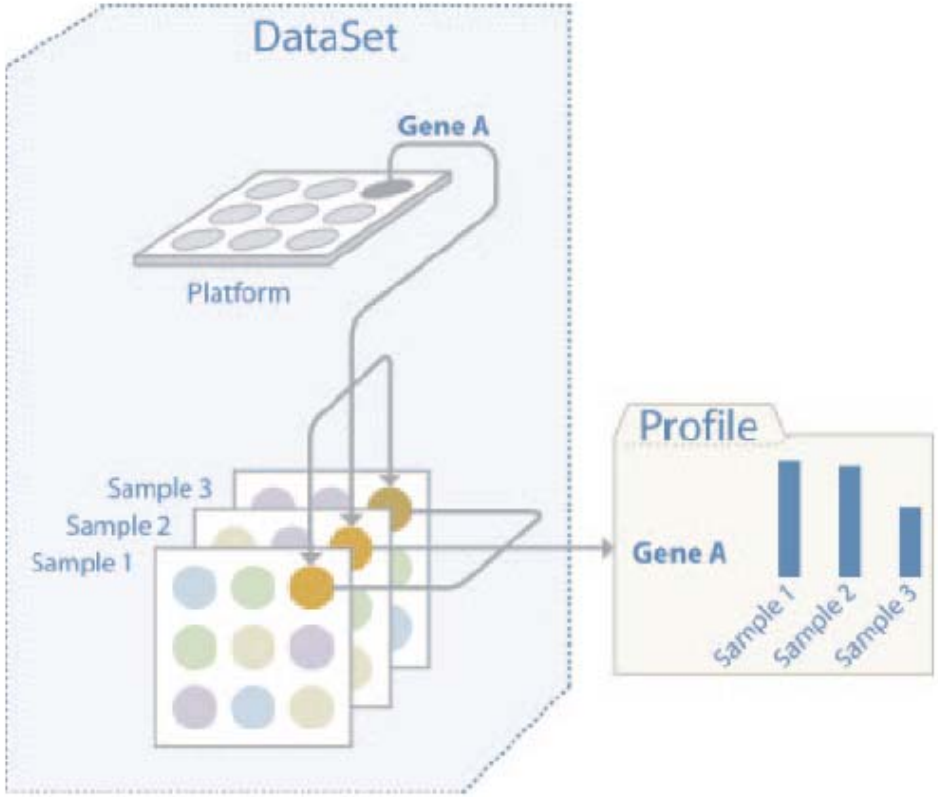
<u>Organized by:</u>

Platform

Sample

Series/DataSet

Profile



http://www.ncbi.nlm.nih.gov/geo/

NCBI

GEO
Gene Expression Omnibus

HOME | SEARCH | SITE MAP | Handout | NAR 2006 Paper | NAR 2002 Paper | FAQ | MIAME | Email GEO

NCBI > GEO

Not logged in | Login

Gene Expression Omnibus: a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

GEO navigation

QUERY
- DataSets [GO]
- Gene profiles [GO]
- GEO accession [GO]
- GEO BLAST

BROWSE
- DataSets
- GEO accessions
  - Platforms
  - Samples
  - Series

SUBMIT
- Direct deposit / update
- Web deposit / update
- Create new account

Public data

GPL	Platforms	5557
GSM	Samples	284951
GSE	Series	11188
Total		301696

Site contents

Documentation
Overview | FAQ | Find
Submission guide
Linking & citing
Journal citations
Programmatic access
DataSet clusters
GEO announce list
Data disclaimer
GEO staff

Query & Browse
Repository browser
Submitter contacts
SAGEmap
FTP site
GEO Profiles
GEO DataSets

Deposit & Update
Direct deposit
Web deposit
New account

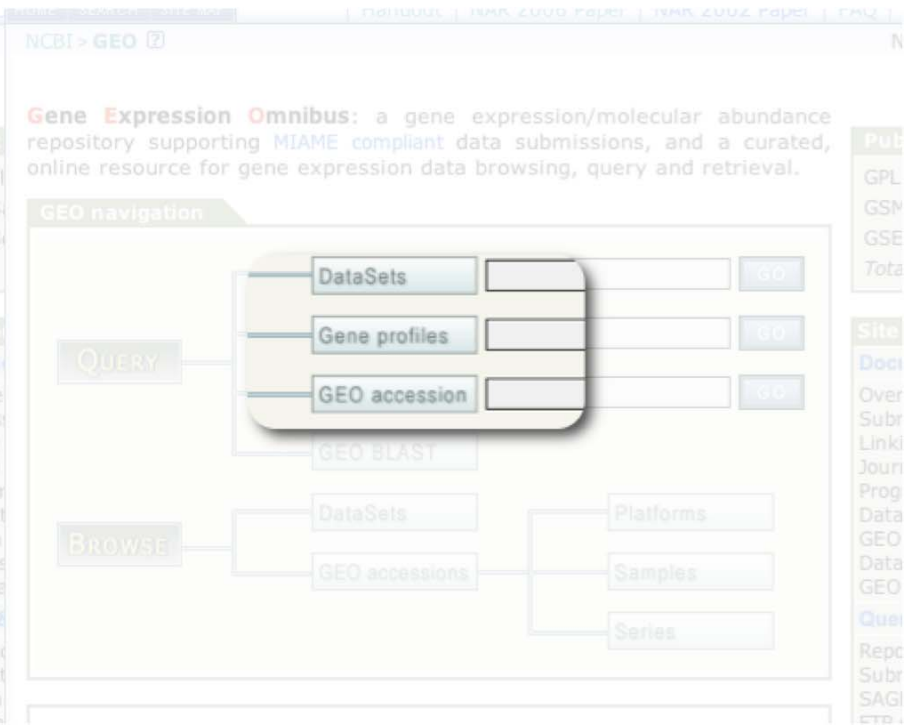
Searching GEO

- Are you interested in a particular type of expt?

✓ GEO DataSets

- Are you looking for your favorite gene?

✓ GEO Profiles



Data in GEO

>120,000 samples
>3.2 billion measurement
200+ organisms
from >2000 labs

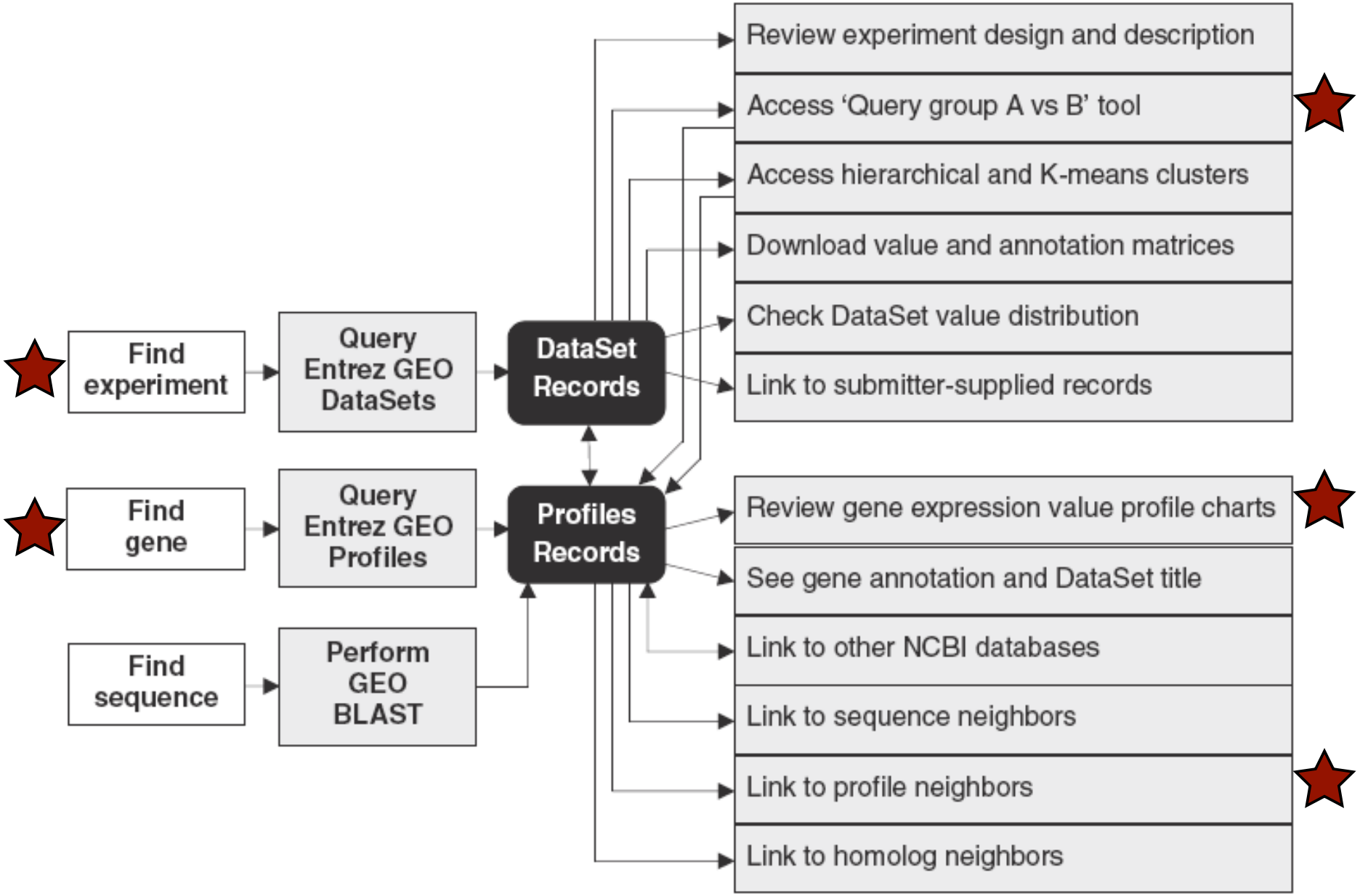
freely available online
ftp downloads

Total holdings

	Public	Unreleased	Total
Platforms	4407	355	**4762**
Samples	201401	45428	**246829**
Series	7883	1623	**9506**

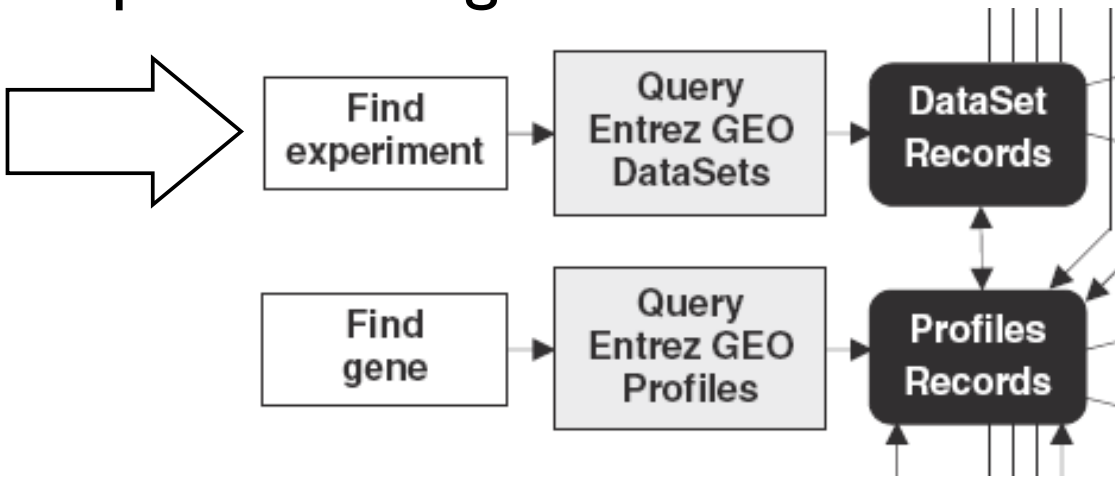
Browse public holdings

- All contacts
- All platforms
  - in situ oligonucleotide (1260)
  - spotted oligonucleotide (1099)
  - spotted DNA/cDNA (1850)
  - antibody (5)
  - tissue (0)
  - MS (10)
  - SARST (1)
  - MPSS (12)
  - RT-PCR (7)
  - oligonucleotide beads (50)
  - mixed spotted oligonucleotide/cDNA (6)
  - spotted protein (4)
  - SAGE (54)
- All samples
  - RNA (167588)
  - genomic (30043)
  - protein (651)
  - SAGE (993)
  - mixed (913)
- All series


Find experiment
Query Entrez GEO DataSets
DataSet Records
Find gene
Query Entrez GEO Profiles
Profiles Records
Find sequence
Perform GEO BLAST
Review experiment design and description
Access 'Query group A vs B' tool
Access hierarchical and K-means clusters
Download value and annotation matrices
Check DataSet value distribution
Link to submitter-supplied records
Review gene expression value profile charts
See gene annotation and DataSet title
Link to other NCBI databases
Link to sequence neighbors
Link to profile neighbors
Link to homolog neighbors

An Example

- Find microarray experiments that look at the expression of genes in cancer



You can use these GEO data mining tools for quick and easy identification of relevant & noteworthy data sets. For serious analyses, you should download the data and use a microarray data analysis software suite.

NCBI

GEO Gene Expression Omnibus

HOME | SEARCH | SITE MAP | Handout | NAR 2006 Paper | NAR 2002 Paper | FAQ | MIAME | Email GEO

NCBI > GEO

Not logged in | Login

Gene Expression Omnibus: a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

GEO navigation

Query
- DataSets — cancer — GO
- Gene profiles — GO
- GEO accession — GO
- GEO BLAST

Browse
- DataSets
- GEO accessions
  - Platforms
  - Samples
  - Series

Submit
- Direct deposit / update
- Web deposit / update
- Create new account

Public data

GPL	Platforms	4407
GSM	Samples	201401
GSE	Series	7883
Total		213691

Site contents

Documentation

Overview | FAQ
Submission guide
Linking & citing
Journal citations
Programmatic access
DataSet clusters
GEO announce list
Data disclaimer
GEO staff

Query & Browse

Repository browser
Submitter contacts
SAGEmap
FTP site
GEO Profiles
GEO DataSets

Deposit & Update

Direct deposit
Web deposit

NCBI

Entrez GEO DataSets

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

Search GEO DataSets for cancer Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

The GEO site

GEO FAQ

List GEO Contents

Entrez
Help | FAQ

Display Summary Show 20 Send to

All: 1885 | DataSets: 339 | Platforms: 206 | Series: 1340

Items 1 - 20 of 1885

Pag

1: GDS2926 record: Megakaryocytic differentiation: time course [Homo sapiens]

Summary: Temporal analysis of phorbol ester-treated CHRF-288-11 megakaryoblastic cells in to undergo megakaryocytic (Mk) differentiation and primary Mk (PriMk) cells deriv cytokine-treated CD34+ peripheral blood cells. Results provide insight into molecu mechanisms underlying megakaryopoiesis.
Parent Platform: GPL887
Reference Series: GSE8914

Type: gene expression array-based, log e ratio

Subsets: 4 agent, 2 cell line, 13 time sets.

Samples: 77

GSM87962: CHRF_Expt3_DMSO_4d_rep1
GSM87963: CHRF_Expt3_DMSO_4d_rep2
GSM87983: CHRF_Expt4_DMSO_4d_rep1
GSM87984: CHRF_Expt4_DMSO_7d_rep1
GSM87961: CHRF_Expt3_DMSO_12d_rep1
GSM87970: CHRF_Expt3_PMA_1h_rep1

NCBI

Entrez GEO DataSets

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

Search GEO DataSets for cancer AND human[Organism] Preview Go Clear

Limits | Preview/Index | History | Clipboard | Details

he GEO site

EO FAQ

ist GEO Contents

ntrez
elp | FAQ

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

Search	Most Recent Queries	Time	Result
#15	Search **cancer AND human[Organism]**	14:25:31	1387
#14	Search **cancer**	14:17:14	1885
#13	Search **all[filter]**	13:08:00	13962

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 human Preview Index

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

NCBI

Entrez GEO DataSets

My NCBI [Sign In] [Register]

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

h GEO DataSets for cancer AND human[Organism] AND in situ oligonucleotide Preview Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

O site
AQ
O Contents
FAQ

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

Search	Most Recent Queries	Time	Result
#16	Search **cancer AND human[Organism] AND in situ oligonucleotide[Platform Technology Type]**	14:28:29	718
#15	Search **cancer AND human[Organism]**	14:27:47	1387
#14	Search **cancer**	14:17:14	1885

Platform Technology Type

All Fields
Author
DataSet Type
Description
Entry Type
Filter
GEO Accession
MeSH Terms
Number of Platform Probes
Number of Samples
Organism
Platform Technology Type
Publication Date
Related Platform
Related Series
Reporter Identifier
Sample Source
Sample Type
Sample Value Type
Submitter Institute

w Index:

ox; use the pull-down menu to specify a search field.
ms to the query box and see the number of search results, or click Index to view terms within a field.

in situ oligonucleotide Preview Index

add a term to the query box

How many "spotted DNA/cDNA" experiments explore cancer in humans?

NCBI

Entrez GEO DataSets

My NCBI [Sign In] [Register]

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

arch GEO DataSets for cancer AND human[Organism] AND spotted DNA/cDNA[Pl Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

GEO site
O FAQ
GEO Contents
rez
p | FAQ

Display Summary Show 20 Send to

All: 530 | DataSets: 60 | Platforms: 109 | Series: 361

Items 1 - 20 of 530

Page 1 of 27 Next

1: GDS2415 record **Breast carcinomas and local recurrence** [Homo sapiens] GEO Profiles, Links

Summary: Analysis of primary breast carcinoma tumors from 50 patients who received breast-conserving therapy (BCT). 19 patients subsequently developed a local recurrence of the carcinoma. 9 recurrent tumors also examined. Compared to mastectomy, BCT is associated with a higher rate of local recurrence.
Parent Platform: GPL3558
Reference Series: GSE4913

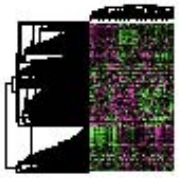
Type: gene expression array-based, log2 ratio

Subsets: 2 disease state, 2 specimen sets.

Supplementary Files: TXT download...

Samples: 59

GSM110395: wsb 1428
GSM110396: wsb 1631
GSM110397: wsb 1642
GSM110398: wsb 1694
GSM110399: wsb 565
GSM110400: wsb 575



CBI

Entrez GEO DataSets

My NCBI [Sign In] [Register]

tabases PubMed Nucleotide Protein Genome Structure PMC Journals Books

O DataSets for cancer AND human[Organism] AND spotted DNA/cDNA[Pl Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Send to

All: 530 DataSets: 60 Platforms: 109 Series: 361

ents

Items 1 - 20 of 60 Page 1 of 3 Next

1: GDS2415 record: Breast carcinomas and local recurrence [Homo sapiens] GEO Profiles, Links

Summary: Analysis of primary breast carcinoma tumors from 50 patients who received breast-conserving therapy (BCT). 19 patients subsequently developed a local recurrence of the carcinoma. 9 recurrent tumors also examined. Compared to mastectomy, BCT is associated with a higher rate of local recurrence.
Parent Platform: GPL3558
Reference Series: GSE4913

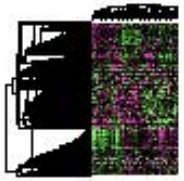
Type: gene expression array-based, log2 ratio

Subsets: 2 disease state, 2 specimen sets.

Supplementary Files: TXT download...

Samples: 59

GSM110395: wsb 1428
GSM110396: wsb 1631
GSM110397: wsb 1642
GSM110398: wsb 1694
GSM110399: wsb 565
GSM110400: wsb 575



NCBI

DATASET BROWSER (CURATED)

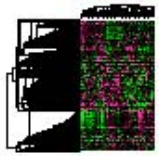
GEO Gene Expression Omnibus

Search for GDS2415[ACCN] Search Clear Show All **Advanced Search**

DataSet Record GDS2415: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Breast carcinomas and local recurrence		
Summary:	Analysis of primary breast carcinoma tumors from 50 patients who received breast-conserving therapy (BCT). 19 patients subsequently developed a local recurrence of the carcinoma. 9 recurrent tumors also examined. Compared to mastectomy, BCT is associated with a higher rate of local recurrence.		
Organism:	*Homo sapiens*		
Platform:	GPL3558: NKI-AVL Homo sapiens 18K cDNA microarray		
Citation:	Kreike B, Halfwerk H, Kristel P, Glas A et al. Gene expression profiles of primary breast carcinomas from patients at high risk for local recurrence after breast-conserving therapy. *Clin Cancer Res* 2006 Oct 1;12(19):5705-12. PMID: 17020974		
Reference Series:	GSE4913	**Sample count:**	59
Value type:	log2 ratio	**Series published:**	2006/09/24

Cluster Analysis



Download

- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

Data Analysis Tools

Find genes

- Compare 2 sets of samples
- Cluster heatmaps
- Experiment design and value distribution

Find gene name or symbol: Go

Find genes that are up/down for this condition(s): ☑ disease state ☑ specimen Go

NLM NIH GEO Help Disclaimer Section 508

NCBI

DATASET BROWSER (CURATED)

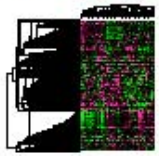
GEO Gene Expression Omnibus

Search for GDS2415[ACCN] | Search | Clear | Show All | Advanced Search

DataSet Record GDS2415: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Breast carcinomas and local recurrence		
Summary:	Analysis of primary breast carcinoma tumors from 50 patients who received breast-conserving therapy (BCT). 19 patients subsequently developed a local recurrence of the carcinoma. 9 recurrent tumors also examined. Compared to mastectomy, BCT is associated with a higher rate of local recurrence.		
Organism:	*Homo sapiens*		
Platform:	GPL3558: NKI-AVL Homo sapiens 18K cDNA microarray		
Citation:	Kreike B, Halfwerk H, Kristel P, Glas A et al. Gene expression profiles of primary breast carcinomas from patients at high risk for local recurrence after breast-conserving therapy. *Clin Cancer Res* 2006 Oct 1;12(19):5705-12. PMID: 17020974		
Reference Series:	GSE4913	**Sample count:**	59
Value type:	log2 ratio	**Series published:**	2006/09/24

Cluster Analysis



Download

- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

Data Analysis Tools

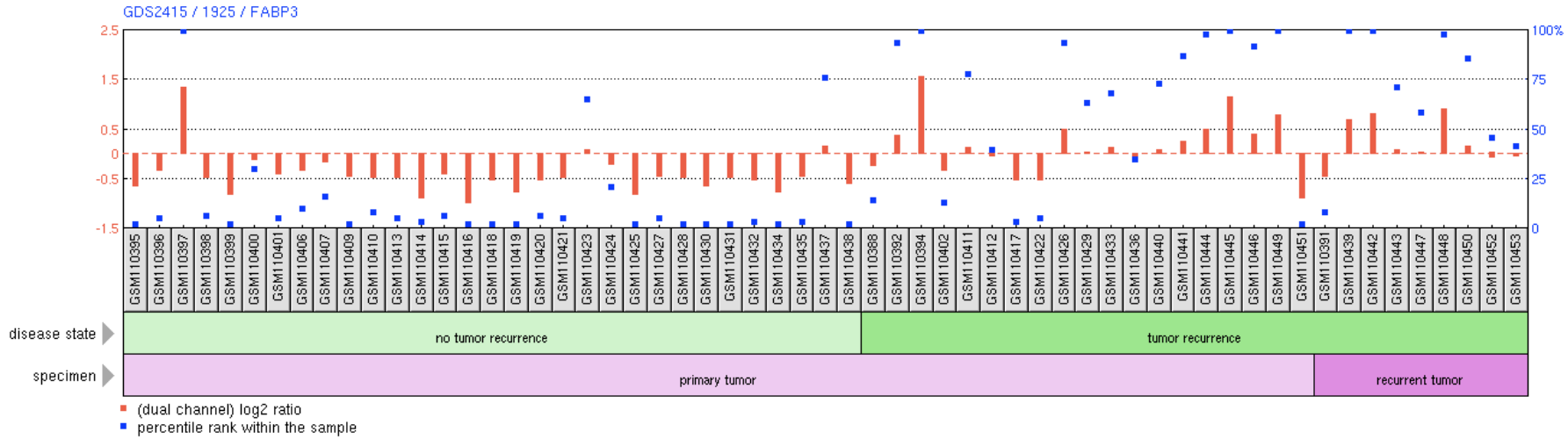
Find genes

- Compare 2 sets of samples
- Cluster heatmaps
- Experiment design and value distribution

Find gene name or symbol: brca1 Go

Find genes that are up/down for this condition(s): ☑ disease state ☑ specimen Go

NLM NIH GEO Help Disclaimer Section 508



- thumbnail image represents the abundance profile for an individual gene across each Sample in a DataSet
- bars at the bottom of the chart represent experimental subsets within the DataSet.
- Red bar: measured level of abundance
- Blue square: indication of where the expression of that gene falls with respect to all other genes on that array

NCBI

DATASET BROWSER (CURATED)

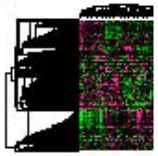
GEO Gene Expression Omnibus

Search for GDS2415[ACCN] | Search | Clear | Show All | Advanced Search

DataSet Record GDS2415: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Breast carcinomas and local recurrence		
Summary:	Analysis of primary breast carcinoma tumors from 50 patients who received breast-conserving therapy (BCT). 19 patients subsequently developed a local recurrence of the carcinoma. 9 recurrent tumors also examined. Compared to mastectomy, BCT is associated with a higher rate of local recurrence.		
Organism:	*Homo sapiens*		
Platform:	GPL3558: NKI-AVL Homo sapiens 18K cDNA microarray		
Citation:	Kreike B, Halfwerk H, Kristel P, Glas A et al. Gene expression profiles of primary breast carcinomas from patients at high risk for local recurrence after breast-conserving therapy. *Clin Cancer Res* 2006 Oct 1;12(19):5705-12. PMID: 17020974		
Reference Series:	GSE4913	**Sample count:**	59
Value type:	log2 ratio	**Series published:**	2006/09/24

Cluster Analysis



Download

- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

Data Analysis Tools

- Find genes
- **Compare 2 sets of samples**
- Cluster heatmaps
- Experiment design and value distribution

Step 1: Select test and significance level

Two-tailed t-test (A vs B) | Significance level: 0.100

Step 2: Select which Samples to put in Group A and Group B

Group A: GSM110388, GSM110451, GSM110449, GSM110446, GSM110444, GSM110445, GSM110441, GSM110436, GSM110440, GSM110433, GSM110429, GSM110426, GSM110422, GSM110412, GSM110417, GSM110411, GSM110402, GSM110394, GSM110392

Group B: GSM110391, GSM110453, GSM110439, GSM110442, GSM110443, GSM110447, GSM110448, GSM110450, GSM110452

Step 3: Query Group A vs. B

NLM NIH GEO Help Disclaimer Section 508

Entrez GEO Profiles

My NCBI [Sign In] [Register]

PubMed Nucleotide Protein Genome Structure PMC Journals Books

iles for Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Select Send to Download profile data

All: 2192

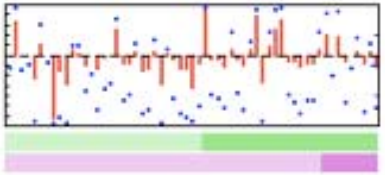
Items 1 - 20 of 2192 Page 1 of 110 Next

1: GDS2415 record | GPL3558 19198 [Homo sapiens] 59 samples Profile Neighbors, Chromosome Neighbors, Links

Annotation: SFRS2IP: Splicing factor, arginine/serine-rich 2, interacting protein

Reporter: H78241

Experiment: Breast carcinomas and local recurrence, gene expression array-based, log2 ratio

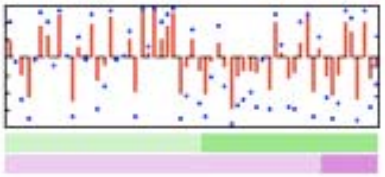


2: GDS2415 record | GPL3558 19188 [Homo sapiens] 59 samples Profile Neighbors, Links

Annotation: yg20e10.s1 Soares infant brain 1NIB Homo sapiens cDNA clone IMAGE:32609 3-, mRNA sequence

Reporter: R43734

Experiment: Breast carcinomas and local recurrence, gene expression array-based, log2 ratio



3: GDS2415 record | GPL3558 19184 [Homo sapiens] 59 samples Chromosome Neighbors, Links

Annotation: LIPC: Lipase, hepatic

Reporter: N68256

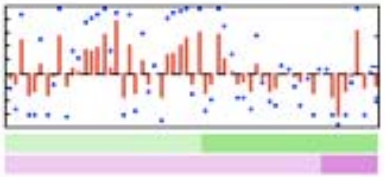
Experiment: Breast carcinomas and local recurrence, gene expression array-based, log2 ratio

4: GDS2415 record | GPL3558 19172 [Homo sapiens] 59 samples Chromosome Neighbors, Links

Annotation: SMARCA1: SWI/SNF related, matrix associated, actin dependent regulator of chromatin, ...

Reporter: AA496809

Experiment: Breast carcinomas and local recurrence, gene expression array-based, log2 ratio



A vs B Query Tool

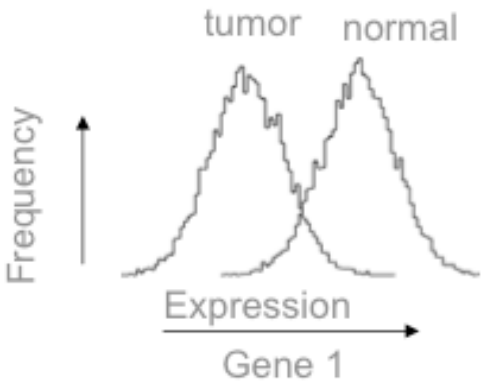
Take home message: GEO data analysis tools are great for quick identification of interesting leads; you download the data to carry out more robust statistical analyses

- **Purpose:** To help identify gene profiles that display marked differences in expression level between two subsets of experimental factors (e.g. tissue, strain, time, dose, etc).
- **Caveats:** The "mean group A vs B" is perhaps the most rudimentary means of filtering data; t-test is well established but comes with a set of basic assumptions.

A Simple Test

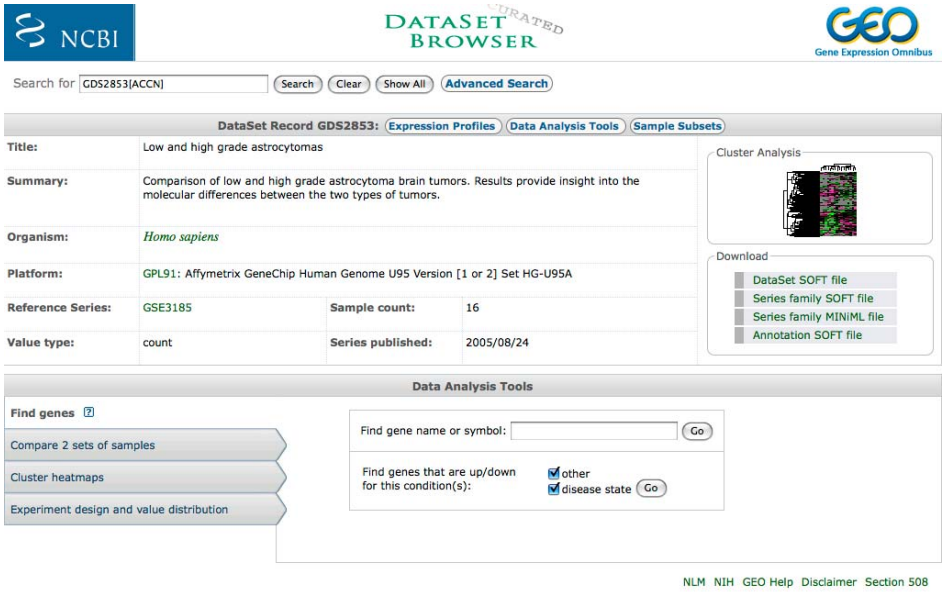
- Student's t-test
  - Assumptions: Normality, equal variance

$$t = \frac{m_{tumor} - m_{normal}}{\sqrt{\frac{s_{tumor}^2}{N_{tumor}} + \frac{s_{normal}^2}{N_{normal}}}}$$



m_i = mean expression value in class i
N_i = number of examples in class i
s_i^2 = variance

Using GEO for differential expression



http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2853

GEO limitations

- Differential expression can only be done for "Datasets"(GDS****)
- T-tests only
- Very little control over parameters
- Output is not that easy to use

Be careful with $p < 0.05$

- In GDS2853 example, 2912 genes met $p<0.05$
- This is 11% of the genes on the array
- Expect 5% by chance (this is what $p < 0.05$ means)
- Probably ~1/2 of the selected 2912 are false positives
- This is the "multiple testing" problem

Download Data Types

SOFT - text based
MiNiML* - xml based

*MIAME Notation in Markup Language

all GEO data are available for bulk download:
ftp://ftp.ncbi.nih.gov/pub/geo/DATA

More Serious Tools

- Free
  - R + Bioconductor
  - TIGR MultiExperimentViewer (MeV)

  ...
- Commerical
  - Genespring–ArrayAssist
  - Rosetta Resolver

  ...

Gene Expression Profiles of Primary Breast Carcinomas from Patients at High Risk for Local Recurrence after Breast-Conserving Therapy

Bas Kreike,[1,3] Hans Halfwerk,[2,3] Petra Kristel,[2,3] Annuska Glas,[2] Hans Peterse,[2] Harry Bartelink,[1] and Marc J. van de Vijver[2]

gene of interest
FABP3

Abstract **Purpose:** Several risk factors for local recurrence of breast cancer after breast-conserving therapy (BCT) have been identified. The identification of additional risk factors would be very useful in guiding optimal therapy and also in improving understanding of the mechanisms underlying local recurrence. We used cDNA microarray analysis to identify gene expression profiles associated with local recurrence.

Experimental Design: Using 18K cDNA microarrays, gene expression profiles were obtained from 50 patients who underwent BCT. Of these 50 patients, 19 developed a local recurrence; the remaining 31 patients were selected as controls as they were free of local recurrence at least 11 years after treatment. For 9 of 19 patients, the local recurrence was also available for gene expression profiling. Unsupervised and supervised methods of classification were used to separate patients in groups corresponding to disease outcome and to study the overall gene expression pattern of primary tumors and their recurrences.

Results: Hierarchical clustering of patients did not show any grouping reflecting local recurrence status. Supervised analysis revealed no significant set of genes that was able to distinguish recurring tumors from nonrecurring tumors. Paired-data analysis of primary tumors and local recurrences showed a remarkable similarity in gene expression profile between primary tumors and their recurrences.

Conclusions: No significant differences in gene expression between primary breast cancer tumors in patients with or without local recurrence after BCT were identified. Furthermore, analyses of primary tumors and local recurrences show a preservation of the overall gene expression pattern in the local recurrence, even after radiotherapy.

Breast-conserving therapy (BCT) has become the therapy of choice for a large proportion of breast cancer patients. Several randomized controlled trials have shown no difference in survival rates after BCT or mastectomy for stage I and II breast cancer (1–4). Studies comparing the psychological effects of BCT with mastectomy have shown that patients treated with

recurrence compared with mastectomy. A local recurrence rate of 10% in 10 years follow-up is generally considered as clinically acceptable for $T_{1-2}N_{0-1}$ breast cancers. However, local recurrence up to 30% have been reported in young patients (7, 8).

Several risk factors for local recurrence after BCT have been

NCBI

DATASET BROWSER (CURATED)

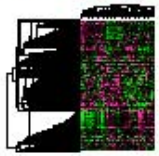
GEO Gene Expression Omnibus

Search for GDS2415[ACCN] | Search | Clear | Show All | Advanced Search

DataSet Record GDS2415: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Breast carcinomas and local recurrence		
Summary:	Analysis of primary breast carcinoma tumors from 50 patients who received breast-conserving therapy (BCT). 19 patients subsequently developed a local recurrence of the carcinoma. 9 recurrent tumors also examined. Compared to mastectomy, BCT is associated with a higher rate of local recurrence.		
Organism:	*Homo sapiens*		
Platform:	GPL3558: NKI-AVL Homo sapiens 18K cDNA microarray		
Citation:	Kreike B, Halfwerk H, Kristel P, Glas A et al. Gene expression profiles of primary breast carcinomas from patients at high risk for local recurrence after breast-conserving therapy. *Clin Cancer Res* 2006 Oct 1;12(19):5705-12. PMID: 17020974		
Reference Series:	GSE4913	**Sample count:**	59
Value type:	log2 ratio	**Series published:**	2006/09/24

Cluster Analysis



Download

- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

Data Analysis Tools

Find genes

- Compare 2 sets of samples
- Cluster heatmaps
- Experiment design and value distribution

Find gene name or symbol: fabp3 | Go

Find genes that are up/down for this condition(s): ☑ disease state ☑ specimen | Go

NLM NIH GEO Help Disclaimer Section 508

NCBI

Entrez GEO Profiles

My NCBI [Sign In] [Register]

Databases PubMed Nucleotide Protein Genome Structure PMC Journals Books

GEO Profiles for "GDS2415"[ACCN] fabp3 Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Subgroup effect Send to Download profile data

All: 2

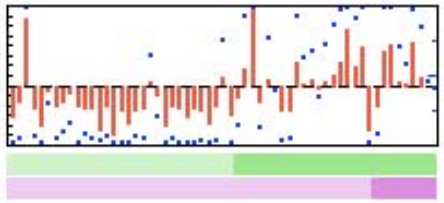
Items 1 - 2 of 2 One page.

1: GDS2415 record | GPL3558 1925 [Homo sapiens] 59 samples Profile Neighbors, Chromosome Neighbors, Links

Annotation: FABP3: Fatty acid binding protein 3, muscle and heart (mammary-derived growth inhib...

Reporter: AA044307

Experiment: Breast carcinomas and local recurrence, gene expression array-based, log2 ratio

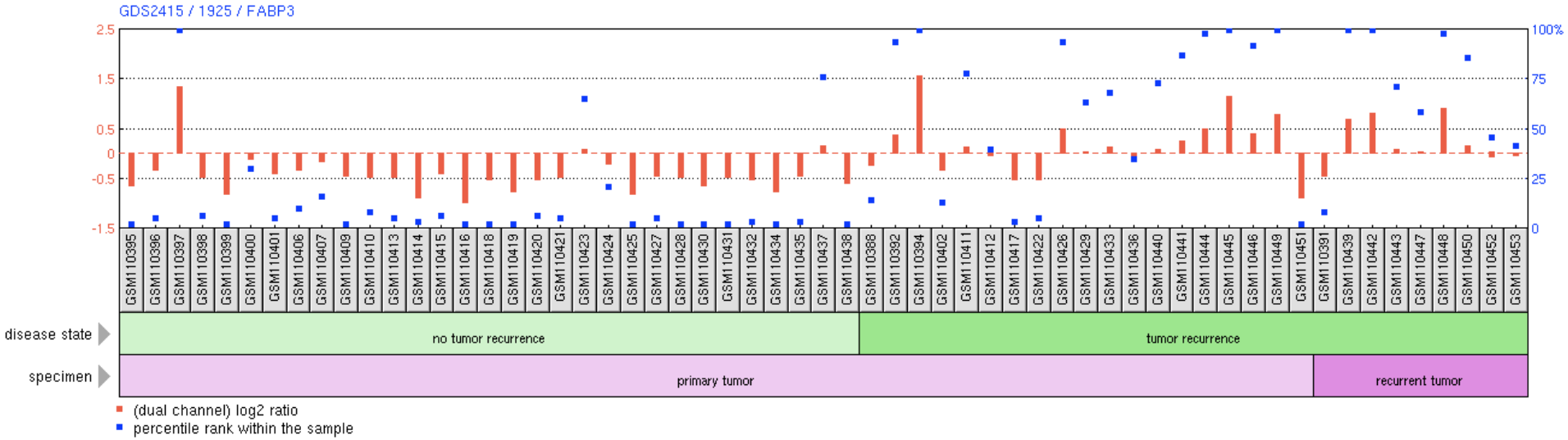


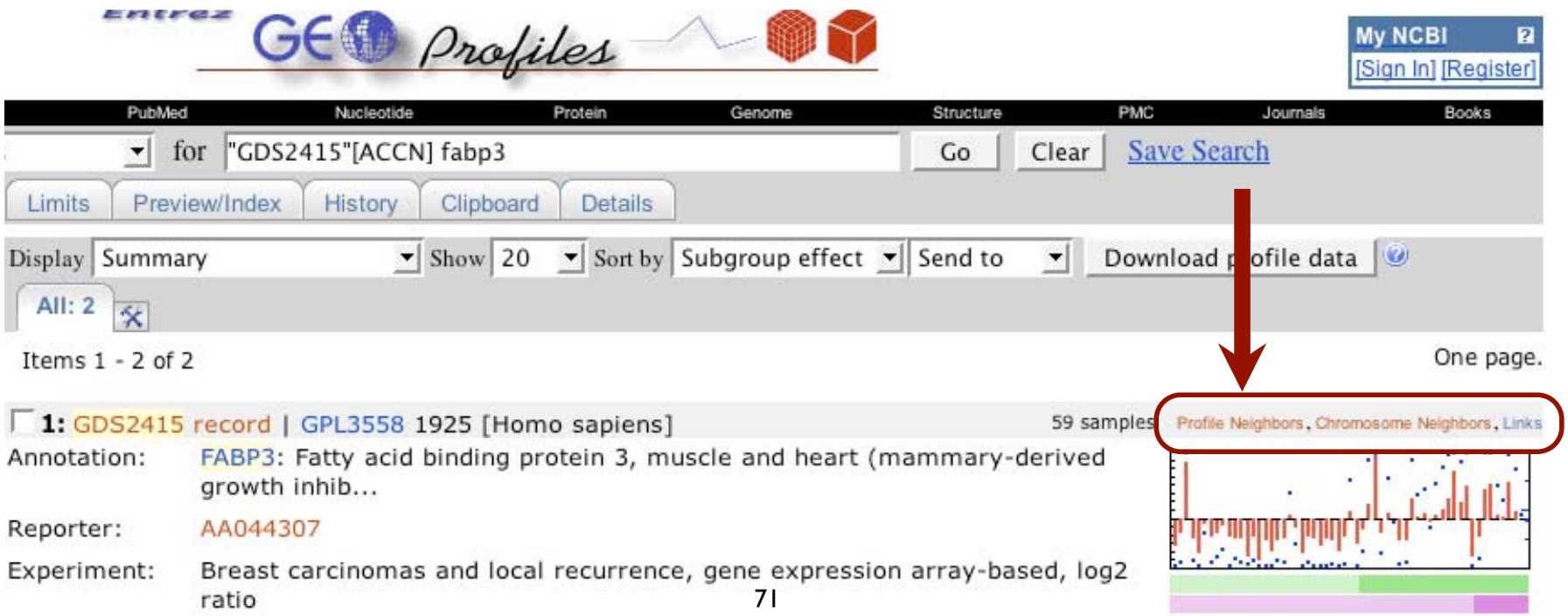
2: GDS2415 record | GPL3558 11434 [Homo sapiens] 59 samples Chromosome Neighbors, Links

Annotation: FABP3: Fatty acid binding protein 3, muscle and heart (mammary-derived growth inhib...

Reporter: AA148548

Experiment: Breast carcinomas and local recurrence, gene expression array-based, log2 ratio


GDS2415 / 1925 / FABP3
2.5
1.5
0.5
0
-0.5
-1.5
100%
75
50
25
0
disease state
no tumor recurrence
tumor recurrence
specimen
primary tumor
recurrent tumor
(dual channel) log2 ratio
percentile rank within the sample


Entrez GEO Profiles
My NCBI
[Sign In] [Register]
PubMed Nucleotide Protein Genome Structure PMC Journals Books
for "GDS2415"[ACCN] fabp3
Go
Clear
Save Search
Limits Preview/Index History Clipboard Details
Display Summary
Show 20
Sort by Subgroup effect
Send to
Download profile data
All: 2
Items 1 - 2 of 2
One page.
1: GDS2415 record | GPL3558 1925 [Homo sapiens]
59 samples
Profile Neighbors, Chromosome Neighbors, Links
Annotation: FABP3: Fatty acid binding protein 3, muscle and heart (mammary-derived growth inhib...
Reporter: AA044307
Experiment: Breast carcinomas and local recurrence, gene expression array-based, log2 ratio

Entrez GEO Profiles

My NCBI [Sign In] [Register]

PubMed Nucleotide Protein Genome Structure PMC Journals Books

for Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Select Send to Download profile data

All: 6

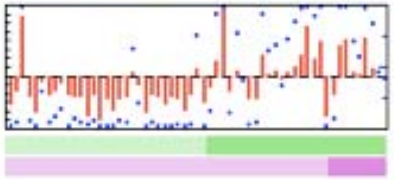
Items 1 - 6 of 6 One page.

1: GDS2415 record | GPL3558 1925 [Homo sapiens] 59 samples Profile Neighbors, Chromosome Neighbors, Links

Annotation: FABP3: Fatty acid binding protein 3, muscle and heart (mammary-derived growth inhib...

Reporter: AA044307

Experiment: Breast carcinomas and local recurrence, gene expression array-based, log2 ratio

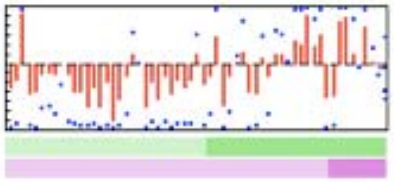


2: GDS2415 record | GPL3558 13888 [Homo sapiens] 59 samples Profile Neighbors, Links

Annotation: Transcribed locus, strongly similar to NP_066407.1 histone family, member B...

Reporter: AI076718

Experiment: Breast carcinomas and local recurrence, gene expression array-based, log2 ratio

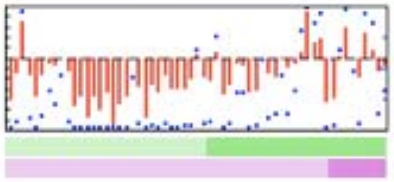


3: GDS2415 record | GPL3558 2942 [Homo sapiens] 59 samples Profile Neighbors, Chromosome Neighbors, Links

Annotation: HIST2H2BE: Histone cluster 2, H2be

Reporter: AA010223

Experiment: Breast carcinomas and local recurrence, gene expression array-based, log2 ratio

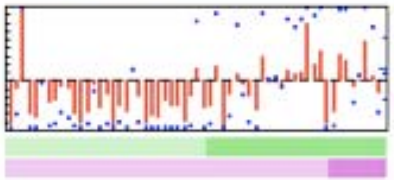


4: GDS2415 record | GPL3558 757 [Homo sapiens] 59 samples Profile Neighbors, Chromosome Neighbors, Links

Annotation: HIST1H2BK: Histone cluster 1, H2bk

Reporter: N71982

Experiment: Breast carcinomas and local recurrence, gene expression array-based, log2 ratio



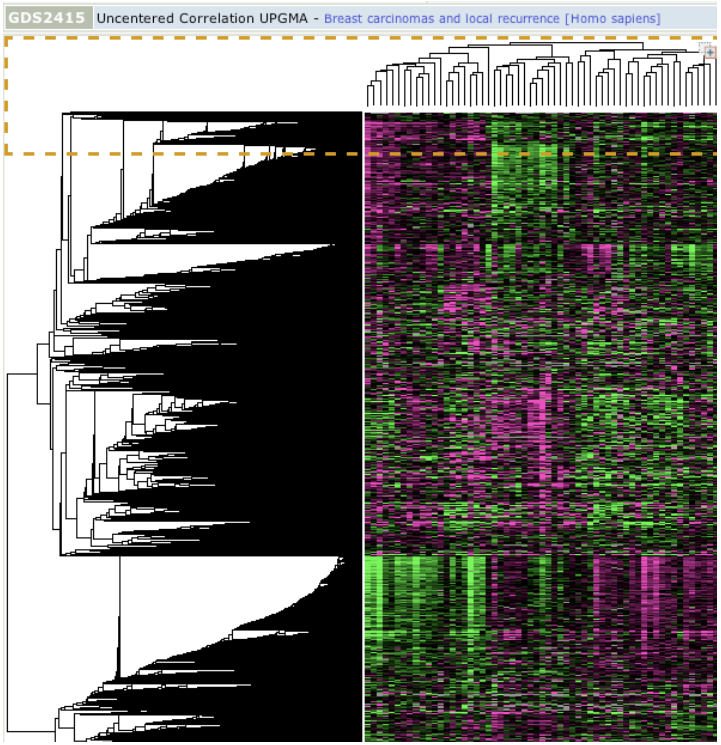
Profile Neighbors

Take home message: GEO data analysis tools are great for quick identification of interesting leads; you download the data to carry out more robust statistical analyses

- Connects groups of genes that have similar expression profiles within a DataSet
- pre-computed
- calculated by Pearson correlation coefficients

Other Features

- Cluster Heat Maps
  - precomputed sample and gene hierarchical cluster heat maps provided
  - different methods available; can select, expand, download
- GEO BLAST
  - retrieve gene expression profiles by sequence similarity



GEO, the gene expression omnibus

- public repository of expression data from many different experimental platforms
- Main uses
  - ✓ search for experiments of interest
  - ✓ search for expression information about gene of interest
- submit, search, analyses tools available
- data standards required MIAME, MiNiML

Credits & References

- NCBI GEO: mining tens of millions of expression profiles–database and tools update. Barrett T, et al. Nucleic Acids Res. 35 (2007) D760-5. [PMID: 17099226]
- GEO: the Gene Expression Omnibus

  http://www.ncbi.nlm.nih.gov/projects/geo/info/GEOHandoutFinal.pdf
- Dr. Paul Pavlidis, UBC Bioinformatics Centre

Bioinformatics

Session 3.2 - Pathway Resources for Systems Biology

Proteomics

- How large is the human proteome, anyway?

Class	Size	Description
Non Redundant Proteins	20,000-25,000	representative protein from every gene locus
Variants	50,000-500,000	different proteins obtained by splicing or proteolysis
Combinatorial Variants	>10,000,000	different proteins generated by somatic DNA rearrangements
Protein Species	>100,000	proteins that differ in chemical composition due to PTM
Protein Alleles	75,000-150,000	proteins that differ by genetic variation (coding SNPs)

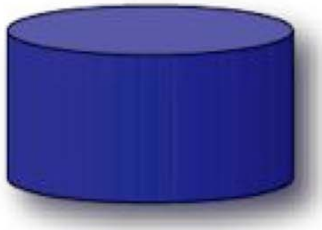
source: Uhlen M et al, Molecular & Cellular Proteomics (2005)

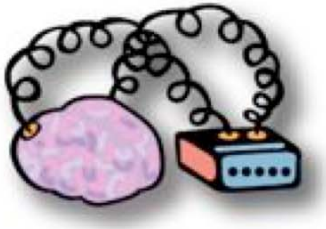
Cellular Pathways

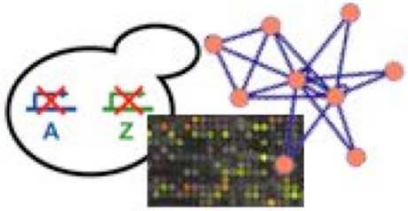
- A striking similarity between intracellular signaling pathways and the Tokyo subway system

Pathway Information

- Databases
  - Fully electronic
  - Easily computer readable
- Literature
  - Increasingly electronic
  - Human readable
- Biologist's brains
  - Richest data source
  - Limited bandwidth access
- Experiments
  - Basis for models







http://www.pathguide.org/

Home | BioPAX | cBio | MSKCC

Pathguide » the pathway resource list

Navigation
- Protein-Protein Interactions
- Metabolic Pathways
- Signaling Pathways
- Pathway Diagrams
- Transcription Factors / Gene Regulatory Networks
- Protein-Compound Interactions
- Genetic Interaction Networks
- Protein Sequence Focused
- Other

Search

Organisms: All

Availability: All

Standards: All

Reset | Search

Statistics

Analyze Pathguide

Contact

Comments, Questions, Suggestions are Always

Complete Listing of All Pathguide Resources

Pathguide contains information about **287** biological pathway resources. Click on a link to go to the resource home page or 'Details' for a description page. Databases that are free and those supporting BioPAX, CellML, PSI-MI or SBML standards are respectively indicated.

If you know of a pathway resource that is not listed here, or have other questions or comments, please send us an e-mail.

News

Major update
All resources were recently reviewed and many new ones were added

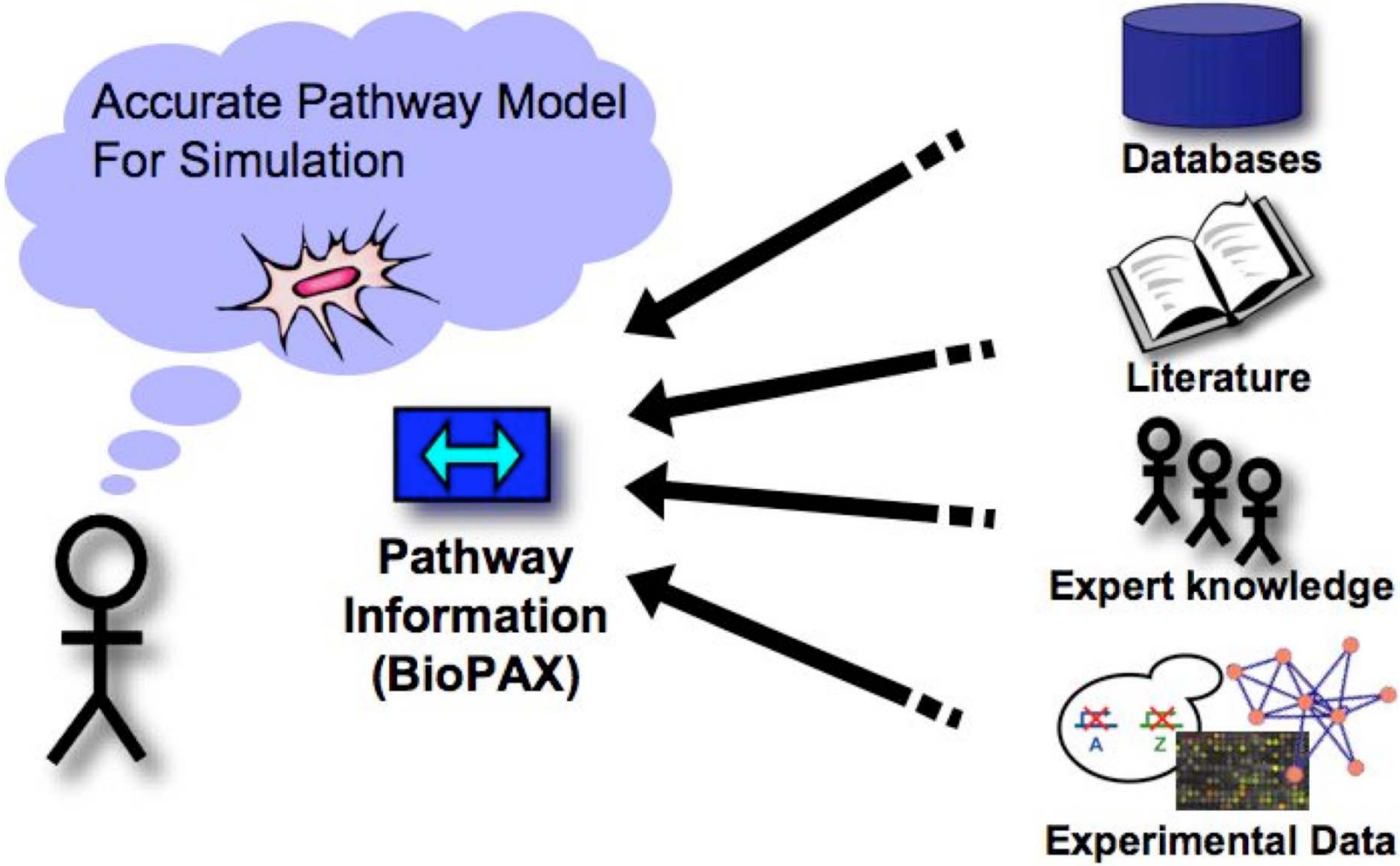
Get the Stats
Detailed Pathguide resource statistics now available

Pathguide Published
Please cite the Pathguide Publication

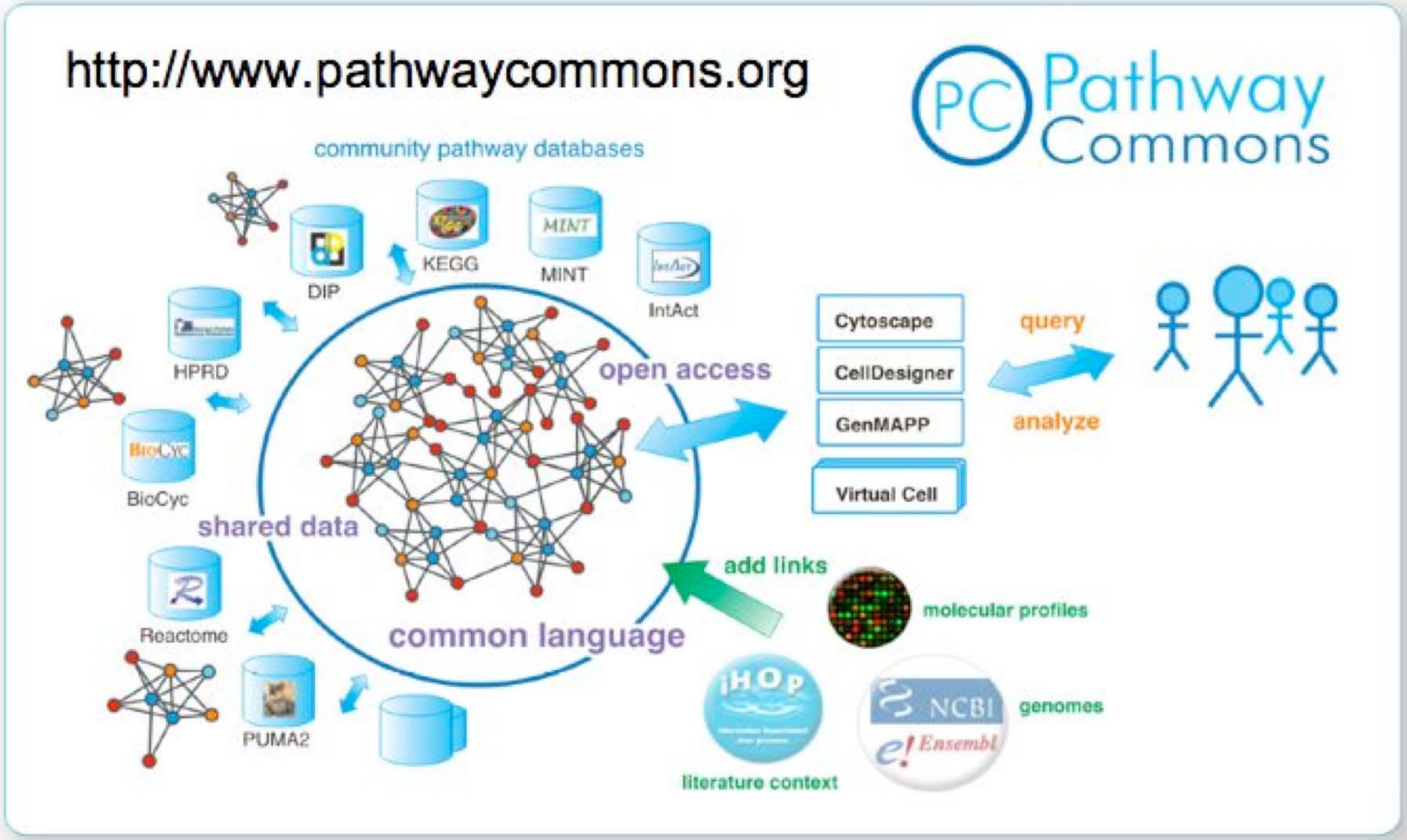
Protein-Protein Interactions

Database Name (Order: alphabetically | by web popularity)	Full Record	Availability	Standards
3DID - 3D interacting domains	Details	Free	
ABCdb - Archaea and Bacteria ABC transporter database	Details	Free	
AfCS - Alliance for Cellular Signaling Molecule Pages Database	Details	Free	
AllFuse - Functional Associations of Proteins in Complete Genomes	Details	X	
aMAZE - Protein Function and Biochemical Pathways Project	Details	Free	
ASEdb - Alanine Scanning Energetics Database	Details	Free	
ASPD - Artificial Selected Proteins/Peptides Database	Details	Free	
BID - Binding Interface Database	Details	X	
BIND - Biomolecular Interaction Network Database	Details	Free	PSI-MI
BioGRID - General Repository for Interaction Datasets	Details		PSI-MI
BRITE - Biomolecular Relations in Information Transmission and Expression	Details	Free	
CA1Neuron - Pathways of the hippocampal CA1 neuron	Details	Free	
Cancer Cell Map - The Cancer Cell Map	Details	Free	BioPAX

Using Pathway Information



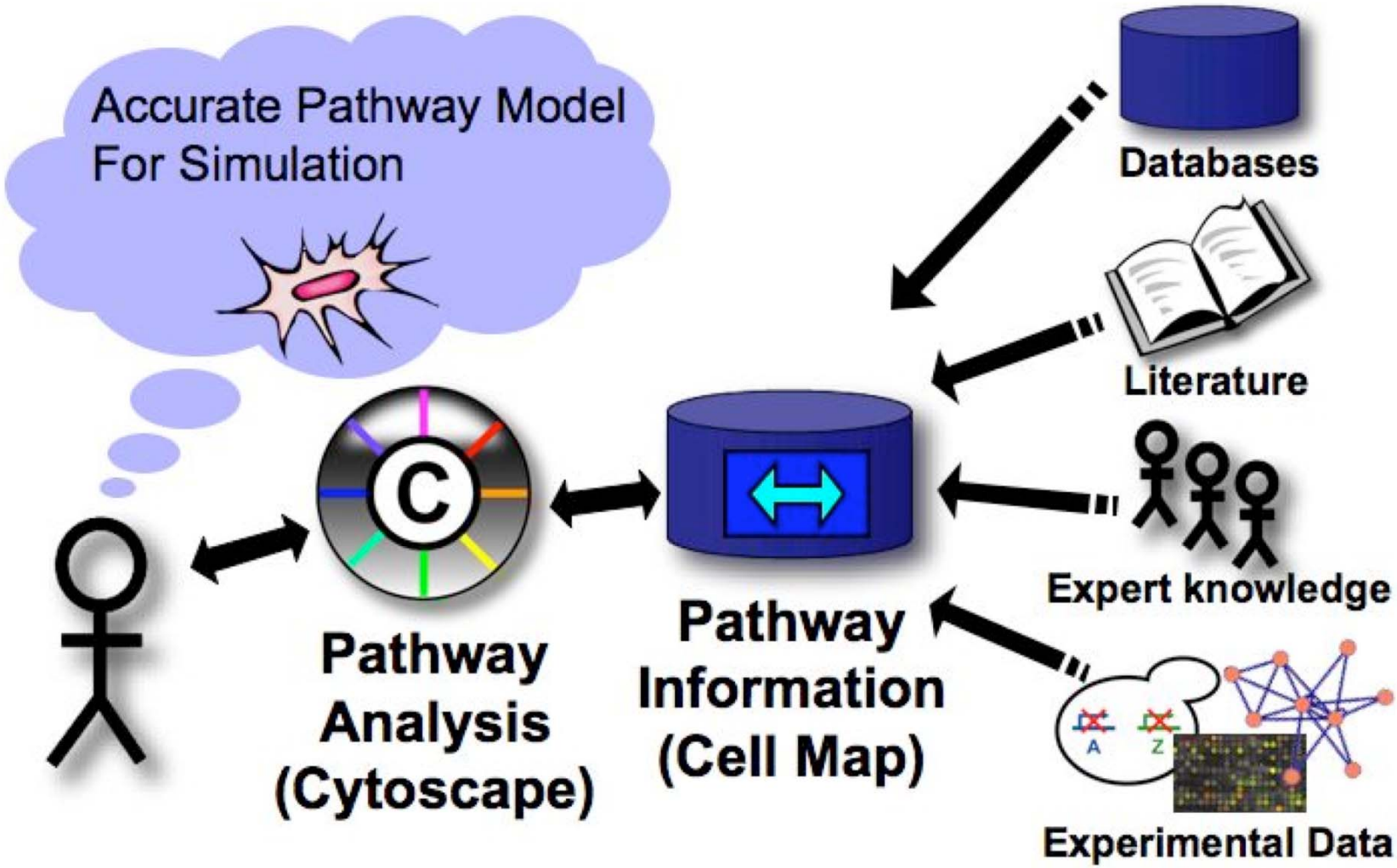
Aim: Convenient Access to Pathway Information



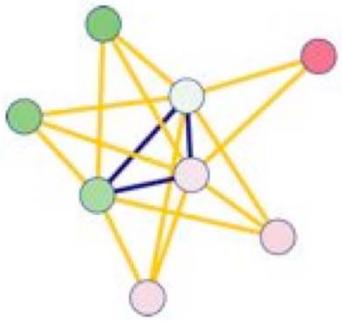
Facilitate creation and communication of pathway data
Aggregate pathway data in the public domain
Provide easy access for pathway analysis

Long term: Converge to integrated cell map

Using Pathway Information



Cytoscape - Network Visualization and Analysis

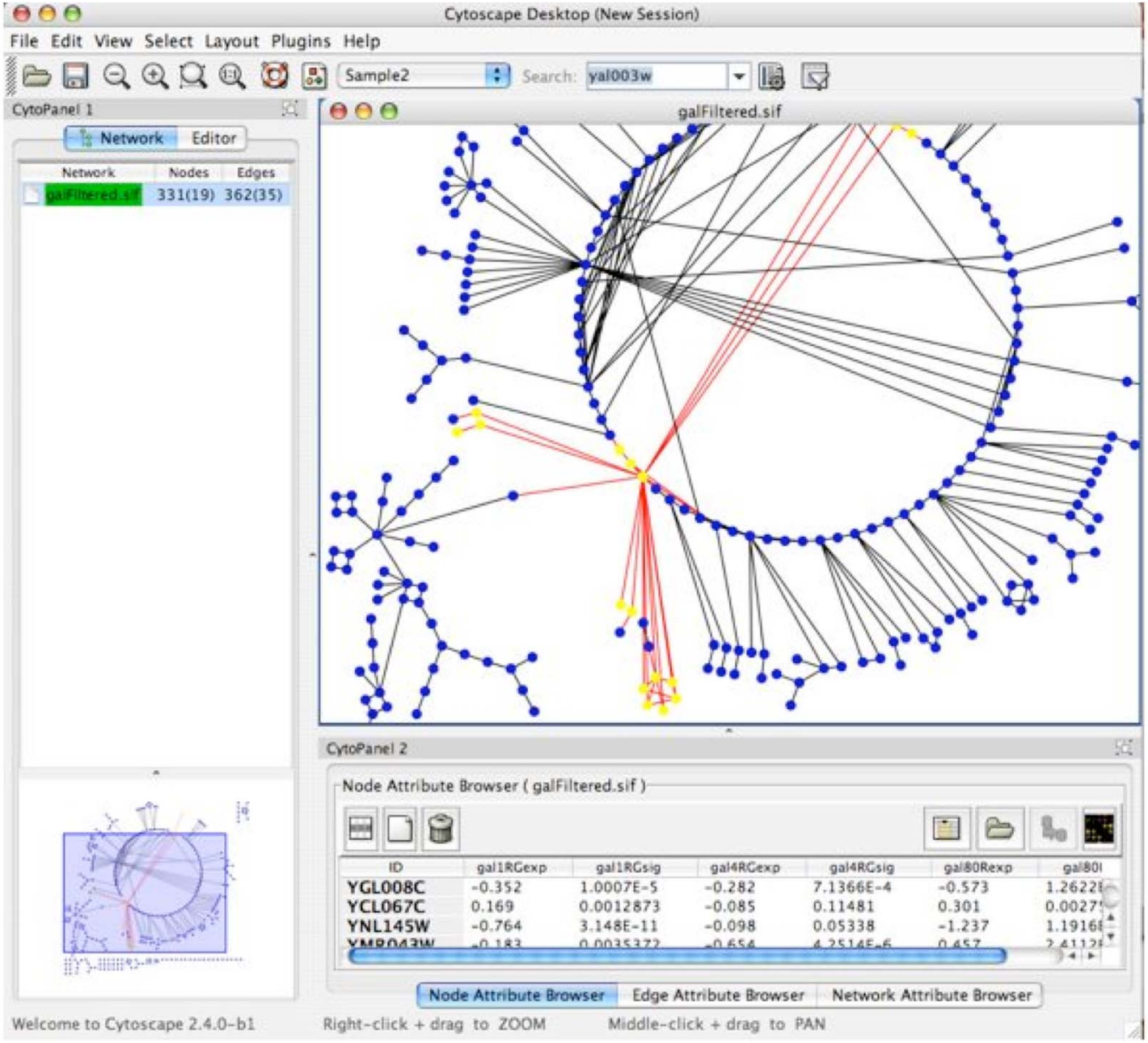


http://cytoscape.org

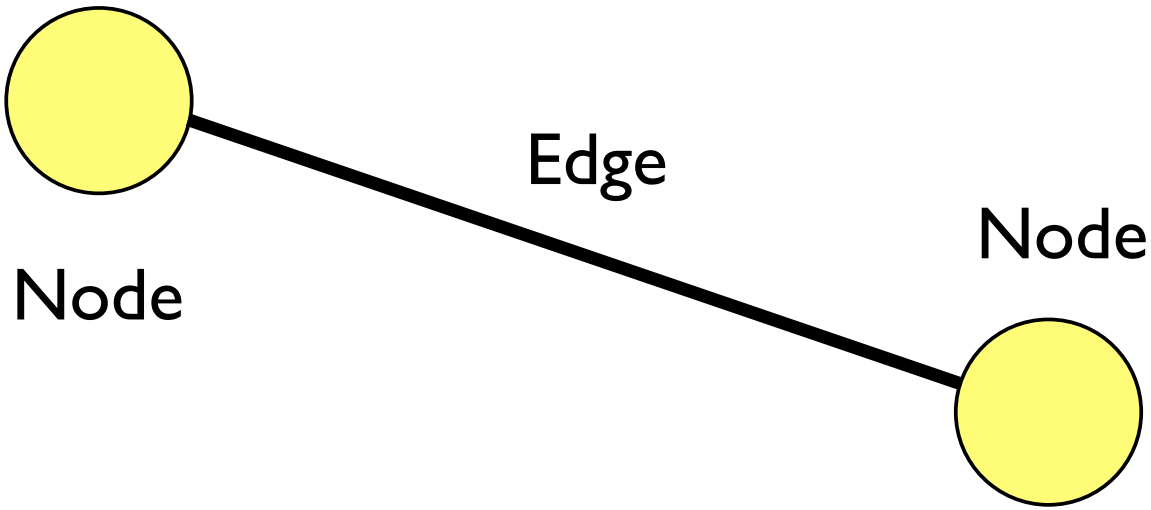
- Freely-available (open-source, java) software
- Visualizing biological networks (e.g. molecular interaction networks)
- Analyzing networks with gene expression profiles and other cell state data

UCSD, ISB, Agilent, MSKCC, Pasteur, UCSF, UToronto

Other software: Osprey, BioLayout, VisANT, Navigator, PIMWalker, ProViz


Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
Sample2
Search: yal003w
CytoPanel 1
Network
Editor
Network
Nodes
Edges
galFiltered.sif
331(19)
362(35)
galFiltered.sif
CytoPanel 2
Node Attribute Browser (galFiltered.sif)
ID
gal1RGexp
gal1RGsig
gal4RGexp
gal4RGsig
gal80Rexp
YGL008C
-0.352
1.0007E-5
-0.282
7.1366E-4
-0.573
YCL067C
0.169
0.0012873
-0.085
0.11481
0.301
YNL145W
-0.764
3.148E-11
-0.098
0.05338
-1.237
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.4.0-b1
Right-click + drag to ZOOM
Middle-click + drag to PAN

Pathway Graphs

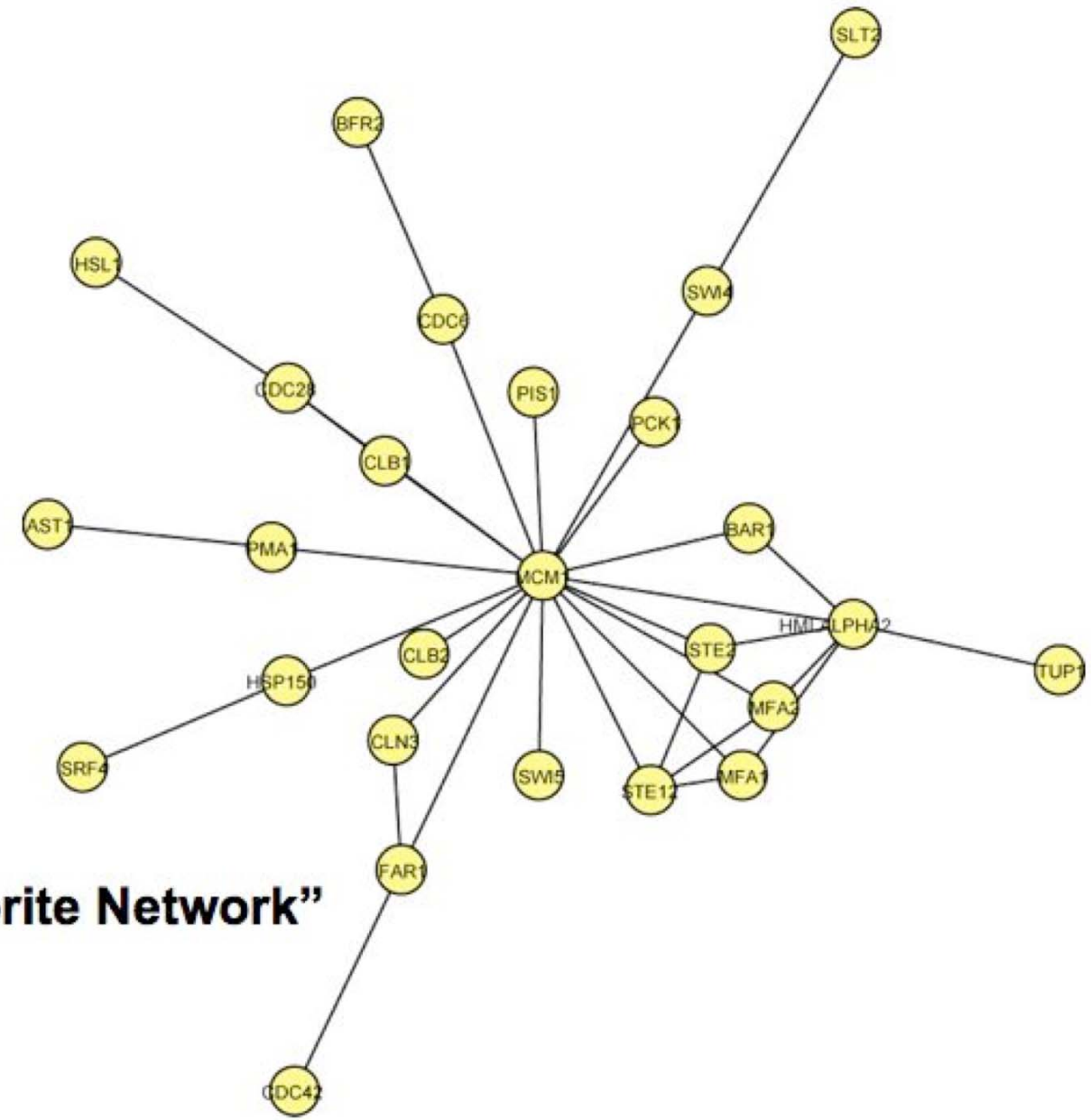


- In addition to describing the network topology, nodes and edges can each have their own attributes

Visual Style

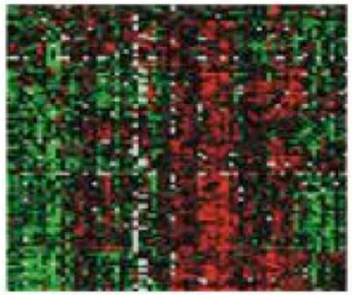
- Customized views of experimental data in a network context
- Network has node and edge attributes
  - E.g. expression data, interaction type, GO function
- Mapped to visual attributes
  - E.g. node/edge size, shape, colour…
- E.g. Visualize gene expression data as node colour gradient on the network

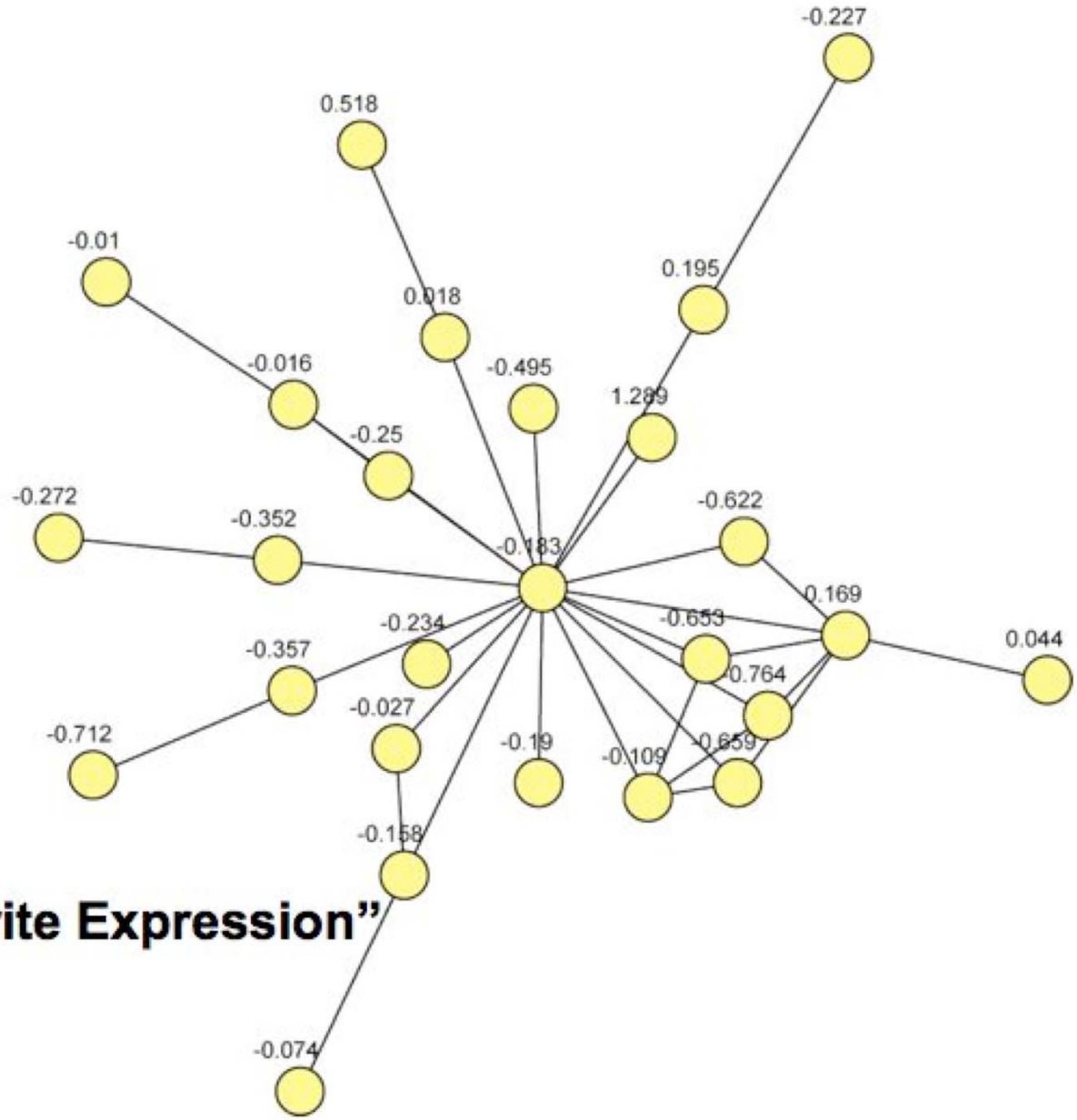
Visual Style



Load "Your Favorite Network"

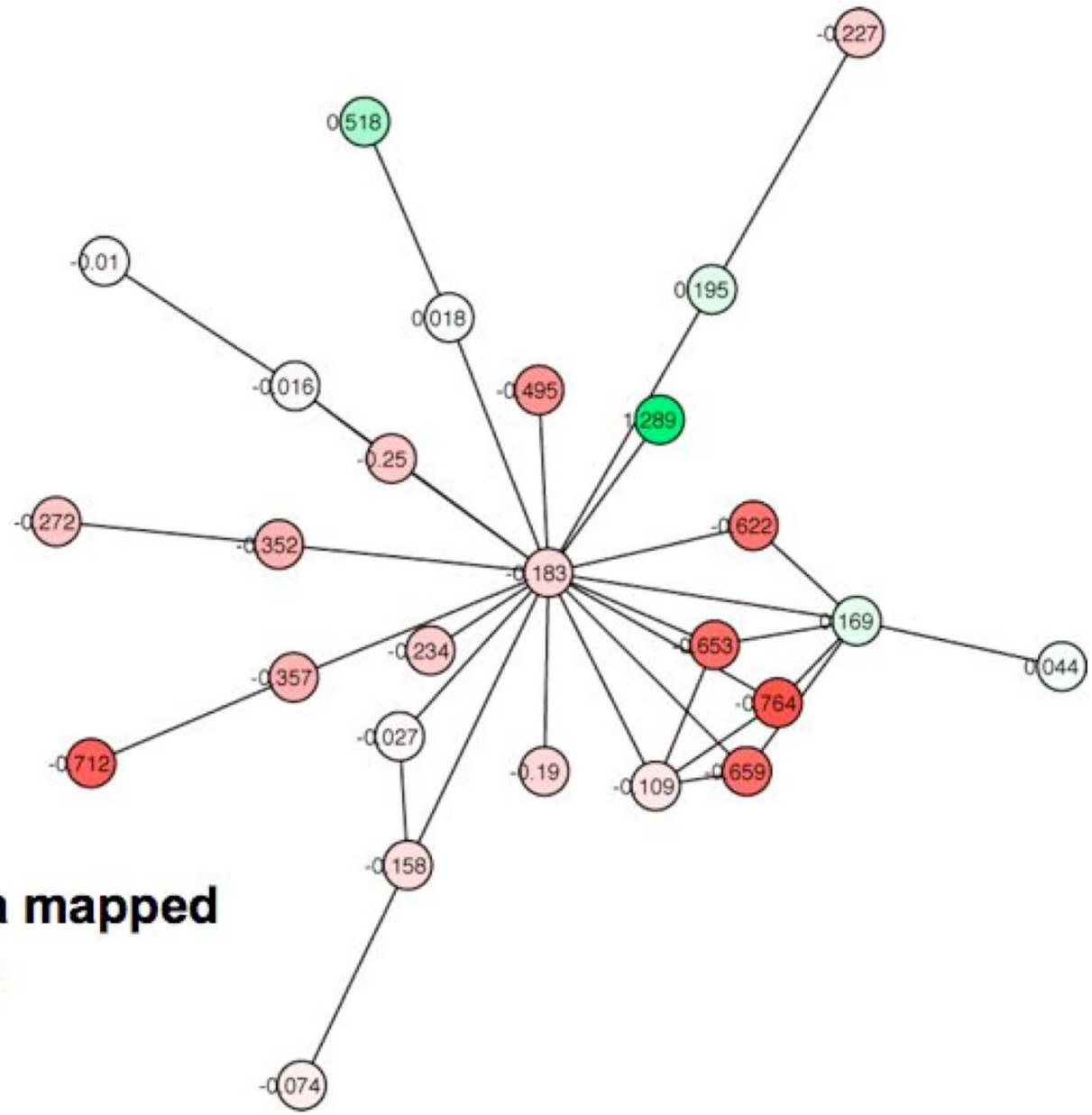
Visual Style





Load "Your Favorite Expression" Dataset

Visual Style



Expression data mapped to node colours

Systems Biology

- Goals:
  - ✓ integrating diverse data types, pathways
  - ✓ cellular simulations
- Community approaches:
  - ✓ pathguide, pathway commons, cytoscape
- Open data exchange key to success

Credits & References

- Dr. Gary Bader, DCCBR, UofT

  slides/images used with permission

- Cary MP, Bader GD, Sander C "Pathway Information for Systems Biology", FEBS Letters (2005)

Bioinformatics Links Directory

Finding online tools & resources for Life Sciences research

Nucleic Acids Research, 2007, Vol. 35, No. Web Server issue **W3–W5**
doi:10.1093/nar/gkm459

Conducting Research on the Web: 2007 Update for the Bioinformatics Links Directory

Joanne A. Fox, Scott McMillan and B. F. Francis Ouellette*

UBC Bioinformatics Centre (http://bioinformatics.ubc.ca/), I
Vancouver, British Columbia, Canada, V6T 1Z4

Received May 18, 2007; Accepted May 22, 2007

ABSTRACT

The Bioinformatics Links Directory, http://bioinformatics.ca/links_directory, is an actively maintained compilation of servers published in this and previous issues of *Nucleic Acids Research* issues together with many other useful tools, databases and resources for life sciences research. The 2007 update includes the 130 websites highlighted in the July 2007 Web Server issue of *Nucleic Acids Research* and brings the total number of servers listed in the Bioinformatics Links Directory to just under 1200 links. In addition to the updated content, the 2007 update of the Bioinformatics Links Directory includes new features for improved navigation, accessibility and open data exchange. A complete listing of all links listed in this *Nucleic Acids Research* 2007 Web Server issue can be accessed online at, http://bioinformatics.ca/links_directory/narweb2007. The 2007 update of the Bioinformatics Links Directory, which includes the Web Server list and summaries is also available online, at the *Nucleic Acids Research* web site, http://nar.oupjournals.org.

COMMENTARY

With the publication of the 2007 *Nucleic Acids Research* Web Server issue, we have a chance to reflect on how the web has transformed the way we conduct scientific

W2–W4 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue*
doi:10.1093/nar/gkn399

Keeping pace with the data: 2008 update on the Bioinformatics Links Directory

Michelle D. Brazas[1], Joanne A. Fox[2], Timothy Brown[1], Scott McMillan[3] and B. F. Francis Ouellette[1,*]

[1]Ontario Institute for Cancer Research, 101 College St, Suite 800, Toronto, Ontario, Canada M5G 0A3, [2]University of British Columbia, Michael Smith Laboratories and [3]University of British Columbia, Office of Learning and Technology, Vancouver, British Columbia, Canada

Received June 3, 2008; Revised and Accepted June 5, 2008

ABSTRACT

The Bioinformatics Links Directory, http://bioinformatics.ca/links_directory/, is an online resource for public access to all of the life science research web servers published in this and previous issues of *Nucleic Acids Research*, together with other useful tools, databases and resources for bioinformatics and molecular biology research. Dependent on community input and development, the Bioinformatics Links Directory exemplifies an open access research tool and resource. The 2008 update includes the 94 web servers featured in the July 2008 Web Server issue of *Nucleic Acids Research*, bringing the total number of servers listed in the Bioinformatics Links Directory to over 1200 links. A complete list of all links listed in this *Nucleic Acids Research* 2008 Web Server issue can be accessed online at http://bioinfomatics.ca/links_directory/narweb2008/. The 2008 update of the Bioinformatics Links Directory, which includes the Web Server list and summaries, is also available online at the *Nucleic Acids Research* website, http://nar.oxfordjournals.org/.

networks at play in a given disease or biological function, or ask questions that explore the commonalities and variations between large data sets from different macromolecules, species or organisms.

Keeping pace with these advances in technology and data output has been the number of specialized web servers and bioinformatic resources developed or upgraded to meet these new data intensive research needs. Since 2004, *Nucleic Acids Research* has peer-reviewed and published in their Web Server issue, a compendium of the latest web servers and freely available online bioinformatic tools to keep researchers abreast of the deluge of bioinformatic resources available to them. This year's Web Server issue introduces an additional 94 bioinformatics and molecular biology web servers, 10 of which are updates (Table 1). Along with the long-standing Database issue (1), the special Web Server issues represent an invaluable source of bioinformatic tools and resources for the international life-science research community. The complete listing of URLs cited in the 2008 Web Server issue can be accessed online at the *Nucleic Acids Research* website, http://nar.oxfordjournals.org/, as well as at http://bioinfomatics.ca/links_directory/narweb2008/.

The Bioinformatics Links Directory, http://bioinformatics.ca/links_directory/, is a public, curated collection of all of these servers together with other useful tools, databases and general purpose resources for bioinformatics and

http://bioinformatics.ca/links_directory/

Bioinformatics Links Directory

The Bioinformatics Links Directory features curated links to molecular resources, tools and databases. The links listed in this directory are selected on the basis of recommendations from bioinformatics experts in the field. We also rely on input from our community of bioinformatics users for suggestions. Starting in 2003, we have also started listing all links contained in the NAR Webserver issue.

Search Directory

Computer Related (64)

This category contains links to resources relating to programming languages often used in bioinformatics. Other tools of the trade, such as web development and database resources, are also included here.

Education (75)

Links to information about the techniques, materials, people, places, and events of the greater bioinformatics community. Included are current news headlines, literature sources, educational material and links to bioinformatics courses and workshops.

Human Genome (128)

This section contains links to draft annotations of the human genome in addition to resources for sequence polymorphisms and genomics. Also included are links related to ethical discussions surrounding the study of the human genome.

Model Organisms (204)

Included in this category are links to resources for various model organisms ranging from mammals to

DNA (441)

This category contains links to useful resources for DNA sequence analyses such as tools for comparative sequence analysis and sequence assembly. Links to programs for sequence manipulation, primer design, and sequence retrieval and submission are also listed here.

Expression (272)

Links to tools for predicting the expression, alternative splicing, and regulation of a gene sequence are found here. This section also contains links to databases, methods, and analysis tools for protein expression, SAGE, EST, and microarray data.

Literature (35)

Links to resources related to published literature, including tools to search for articles and through literature abstracts. Additional text mining resources, open access resources, and literature goldmines are also listed.

Other Molecules (15)

Bioinformatics tools related to molecules other than DNA, RNA, and protein. This category will include resources

- Main Page
- Citations
- Acknowledgements
- News
- Suggest URL
- NAR Collaboration
- RSS Feeds