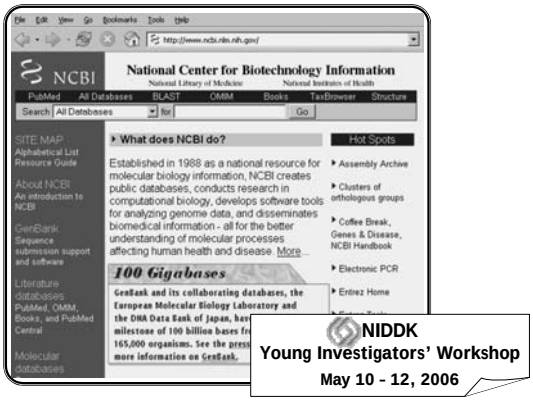


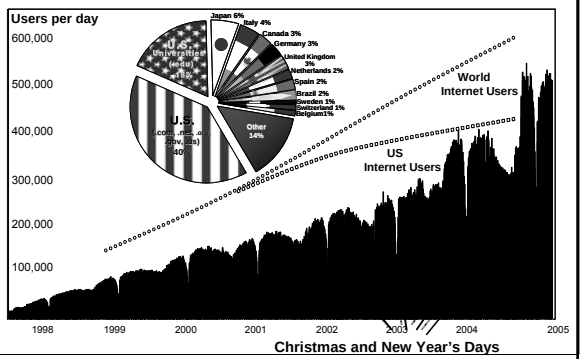
NIDDK
Young Investigators' Workshop
May 10 – 12, 2006

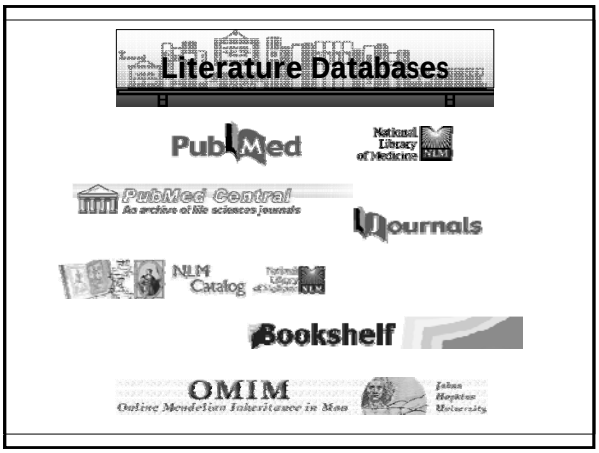
The National Center for Biotechnology Information

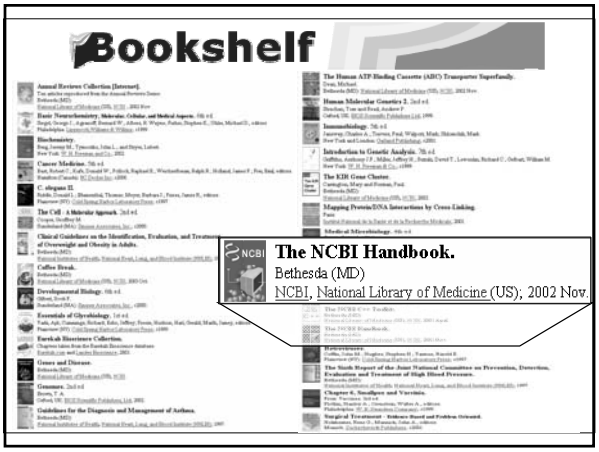
Created in 1988 as a part of the National Library of Medicine at NIH

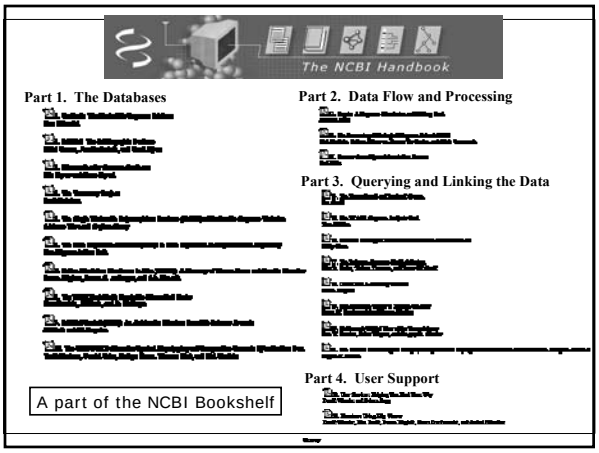
- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

NCBI Web Traffic

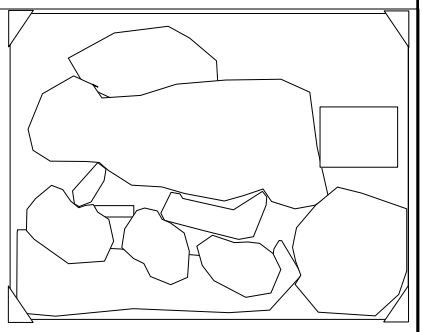


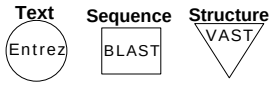

Literature Databases
PubMed
National Library of Medicine
PubMed Central
An archive of life sciences journals
Journals
NLM Catalog
Bookshelf
OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University


Bookshelf
The NCBI Handbook.
Bethesda (MD)
NCBI, National Library of Medicine (US); 2002 Nov.


The NCBI Handbook
Part 1. The Databases
Part 2. Data Flow and Processing
Part 3. Querying and Linking the Data
Part 4. User Support
A part of the NCBI Bookshelf

- **NCBI Databases**
  - **Literature**
  - **Sequence**
  - **Genome**
  - **Expression**
  - **Structure**
- **Summary pages for Information**
  - **Global Query**
  - **Taxonomy**
  - **Genome Project**
  - **Gene**
  - **UniGene**
  - **HomoloGene**
- **Search Methods**
  - **Entrez: by Text**
  - **BLAST: by Sequence**
  - **VAST: by Structure**
- **Pre-Computed Links**
  - **An Integrated Example**





Types of Molecular Databases

- **Primary Databases**
  - Raw and redundant Data…..submitted, "owned" and updated by experimentalists
    - Examples: GenBank, SNP, GEO, PubChem Substance & BioAssay
- **Derivative Databases**
  - *Human-curated* (compilation and curation of data)
    - Examples: GEO Datasets & Literature databases
  - *Computationally-Derived*
    - Examples: UniGene, HomoloGene, PubChem Compound
  - *Combination*
    - Examples: RefSeq, Gene, Genome Assembly, Conserved Domain and Structure databases

GenBank

Release 153	April 2006
56,620,500	Records
61,582,143,971	Nucleotides
>140,000	Species
224 Gigabytes	*944 files*

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

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

GenBank Divisions

PRI (30)	Primate
ROD (24)	Rodent
PLN (17)	Plant and Fungal
BCT (14)	Bacterial/Archeal
VRT (10)	Other Vertebrate
INV (8)	Invertebrate
VRL (5)	Viral
MAM (2)	Mammalian
PHG (1)	Phage
SYN (1)	Synthetic
UNA (1)	Unannotated

Traditional

- Direct Submissions (Sequin/Bankit)
- Accurate (~1 error per 10,000 bp)
- Well characterized
- Organized by taxonomy

EST (512)	Expressed Sequence Tag
GSS (174)	Genome Survey Sequence
HTG (81)	High Throughput Genomic
PAT (20)	Patent sequences
STS (14)	Sequence Tagged Site
HTC (10)	High Throughput cDNA
ENV (3)	Environmental Samples

Bulk

- From sequencing projects
- Batch submissions (ftp/Email)
- Inaccurate
- Poorly Characterized
- Organized by sequence type

Derivative Sequence Database

RefSeq

genomes·transcripts·proteins

RELEASE 18 IS NOW AVAILABLE ON THE FTP SITE!

- The curated "best representative" sequences
- Standardized nomenclature and record structure
- Added annotation (references, sequence features)

Genomes

mRNAs

Proteins

Genome

Entrez Genes

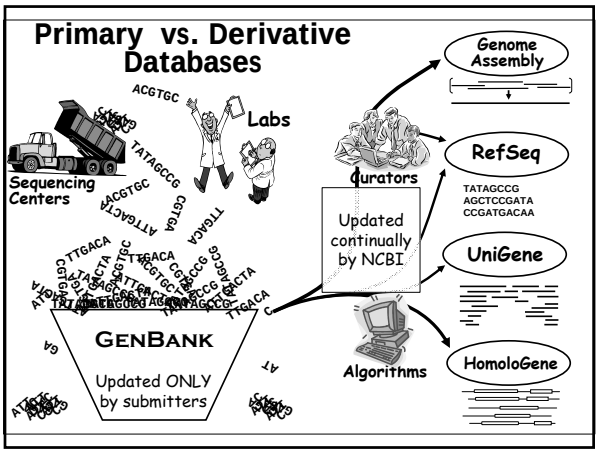
RefSeq Curation Processes

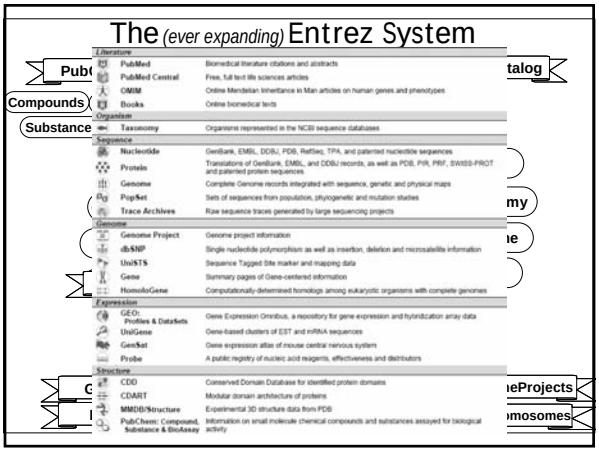
Curated genomic DNA (NC, NT, NW)

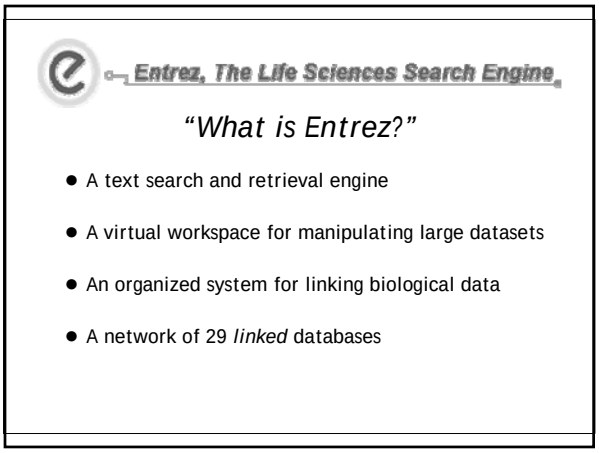
Scanning....

Curated Model mRNA (XM) (XR) → Model protein (XP)

Curated mRNA (NM) (NR) → Protein (NP)


Primary vs. Derivative Databases
Sequencing Centers
Labs
Curators
Updated continually by NCBI
GENBANK
Updated ONLY by submitters
Algorithms
Genome Assembly
RefSeq
TATAGCCG
AGCTCCGATA
CCGATGACAA
UniGene
HomoloGene


The (ever expanding) Entrez System
Compounds
Substance
Literature
PubMed
PubMed Central
OMIM
Books
Organism
Taxonomy
Sequence
Nucleotide
Protein
Genome
PopSet
Trace Archives
Genome
Genome Project
dbSNP
UniSTS
Gene
HomoloGene
Expression
GEO: Profiles & DataSets
UniGene
GenSat
Probe
Structure
CDD
CDART
MMDB/Structure
PubChem: Compound, Substance & BioAssay


Entrez, The Life Sciences Search Engine
"What is Entrez?"
A text search and retrieval engine
A virtual workspace for manipulating large datasets
An organized system for linking biological data
A network of 29 linked databases

Essential Components of Entrez Databases

- All Molecular Database entries are organized by organism *(Taxonomy Database)*.
- Each record is assigned a UID.
  - A "unique integer identifier" for internal tracking
- Each record is indexed by data fields.
  - [author], [title], [organism], and many others
- Each record is given a Document Summary.
  - a summary of the record's content (DocSum)
- Each record is manually or computationally assigned links to biologically related UIDs in and across databases.

Searching the NCBI Databases

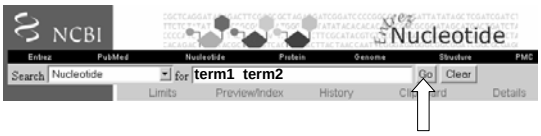
Text — Entrez

Sequence — BLAST

Structure — VAST

Entrez, The Life Sciences Search Engine

How to Query a Particular Database



(term1[*tag delimiter*] op term2[*tag delimiter*] op …)

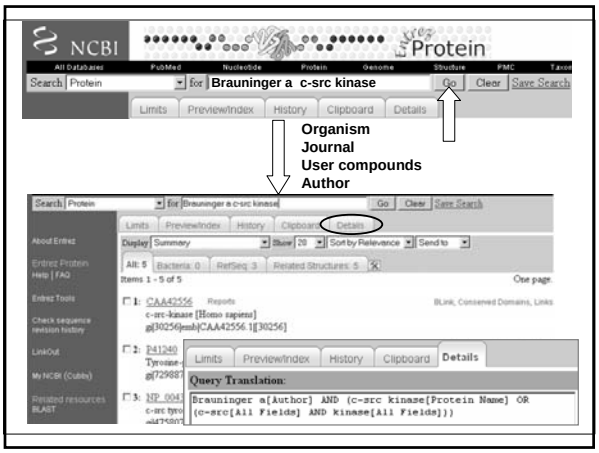
op = AND, OR, NOT

Boolean operators MUST be in ALL CAPS!

tag delimiter = Entrez indexing field

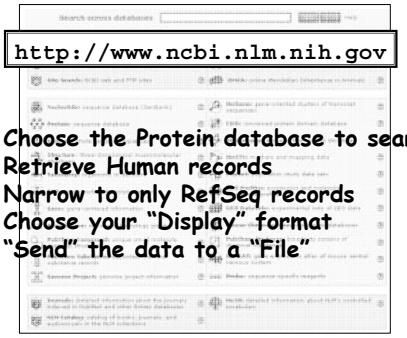
Organism
Journal
User compounds
Author

Examples of tag delimiters



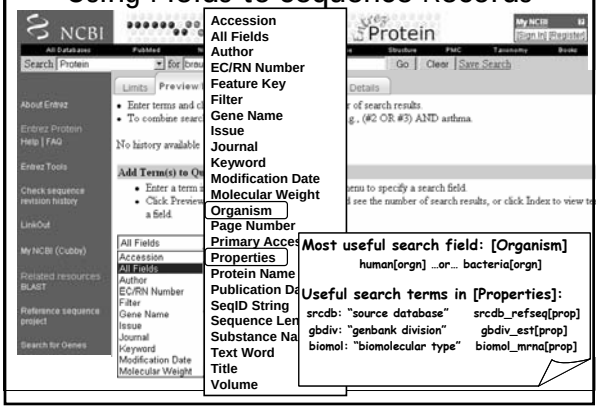
EXAMPLE: Searching with Entrez

"How do I download the Human Proteome?"



1. **Choose the Protein database to search**
2. **Retrieve Human records**
3. **Narrow to only RefSeq records**
4. **Choose your "Display" format**
5. **"Send" the data to a "File"**

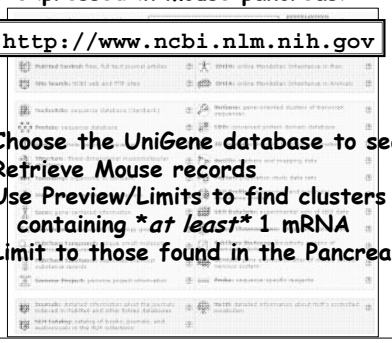
Using Fields to Sequence Records



EXAMPLE: More Complicated Entrez Searching

"How do I find a information on genes expressed in mouse pancreas?"

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



1. **Choose the UniGene database to search**
2. **Retrieve Mouse records**
3. **Use Preview/Limits to find clusters containing **at least** 1 mRNA**
4. **Limit to those found in the Pancreas**

Other Advanced Queries

Nucleotide: Non-genomic sequences from the PLN division of Genbank

gbdiv_pln [properties] NOT biomol_genomic [properties]

Protein: RefSeq sequences with molecular weights of 80 to 100 kDa

srcdb_refseq [properties] AND 080000:100000 [Molecular Weight]

SNP: True SNPs that are uniquely mapped on the mouse genome

Snp [SNP Class] AND 1 [Map Weight] AND mouse [organism]

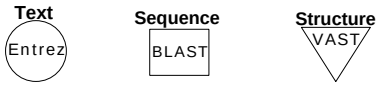
UniSTS: Markers on the Genethon map of human chromosome 12

Genethon [Map Name] AND human [organism] AND 12 [chromosome]

Structure: Structures of bacterial kinases with resolutions below 2 Å

Bacteria [organism] AND kinase AND 000.00:002.00 [resolution]

Searching the NCBI Databases



BLAST

Searching with Sequences

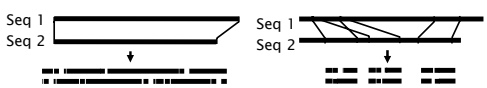
Why do we need sequence similarity searching?

- To identify and annotate sequences with…
  - incomplete (or no) annotations (GenBank)
  - incorrect annotations
- To assemble genomes
- To explore evolutionary relationships by…
  - finding homologous molecules
  - developing phylogenetic trees

NOTE: Similar sequences may NOT have similar function!

Basic Local Alignment Search Tool

- Calculates similarity for biological sequences
- Finds best *local* alignments



Global Alignment ***Local Alignment***

- A Heuristic approach *based on* the Smith-Waterman algorithm
  - Searches for matching "words" (W) rather than individual residues
  - Uses statistical theory to determine *if a match might have occurred by chance*

How BLAST Works

Query:

GTACTGGACATGGACCCTACAGGAA

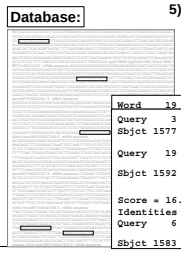
1) Break query sequence up into "Words" and create a "Lookup Table".

2) Compare "Words" to those in pre-formatted Database files creating a collection of "Word Matches".

Lookup Table:

GTACTGGACAT	1..12
TACTGGACATG	2..13
ACTGGACATGG	3..14
CTGGACATGGA	4..15
TGGACATGGAC	5..16
GGACATGGACC	6..17
GACATGGACCC	7..18
ACATGGACCCT	8..19
CATGGACCCTA	9..20
ATGGACCCTAC	10..21
TGGACCCTACA	11..22
GGACCCTACAG	12..23
GACCCTACAGG	13..24
ACCCTACAGGA	14..25
CCCTACAGGAA	15..26

Database:



3) Individually compare the "Word Matches" in an "ungapped extension" until similarities begin to "drop off", then calculate bit score. Save a list of "High Scoring Pairs" (HSPs) with their positions and bit scores.

4) Repeat comparison with "gapped extension". Save a new list of HSPs with their positions and bit scores.

5) Finally, repeat comparison through a last "gapped extension phase with traceback" where actual sequence alignments are generated, the final bit score calculated, then the E-value is generated.

```
Word     19                ACTGGACATGG                    29
Query     3 ATCTGTGCTCTGTGAACTGGACATGGAACCTACAAAAGCTGAA 46
Sbjct 1577 GAAGCGTCGGTCTGTAACTGGACATGGACCCAATTGTTTATGGA 1620

Query    19                ACTGGACATGGAACCTA             35
                           ||||||||||| || |
Sbjct 1592                 ACTGGACATGGACCCAA             1608

Score = 16.6 bits (12), Expect = 0.052
Identities = 24/29 (82.8%), Gaps = 2/29 (6.9%)
Query     6    GTTGCTCTGTGAACTGGACATGGAACCTA             35
               || |||||||  ||||||||||| || |
Sbjct 1583     GTCGGTCTGT--ACTGGACATGGACCCAA             1608
```

Minimum Requirements for a Hit

Nucleotide-based BLAST

```
                    ATCGCCATGCTTAATTGGGCTT
exact word match    CATGCTTAATT
                    one word match
```

Minimum word size = 7
BLASTn default = 11
megaBLAST default = 28

Protein-based BLAST

```
                    GTQITVEDLFYNI
neighborhood words  SEI      YYN
                    two word matches within 40 residues
```

Word size = 2 or 3 *(default = 3)*

An alignment that BLAST can't find

```
  1 GAATATATGAAGACCAAGATTGCAGTCCTGCTGGCCTGAACCACGCTATTCTTGCTGTTG
    || | || || || |  || || ||   ||  |  ||| |||||| | | || | ||| |
  1 GAGTGTACGATGAGCCCGAGTGTAGCAGTGAAGATCTGGACCACGGTGTACTCGTTGTCG

 61 GTTACGGAACCGAGAATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAGCCAGTT
    | || ||    ||  ||| ||  | |||||| || | |||||| |||||  |     |
 61 GCTATGGTGTTAAGGGTGGGAAGAAGTACTGGCTCGTCAAGAACAGCTGGGCTGAATCCT

121 GGGGTGAACAAGGTTATTTCAGGCTTGCTCGTGGTAAAAAC
    |||| || ||||| ||  ||   |  | ||||  || |||
121 GGGGAGACCAAGGCTACATCCTTATGTCCCGTGACAACAAC
```

Here there are no words longer than 6…

...for nucleotides
there must be an exact match of at least 7.

Scoring Systems - Nucleotides

Identity matrix

	A	G	C	T
A	+1	–3	–3	-3
G	–3	+1	–3	-3
C	–3	–3	+1	-3
T	–3	–3	–3	+1

Simply add the scores
for each pair of aligned residues
and, as necessary, factor in the gaps!

```
CAGGTAGCAAGCTTGCATGTCA
|| |||||||||||  |||||    raw score = 19-9 = 10
CACGTAGCAAGCTTG-GTGTCA
```

Scoring Systems - Proteins

Position Independent Matrices

PAM Matrices (Percent Accepted Mutation)

- **Derived from observation; small dataset of alignments**
- **Implicit model of evolution**
- **All calculated from PAM1**
- **PAM250 widely used**

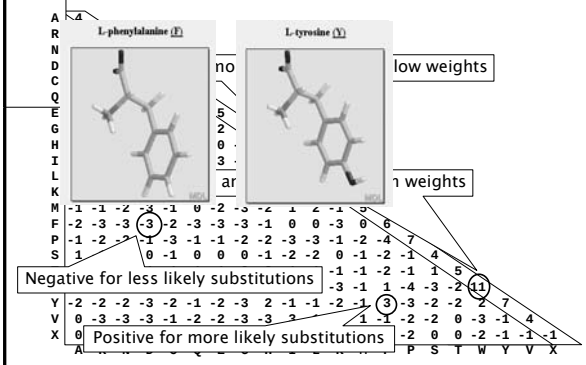
BLOSUM Matrices (BLOck SUbstitution Matrices)

- **Derived from observation; large dataset of highly conserved blocks**
- **Each matrix derived separately from blocks with a defined percent identity cutoff**
- ***BLOSUM62* - default matrix for BLAST**

Position Specific Scoring Matrices (PSSMs)

PSI- and RPS-BLAST

BLOSUM62 Substitution Matrix



Scores

Simply add the scores for each pair of aligned residues and, as necessary, factor in the gaps!

	V	D	S	–	C	Y		
	V	E	T	L	C	F		
BLOSUM62	+4	+2	+1	-12	+9	+3	=	7
PAM30	+7	+2	0	-10	+10	+2	=	11

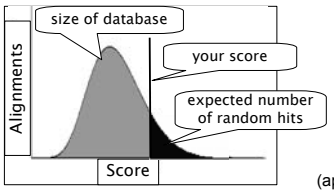
Different matrices produce different scores!

Local Alignment Statistics

High scores of local alignments between two random Sequences follow the Extreme Value Distribution.

Expect Value

E = number of database hits you expect to find by chance



$E = Kmne^{-\lambda S}$

$E = mn2^{-S'}$

K = scale for search space
λ = scale for scoring system
S' = bitscore = (λS - lnK)/ln2

(applies to ungapped alignments)

Nucleotide vs. Protein BLAST

Comparing ADSS from *H. sapiens* and *A. thaliana*

```
            aaccggtgacggtggtgctcggtgcgcagtggggcgacgaaggc
H.sapiens:   N  R  V  T  V  V  L  G  A  Q  W  G  D  E  G

A.thaliana: S  Q  V  S  G  V  L  G  C  Q  W  G  D  E  G
            agtcaagtatctggtgtactcggttgccaatggggagatgaaggt
```

BLASTn finds no match, because there are no 7 bp words
BLASTp finds three matching words

Protein searches are generally more sensitive than nucleotide searches.

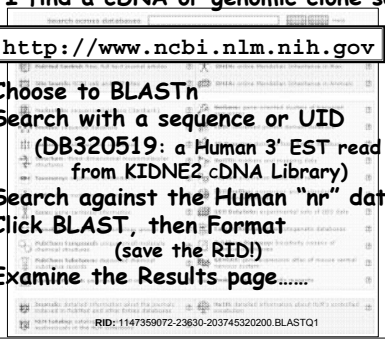
N ucleotide P rotein

Some Flavors of BLAST

Program	Query	Database
blastn	N	N
blastp	P	P
blastx	N ⇨ PPP / ddd	P
tblastn	P	N ⇨ PPP / ddd
tblastx	N ⇨ PPP / ddd	N ⇨ PPP / ddd

EXAMPLE: Searching with BLASTn

"How do I figure out what an EST sequence is and can I find a cDNA or genomic clone source?"

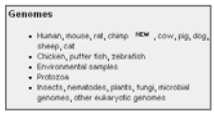


`http://www.ncbi.nlm.nih.gov`

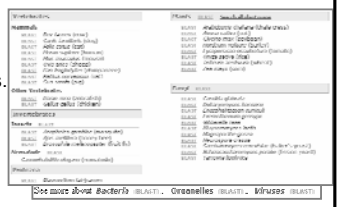
1. **Choose to BLASTn**
2. **Search with a sequence or UID (DB320519: a Human 3' EST read from KIDNE2 cDNA Library)**
3. **Search against the Human "nr" database**
4. **Click BLAST, then Format (save the RID!)**
5. **Examine the Results page……**

Genomic BLAST

- The BLAST homepage links to the Genome BLAST pages provide customized nucleotide and protein databases for each genome.

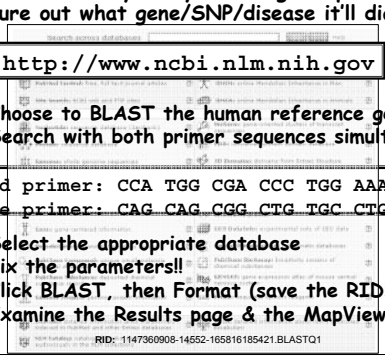


- If a Map Viewer is available, the BLAST hits can be viewed on the maps.



EXAMPLE: Searching with genome BLASTn

"How do I test/analyze my human gPCR primer pair? Can I figure out what gene/SNP/disease it'll diagnose?"



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

1. **Choose to BLAST the human reference genome**
2. **Search with both primer sequences simultaneously**

```
Forward primer: CCA TGG CGA CCC TGG AAA AGC
Reverse primer: CAG CAG CGG CTG TGC CTG CGG
```

3. **Select the appropriate database**
4. **Fix the parameters!!**
5. **Click BLAST, then Format (save the RID!)**
5. **Examine the Results page & the MapView……**

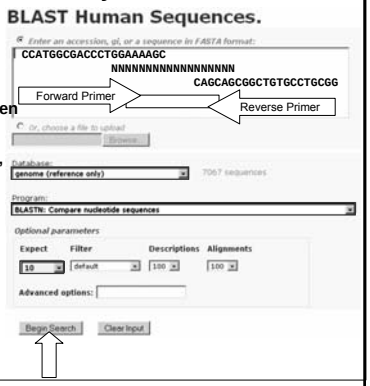
PCR Primer Set Analysis with BLAST

1) Type in primer sequences with an "N" spacer between

2) Choose your database: "genome" or "Ref mRNA"

3) Select BLASTN! *(Advanced options: "-W 7")*

4) Increase the Expect value threshold to >1.



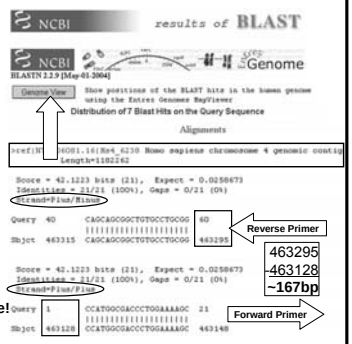
PCR Primer Set Analysis with BLAST

1) View the graphic:
   Find primer "annealing" sites.
   Find hits to a single record.

2) Examine the alignments.

3) Identify potential amplicon:
   Record's accession number
   Verify forward & reverse priming
   Calculate approx. product size from hit positions

4) Look at the hits on the Genome!

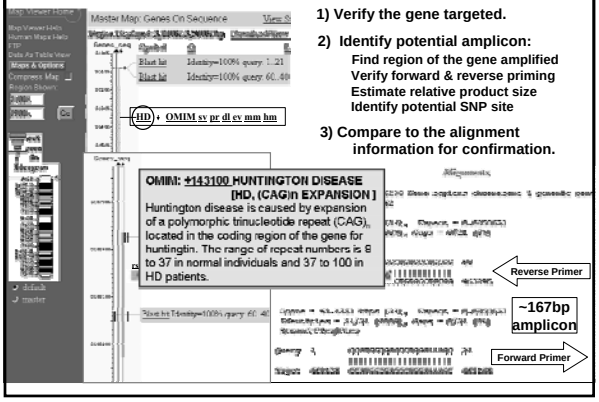


PCR Primer Set Analysis with BLAST

1) Verify the gene targeted.

2) Identify potential amplicon:
   Find region of the gene amplified
   Verify forward & reverse priming
   Estimate relative product size
   Identify potential SNP site

3) Compare to the alignment information for confirmation.

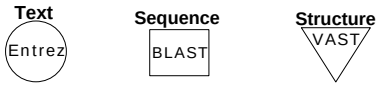


Other Specialized BLASTing

- PSI BLAST
- Reverse-PSI BLAST (CDD Search)
- PHI BLAST (skip)
- VecScreen
- BLAST2Seq
- GEO BLAST
- SNP BLAST
- ePCR (?)

WORK IT UP!!

Searching the NCBI Databases



In-lieu-of BLAST

Precomputed BLASTn&p Services

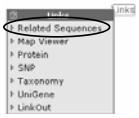
If you don't want to run a BLAST search…
We've already done some for you!

- Precomputed Entrez Links:
  "Related Sequences"
- The Transcript Cluster Database:
  "UniGene"
- The Protein Homolog Database:
  "HomoloGene"
- The **"Conserved Domain"** Database &
  CDART a.k.a. **"Domain Relatives"**
- The Multifunctional Blast Link:
  "BLink"

Related Sequences:
Precomputed BLASTn & BLASTp Lists

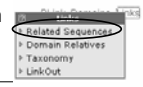
☐ 1: AF395536
Homo sapiens sorting nexin 18 (SNX18) mRNA, complete cds
gi|15042690|gb|AF395536.1|AF395536[15042690]

Nucleotide



☐ 1: Q96RF0
Sorting nexin 18 (Sorting nexin associated Golgi protein 1)
gi|20140146|sp|Q96RF0|SNXI_HUMAN[20140146]

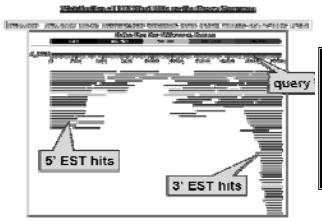
Protein



The "Related Sequences" link retrieves GenBank/GenPept sequences sorted by BLAST score, but with no alignment details.

UniGene
ORGANISED VIEW OF THE TRANSCRIPTOME

Summary pages of sequence and expression information for sets of expressed sequences clustered by BLAST

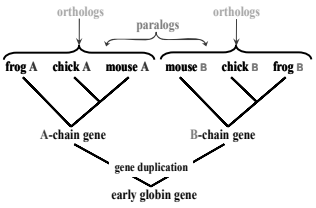


UniGene Cluster Hs.XXXXX *Homo sapiens*
SEE ALSO: LocusLink | OMIM | HomoloGene
MODEL ORGANISM PROTEIN SIMILARITIES
MAPPING INFORMATION
EXPRESSION INFORMATION
mRNA SEQUENCES
EST SEQUENCES
PC Download sequences

HoMologεNε

Clusters of homologous protein sequences based on BLASTp (also guided by taxonomic tree information)

- Includes orthologs and paralogs -

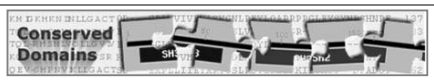


HomoloGene Release Statistics

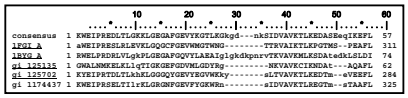
Initial numbers of genes from complete genomes, numbers of genes placed in a homology group, and the numbers of groups for each species

HomoloGene Build 36

Species	Number of Genes input	Number of Genes grouped	HomoloGene Groups
H.sapiens	22,827	18,055	16,782
M.musculus	24,019	19,996	18,036
R.norvegicus	20,913	17,429	16,042
D.melanogaster	12,918	8,717	7,683
A.gambiae	12,012	8,543	7,577
C.elegans	19,109	6,502	5,260
S.pombe	4,947	3,625	3,359
S.cerevisiae	5,863	3,612	3,146
N.crassa	10,079	6,156	6,049
M.grisea	11,109	6,307	6,028
A.thaliana	26,281	8,022	4,791
P.falciparum	5,222	1,770	1,589



Conserved sequence elements that perform common functions curated from multiple sequence alignments with similar function



(Position-Specific Scoring Matrices)

CD	NCBI
Pfam	Sanger
SMART	EMBL
COG	NCBI

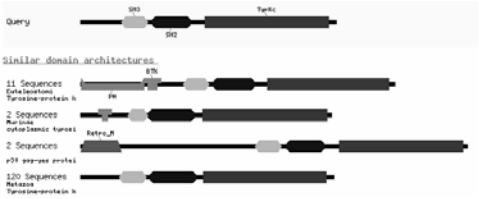
"Reverse-Position Specific" Sequence Comparisons (RPS-BLAST**)**
a.k.a. **"Conserved Domain Database" (**CDD**) Search**

CDART:
Conserved Domain Architecture Retrieval Tool

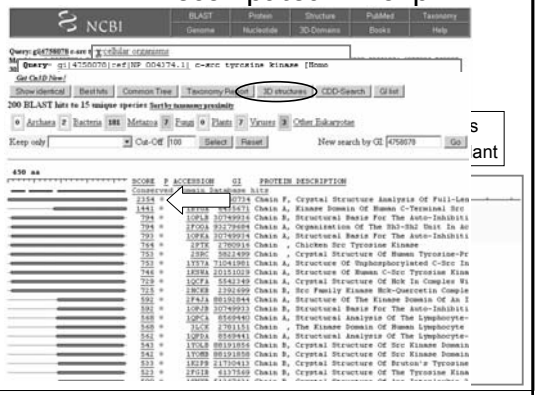
Modular Architecture of Domains

Cartoon descriptions of relative protein domain organization

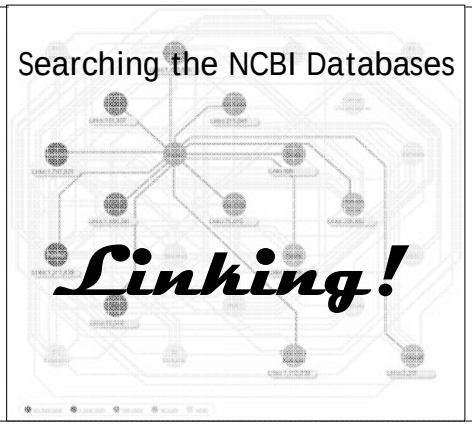
Allows for comparison with other proteins with the same domain



BLink: Precomputed BLASTp

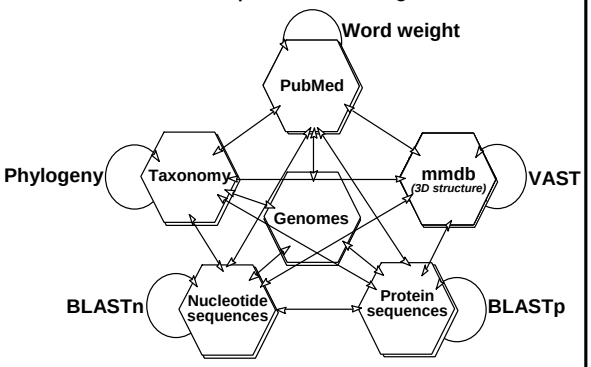


Searching the NCBI Databases

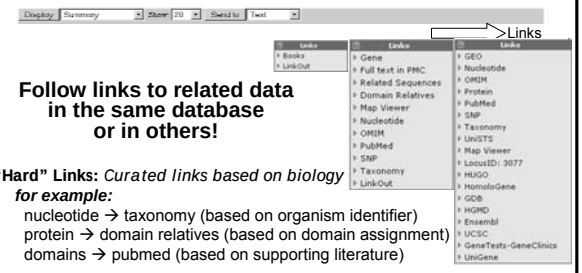


Linking!

Examples of Linking



Following Links



Follow links to related data in the same database or in others!

"Hard" Links: *Curated links based on biology*
for example:
- nucleotide → taxonomy (based on organism identifier)
- protein → domain relatives (based on domain assignment)
- domains → pubmed (based on supporting literature)

"Soft" Links: *Pre-computed analyses*
for example:
- nucleotide → related sequences (BLAST neighbors)
- protein → conserved domains (CDD/RPS-BLAST search)
- gene → map viewer (map position of annotated gene)

Entrez Gene

Summary pages of curated information (LINKS!) for genetic loci of RefSeq organisms.

- ►Graphics
- ►Gene information
- ►Bibliography (PubMed links)
- ►General gene information
- ►NCBI Reference Sequences
- ►Related sequences
- ►Additional Links

Entrez Genes

H.sapiens & B.taurus
G6PD

R.norvegicus
G6pdx

M.musculus
G6pd1, G6pdx

D.melanogaster
Zw

A.thalia
At5g35970

S.pombe
SPAC3C7.13c
SPAC9.01
SPCC794.01c

B.anthracis
BA_3932

H.pylori
HP1101

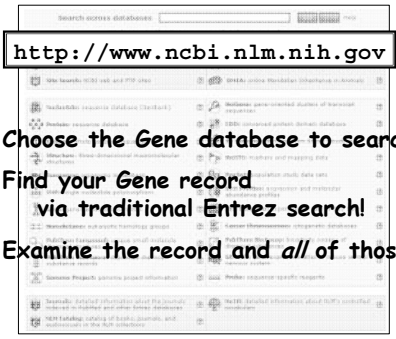
zwf
E.coli, Salmonella, Shigella, Yersinia, Neisseria….

glucose-6-phosphate dehydrogenase

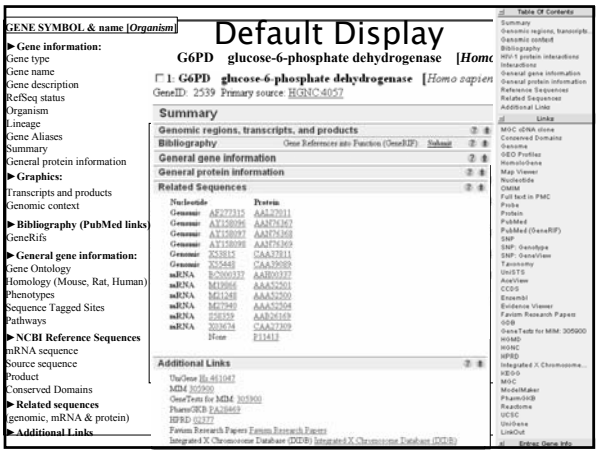
EXAMPLE: Finding Information on your Gene

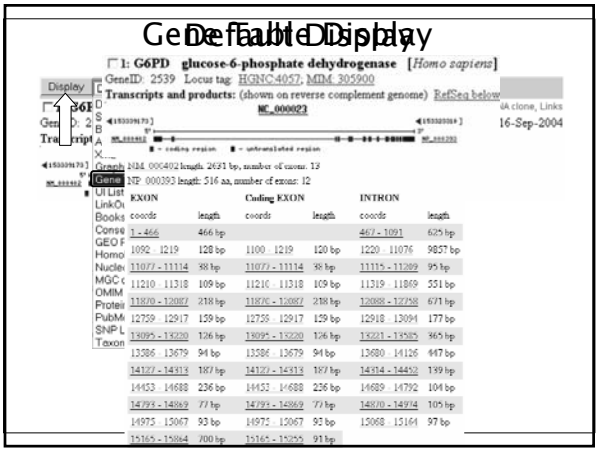
"How do I download the Human Proteome?"

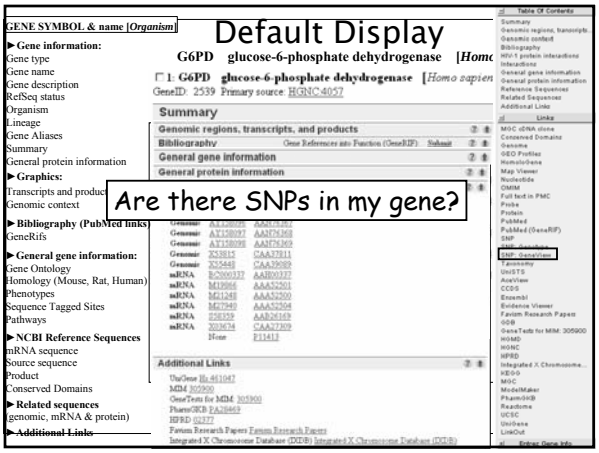
http://www.ncbi.nlm.nih.gov

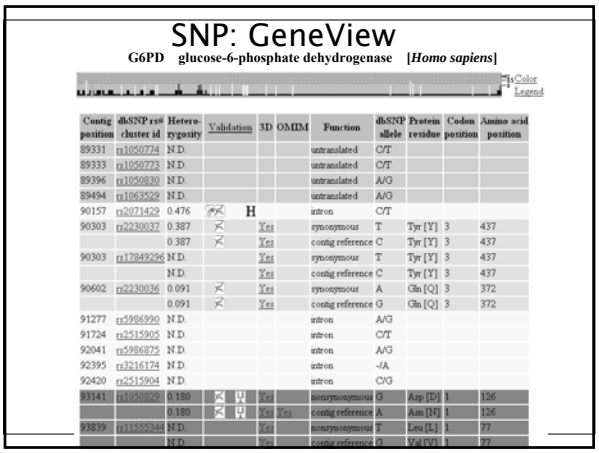


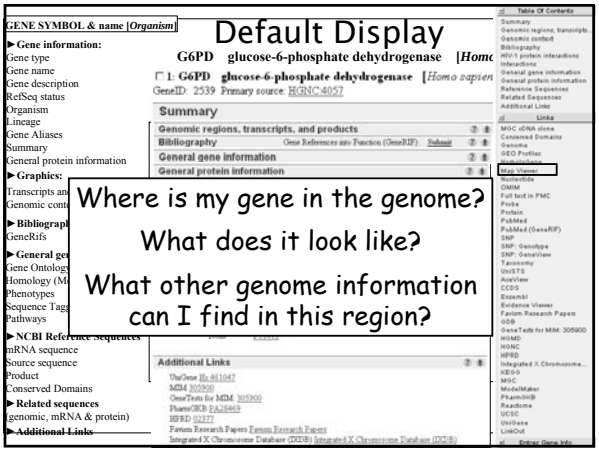
1. **Choose the Gene database to search**
2. **Find your Gene record via traditional Entrez search!**
3. **Examine the record and *all* of those links!**

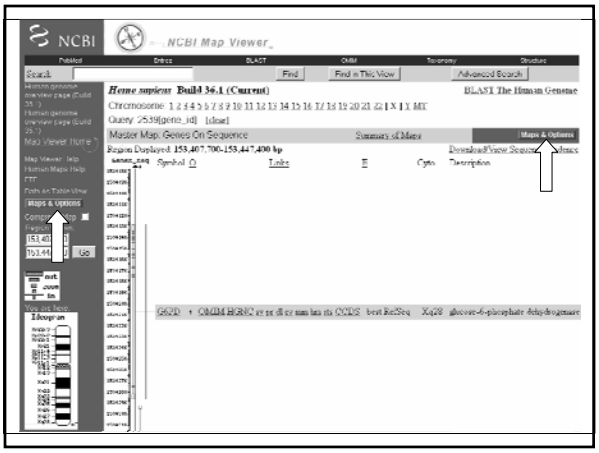

GENE SYMBOL & name [Organism]
Default Display
G6PD glucose-6-phosphate dehydrogenase [Homo
1: G6PD glucose-6-phosphate dehydrogenase [Homo sapien
GeneID: 2539 Primary source: HGNC:4057
Summary
Genomic regions, transcripts, and products
Bibliography
General gene information
General protein information
Related Sequences
Additional Links
Gene information:
Gene type
Gene name
Gene description
RefSeq status
Organism
Lineage
Gene Aliases
Summary
General protein information
Graphics:
Transcripts and products
Genomic context
Bibliography (PubMed links)
GeneRifs
General gene information:
Gene Ontology
Homology (Mouse, Rat, Human)
Phenotypes
Sequence Tagged Sites
Pathways
NCBI Reference Sequences
mRNA sequence
Source sequence
Product
Conserved Domains
Related sequences
(genomic, mRNA & protein)
Additional Links


Default Display
1: G6PD glucose-6-phosphate dehydrogenase [Homo sapiens]
GeneID: 2539 Locus tag: HGNC:4057, MIM: 305900
Transcripts and products: (shown on reverse complement genome) RefSeq below
NC_000023
16-Sep-2004
Display
Graph
Gene
EXON
Coding EXON
INTRON
coords
length

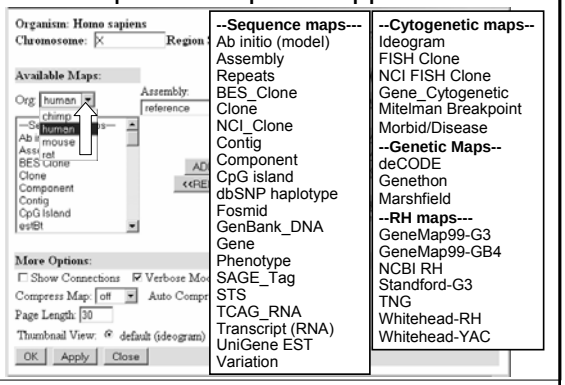

GENE SYMBOL & name [Organism]
Default Display
G6PD glucose-6-phosphate dehydrogenase [Homo
1: G6PD glucose-6-phosphate dehydrogenase [Homo sapien
GeneID: 2539 Primary source: HGNC:4057
Summary
Genomic regions, transcripts, and products
Bibliography
General gene information
General protein information
Are there SNPs in my gene?
Additional Links
Gene information:
Gene type
Gene name
Gene description
RefSeq status
Organism
Lineage
Gene Aliases
Summary
General protein information
Graphics:
Transcripts and products
Genomic context
Bibliography (PubMed links)
GeneRifs
General gene information:
Gene Ontology
Homology (Mouse, Rat, Human)
Phenotypes
Sequence Tagged Sites
Pathways
NCBI Reference Sequences
mRNA sequence
Source sequence
Product
Conserved Domains
Related sequences
(genomic, mRNA & protein)
Additional Links


SNP: GeneView
G6PD glucose-6-phosphate dehydrogenase [Homo sapiens]
Color
Legend
Contig position | dbSNP rs# cluster id | Hetero-zygosity | Validation | 3D | OMIM | Function | dbSNP allele | Protein residue | Codon position | Amino acid position
89331 rs1050774 N.D. untranslated C/T
89333 rs1050773 N.D. untranslated C/T
89396 rs1050830 N.D. untranslated A/G
89494 rs1063529 N.D. untranslated A/G
90157 rs2071429 0.476 H intron C/T
90303 rs2230037 0.387 Yes synonymous T Tyr [Y] 3 437
0.387 Yes contig reference C Tyr [Y] 3 437
90303 rs17849296 N.D. Yes synonymous T Tyr [Y] 3 437
N.D. Yes contig reference C Tyr [Y] 3 437
90602 rs2230036 0.091 Yes synonymous A Gln [Q] 3 372
0.091 Yes contig reference G Gln [Q] 3 372
91277 rs5986990 N.D. intron A/G
91724 rs2515905 N.D. intron C/T
92041 rs5986875 N.D. intron A/G
92395 rs3216174 N.D. intron -/A
92420 rs2515904 N.D. intron C/G
93141 rs1050829 0.180 Yes nonsynonymous G Asp [D] 1 126
0.180 Yes Yes contig reference A Asn [N] 1 126
93839 rs11555344 N.D. Yes nonsynonymous T Leu [L] 1 77
N.D. Yes contig reference G Val [V] 1 77


GENE SYMBOL & name [Organism]
Default Display
G6PD glucose-6-phosphate dehydrogenase [Homo
1: G6PD glucose-6-phosphate dehydrogenase [Homo sapien
GeneID: 2539 Primary source: HGNC:4057
Summary
Genomic regions, transcripts, and products
Bibliography
General gene information
General protein information
Additional Links
Table Of Contents
Where is my gene in the genome?
What does it look like?
What other genome information can I find in this region?


NCBI
NCBI Map Viewer
Home sapiens Build 36.1 (Current)
BLAST The Human Genome
Master Map: Genes On Sequence
Maps & Options

Examples of Maps & Mapped Data



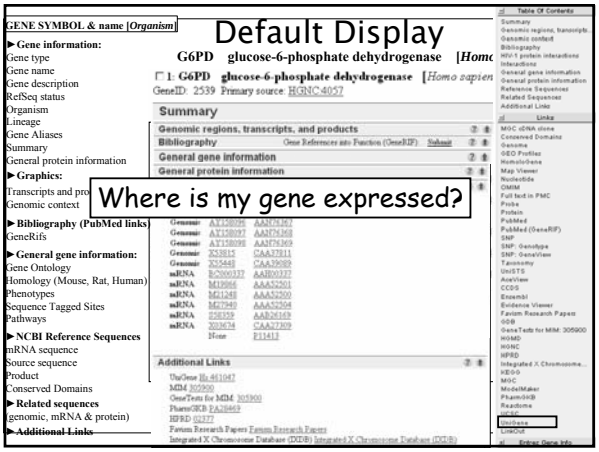
--Sequence maps---
Ab initio (model)
Assembly
Repeats
BES_Clone
Clone
NCI_Clone
Contig
Component
CpG island
dbSNP haplotype
Fosmid
GenBank_DNA
Gene
Phenotype
SAGE_Tag
STS
TCAG_RNA
Transcript (RNA)
UniGene EST
Variation

--Cytogenetic maps--
Ideogram
FISH Clone
NCI FISH Clone
Gene_Cytogenetic
Mitelman Breakpoint
Morbid/Disease

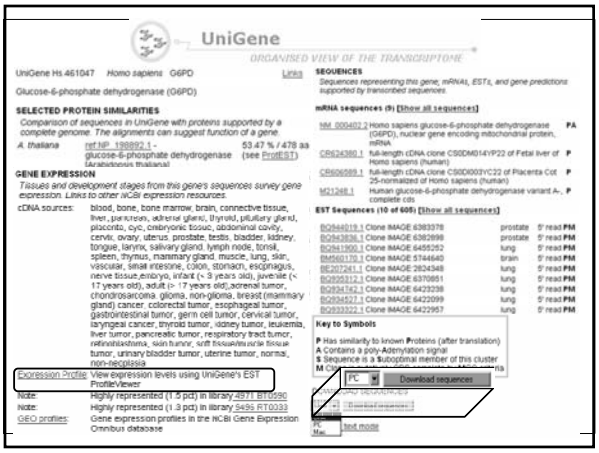
--Genetic Maps--
deCODE
Genethon
Marshfield

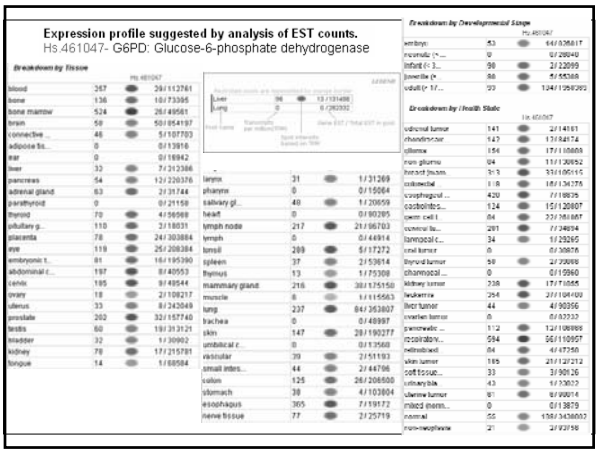
--RH maps---
GeneMap99-G3
GeneMap99-GB4
NCBI RH
Standford-G3
TNG
Whitehead-RH
Whitehead-YAC

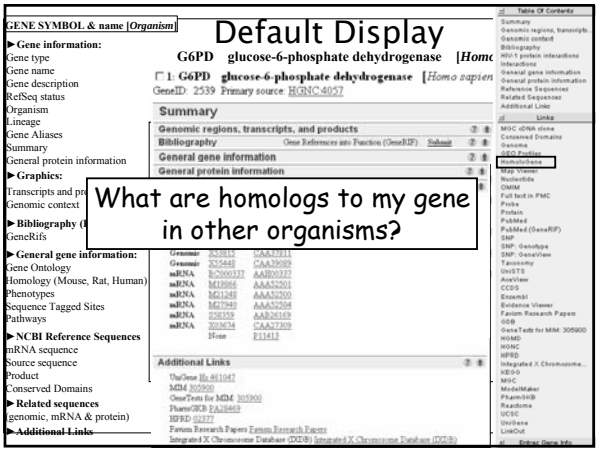
Default Display

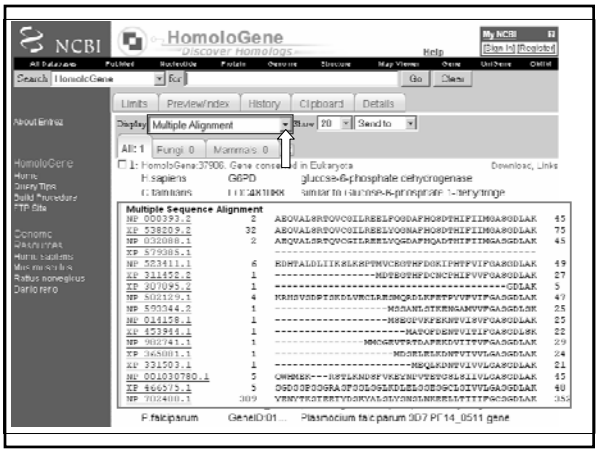


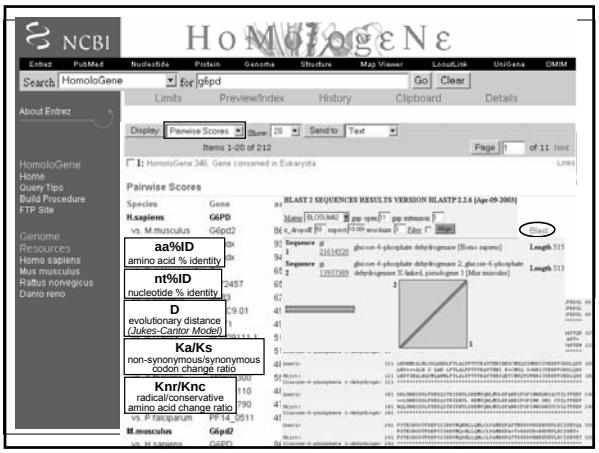
Where is my gene expressed?




Expression profile suggested by analysis of EST counts.
Hs.461047- G6PD: Glucose-6-phosphate dehydrogenase


Default Display
GENE SYMBOL & name [Organism]
G6PD glucose-6-phosphate dehydrogenase
Summary
Genomic regions, transcripts, and products
Bibliography
General gene information
General protein information
Additional Links
What are homologs to my gene in other organisms?

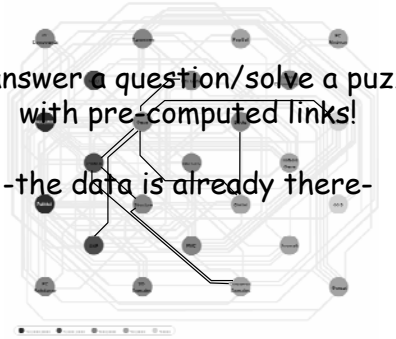

NCBI
HomoloGene
Discover Homologs
Multiple Sequence Alignment



EXAMPLE: Finding information on the Etiology of Hemochromatosis

To answer a question/solve a puzzle with pre-computed links!

-the data is already there-



What causes Hemochromatosis?

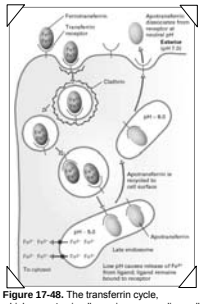


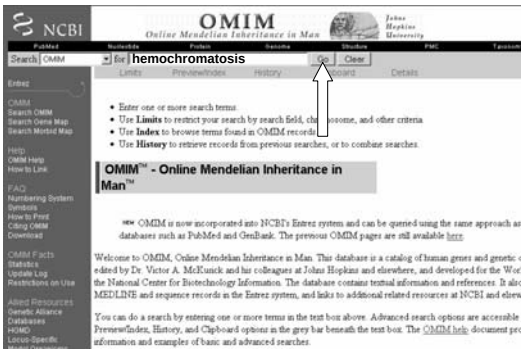
Figure 17-48. The transferrin cycle, which operates in all growing mammalian cells

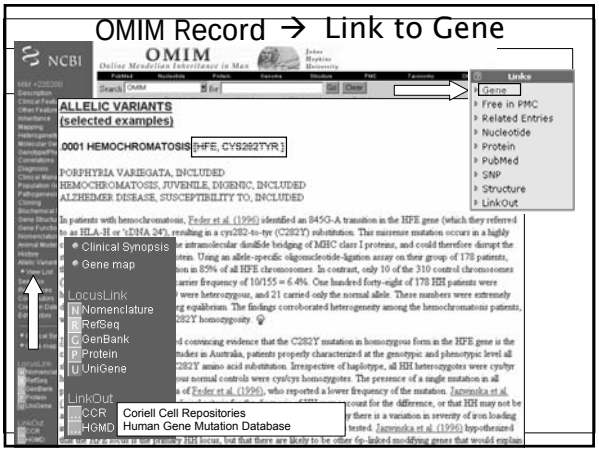
Molecular Cell Biology by Lodish, Berk, Zipursky, Matsudaira, Baltimore & Darnell © 2000 by W.H. Freeman & Co.

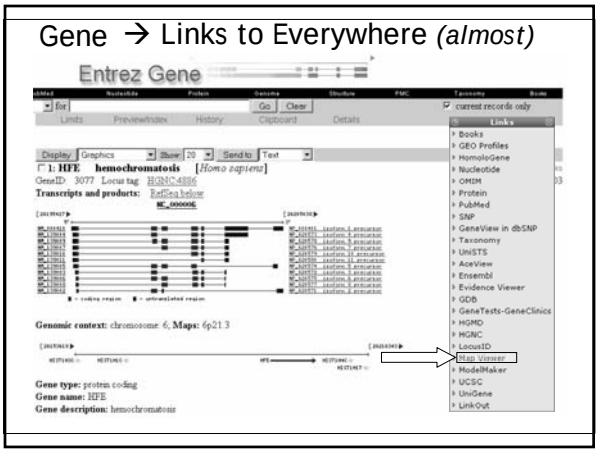
A Bookshelf member!

Classic hemochromatosis (HFE), an autosomal recessive disorder, is an iron overload disease.

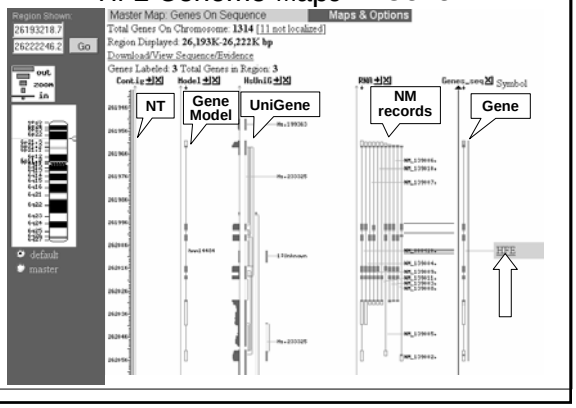
- **Iron is an essential nutrient required for the synthesis of hemoglobin, cytochromes, and many other proteins**
- **Iron is transported around the body by Transferrin which binds to the extracellular Transferrin Receptor, causing internalization and cellular uptake of the complex.**
- **High intracellular iron levels causes free-radical damage to intracellular proteins, lipids, and nucleic acids.**
- **HFE protein is a competitive inhibitor with Transferrin for binding to the Transferrin Receptor, thereby inhibiting cellular uptake of iron and preventing intracellular damage.**
- ***Medical complications of Hemochromatosis:* hypermelanotic pigmentation of the skin, arthritis, cirrhosis of the liver, diabetes, heart failure and primary hepatocellular carcinoma**


NCBI
OMIM
Online Mendelian Inheritance in Man
hemochromatosis
OMIM™ - Online Mendelian Inheritance in Man™

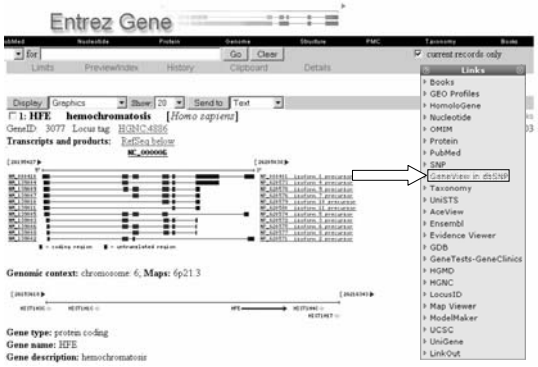

OMIM Record → Link to Gene
NCBI
OMIM
Online Mendelian Inheritance in Man
ALLELIC VARIANTS
(selected examples)
.0001 HEMOCHROMATOSIS [HFE, CYS282TYR]
PORPHYRIA VARIEGATA, INCLUDED
HEMOCHROMATOSIS, JUVENILE, DIGENIC, INCLUDED
ALZHEIMER DISEASE, SUSCEPTIBILITY TO, INCLUDED
Links
Gene
Free in PMC
Related Entries
Nucleotide
Protein
PubMed
SNP
Structure
LinkOut
Clinical Synopsis
Gene map
LocusLink
N Nomenclature
R RefSeq
G GenBank
P Protein
U UniGene
LinkOut
CCR
HGMD
Coriell Cell Repositories
Human Gene Mutation Database


Gene → Links to Everywhere (almost)
Entrez Gene
1: HFE hemochromatosis [Homo sapiens]
Links
Books
GEO Profiles
HomoloGene
Nucleotide
OMIM
Protein
PubMed
SNP
GeneView in dbSNP
Taxonomy
UniSTS
AceView
Ensembl
Evidence Viewer
GDB
GeneTests-GeneClinics
HGMD
HGNC
LocusID
Map Viewer
ModelMaker
UCSC
UniGene
LinkOut

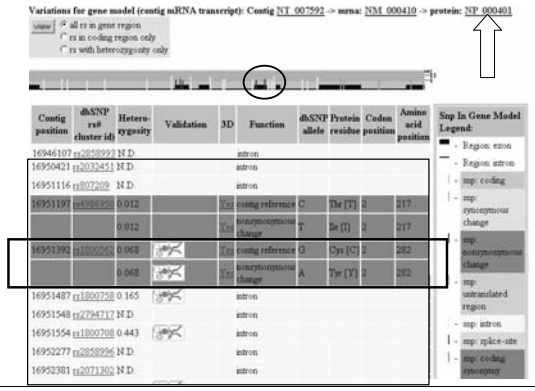
HFE Genome Maps → Gene



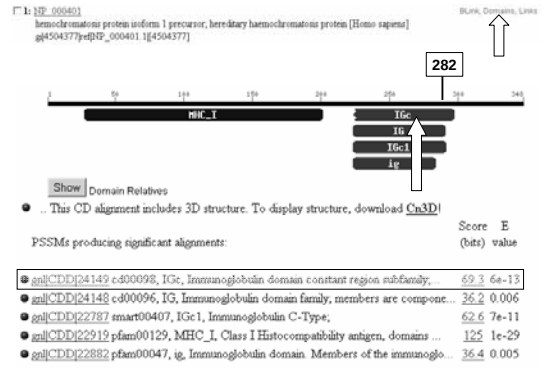
Gene → Links to Everywhere *(almost)*



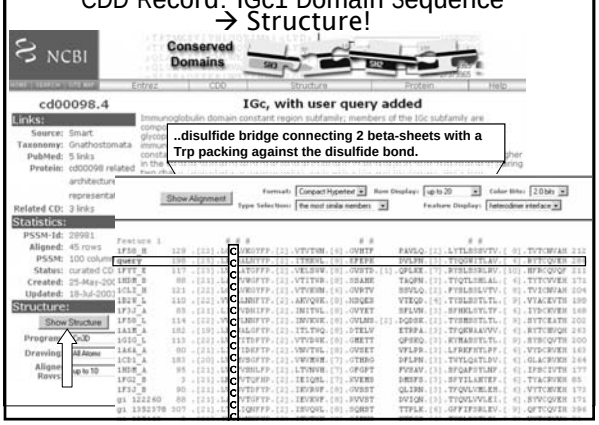
GeneView in SNP → Protein



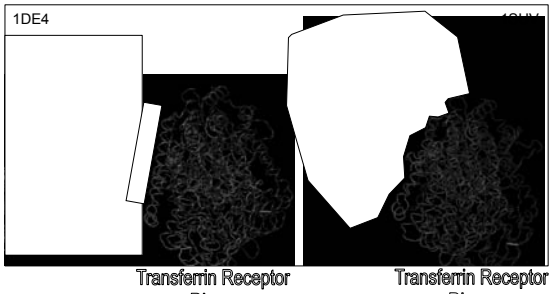
Protein Record →Domain (CDD) Search



CDD Record: IGc1 Domain Sequence → Structure!



Human Transferrin Receptor Complexes



Additional Courses

A Field Guide to NCBI Resources

Enhanced Field Guide

Exploring sequences, genomes and structures

Course Description

The National Center for Biotechnology Information presents an expanded, on-site, two-day version of the popular Field Guide.

Online Registration

July 31, 2006
December, 2006

NCBI Structures

Mini Courses

1. Making Sense of DNA and Protein Sequences
2. Unmasking Genes in Human DNA
3. Identification of Disease Genes
4. Correlating Disease Genes and Phenoypes
5. BLAST Quick Start
6. EntrezGene Quick Start
7. Structure Analysis Quick Start
8. MapViewer Quick Start
9. GenBank Quick Start
10. Entrez Quick Start

PowerTools NCBI Technical Workshop Series

For More Information…

E-mail addresses:

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

The *(free!)* NCBI Newsletter

http://www.ncbi.nih.gov/About/newsletter.html

The NCBI Handbook

Follow the link from the NCBI Home Page under "Hot Spots"

The NCBI Education Page

http://www.ncbi.nih.gov/Education/index.html