NCBI Resources: from Sequence to Function

Medha Bhagwat, NCBI

Current Topics in Genome Analysis
January 18, 2005

NCBI

Outline

About NCBI
NCBI databases and tools
The Entrez- search and retrieval system
Training at NCBI

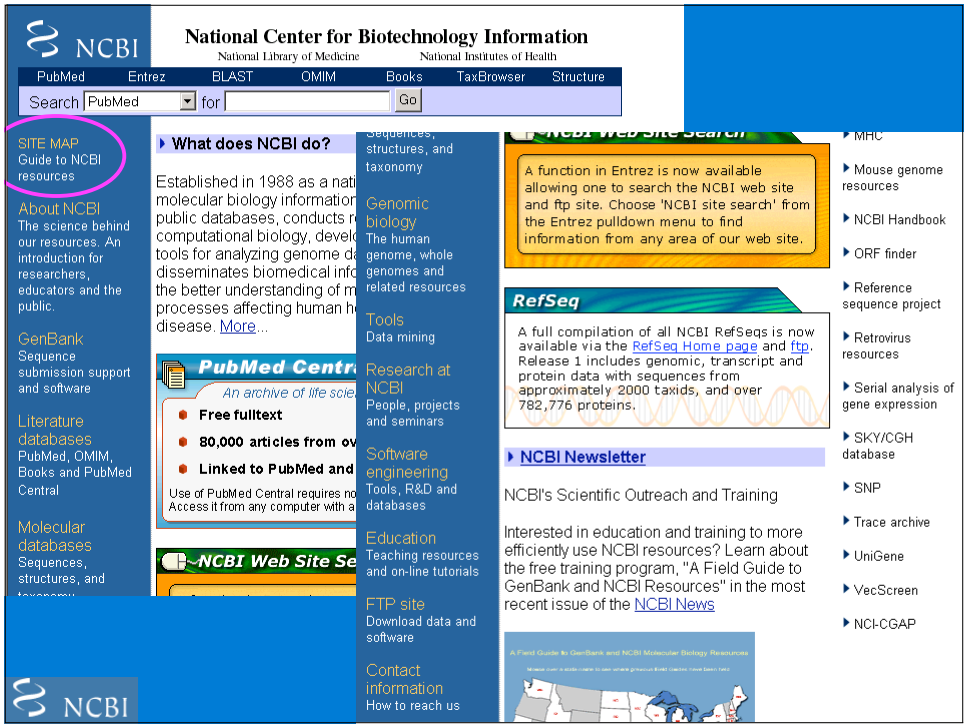
NCBI

National Center for Biotechnology Information

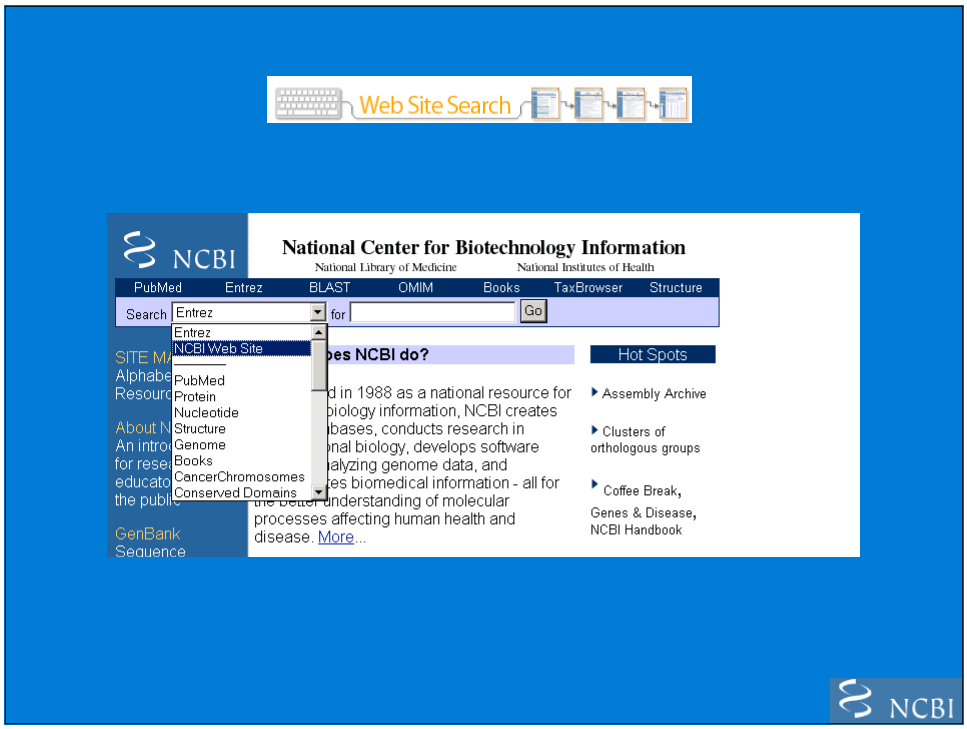
http://www.ncbi.nlm.nih.gov/

Created as a part of NLM in 1988

- To establish public databases
  GenBank and others
- To perform research in computational biology
- To develop software tools for sequence analysis
- To disseminate biomedical information



NCBI
Alphabetical Quicklinks Table
PubMed Entrez BLAST OMIM Taxonomy Structure
ALPHABETICAL QUICKLINKS TABLE
(To view resource descriptions and a complete list of services, see the NCBI Resource Guide.
To view resources by category, see the graphical Site Map.)
About NCBI | Education | LocusLink | SAGEmap
Announcements | e-PCR | Malaria | Science Primer
ASN.1 | Entrez | Map Viewer | Seminars
BankIt | Entrez Utilities | MeSH | Sequin
BLAST | Expression | MGC | Site Search
BLink | FTP | Microbial Genomes | SKY/M-FISH & CGH Database
Books | GenBank | MMDB | Software Engineering
Cancer Chromosomes | GenBank sample record | Model Maker | Spidey
CDART | Genes NEW | Mutation Databases (external) | Statistics
CDD | Genes and Disease | NCBI Home | Structures
CGAP | Genomes | NCBI News | Submit Data
Clones | GENSAT NEW | Nucleotide Sequences (Entrez) | Taxonomy
Cn3D | GEO (Expression) | OMIM | Tools
Coffee Break | Glossary | ORF Finder | TPA
COGs | Handbook | Plant Genomes | Trace Archive
Computational Biology Branch | HIV Interactions NEW | Protein Sequences (Entrez) | UniGene
Data Submissions | HTGs | PubChem NEW | UniSTS
dbEST | HomoloGene | PubMed | VAST
dbGSS | Human Genome Resources | PubMed Central | VecScreen
dbMHC | Human-Mouse Homology Maps | RefSeq | Viruses
dbSNP | Journals | Research at NCBI | WGS
dbSTS | LinkOut | Retroviruses | What's New NEW
NEW indicates a resource which has become available in the last 12 months.

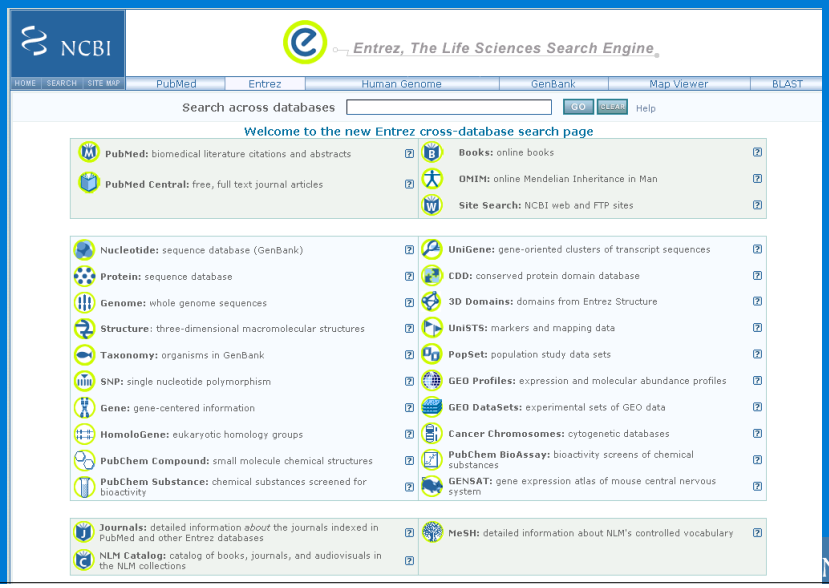

Web Site Search
NCBI
National Center for Biotechnology Information
National Library of Medicine National Institutes of Health
PubMed Entrez BLAST OMIM Books TaxBrowser Structure
Search Entrez for Go
Entrez
NCBI Web Site
PubMed
Protein
Nucleotide
Structure
Genome
Books
CancerChromosomes
Conserved Domains
Hot Spots
Assembly Archive
Clusters of orthologous groups
Coffee Break, Genes & Disease, NCBI Handbook
GenBank
Sequence
NCBI

NCBI Databases and Sequence Analysis Tools

NCBI

Entrez: Search and Retrieval System

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



NCBI

NCBI

Databases

Nucleotide sequences
Protein sequences
Structures
Taxonomy
Genomes
Expression
Chemical
Literature

An Array of Sequence Analysis Tools

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

Nucleotide sequence analysis
Protein sequence analysis
Genome analysis
Structure
Gene expression

NCBI

NCBI

Databases

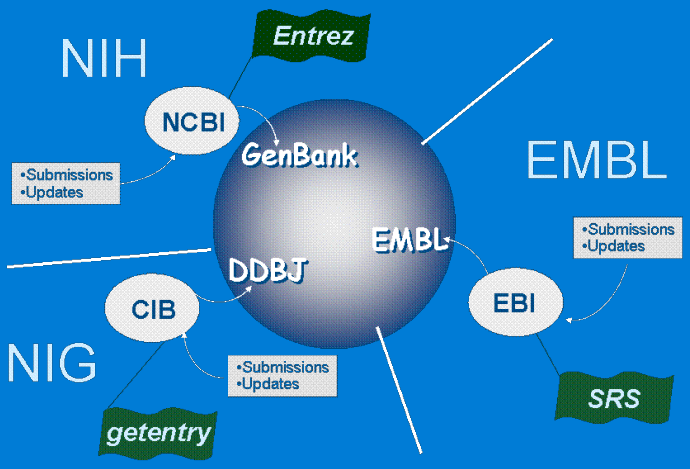
Nucleotide Databases

GenBank
- Individual submissions
- Bulk submissions
  - EST, GSS, HTGS, WGS

Derived database
- RefSeq

International Nucleotide Sequence Database Collaboration

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



NCBI

NCBI Databases

Primary	Derived
Redundant	Non-redundant
Archival/repository	Curated
Submitter owner	NCBI owner
Sequenced	Combined/edited
Ex: GenBank	Ex: RefSeq

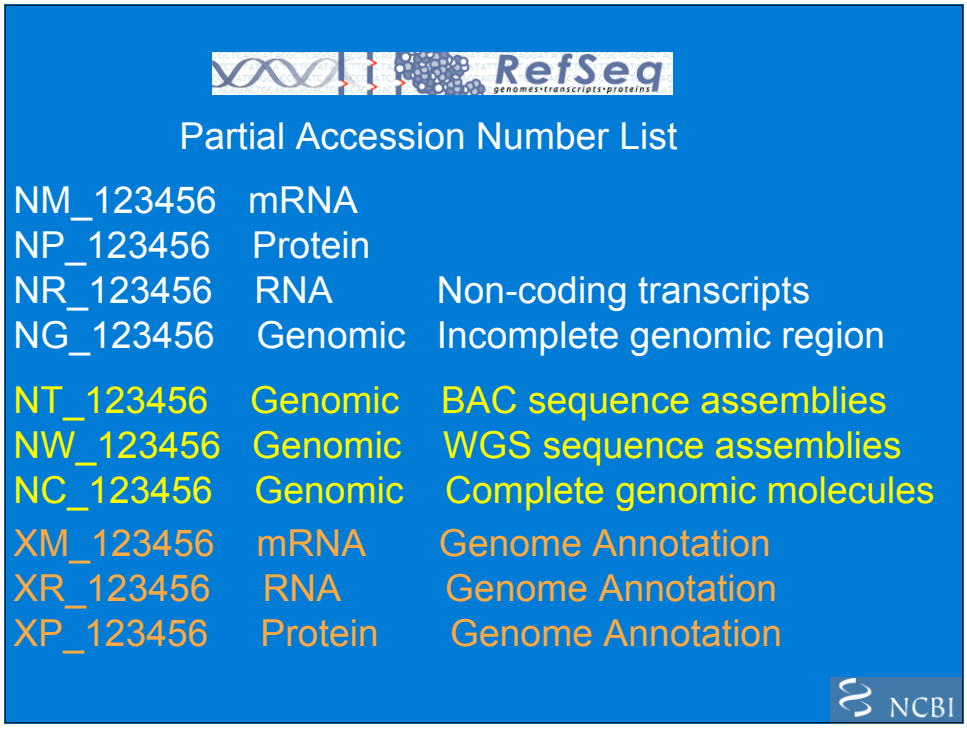
NCBI

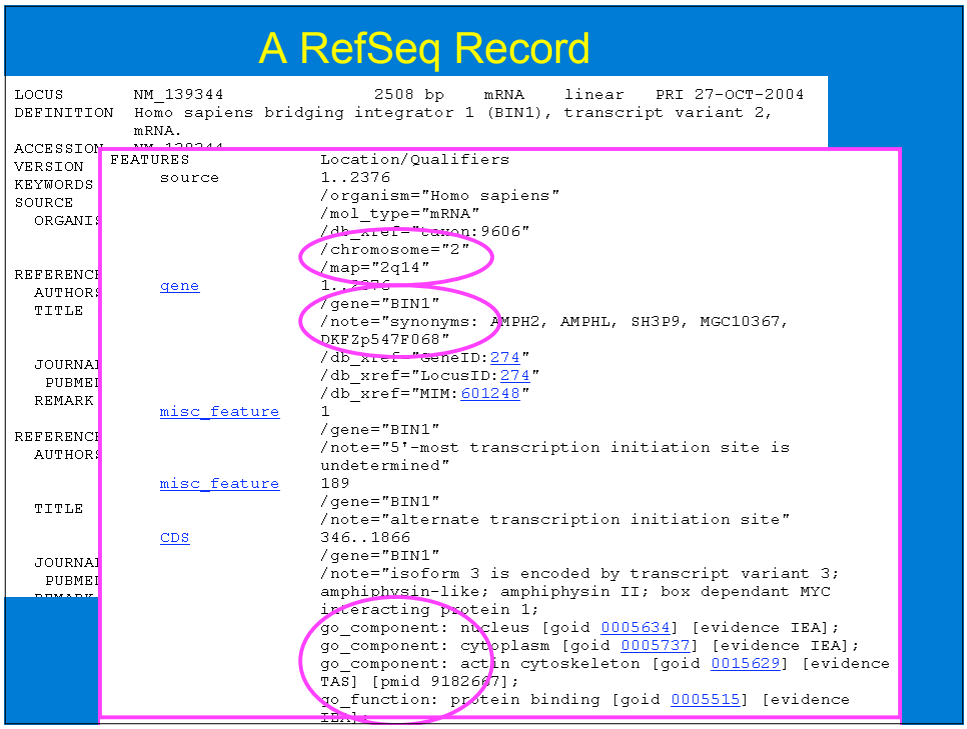
http://www.ncbi.nlm.nih.gov/RefSeq/

- best, comprehensive, non-redundant set of sequences
- for genomic DNA, transcript (RNA), and protein
- for major research organisms
  2645 organisms

- based on GenBank derived sequences
- ongoing curation by NCBI staff and collaborators, with review status indicated on each record
- updates to reflect current knowledge of sequence data and biology

NCBI


RefSeq
genomes·transcripts·proteins
Partial Accession Number List
NM_123456 mRNA
NP_123456 Protein
NR_123456 RNA Non-coding transcripts
NG_123456 Genomic Incomplete genomic region
NT_123456 Genomic BAC sequence assemblies
NW_123456 Genomic WGS sequence assemblies
NC_123456 Genomic Complete genomic molecules
XM_123456 mRNA Genome Annotation
XR_123456 RNA Genome Annotation
XP_123456 Protein Genome Annotation
NCBI


A RefSeq Record
LOCUS NM_139344 2508 bp mRNA linear PRI 27-OCT-2004
DEFINITION Homo sapiens bridging integrator 1 (BIN1), transcript variant 2, mRNA.
ACCESSION
VERSION
KEYWORDS
SOURCE
ORGANI
REFERENCE
AUTHORS
TITLE
JOURNAL
PUBMED
REMARK
REFERENCE
AUTHORS
TITLE
JOURNAL
PUBMED
FEATURES Location/Qualifiers
source 1..2376
/organism="Homo sapiens"
/mol_type="mRNA"
/db_xref="taxon:9606"
/chromosome="2"
/map="2q14"
gene 1..2376
/gene="BIN1"
/note="synonyms: AMPH2, AMPHL, SH3P9, MGC10367, DKFZp547F068"
/db_xref="GeneID:274"
/db_xref="LocusID:274"
/db_xref="MIM:601248"
misc_feature 1
/gene="BIN1"
/note="5'-most transcription initiation site is undetermined"
misc_feature 189
/gene="BIN1"
/note="alternate transcription initiation site"
CDS 346..1866
/gene="BIN1"
/note="isoform 3 is encoded by transcript variant 3; amphiphysin-like; amphiphysin II; box dependant MYC interacting protein 1;
go_component: nucleus [goid 0005634] [evidence IEA];
go_component: cytoplasm [goid 0005737] [evidence IEA];
go_component: actin cytoskeleton [goid 0015629] [evidence TAS] [pmid 9182607];
go_function: protein binding [goid 0005515] [evidence

NCBI Databases

Protein

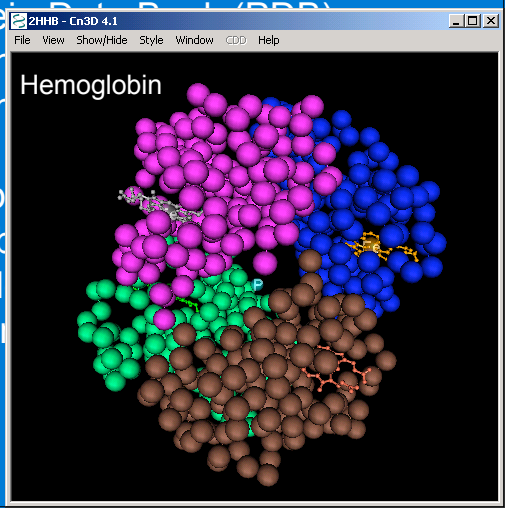
- Conceptual translations of GenBank and RefSeq records
- SwissProt, PIR, PRF, PDB

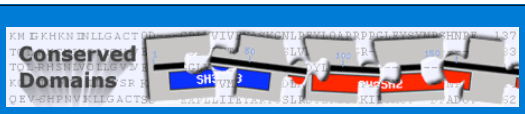
Molecular Modeling DataBase (MMDB)

http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml

- obtained from the Protei
- experimentally determin
- can be viewed using Cr

- sequences also availab
  Entrez protein datab
- useful for finding homol
  structures for a protei





http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

Conserved Domain

- recurring unit in molecular evolution, whose extents can be determined by sequence and structure analysis
- performs a particular function
- represented as a multiple local sequence alignment of proteins containing the domain

NCBI

Conserved Domain Database

Pfam (5252)

SMART (575)

COGs (4101)

CDD (10573)

Curated CDs (645)

- A position-specific scoring matrix (PSSM) is calculated
- CD-Search can be used to search against the PSSMs
- Manual curation of CDs has begun

NCBI

Conserved Domain in Beta Globin

cd01040.1 **globin**

Links:

Source: CDD
Taxonomy: cellular organisms
Pubmed: 5 links
Book: 3 book links
Proteins: cd01040 related
Related CD: 4 links

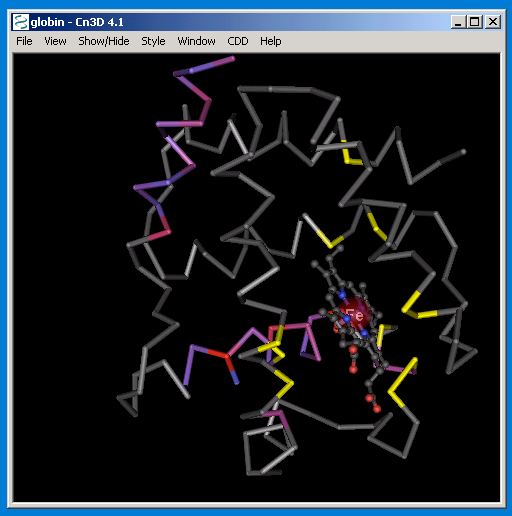
Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen in the bloodstream, (2) microorganismal flavohemoglobins, which are linked to C-terminal FAD-dependend reductase domains, (3) homodimeric bacterial hemoglobins, such as from Vitreoscilla, (4) plant leghemoglobins (symbiotic hemoglobins, involved in nitrogen metabolism in plant rhizomes), (5) plant non-symbiotic hexacoordinate globins and hexacoordinate globins from bacteria and animals, such as neuroglobin, (6) invertebrate hemoglobins, which may occur in tandem-repeat arrangements, and (7) monomeric myoglobins found in animal muscle tissue.

Feature 1: heme-binding site
Evidence: Structure: Ascaris hemoglobin with bound heme and oxygen molecule - View structure with Cn3D 4.1
Comment: Ascaris hemoglobin exhibits strong affinity to oxygen
Citation: PMID 7753786
Structure: Bovine deoxy-hemoglobin A with bound heme - View structure with Cn3D 4.1
Citation: PMID 8411160

Show Alignment
Format: Hypertext
Row Display: up to 5
Color Bits: 2.0 bits
Type Selection: the most similar members
Feature Display: heme-binding site

```
                    10        20        30        40        50        60        70        80
                ....*....|....*....|....*....|    *    |    *    |    *    |    *    |....*....|
Feature 1                                    #         #  ##                   #  #
consensus    1 SAEEKKLVKASWAKLk----aDREEIGLEFYERLFKAHPETRALFSRFGGLSA--ALKGSPKFKAHGKRVLNALDEAIKN 74
query        5 TPEEKSAVTALWGKV------NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTpdAVMGNPKVKAHGKKVLGAFSDGLAH 78
1ASH         1 ANKTRELCMKSLehakvdtsnearQDGIDLYKHMFENYppLRKYFksreeyta-edvqndpfFAKQGQKILLACHVLCAt 79
1FDH_G       5 TEEDKATITSLWgkv------nveDAGGETLGRLLVVYpwTQRFFdsfgnlssasaimgnpkVKAHGKKVLTSLGDAIKh 78
1PBX_B       4 TDKERSIISDIFshm------dydDIGPKALSRCLIVYpwTQRHFsgfgnlynaeaiignanVAAHGIKVLHGLDRGVKn 77
gi 122300    4 SAEEKALVVGLCgkis----ghcdALGGEALDRLFASFgqTRTYFshfdls------pgsadVKRHGGKVLSAIGEAAKh 73
gi 122536    5 TAEEKAAITSVWqkv------nveHDGHDALGRLLIVYpwTQRYFsnfgnlsnsaavagnakVQAHGKKVLSAVGNAISh 78
gi 122542    1 GGSDVSAFLAKVdk---------rAVGGEALARLLIVYpwTQRYFstfgnlgsadaishnskVLAHGQRVLDSIEEGLKh 71
gi 122690    4 TGEEKALVNAVWtkt------dhqAVVAKALERLFVVYpwTKTYFvkfngkf----hasdstVQTHAGKVVSALTVAYNh 73
gi 229556    3 SIADKTSLKNAWgkis----tdttEIGTEALERLHLSFp-TQKFLshg-----------lahVKAHGSKVAGALTSILGp 66
                    90        100       110       120       130       140
```

Conserved Domain in Beta Globin

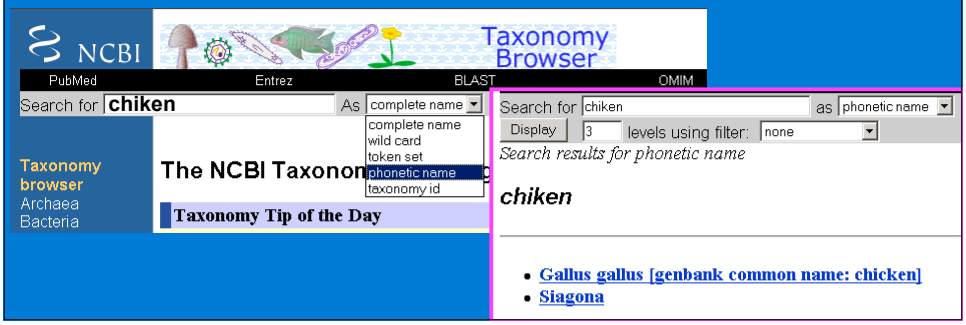


NCBI Databases

Organisms

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

incorporates phylogenetic and taxonomic knowledge from a variety of sources



Taxonomy Browser

Gallus gallus

Taxonomy ID: 9031
Genbank common name: **chicken**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:

common name: **chickens**
includes: **dwarf Leghorn chickens**
includes: **red junglefowl**
misnomer: **Gallus domesticus**
misnomer: **Gallus gallus domesticus**

Lineage (full)
cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Sauropsida; Sauria; Archosauria; Aves; Neognathae; Galliformes; Phasianidae; Phasianinae; Gallus

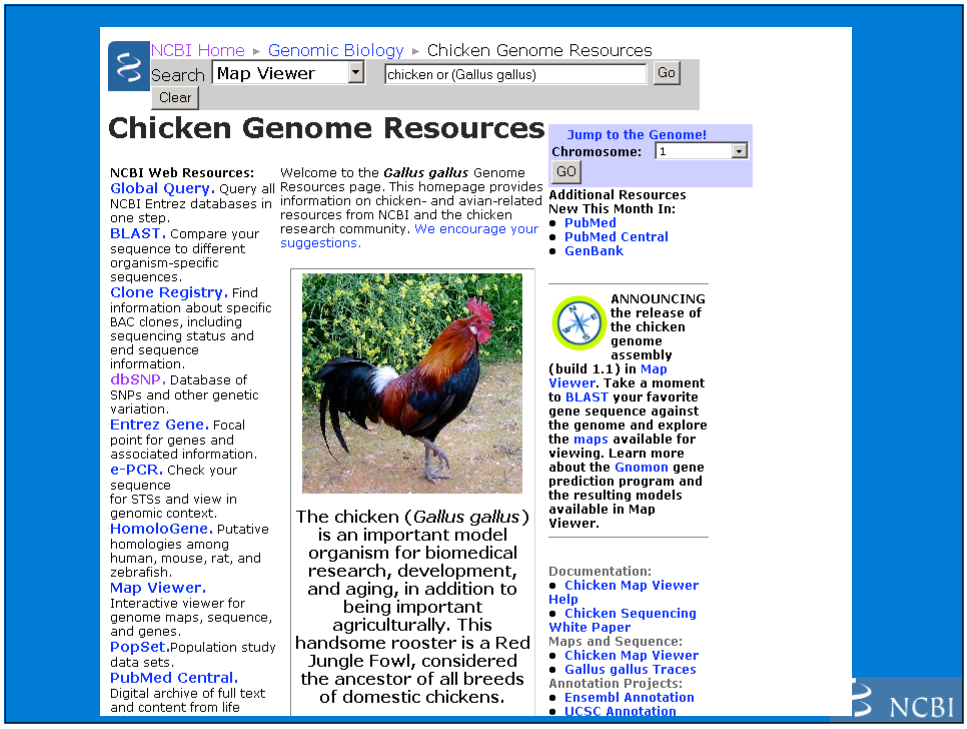
Entrez records		
Database name	Subtree links	Direct links
Nucleotide	871,867	871,825
Protein	29,194	29,192
Structure	453	453
Genome	31	31
Popset	31	31
3D Domains	1,967	1,967
Domains	1	1
UniGene	21,447	21,447
UniSTS	1,958	1,958
PubMed Central	243	242
Gene	18,505	18,505
HomoloGene	9,700	9,700
Taxonomy	3	1

Genome Information

See the NCBI Genome homepage

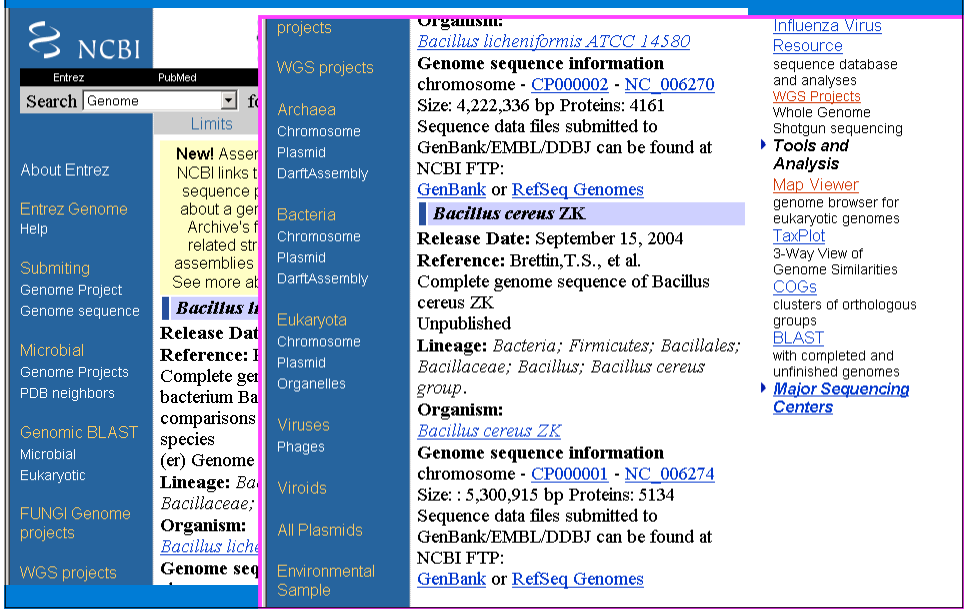
Trace records (raw single-pass reads of DNA sequence)				
Center name	Record counts per type			
	FINISHING	SHOTGUN	WGS	ALL
JGI - Joint Genome Institute, U.S. Department of Energy	0	0	2,477,710	2,477,710
UOKNOR - University of Oklahoma Norman Campus, Advanced Center for Genome Technology	223	13,662	0	13,885
Total	223	13,662	2,477,710	**2,491,595**

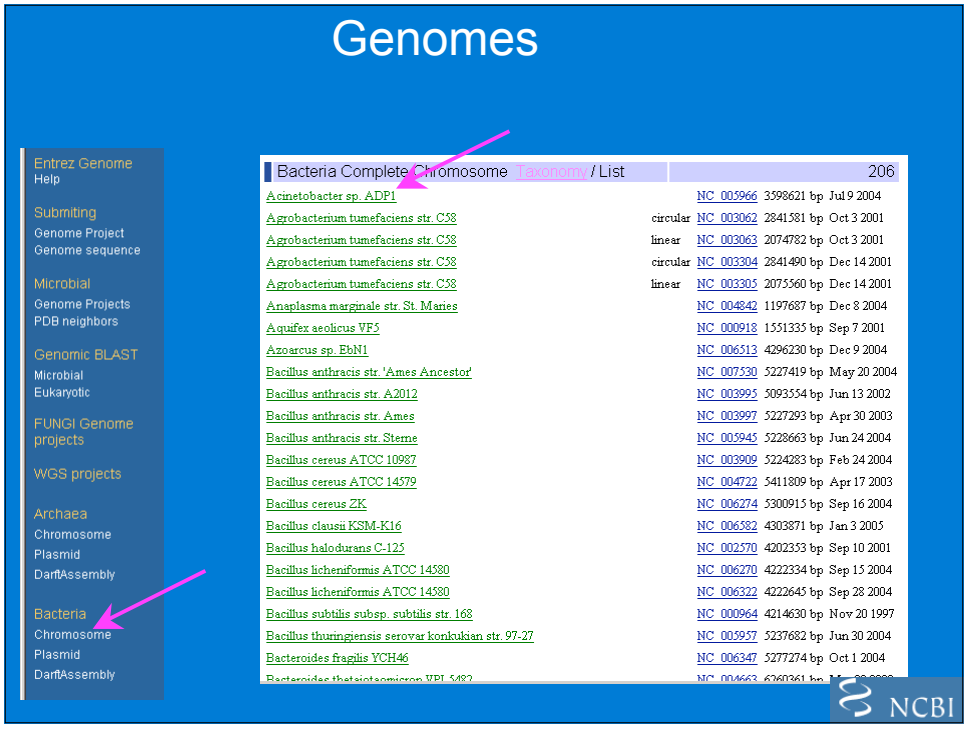
NCBI



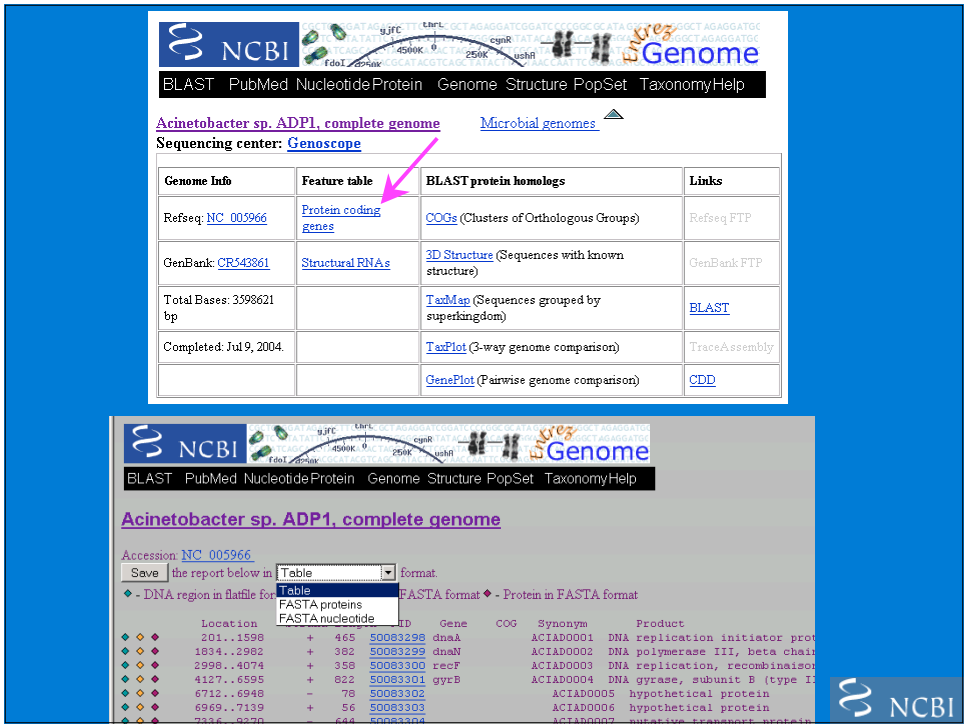
Genomes

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome




Genomes
Entrez Genome
Help
Submiting
Genome Project
Genome sequence
Microbial
Genome Projects
PDB neighbors
Genomic BLAST
Microbial
Eukaryotic
FUNGI Genome projects
WGS projects
Archaea
Chromosome
Plasmid
DarftAssembly
Bacteria
Chromosome
Plasmid
DarftAssembly
Bacteria Complete Chromosome Taxonomy / List 206
Acinetobacter sp. ADP1 NC_005966 3598621 bp Jul 9 2004
Agrobacterium tumefaciens str. C58 circular NC_003062 2841581 bp Oct 3 2001
Agrobacterium tumefaciens str. C58 linear NC_003063 2074782 bp Oct 3 2001
Agrobacterium tumefaciens str. C58 circular NC_003304 2841490 bp Dec 14 2001
Agrobacterium tumefaciens str. C58 linear NC_003305 2075560 bp Dec 14 2001
Anaplasma marginale str. St. Maries NC_004842 1197687 bp Dec 8 2004
Aquifex aeolicus VF5 NC_000918 1551335 bp Sep 7 2001
Azoarcus sp. EbN1 NC_006513 4296230 bp Dec 9 2004
Bacillus anthracis str. 'Ames Ancestor' NC_007530 5227419 bp May 20 2004
Bacillus anthracis str. A2012 NC_003995 5093554 bp Jun 13 2002
Bacillus anthracis str. Ames NC_003997 5227293 bp Apr 30 2003
Bacillus anthracis str. Sterne NC_005945 5228663 bp Jun 24 2004
Bacillus cereus ATCC 10987 NC_003909 5224283 bp Feb 24 2004
Bacillus cereus ATCC 14579 NC_004722 5411809 bp Apr 17 2003
Bacillus cereus ZK NC_006274 5300915 bp Sep 16 2004
Bacillus clausii KSM-K16 NC_006582 4303871 bp Jan 3 2005
Bacillus halodurans C-125 NC_002570 4202353 bp Sep 10 2001
Bacillus licheniformis ATCC 14580 NC_006270 4222334 bp Sep 15 2004
Bacillus licheniformis ATCC 14580 NC_006322 4222645 bp Sep 28 2004
Bacillus subtilis subsp. subtilis str. 168 NC_000964 4214630 bp Nov 20 1997
Bacillus thuringiensis serovar konkukian str. 97-27 NC_005957 5237682 bp Jun 30 2004
Bacteroides fragilis YCH46 NC_006347 5277274 bp Oct 1 2004
NCBI

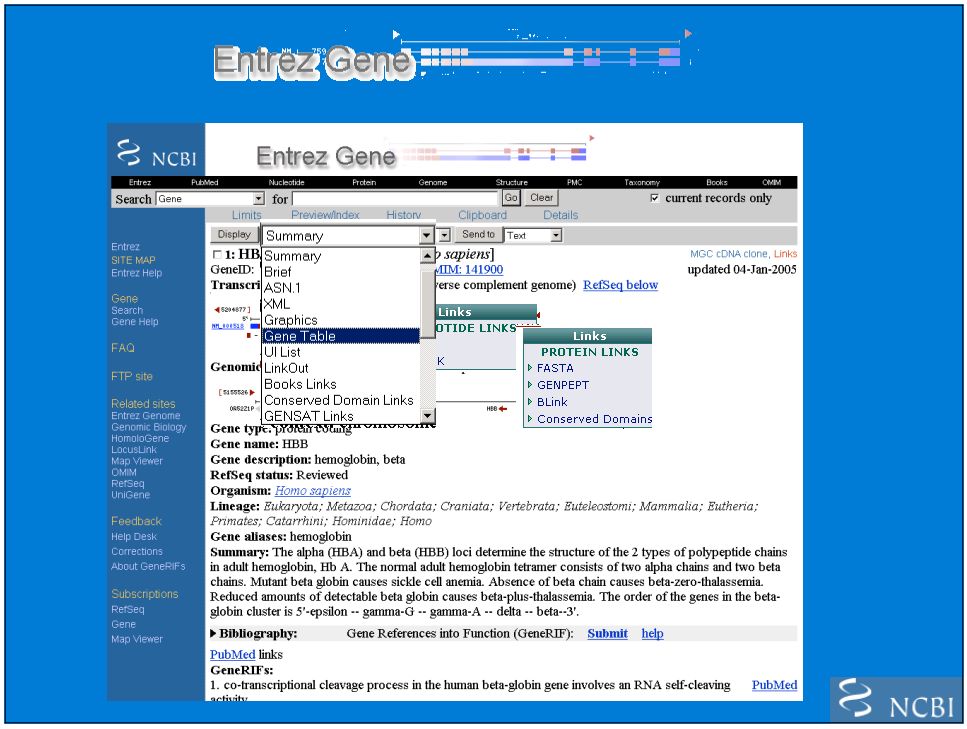
NCBI
Taxonomy Browser
Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
Search for as complete name lock Go Clear
Display 3 levels using filter: none
Acinetobacter sp. ADP1
Taxonomy ID: 62977
Rank: species
Genetic code: Translation table 11 (Bacterial and Plant Plastid)
Other names:
synonym: Acinetobacter calcoaceticus ADP1
Lineage (full)
cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter
Entrez records
Database name Direct links
Nucleotide 20
Protein 6,891
Structure 9
Genome 1
3D Domains 33
PubMed Central 4
Gene 3,425
Taxonomy 1
NCBI

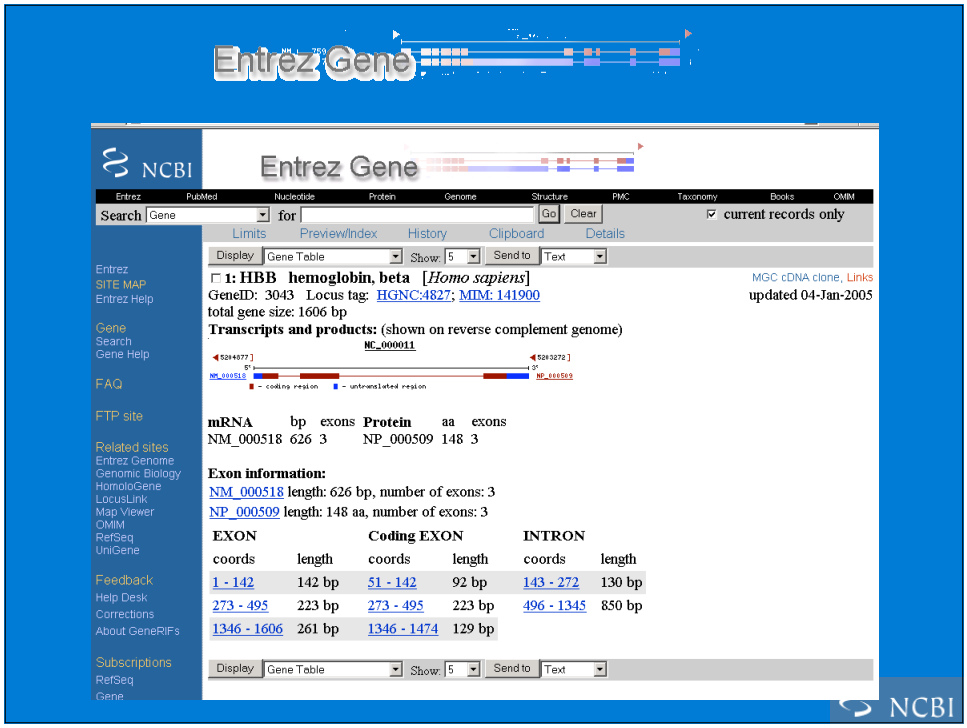


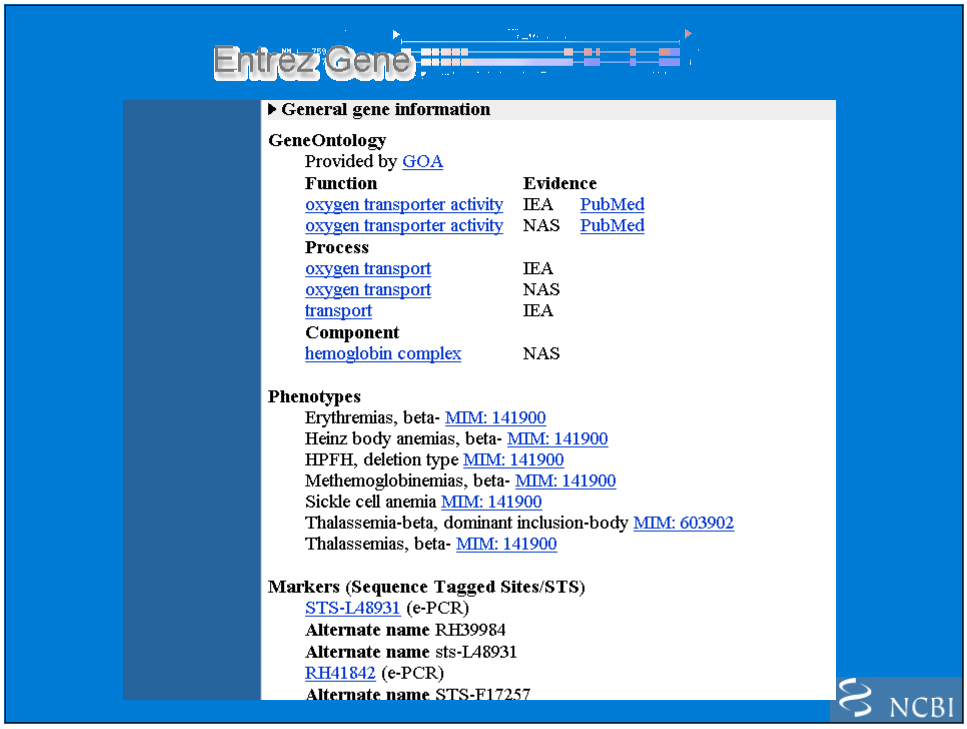
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

- A record represents a single gene from an organism
- A gene-specific information such as
  map, sequence, expression, structure,
  function, homology and publications
- Includes data for all organisms that have
  RefSeq genome records
- Successor to LocusLink
  - more organisms
  - efficient searching options

NCBI

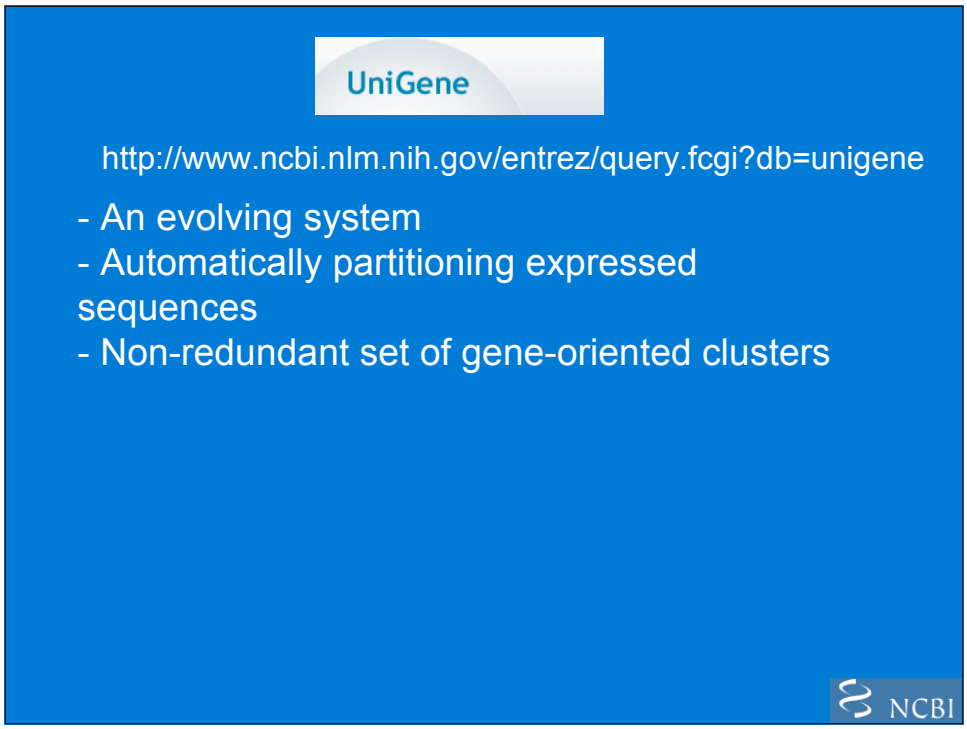

Entrez Gene
NCBI
Entrez Gene
Search Gene for Go Clear current records only
Limits Preview/Index History Clipboard Details
Display Summary Send to Text
Summary
Brief
ASN.1
XML
Graphics
Gene Table
UI List
LinkOut
Books Links
Conserved Domain Links
GENSAT Links
Links
PROTEIN LINKS
FASTA
GENPEPT
BLink
Conserved Domains
Gene type:
Gene name: HBB
Gene description: hemoglobin, beta
RefSeq status: Reviewed
Organism: Homo sapiens
Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo
Gene aliases: hemoglobin
Summary: The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta--3'.
Bibliography: Gene References into Function (GeneRIF): Submit help
PubMed links
GeneRIFs:
1. co-transcriptional cleavage process in the human beta-globin gene involves an RNA self-cleaving
PubMed
NCBI

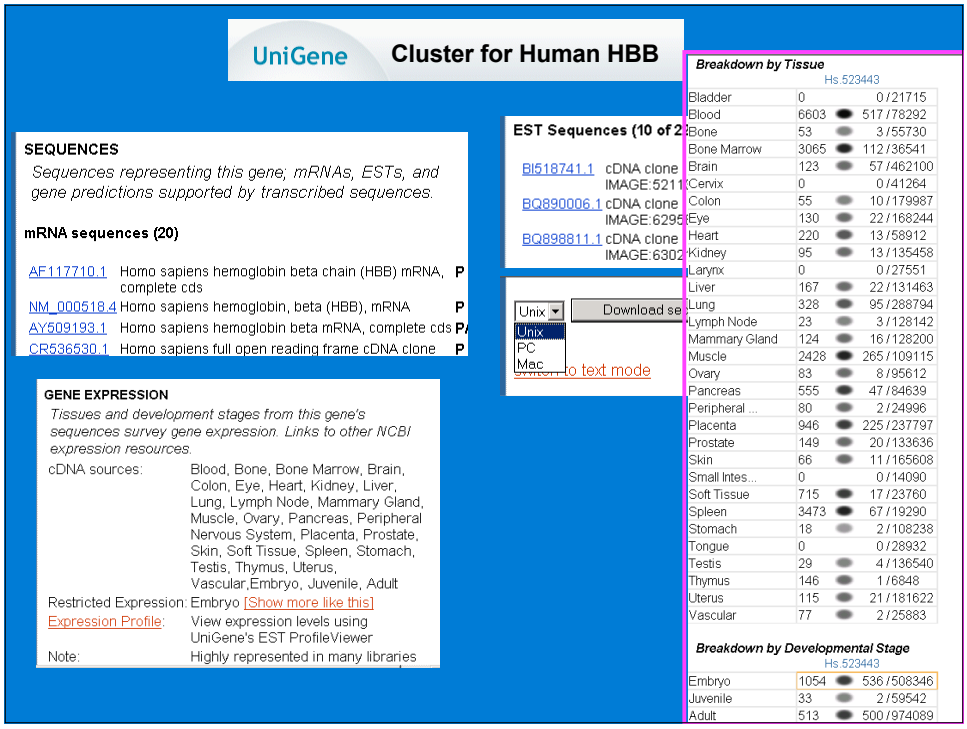

Entrez Gene
NCBI
Entrez Gene
Search Gene for Go Clear current records only
Limits Preview/Index History Clipboard Details
Display Gene Table Show: 5 Send to Text
1: HBB hemoglobin, beta [Homo sapiens]
GeneID: 3043 Locus tag: HGNC:4827; MIM: 141900
MGC cDNA clone, Links
updated 04-Jan-2005
total gene size: 1606 bp
Transcripts and products: (shown on reverse complement genome)
NC_000011
mRNA bp exons Protein aa exons
NM_000518 626 3 NP_000509 148 3
Exon information:
NM_000518 length: 626 bp, number of exons: 3
NP_000509 length: 148 aa, number of exons: 3
EXON Coding EXON INTRON
coords length coords length coords length
1 - 142 142 bp 51 - 142 92 bp 143 - 272 130 bp
273 - 495 223 bp 273 - 495 223 bp 496 - 1345 850 bp
1346 - 1606 261 bp 1346 - 1474 129 bp
Display Gene Table Show: 5 Send to Text
NCBI

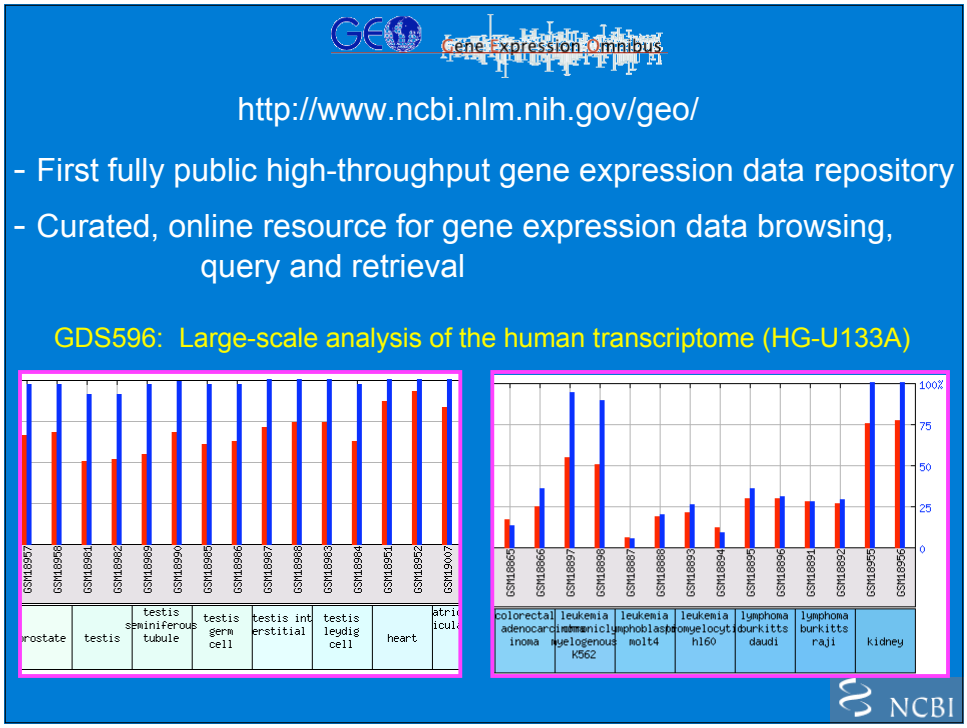

Entrez Gene
General gene information
GeneOntology
Provided by GOA
Function
Evidence
oxygen transporter activity IEA PubMed
oxygen transporter activity NAS PubMed
Process
oxygen transport IEA
oxygen transport NAS
transport IEA
Component
hemoglobin complex NAS
Phenotypes
Erythremias, beta- MIM: 141900
Heinz body anemias, beta- MIM: 141900
HPFH, deletion type MIM: 141900
Methemoglobinemias, beta- MIM: 141900
Sickle cell anemia MIM: 141900
Thalassemia-beta, dominant inclusion-body MIM: 603902
Thalassemias, beta- MIM: 141900
Markers (Sequence Tagged Sites/STS)
STS-L48931 (e-PCR)
Alternate name RH39984
Alternate name sts-L48931
RH41842 (e-PCR)
Alternate name STS-F17257
NCBI

Entrez Gene
NCBI Reference Sequences (RefSeq)
Reference NG_000007
mRNA Sequence NM_000518
Source Sequence L48217
Product NP_000509 beta globin
Conserved Domains (1) summary
cd01040: globin; Globins are heme proteins, which bind and transport oxygen
Location: 5 - 142 Blast Score: 278
Related Sequences
Nucleotide
Protein
Genomic A01592 CAA00182
Genomic AF007546 AAB62944
Genomic AF104901 AAC97372
Genomic AF105973 AAC97959
Genomic AF186606 AAF08258
Genomic AF186607 AAF08259
Genomic AF186608 AAF08260
Genomic AF186609 AAF08261
Genomic AF186610 AAF08262
Genomic AF186611 AAF08263
Genomic AF186612 AAF08264
Genomic AF186613 AAF08265
Genomic AF186614 AAF08266
Genomic AF186615 AAF08267
NCBI

Entrez Gene
Entrez
SITE MAP
Entrez Help
Gene
Search
Gene Help
FAQ
FTP site
Related sites
Entrez Genome
Genomic Biology
HomoloGene
LocusLink
Map Viewer
OMIM
RefSeq
UniGene
Feedback
Help Desk
Corrections
About GeneRIFs
Subscriptions
RefSeq
Display
Gene Table
Show: 5
Send to
Text
1: HBB hemoglobin, beta [Homo sapiens]
GeneID: 3043 Locus tag: HGNC:4827; MIM: 141900
Transcripts and products: (shown on reverse complement genome) RefSeq belov
NC_000011
Genomic context: chromosome: 11; Maps: 11p15.5
Gene type: protein coding
Gene name: HBB
Gene description: hemoglobin, beta
RefSeq status: Reviewed
Organism: Homo sapiens
Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Primates; Catarrhini; Hominidae; Homo
Gene aliases: hemoglobin
Summary: The alpha (HBA) and beta (HBB) loci determine the structure of the 2 ty
in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two al
chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia.
Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-
globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta--3'.
Links
Conserved Domains
GEO Profiles
HomoloGene
Map Viewer
Nucleotide
OMIM
Protein
PubMed
SNP
GeneView in dbSNP
Taxonomy
UniSTS
AceView
Ensembl
Evidence Viewer
GDB
GeneTests for MIM: 141900
Globin Gene Server
HGNC
LocusID
MGC
ModelMaker
PharmGKB
UCSC
UniGene
LinkOut
NCBI


UniGene
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene
- An evolving system
- Automatically partitioning expressed sequences
- Non-redundant set of gene-oriented clusters
NCBI


UniGene
Cluster for Human HBB
SEQUENCES
Sequences representing this gene; mRNAs, ESTs, and gene predictions supported by transcribed sequences.
mRNA sequences (20)
AF117710.1 Homo sapiens hemoglobin beta chain (HBB) mRNA, complete cds
NM_000518.4 Homo sapiens hemoglobin, beta (HBB), mRNA
AY509193.1 Homo sapiens hemoglobin beta mRNA, complete cds
CR536530.1 Homo sapiens full open reading frame cDNA clone
EST Sequences (10 of 2
BI518741.1 cDNA clone IMAGE:5211
BQ890006.1 cDNA clone IMAGE:6295
BQ898811.1 cDNA clone IMAGE:6302
Unix
PC
Mac
Download se
to text mode
GENE EXPRESSION
Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.
cDNA sources: Blood, Bone, Bone Marrow, Brain, Colon, Eye, Heart, Kidney, Liver, Lung, Lymph Node, Mammary Gland, Muscle, Ovary, Pancreas, Peripheral Nervous System, Placenta, Prostate, Skin, Soft Tissue, Spleen, Stomach, Testis, Thymus, Uterus, Vascular,Embryo, Juvenile, Adult
Restricted Expression: Embryo [Show more like this]
Expression Profile: View expression levels using UniGene's EST ProfileViewer
Note: Highly represented in many libraries
Breakdown by Tissue
Hs.523443
Bladder 0 0 / 21715
Blood 6603 517 / 78292
Bone 53 3 / 55730
Bone Marrow 3065 112 / 36541
Brain 123 57 / 462100
Cervix 0 0 / 41264
Colon 55 10 / 179987
Eye 130 22 / 168244
Heart 220 13 / 58912
Kidney 95 13 / 135458
Larynx 0 0 / 27551
Liver 167 22 / 131463
Lung 328 95 / 288794
Lymph Node 23 3 / 128142
Mammary Gland 124 16 / 128200
Muscle 2428 265 / 109115
Ovary 83 8 / 95612
Pancreas 555 47 / 84639
Peripheral ... 80 2 / 24996
Placenta 946 225 / 237797
Prostate 149 20 / 133636
Skin 66 11 / 165608
Small Intes... 0 0 / 14090
Soft Tissue 715 17 / 23760
Spleen 3473 67 / 19290
Stomach 18 2 / 108238
Tongue 0 0 / 28932
Testis 29 4 / 136540
Thymus 146 1 / 6848
Uterus 115 21 / 181622
Vascular 77 2 / 25883
Breakdown by Developmental Stage
Hs.523443
Embryo 1054 536 / 508346
Juvenile 33 2 / 59542
Adult 513 500 / 974089


GEO
Gene Expression Omnibus
http://www.ncbi.nlm.nih.gov/geo/
- First fully public high-throughput gene expression data repository
- Curated, online resource for gene expression data browsing, query and retrieval
GDS596: Large-scale analysis of the human transcriptome (HG-U133A)
GSM18957
GSM18958
GSM18961
GSM18962
GSM18989
GSM18990
GSM18985
GSM18986
GSM18987
GSM18988
GSM18963
GSM18984
GSM18951
GSM18952
GSM19007
prostate
testis
testis seminiferous tubule
testis germ cell
testis interstitial
testis leydig cell
heart
GSM18865
GSM18866
GSM18897
GSM18898
GSM18887
GSM18888
GSM18893
GSM18894
GSM18895
GSM18896
GSM18891
GSM18892
GSM18955
GSM18956
100
75
50
25
0
colorectal adenocarcinoma
leukemia myelogenous K562
leukemia lymphoblastic molt4
leukemia promyelocytic hl60
lymphoma burkitts daudi
lymphoma burkitts raji
kidney
NCBI

Homologene
Discover Homologs

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene

- An automated system
- Detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes

NCBI

Homologene for Human HBB
Discover Homologs

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

- H.sapiens HBB
  hemoglobin, beta.
- P.troglodytes LOC450978
  similar to beta globin; hemoglobin beta chain; beta globin mutant; beta globin chain.
- M.musculus LOC436003
  similar to hemoglobin beta chains - white rhinoceros.
- R.norvegicus Hbb
  hemoglobin beta chain complex.

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

- NP_000509.1 147 aa
- XP_508242.1 147 aa
- XP_488069.1 147 aa
- NP_150237.1 147 aa
- XP_215033.1 147 aa

Species	Gene	aa%ID	nt%ID	D	Ka/Ks	Knr/Knc	
H.sapiens	**HBB**						
vs. M.musculus	LOC436003	57.1	71.7	0.356	0.555	0.820	Blast
vs. R.norvegicus	Hbb	81.6	82.8	0.196	0.263	0.480	Blast
vs. R.norvegicus	LOC293265	78.9	81.0	0.220	0.271	0.694	Blast
vs. P.troglodytes	LOC450978	100.0	99.8	0.002	0	0	Blast

D	evolutionary distance
Ka/Ks	non-synonymous/synonymous changes
Knr/Knc	conserved/non-conserved changes

NCBI

OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University

- A catalog of human genes and genetic disorders at John Hopkins
- Developed for the World Wide Web by NCBI

+141900
HEMOGLOBIN--BETA LOCUS; HBB

ALLELIC VARIANTS
(selected examples)

mature protein

- 0001 HEMOGLOBIN AALBORG [HBB, GLY74ARG]
- 0002 HEMOGLOBIN ABRUZZO [HBB, HIS143ARG]
- 0003 HEMOGLOBIN AGENOGI [HBB, GLU90LYS]
- 0004 HEMOGLOBIN ALABAMA [HBB, GLN39LYS]
- 0005 HEMOGLOBIN ALAMO [HBB, ASN19ASP]
- 0242 HEMOGLOBIN RUSH [HBB, GLU101GLN]
- 0243 HEMOGLOBIN S [HBB, GLU6VAL]
- 0244 HEMOGLOBIN S (ANTILLES) [HBB, GLU6VAL AND VAL23ILE]
- 0245 HEMOGLOBIN S (OMAN) [HBB, GLU6VAL AND GLU121LYS]
- 0246 HEMOGLOBIN S (PROVIDENCE) [HBB, GLU6VAL AND LYS82.
- 0247 HEMOGLOBIN S (TRAVIS) [HBB, GLU6VAL AND ALA142VAL]
- 0248 HEMOGLOBIN SABINE [HBB, LEU91PRO]
- 0249 HEMOGLOBIN SAINT JACQUES [HBB, ALA140THR]

Glu7Val in the precursor
Hemoglobin S
sickle cell anemia

NCBI

Entrez Gene

PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books OMIM

for Go Clear current records only

Limits Preview/Index History Clipboard Details

Display Graphics Show: 5 Send to Text

1: HBB hemoglobin, beta [*Homo sapiens*]
GeneID: 3043 Locus tag: HGNC:4827; MIM: 141900
Transcripts and products: (shown on reverse complement genome) RefSeq belov

NC_000011

5204877 5203272
NM_000518 NP_000509
- coding region - untranslated region

Genomic context: chromosome: 11; **Maps:** 11p15.5

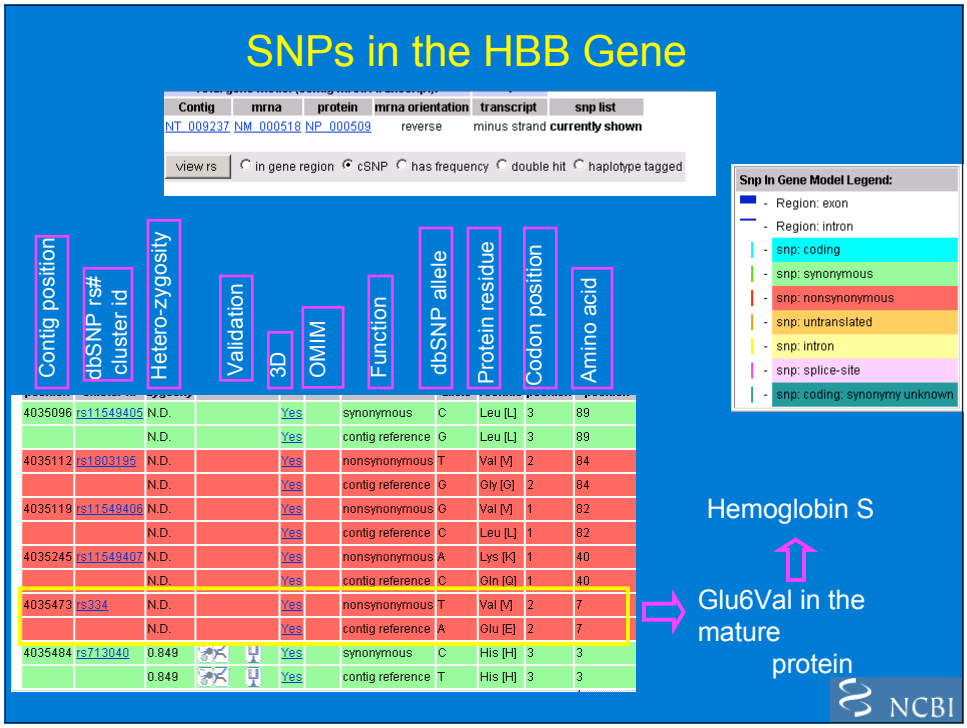
5155526 5221399
OR5221P OR51V1 HBB HBD HBBP1

Gene type: protein coding
Gene name: HBB
Gene description: hemoglobin, beta
RefSeq status: Reviewed
Organism: *Homo sapiens*
Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; M*
Primates; Catarrhini; Hominidae; Homo
Gene aliases: hemoglobin
Summary: The alpha (HBA) and beta (HBB) loci determine the structure of the 2 ty
in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two al
chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia.

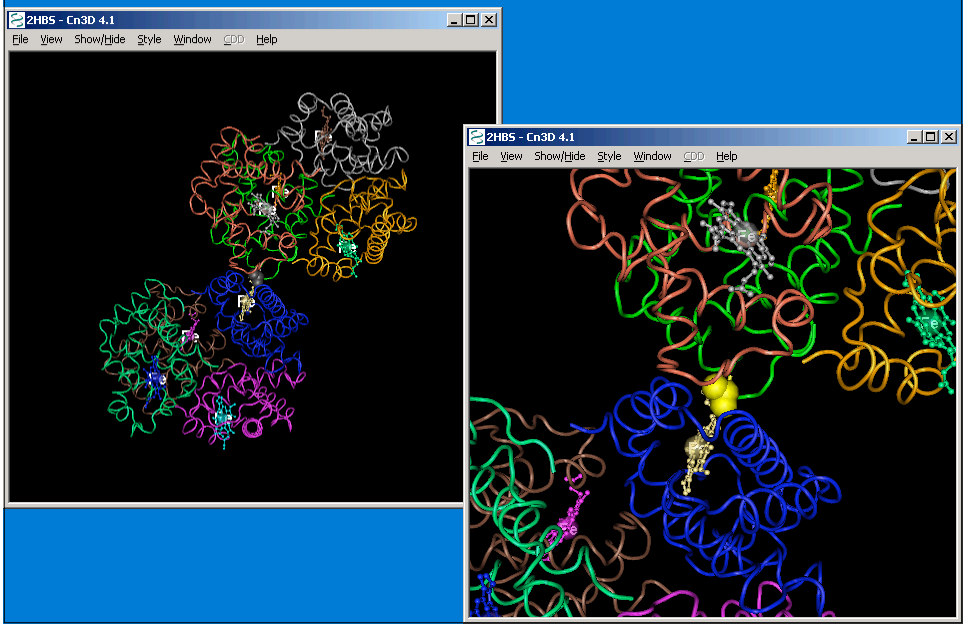
Links
- Conserved Domains
- GEO Profiles
- HomoloGene
- Map Viewer
- Nucleotide
- OMIM
- Protein
- PubMed
- SNP
- GeneView in dbSNP
- Taxonomy
- UniSTS
- AceView
- Ensembl
- Evidence Viewer
- GDB
- GeneTests for MIM: 141900
- Globin Gene Server
- HGNC
- LocusID
- MGC
- ModelMaker
- PharmGKB
- UCSC
- UniGene
- LinkOut

NCBI

SNPs in the HBB Gene



Structure of Deoxyhemoglobin S



NCBI

Literature Databases

Pubmed — Biomedical literature

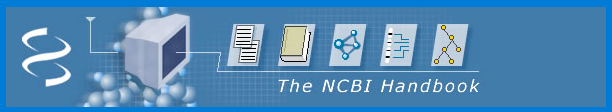
PubMed Central — Free online journals

http://www.pubmedcentral.gov

Books — Free online textbooks

Online Books

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=books



NCBI

Other Databases in Entrez

Cancer Choromosomes — chromosomal aberrations

NCI/NCBI SKY/M-FISH & CGH Database
NCI Mitelman Database of Chromosome Aberrations in Cancer
NCI Recurrent Aberrations in Cancer

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=CancerChromosomes

PubChem — catalog of small organic molecules

To support the Molecular Libraries and Imaging component of the NIH Roadmap Initiative

- chemical structures
- information on their biological activities

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

NCBI

NCBI Databases and Sequence Analysis Tools

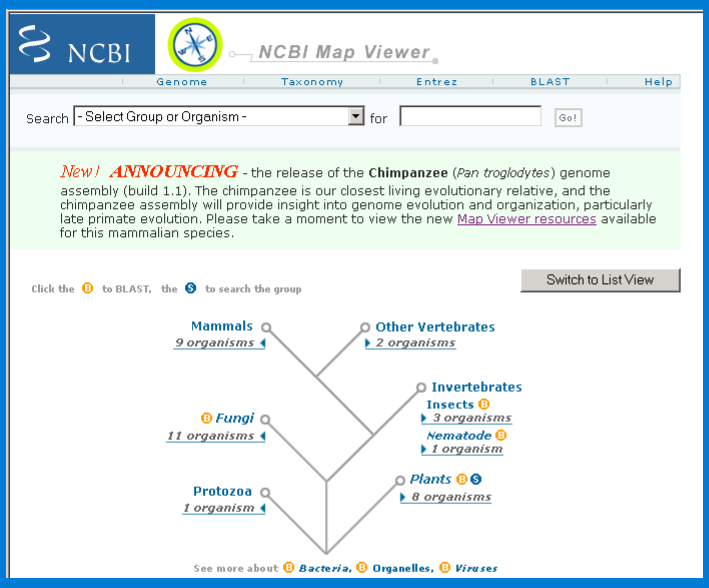
NCBI

An Array of Sequence Analysis Tools

http://www.ncbi.nlm.nih.gov/Tools/index.html

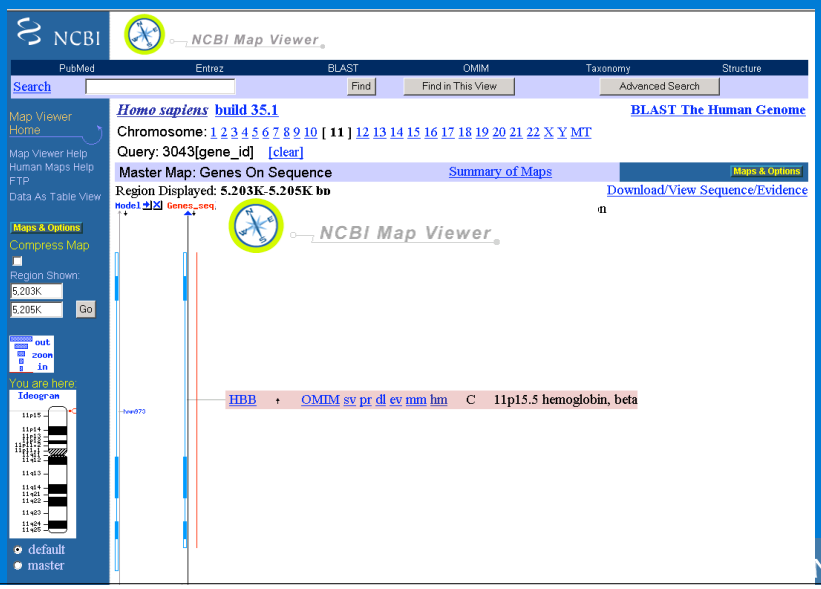
Nucleotide sequence analysis
Protein sequence analysis
Genome analysis
Structure
Gene expression

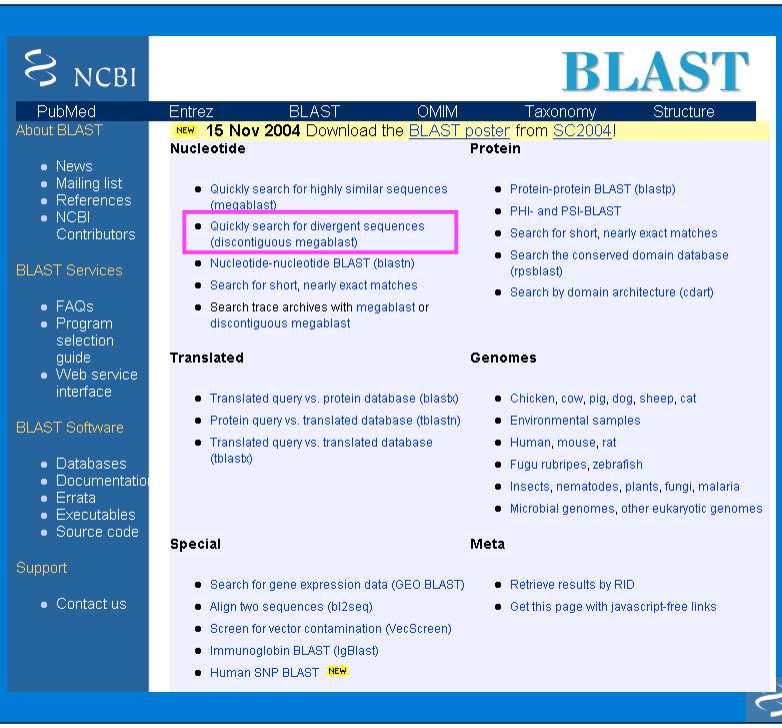
NCBI

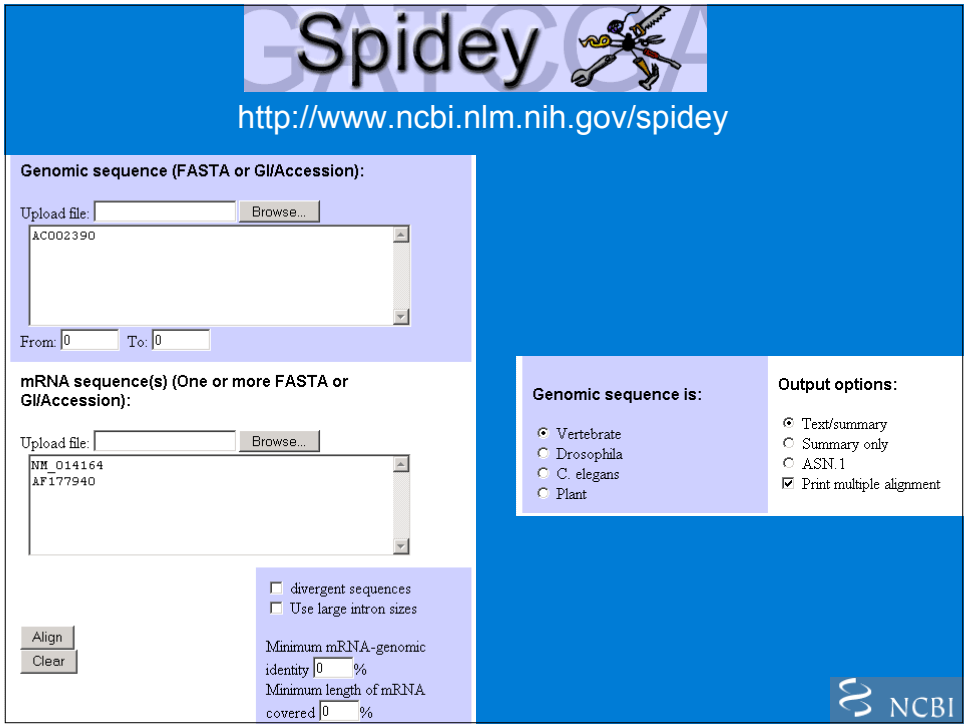


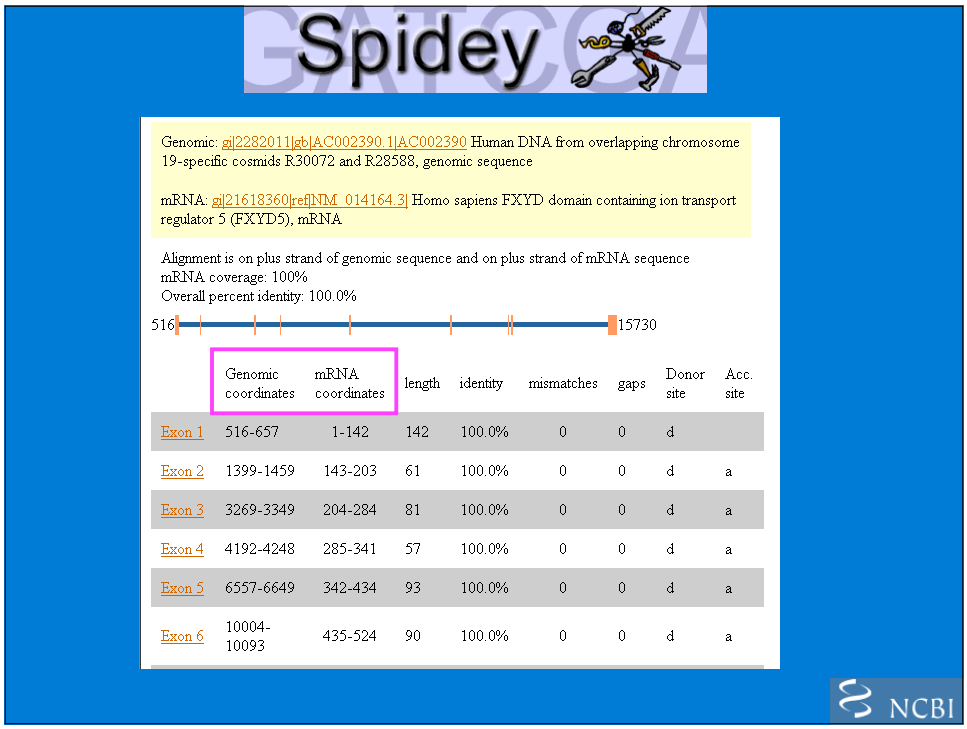
NCBI

HBB Gene in the Human Map Viewer





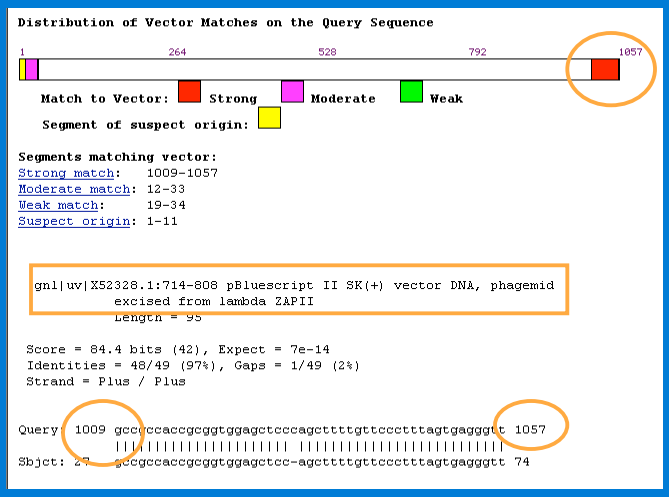




	Genomic coordinates	mRNA coordinates	length	identity	mismatches	gaps	Donor site	Acc. site
Exon 1	516-657	1-142	142	100.0%	0	0	d	
Exon 2	1399-1459	143-203	61	100.0%	0	0	d	a
Exon 3	3269-3349	204-284	81	100.0%	0	0	d	a
Exon 4	4192-4248	285-341	57	100.0%	0	0	d	a
Exon 5	6557-6649	342-434	93	100.0%	0	0	d	a
Exon 6	10004-10093	435-524	90	100.0%	0	0	d	a

VecScreen

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



NCBI

Outline

About NCBI
NCBI databases and tools
The Entrez- search and retrieval system
Training at NCBI

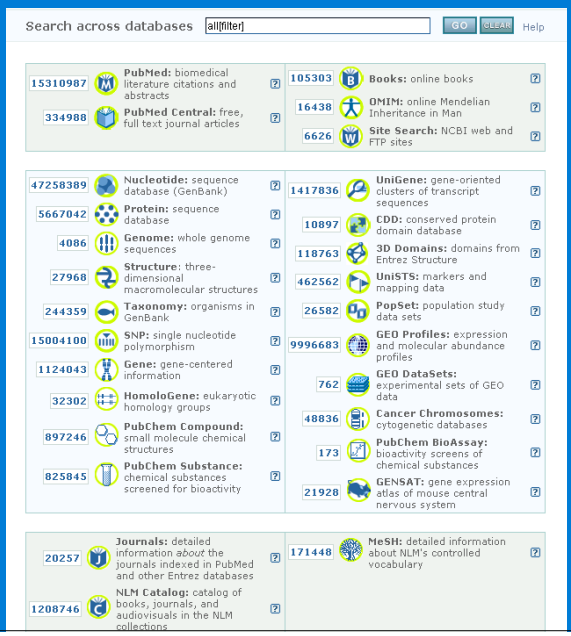
NCBI

Entrez, The Life Sciences Search Engine

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

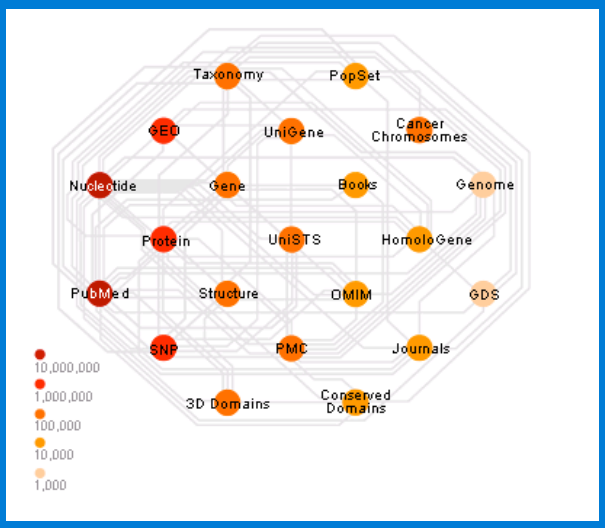
NCBI

Entrez: Search and Retrieval System



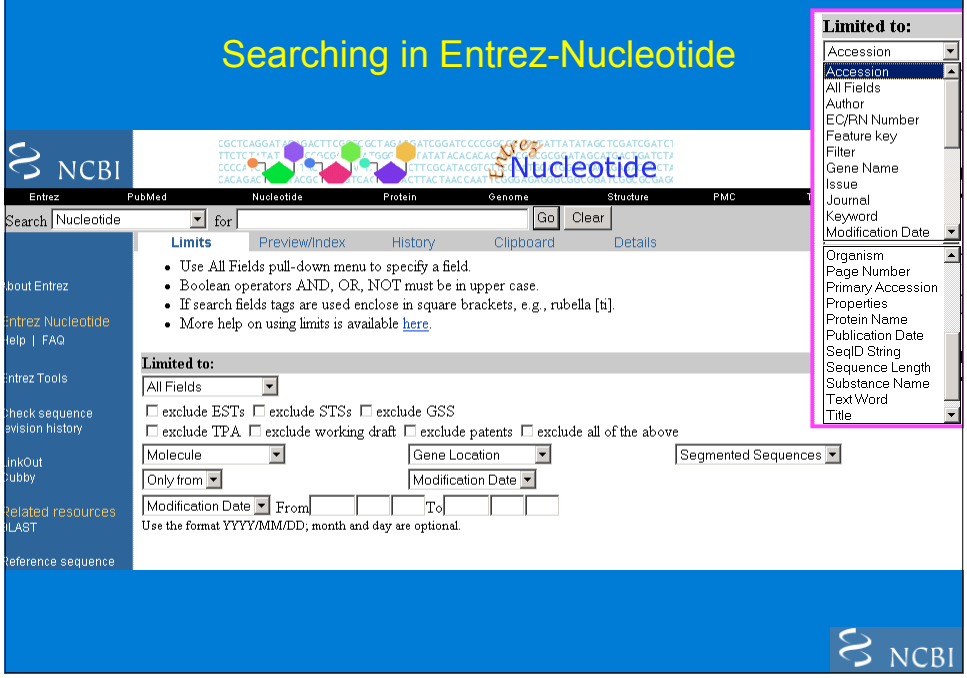
NCBI

Linking within Databases in Entrez



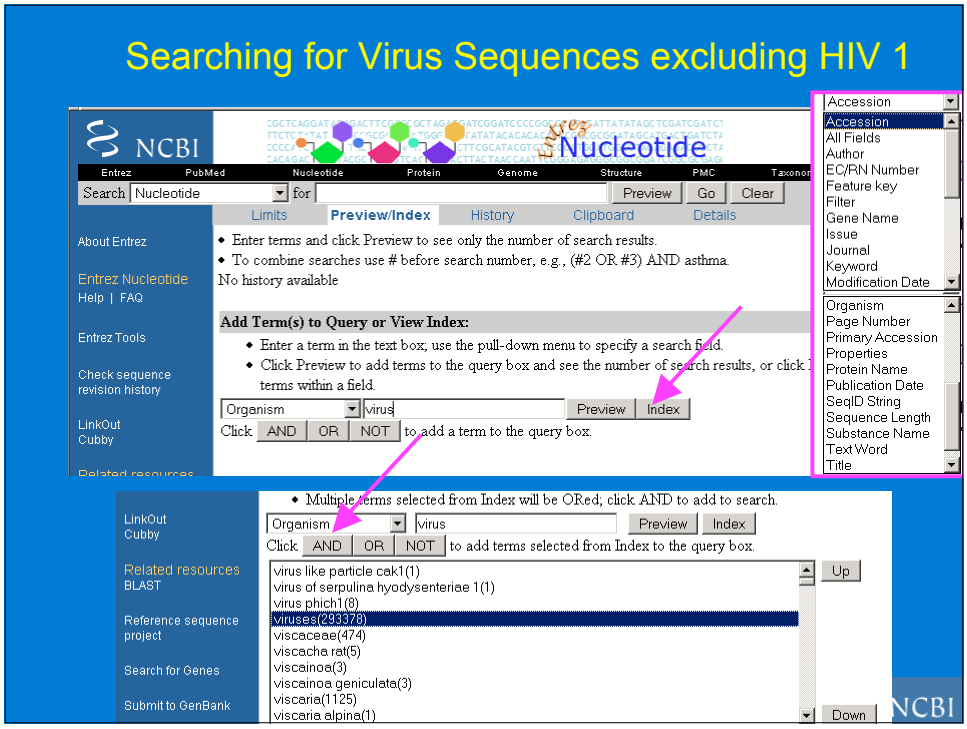
NCBI

Searching in Entrez-Nucleotide

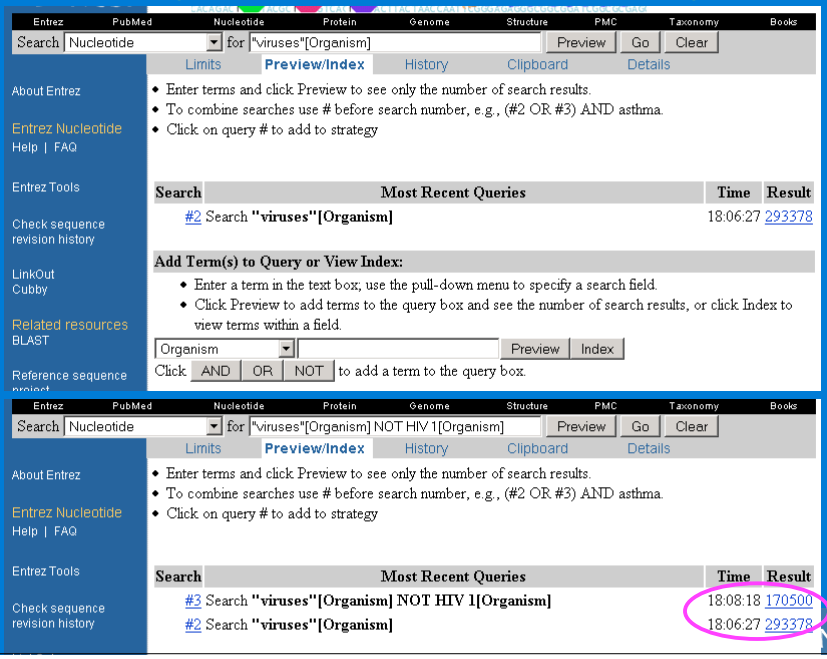


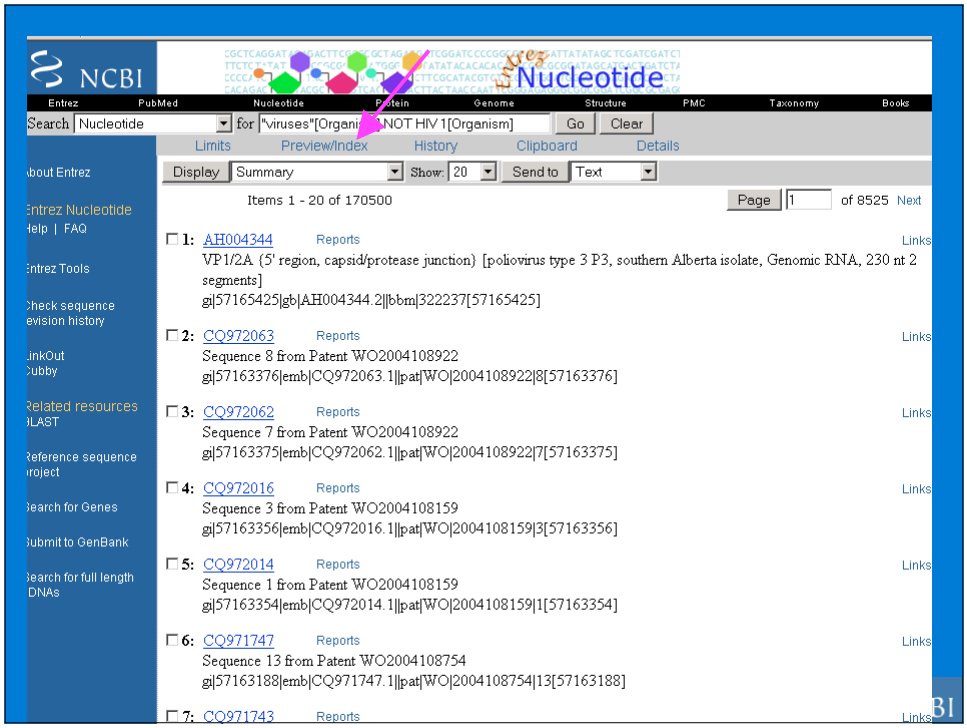
NCBI

Searching for Virus Sequences excluding HIV 1

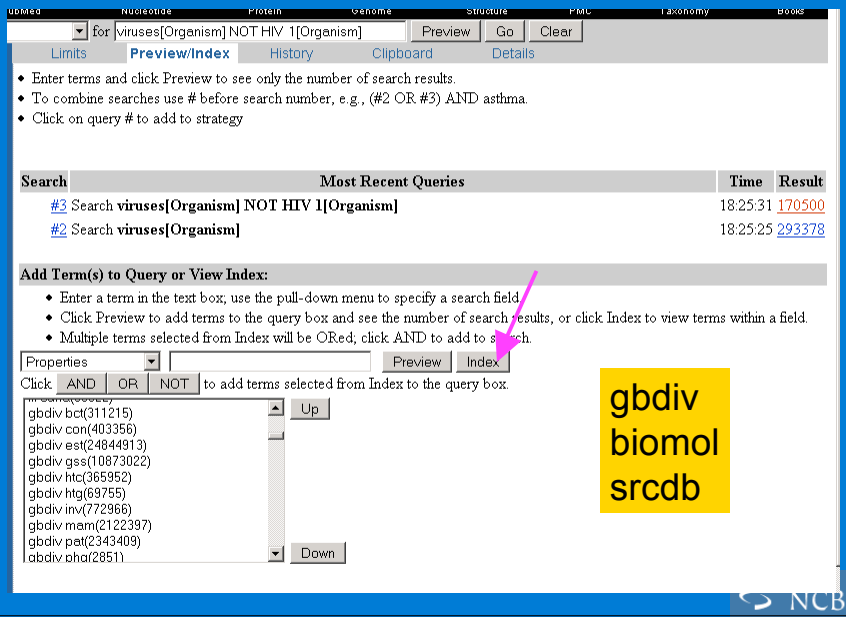


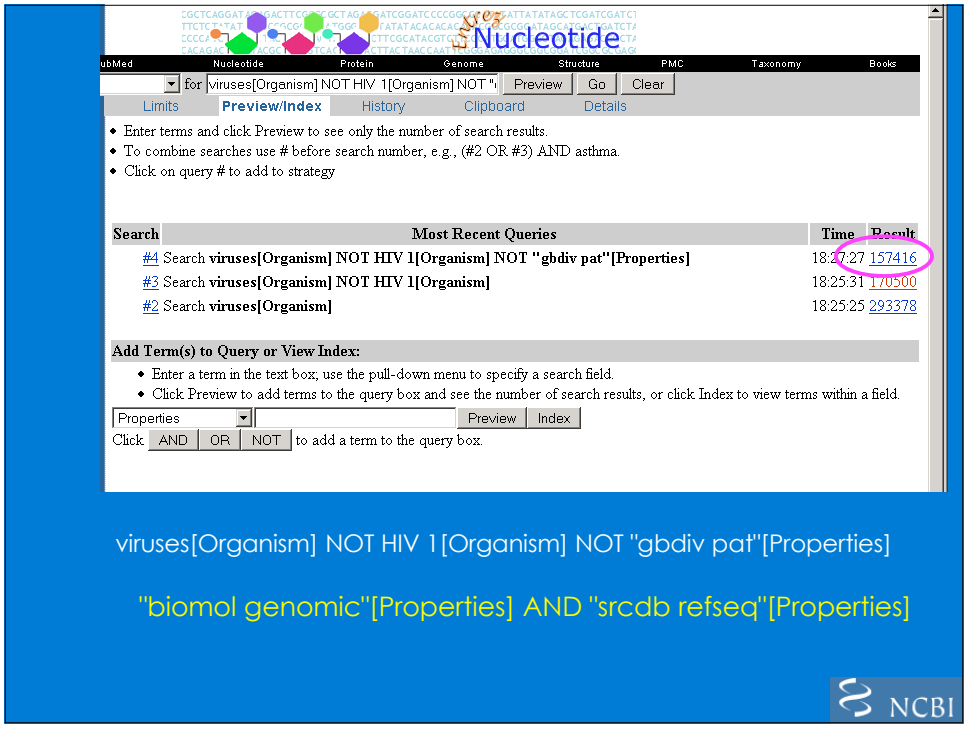
Searching for Virus Sequences excluding HIV 1

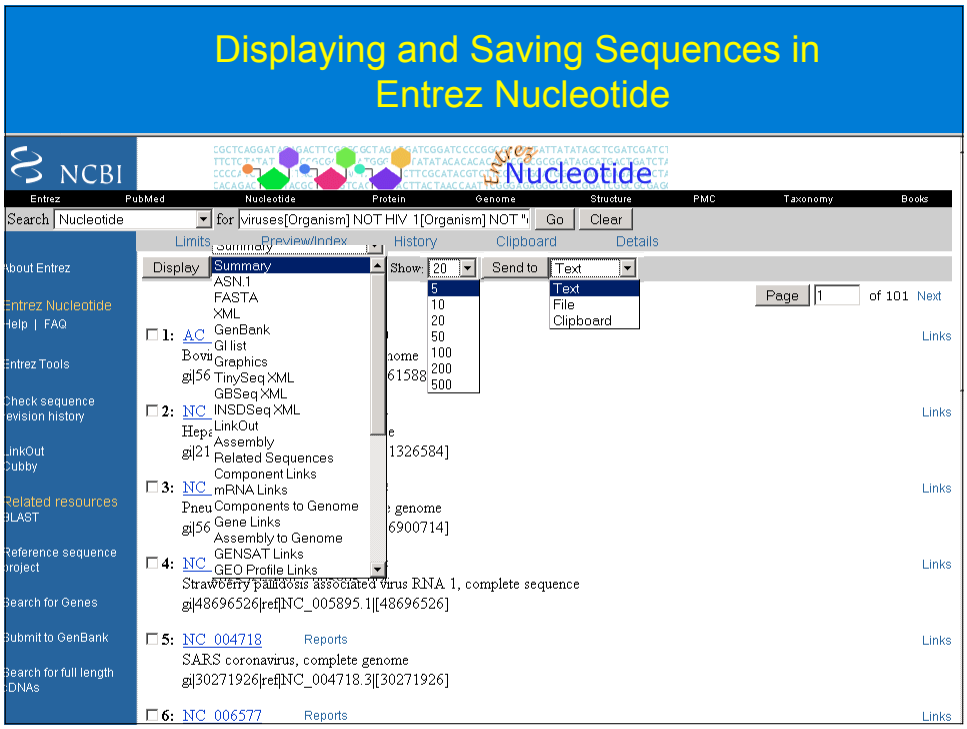




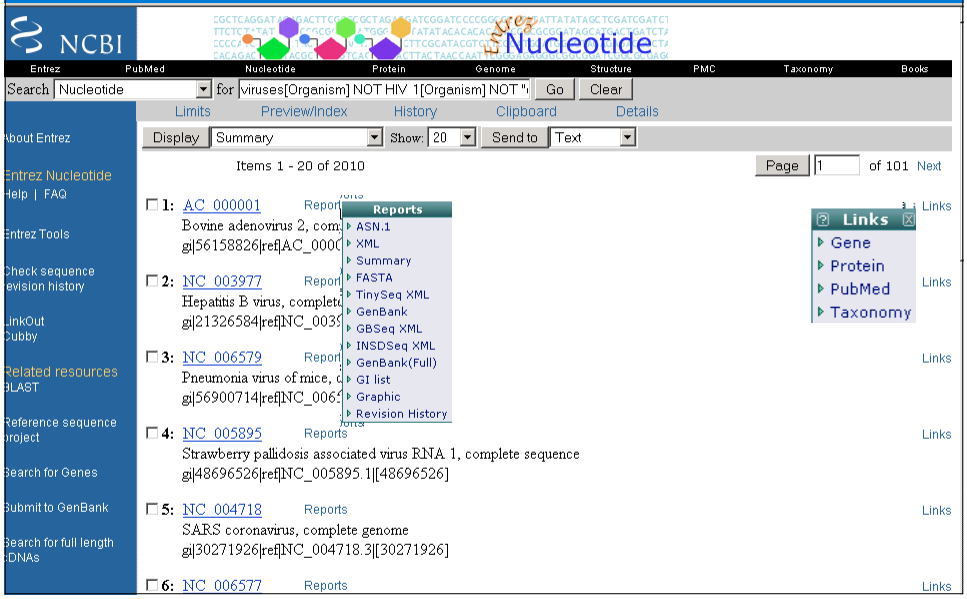
Searching in Entrez Nucleotide Properties Field

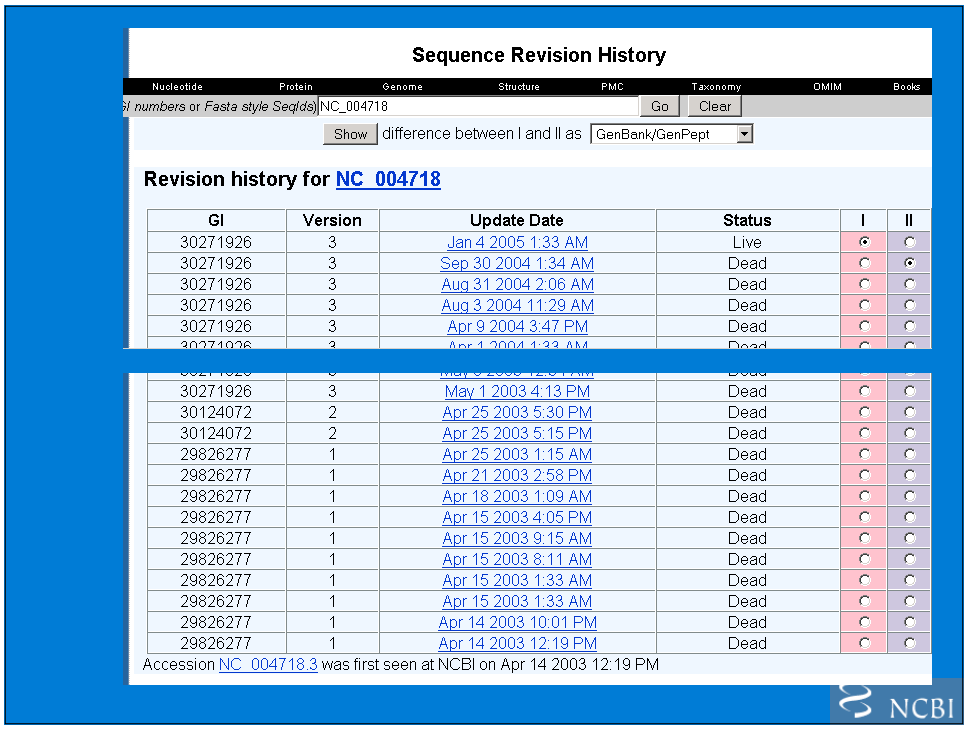



Nucleotide
PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
for viruses[Organism] NOT HIV 1[Organism] NOT "
Preview Go Clear
Limits Preview/Index History Clipboard Details
Enter terms and click Preview to see only the number of search results.
To combine searches use # before search number, e.g., (#2 OR #3) AND asthma.
Click on query # to add to strategy
Search Most Recent Queries Time Result
#4 Search viruses[Organism] NOT HIV 1[Organism] NOT "gbdiv pat"[Properties] 18:27:27 157416
#3 Search viruses[Organism] NOT HIV 1[Organism] 18:25:31 170500
#2 Search viruses[Organism] 18:25:25 293378
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.
Properties Preview Index
Click AND OR NOT to add a term to the query box.
viruses[Organism] NOT HIV 1[Organism] NOT "gbdiv pat"[Properties]
"biomol genomic"[Properties] AND "srcdb refseq"[Properties]
NCBI

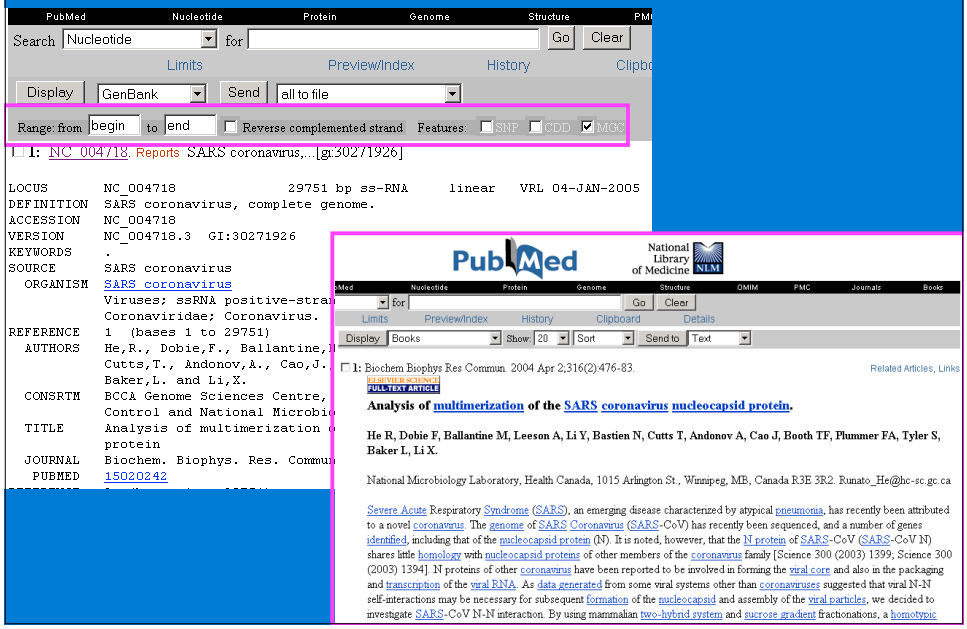

Displaying and Saving Sequences in Entrez Nucleotide
NCBI
Nucleotide
Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
Search Nucleotide for viruses[Organism] NOT HIV 1[Organism] NOT "
Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Entrez Nucleotide
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut Cubby
Related resources BLAST
Reference sequence project
Search for Genes
Submit to GenBank
Search for full length cDNAs
Display Summary Show: 20 Send to Text
Summary
ASN.1
FASTA
XML
GenBank
GI list
Graphics
TinySeq XML
GBSeq XML
INSDSeq XML
LinkOut
Assembly
Related Sequences
Component Links
mRNA Links
Components to Genome
Gene Links
Assembly to Genome
GENSAT Links
GEO Profile Links
5
10
20
50
100
200
500
Text
File
Clipboard
Page 1 of 101 Next
1: AC_ Links
2: NC_ Links
3: NC_ Links
4: NC_ Links
Strawberry pallidosis associated virus RNA 1, complete sequence
gi|48696526|ref|NC_005895.1|[48696526]
5: NC_004718 Reports Links
SARS coronavirus, complete genome
gi|30271926|ref|NC_004718.3|[30271926]
6: NC_006577 Reports Links

Searching in Entrez Nucleotide





Accessing the Sequence and Annotation Information



Examples of Searching in Entrez

Nucleotide:

Mouse EST sequences
mouse[Organism] AND "gbdiv est"[Properties]

DNA barcode sequences
"barcode"[Properties]

Protein:

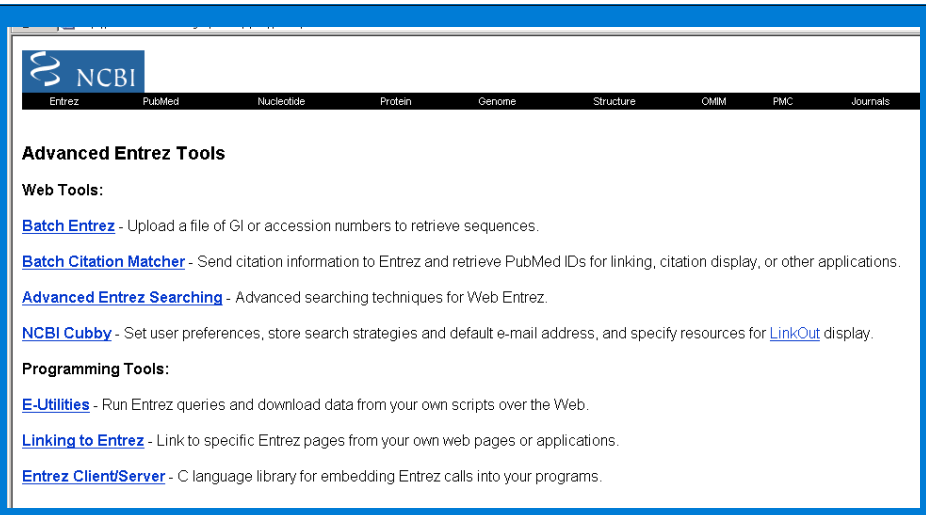
Peptide sequences of length between 40 and 50
40:50[Sequence Length]

Proteins with links to PubChem Compound
"protein pccompound"[Filter]

Homologene:

Entries for human disease genes
"link phenotype omim"[Properties]

NCBI



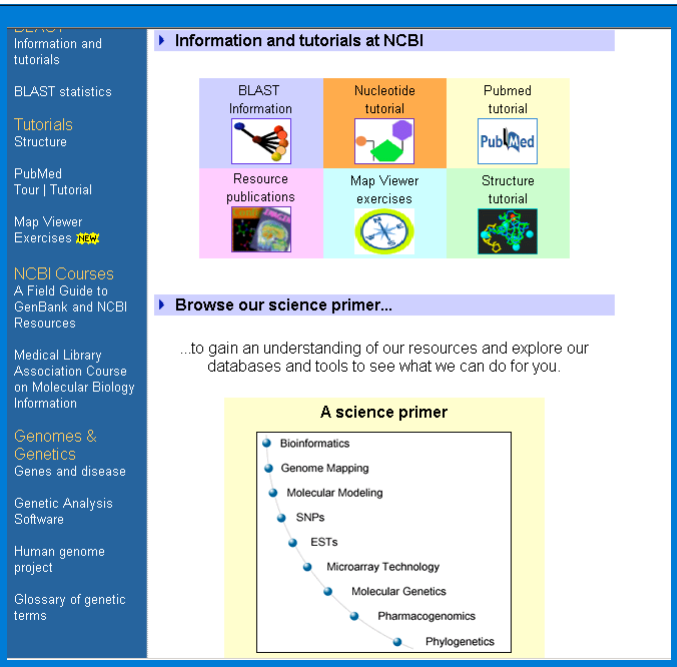
http://www.ncbi.nlm.nih.gov/entrez/query/static/advancedentrez.html

NCBI

Outline

About NCBI
NCBI databases and tools
The Entrez- search and retrieval system
Training at NCBI

NCBI



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

NCBI

NCBI Training

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

NCBI Field Guide

A Field Guide to GenBank and NCBI Molecular Biology Resources

3 hour lecture and 2 hour hands-on

Mini Courses

on specific topics 2 hour lecture and hands-on

POWER NCBI ENTREZ E-utilities course tools

Three day workshops at NCBI

NCBI

NCBI Core Bioinformatics Facility

- Supports a network of bioinformatics specialists serving individual institutes at NIH
- Trains Core Members in the use of NCBI tools
- The Core Members, in turn, support the use of NCBI's tools and databases by researchers in their institutes
- Currently 18 Members from 14 institutes

Refer to the handout for the Core Member from your institute

NCBI

Access More Information at

1.

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books

2. Database Resources of the National Center for Biotechnology Information
Nucleic Acids Res. 2005 Jan 1;33 Database Issue:D39-45

3. GenBank
Nucleic Acids Res. 2005 Jan 1;33 Database Issue:D34-8

NCBI

Outline

About NCBI
NCBI databases and tools
The Entrez- search and retrieval system
Training at NCBI