NCBI web resources I: databases and Entrez

Yanbin Yin

Most materials are downloaded from ftp://ftp.ncbi.nih.gov/pub/education/

Homework assignment 1

- Two parts:
- 1: Extract the gene IDs reported in table 1 of http://www.ncbi.nlm.nih.gov/pmc/articles/PMC523881/ Using NCBI batch Entrez to download all refseq protein fasta sequences from Arabidopsis thaliana
- 2: Read https://academic.oup.com/nar/article/40/D1/D136/2903327 and use your own language to describe what is NCBI taxonomy database and what it can do.

- Write a report (in word or ppt) to explain all the operations and include screen shots; save file with your last name in the file name.

Due on 9/12 (send by email)

References

- NCBI mcbios workshop
  - ftp://ftp.ncbi.nih.gov/pub/education/mcbios2012/
- NCBI web resource tutorials
  - ftp://ftp.ncbi.nih.gov/pub/education/tutorials/
- NCBI discovery workshops
  - ftp://ftp.ncbi.nih.gov/pub/education/discovery_workshops/NLM/2012/Sept2012/
- NCBI Help Manual
  - http://www.ncbi.nlm.nih.gov/books/NBK3831/

Youtube

- http://www.youtube.com/ncbinlm

- Go to www.youtube.com
- Search "NCBI tutorial general"

Topics

- Intro. to NCBI
- Selected NCBI Databases
- The Entrez system
- Hands on practice

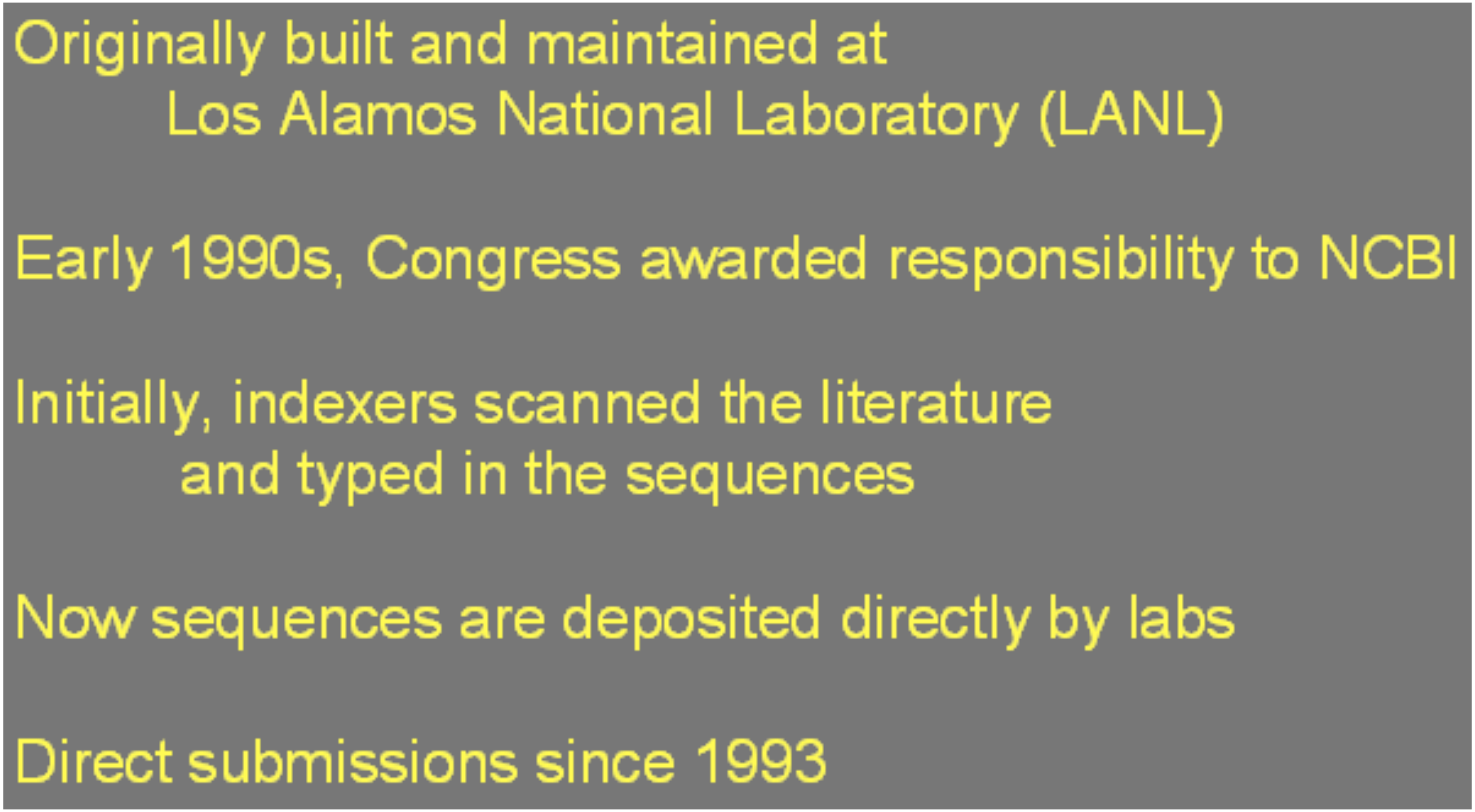
The National Center for Biotechnology Information

Bethesda,MD

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

GenBank history



Originally built and maintained at
Los Alamos National Laboratory (LANL)

Early 1990s, Congress awarded responsibility to NCBI

Initially, indexers scanned the literature
and typed in the sequences

Now sequences are deposited directly by labs

Direct submissions since 1993

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

Molecular Data

- Sequences
- Expression
- Genome Maps
- 3D Structures
- Protein Domains
- Homologous Genes, Proteins, Structures
- Pathways
- Genetic Variation

Selected NCBI Databases

- Biomedical literature
  - PubMed free Medline
  - PubMed Central full text online access
  - NCBI Bookshelf online biomedical textbooks
- Biomolecular Databases
  - Nucleotide
    - GenBank submitted sequence records
    - RefSeq curated NCBI reference sequences
  - Protein GenBank and RefSeq translations, outside protein
  - dbSNP small scale genetic variations
  - Structure biomolecular 3-D structures
    - MMDB NCBI's 3D structure database
  - GEO microarray expression data
  - SRA next-generation sequence data

Information Hubs: Aggregators

- Taxonomy access to NCBI data through source organism classification
- BioProjects molecular data and literature related to large scale molecular projects (genomes, transcriptomes, metagenomes)
- Genome specialized displays for complete genomes and access to microbial genome analysis tools
- Gene molecular data and literature related to genes
- BioSystems biochemical pathways and processes linked to NCBI genes, gene products, small molecules, and structures

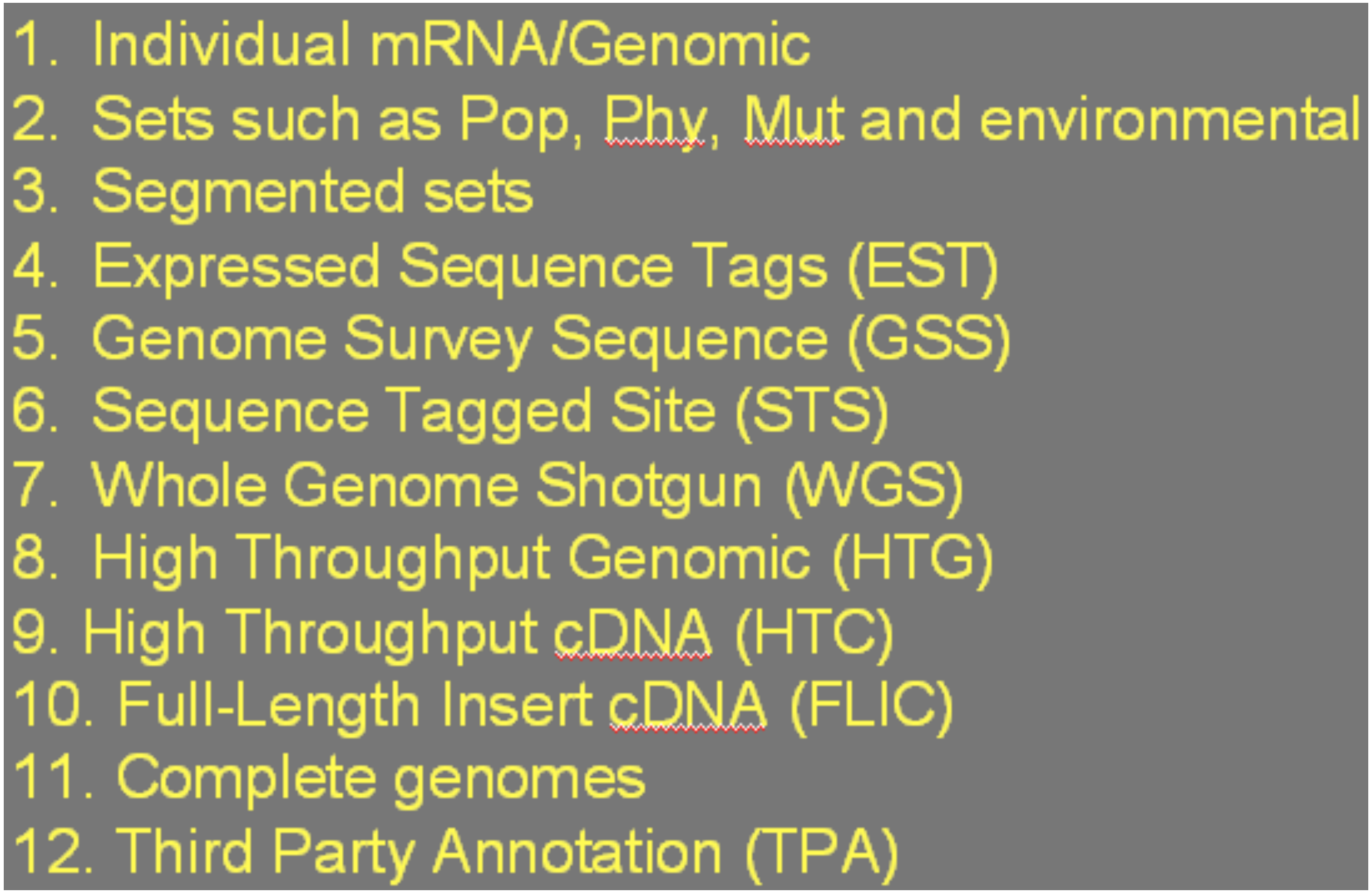
Information Hubs: Analyses

- Analysis Results
  - HomoloGene homologous genes from selected eukaryotes
  - Protein Clusters homologs (proteins) from microbial genomes
  - UniGene sequence-based gene catalog (eukaryotes)
  - GEO Datasets microarray experiments and analyses

Sequence Databases at NCBI

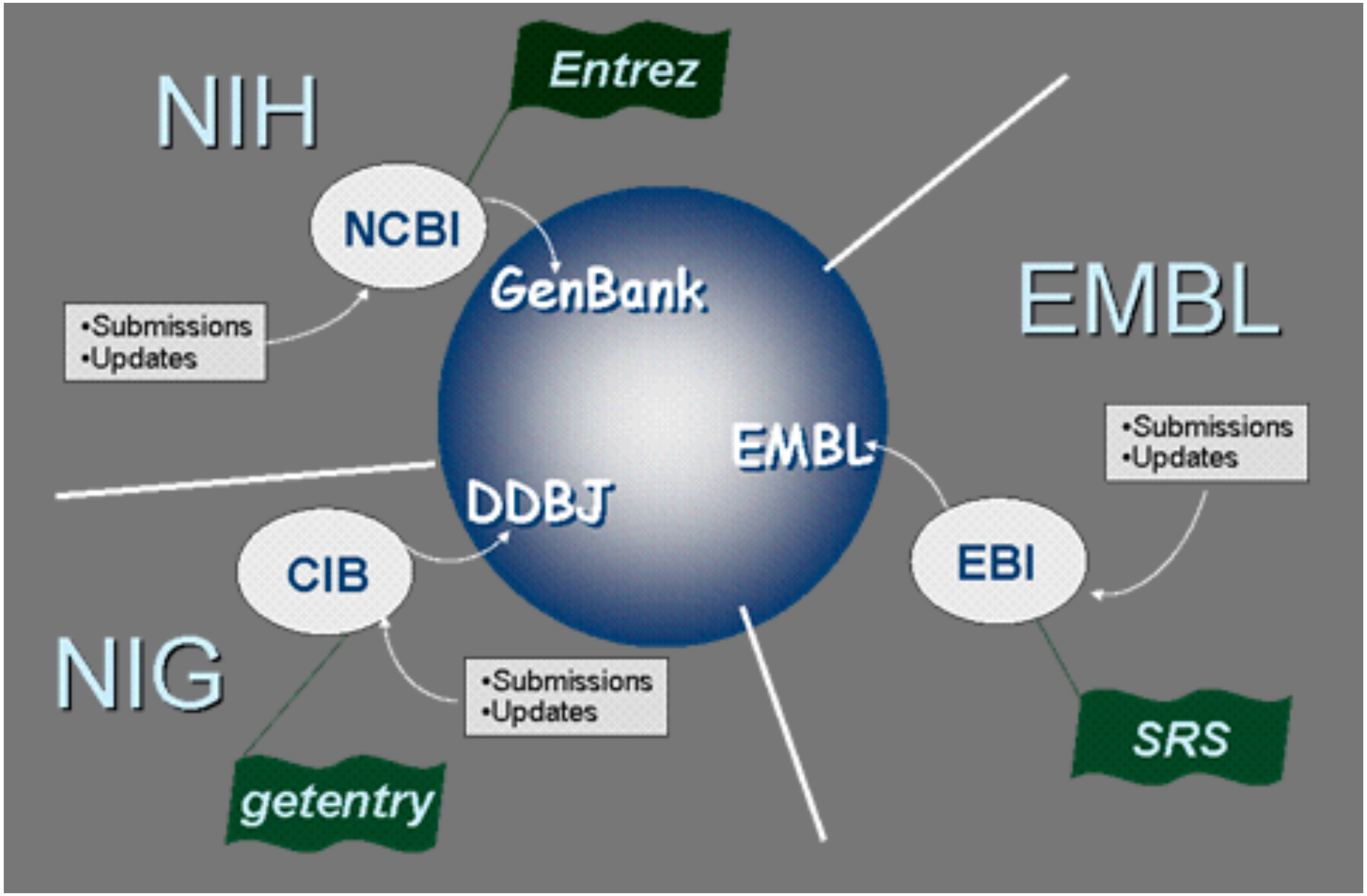
- Primary
  - GenBank: NCBI's primary sequence database
  - Trace Archive: reads from capillary sequencers
  - Sequence Read Archive: next generation data
- Derivative
  - GenPept (GenBank translations)
  - Outside Protein (UniProt—Swiss-Prot, PDB)
  - NCBI Reference Sequences (RefSeq)

GenBank types of entries



1. Individual mRNA/Genomic
2. Sets such as Pop, Phy, Mut and environmental
3. Segmented sets
4. Expressed Sequence Tags (EST)
5. Genome Survey Sequence (GSS)
6. Sequence Tagged Site (STS)
7. Whole Genome Shotgun (WGS)
8. High Throughput Genomic (HTG)
9. High Throughput cDNA (HTC)
10. Full-Length Insert cDNA (FLIC)
11. Complete genomes
12. Third Party Annotation (TPA)

Three international nucleotide sequence databases

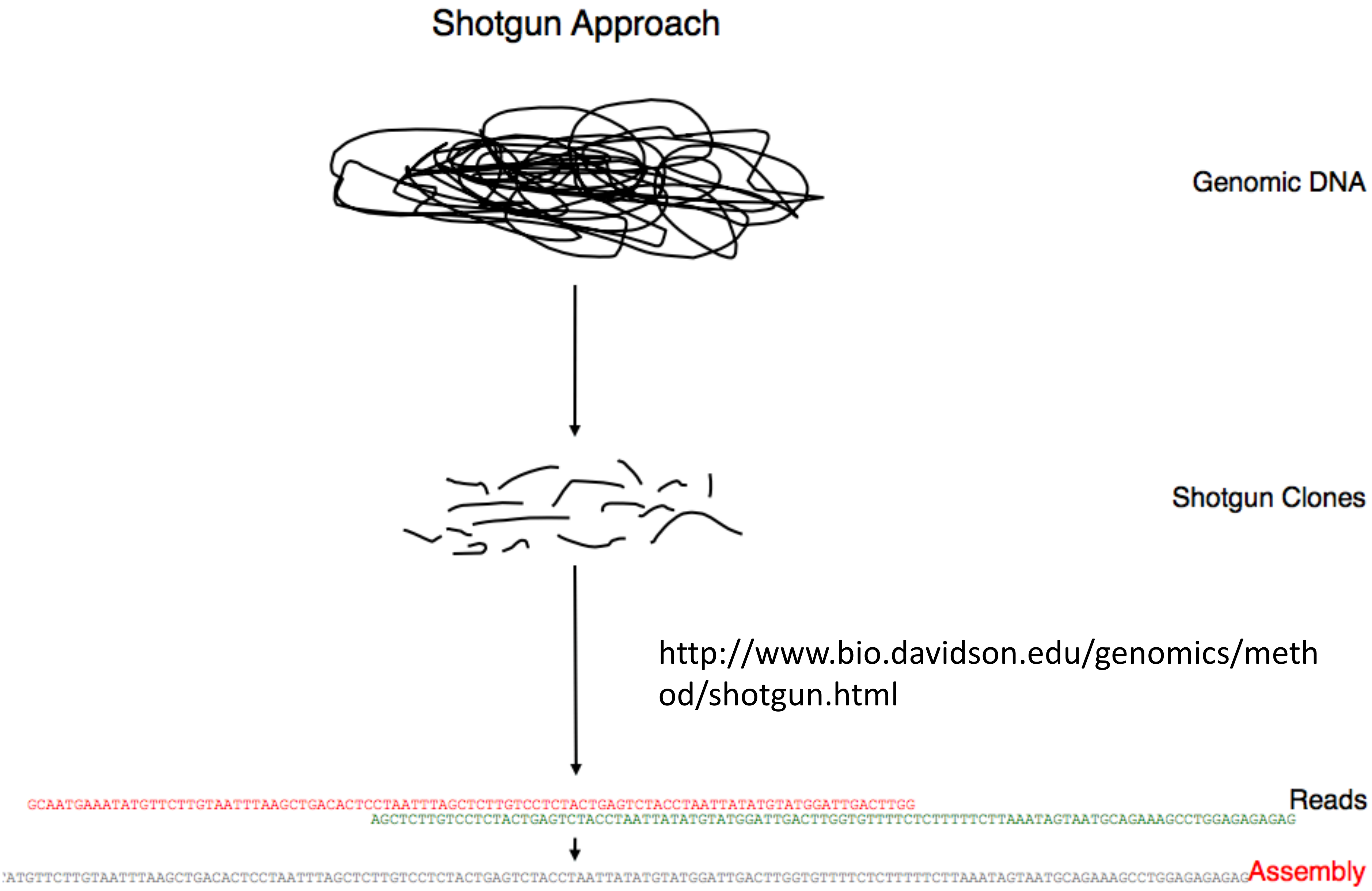


RefSeq: NCBI's Derivative Sequence Database

- **Experimentally verified / curated transcripts and proteins**
  - **NM_, NP_ accession numbers**
- **Model transcripts and proteins**
  - **XM_, XP_ accession numbers**
- **Assembled Genomic Regions (contigs)**
  - **NT_, NW_ accession numbers**
- **Chromosome records**
  - **NC_, AC_ accession numbers**
- **RefSeqGene Records**
  - **NG_ accession numbers (NG_ also used pseudo genes and other fixed genomic sequences)**
- **Draft whole genome shotgun assemblies (microbial)**
  - **NZ_ accession numbers**
- **Microbial proteins**
  - **NP_, YP_, ZP_ accessions**

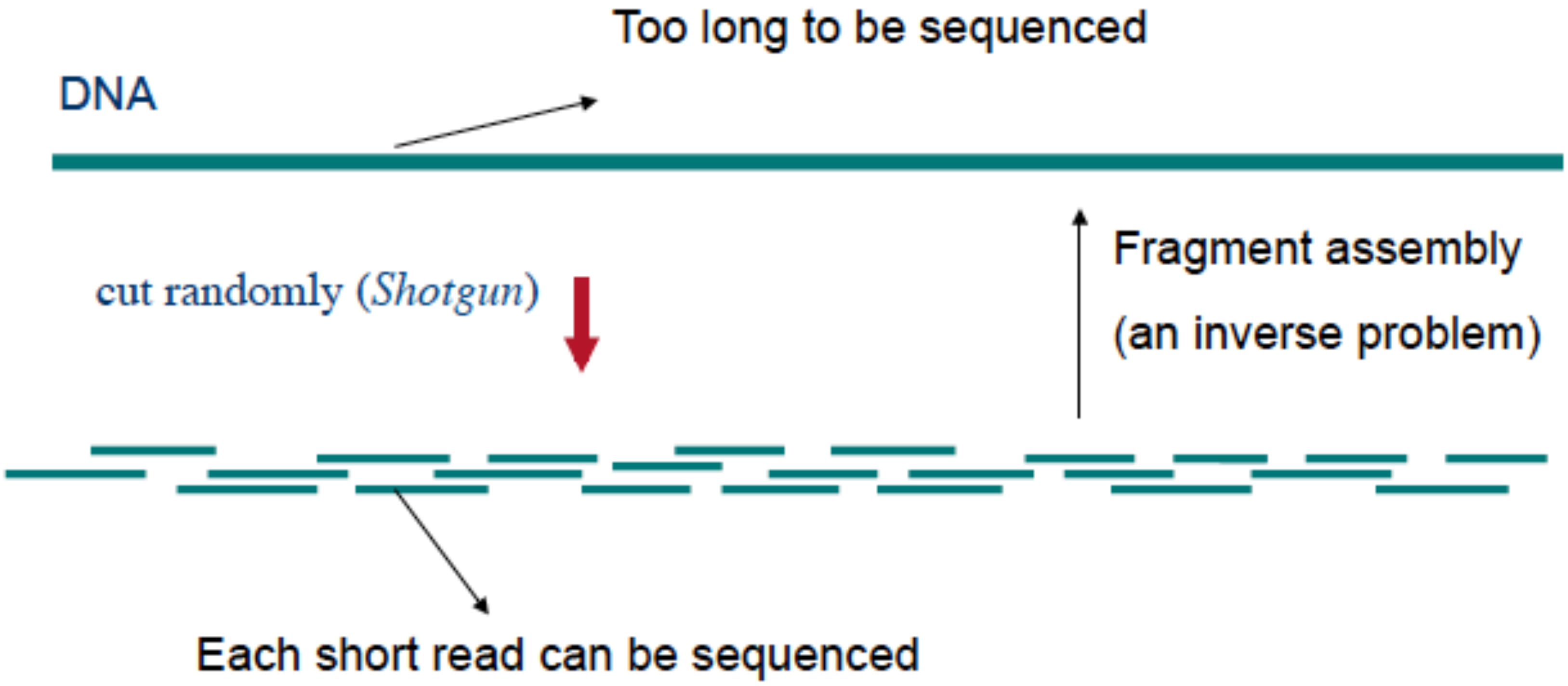
ftp://ftp.ncbi.nih.gov/refseq/release/

Whole Genome Sequencing Approaches



http://www.bio.davidson.edu/genomics/method/shotgun.html

Shotgun sequencing



Too long to be sequenced

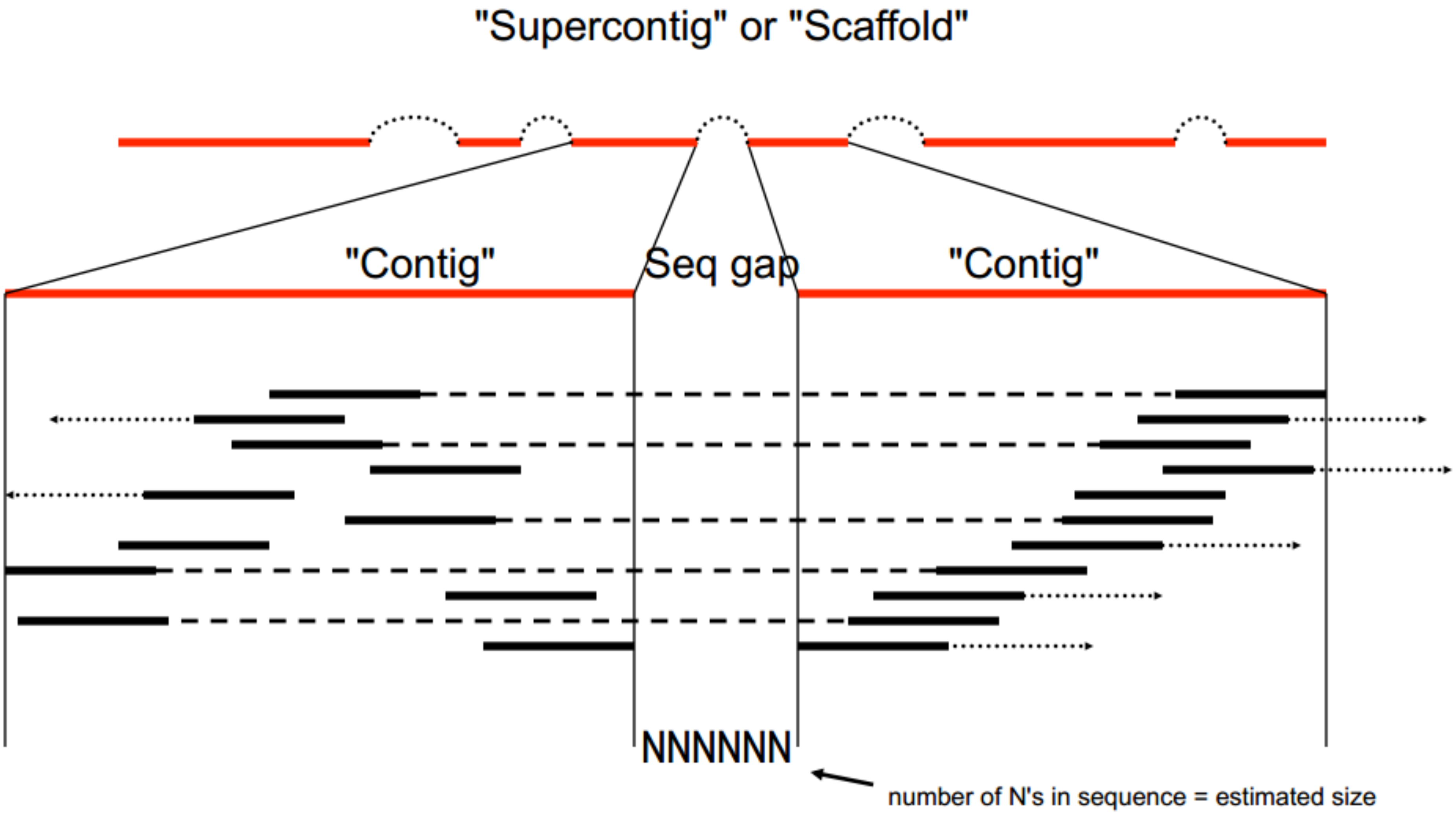
DNA

cut randomly (*Shotgun*)

Fragment assembly

(an inverse problem)

Each short read can be sequenced



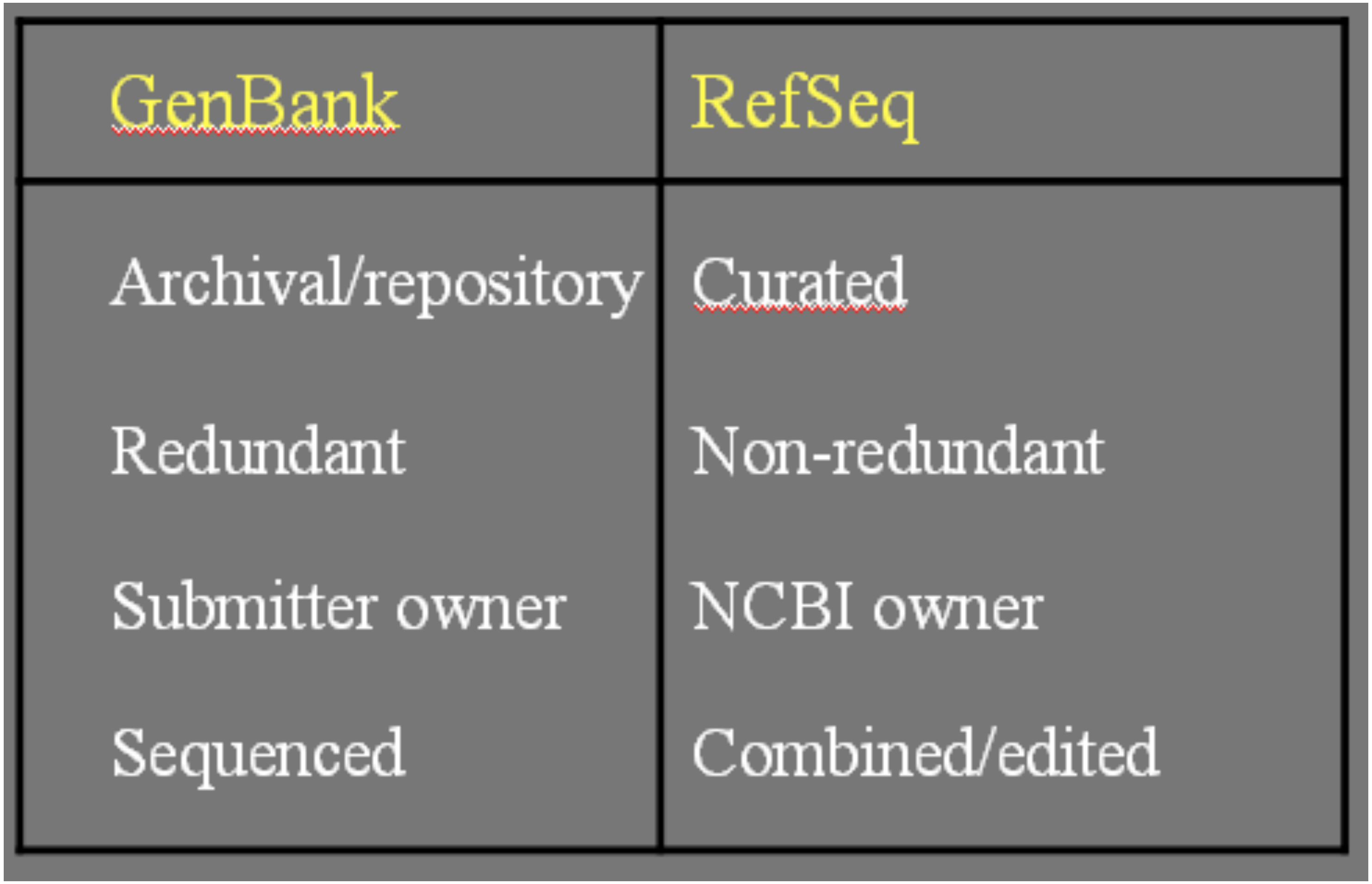
Sequence assembly is the problem of merging and ordering shorter fragments, termed “reads,” sampled from a set of larger sequences in order to reconstruct the larger sequences. The output of an assembly is typically a set of “contigs,” which are contiguous sequence fragments, ordered and oriented into “scaffold” sequences, with gaps between contigs within scaffolds representing regions of uncertainty

Genome assemblies are composed of scaffolds and contigs.

Contigs are contiguous consensus sequences that are derived from collections of overlapping reads (no gaps).

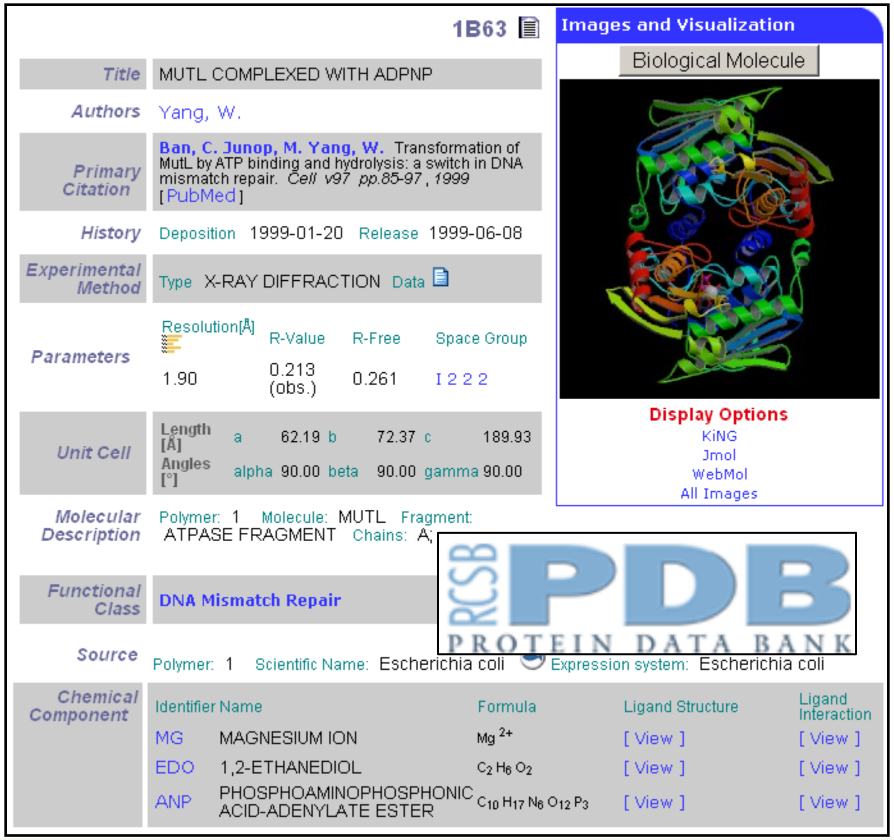
Scaffolds are ordered and orientated sets of contigs that are linked to one another by mate pairs of sequencing reads (have gaps).

GenBank & RefSeq



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

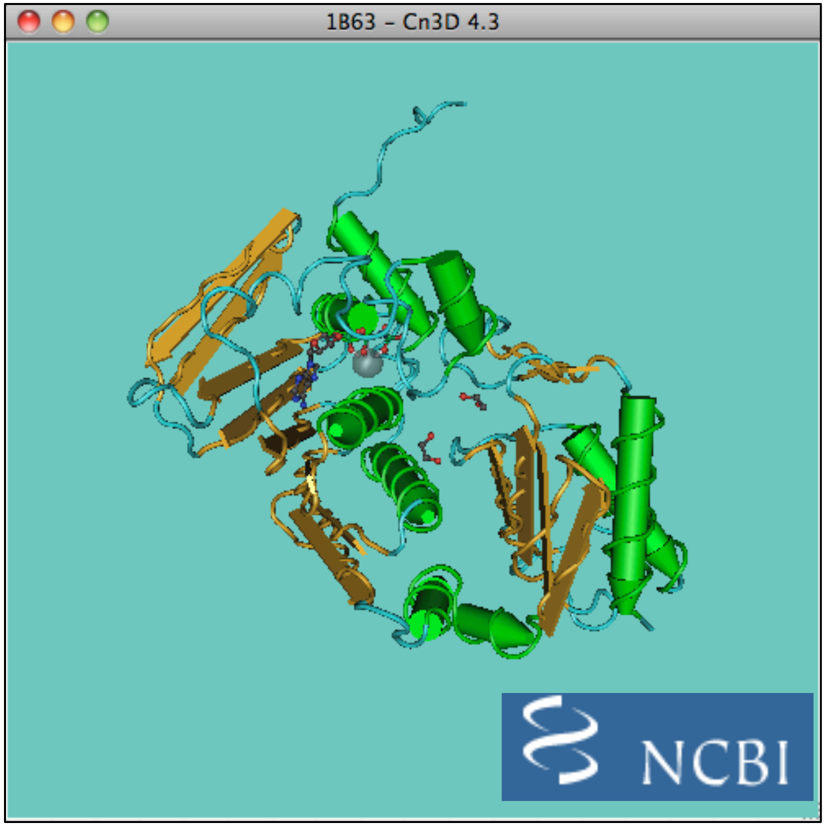
Protein Sequences from Structures


1B63

Title	MUTL COMPLEXED WITH ADPNP
Authors	Yang, W.
Primary Citation	**Ban, C. Junop, M. Yang, W.** Transformation of MutL by ATP binding and hydrolysis: a switch in DNA mismatch repair. *Cell v97 pp.85-97 , 1999* [PubMed]
History	Deposition 1999-01-20 Release 1999-06-08
Experimental Method	Type X-RAY DIFFRACTION Data
Parameters	Resolution[Å] 1.90 · R-Value 0.213 (obs.) · R-Free 0.261 · Space Group I 2 2 2
Unit Cell	Length [Å] a 62.19 b 72.37 c 189.93 Angles [°] alpha 90.00 beta 90.00 gamma 90.00
Molecular Description	Polymer: 1 Molecule: MUTL Fragment: ATPASE FRAGMENT Chains: A;
Functional Class	**DNA Mismatch Repair**
Source	Polymer: 1 Scientific Name: Escherichia coli Expression system: Escherichia coli

Chemical Component

Identifier	Name	Formula	Ligand Structure	Ligand Interaction
MG	MAGNESIUM ION	Mg^{2+}	[View]	[View]
EDO	1,2-ETHANEDIOL	$C_2H_6O_2$	[View]	[View]
ANP	PHOSPHOAMINOPHOSPHONIC ACID-ADENYLATE ESTER	$C_{10}H_{17}N_6O_{12}P_3$	[View]	[View]



```
>gi|5542073|pdb|1B63|A Chain A, Mutl Complexed With Adpnp
SHMPIQVLPPQLANQIAAGEVVERPASVVKELVENSLDAGATRIDIDIERGGAKLIRIRDNGCGIKKDEL
ALALARHATSKIASLDDLEAIISLGFRGEALASISSVSRLTLTSRTAEQQEAWQAYAEGRDMNVTVKPAA
HPVGTTLEVLDLFYNTPARRKFLRTEKTEFNHIDEIIRRIALARFDVTINLSHNGKIVRQYRAVPEGGQK
ERRLGAICGTAFLEQALAIEWQHGDLTLRGWVADPNHTTPALAEIQYCYVNGRMMRDRLINHAIRQACED
KLGADQQPAFVLYLEIDPHQVDVNVHPAKHEVRFHQSRLVHDFIYQGVLSVLQ
```

MMDB: **M**olecular **M**odeling **D**ata **B**ase

- Derived from experimentally determined PDB records
- Value added to PDB records including:
  - Addition of explicit chemical graph information
  - Validation (secondary structure elements)
  - Inclusion of Taxonomy, Citation
  - Conversion to ASN.1 data description language
- Structure neighbors determined by Vector Alignment Search Tool (VAST)

Protein Domains

- Structural Domain
  - Discrete independently folding unit of a protein
- Conserved Domain (sequence-based)
  - Protein region with recognizable position-specific pattern of sequence conservation
- Sequence-based domains often roughly correspond to structural domains
- Domains often have distinct, identifiable functions

NCBI's Conserved Domain Database

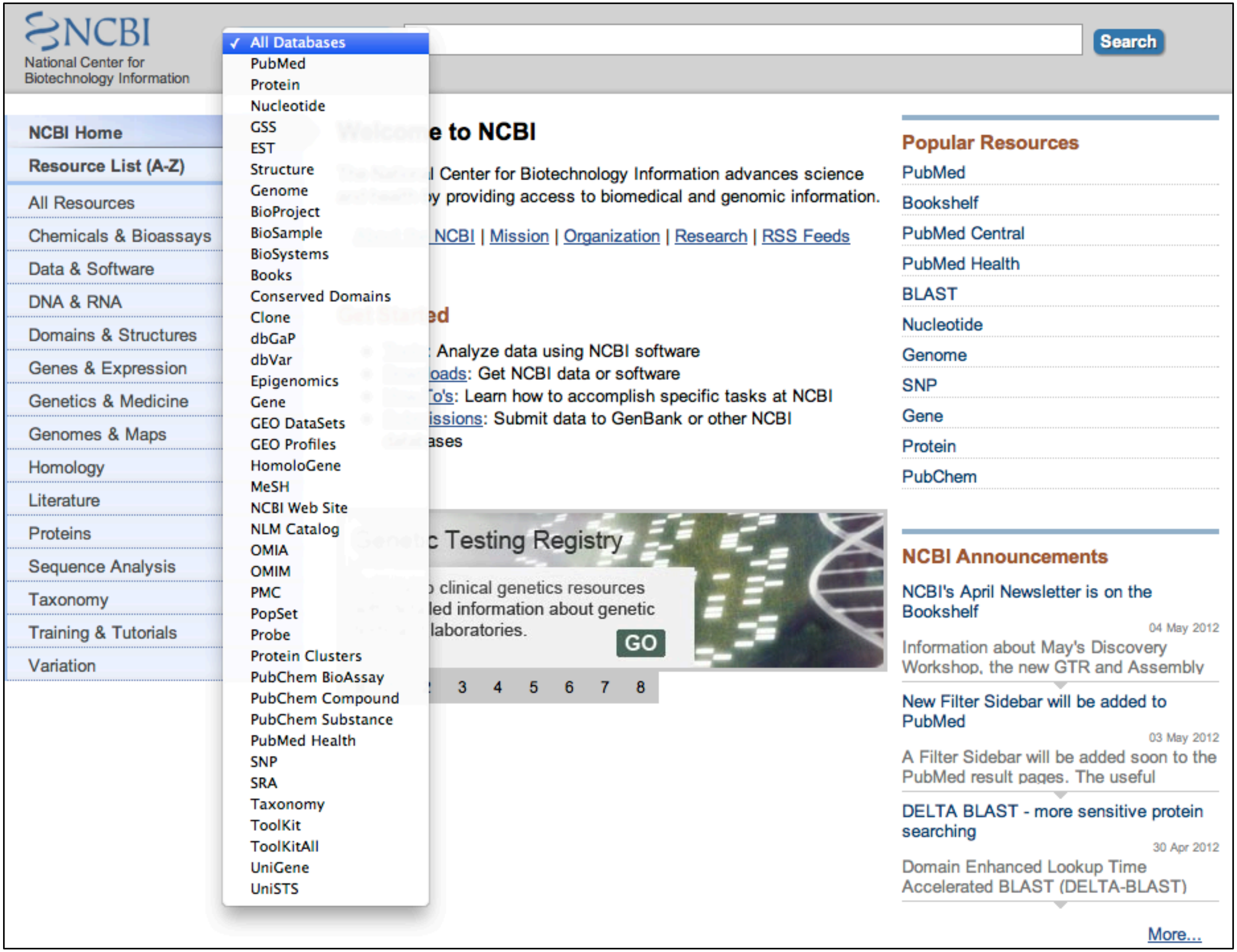
- Searchable with RPS-BLAST
- Sources
  - SMART
  - PFAM
  - COGs
  - NCBI curated domains
    - structure-informed alignments

NCBI Search Services and Tools

- Entrez integrated literature and molecular databases
  - BLink protein similarities
  - Graphical Sequence Viewer incipient genome browser
- BLAST highest volume sequence search service
- VAST structure similarity searches
- Map Viewer graphical genome map display (assembled eukaryotic genomes only)
- Cn3D 3D structure viewer
- Genome Workbench standalone sequence analysis annotation platform

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

Entrez: Integrated Molecular and Sequence Databases



e to NCBI

l Center for Biotechnology Information advances science y providing access to biomedical and genomic information.

NCBI | Mission | Organization | Research | RSS Feeds

ed

- : Analyze data using NCBI software
- oads: Get NCBI data or software
- o's: Learn how to accomplish specific tasks at NCBI
- issions: Submit data to GenBank or other NCBI ases

c Testing Registry

clinical genetics resources led information about genetic laboratories.

NCBI's April Newsletter is on the Bookshelf

04 May 2012

Information about May's Discovery Workshop, the new GTR and Assembly

New Filter Sidebar will be added to PubMed

03 May 2012

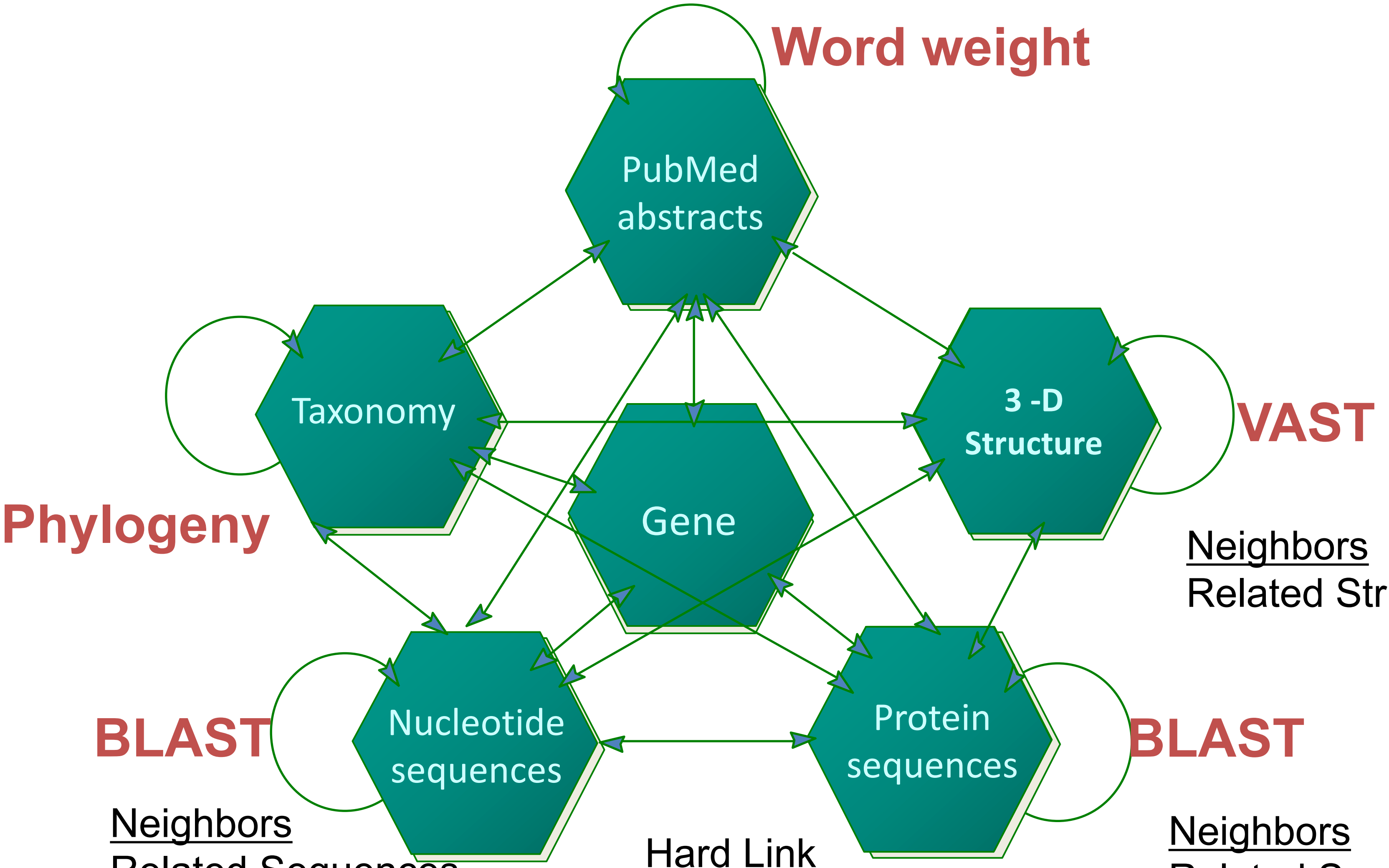
A Filter Sidebar will be added soon to the PubMed result pages. The useful

DELTA BLAST - more sensitive protein searching

30 Apr 2012

Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST)

Entrez: A Discovery System



PubMed abstracts

Word weight

Taxonomy

Phylogeny

Gene

3 -D Structure

VAST

Nucleotide sequences

BLAST

Protein sequences

BLAST

Hard Link

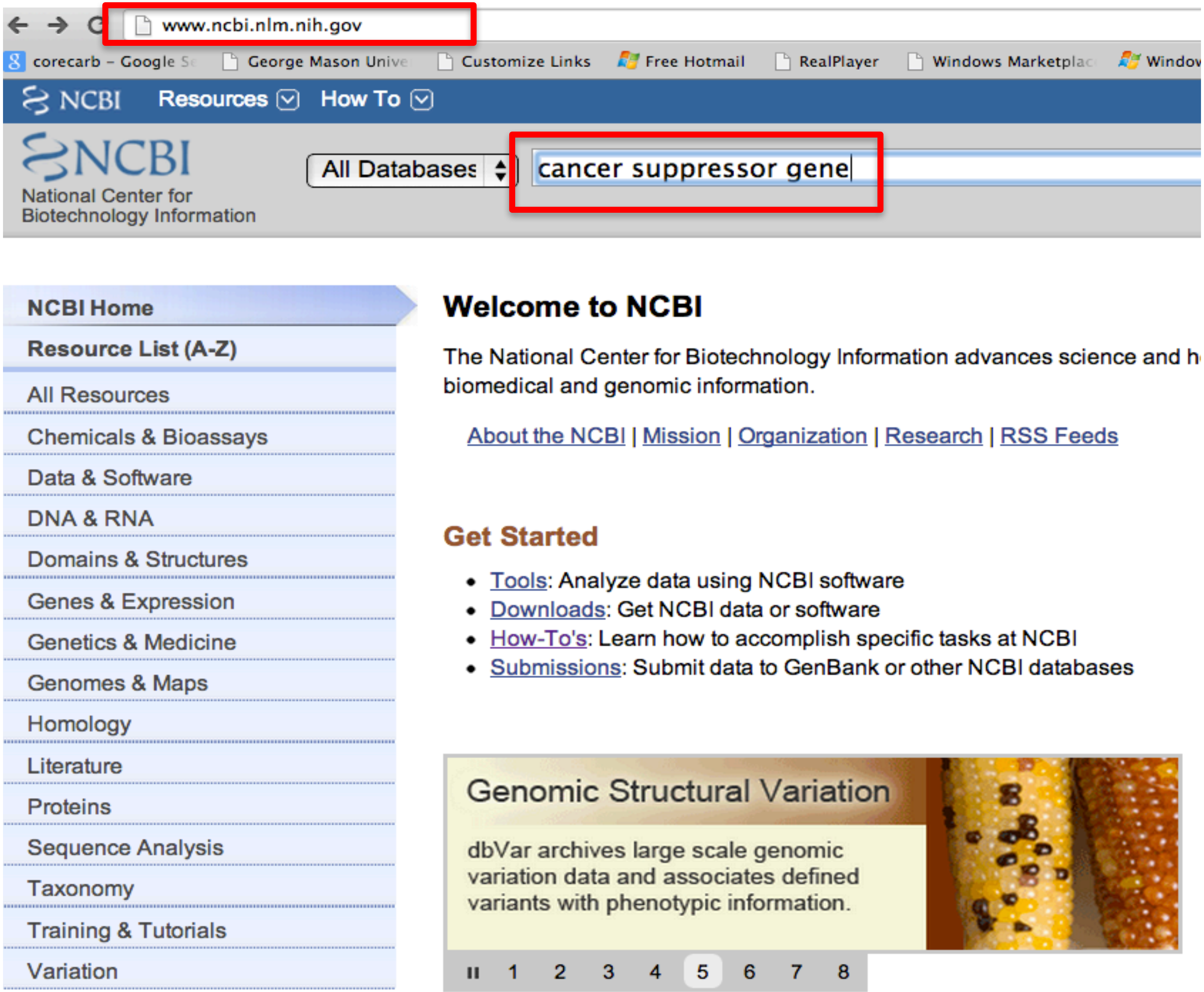
<u>Neighbors</u>
Related Structures

<u>Neighbors</u>
Related Sequences

<u>Neighbors</u>
Related Sequences
BLink
Domains

Hands-on exercise 1

Cancer related genes



Welcome to NCBI

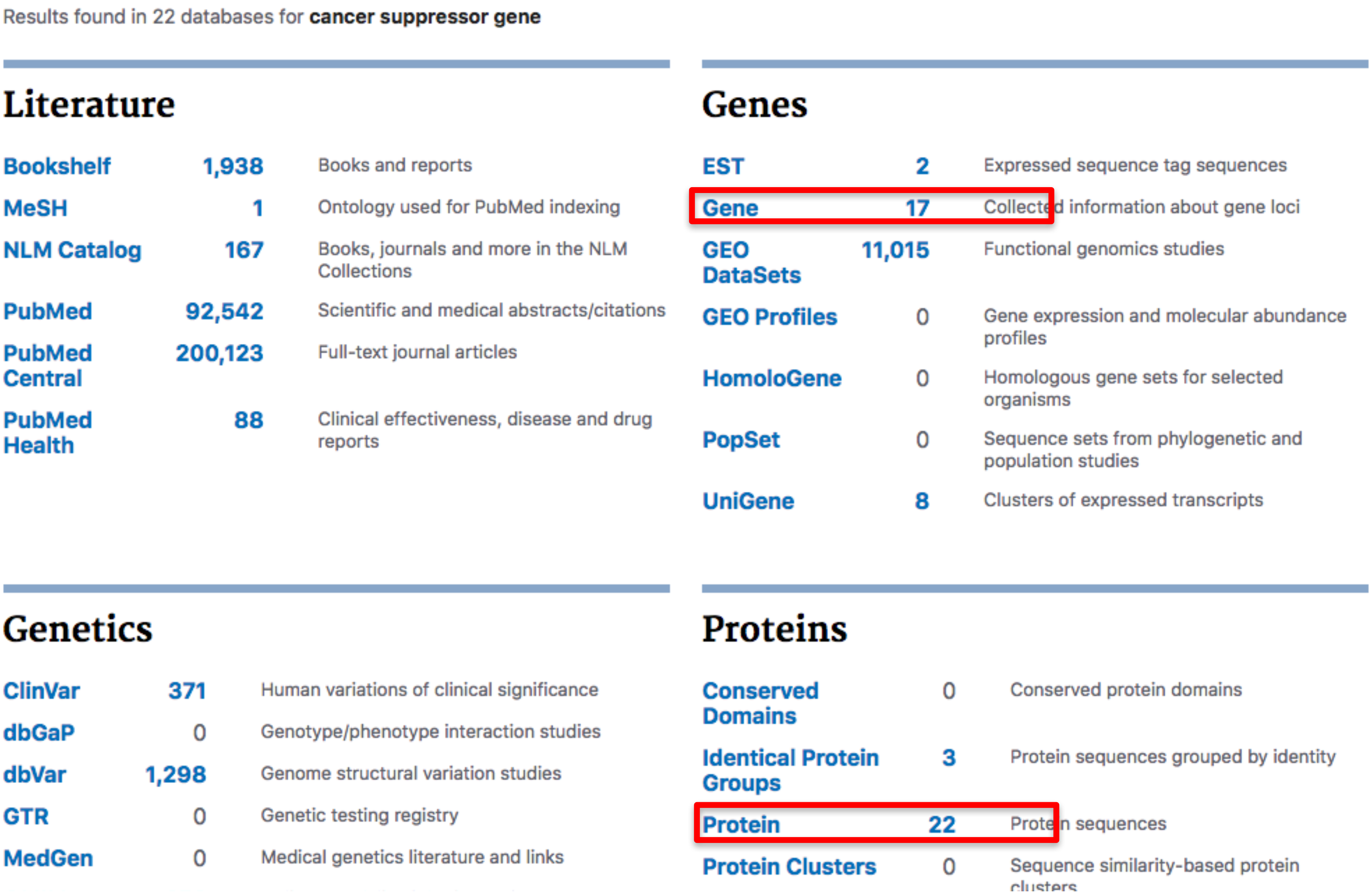
The National Center for Biotechnology Information advances science and h biomedical and genomic information.

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Genomic Structural Variation

dbVar archives large scale genomic variation data and associates defined variants with phenotypic information.



Results found in 22 databases for **cancer suppressor gene**

Literature

Bookshelf	1,938	Books and reports
MeSH	1	Ontology used for PubMed indexing
NLM Catalog	167	Books, journals and more in the NLM Collections
PubMed	92,542	Scientific and medical abstracts/citations
PubMed Central	200,123	Full-text journal articles
PubMed Health	88	Clinical effectiveness, disease and drug reports

Genes

EST	2	Expressed sequence tag sequences
Gene	17	Collected information about gene loci
GEO DataSets	11,015	Functional genomics studies
GEO Profiles	0	Gene expression and molecular abundance profiles
HomoloGene	0	Homologous gene sets for selected organisms
PopSet	0	Sequence sets from phylogenetic and population studies
UniGene	8	Clusters of expressed transcripts

Genetics

ClinVar	371	Human variations of clinical significance
dbGaP	0	Genotype/phenotype interaction studies
dbVar	1,298	Genome structural variation studies
GTR	0	Genetic testing registry
MedGen	0	Medical genetics literature and links

Proteins

Conserved Domains	0	Conserved protein domains
Identical Protein Groups	3	Protein sequences grouped by identity
Protein	22	Protein sequences
Protein Clusters	0	Sequence similarity-based protein clusters

Species
Animals (22)
Customize ...

Source databases
RefSeq (8)
UniProtKB / Swiss-Prot (5)
Customize ...

Sequence length
Custom range...

Molecular weight
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾

See the results of this search (3 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 22

<< First < Prev Page 1 of 2 Next > Last >>

☐ 1. RecName: Full=Mitogen-activated protein kinase kinase kinase 20; AltName: Full=Human cervical **cancer suppressor gene** 4 protein; Short=HCCS-4; AltName: Full=Leucine zipper- and sterile alpha motif-containing kinase; AltName: Full=MLK-like mitogen-activated protein triple kinase; AltName: Full=Mitogen-activated protein kinase kinase kinase MLT; AltName: Full=Mixed lineage kinase-related kinase; Short=MLK-related kinase; Short=MRK; AltName: Full=Sterile alpha motif- and leucine zipper-containing kinase AZK
800 aa protein
Accession: Q9NYL2.3 GI: 313104215
GenPept Identical Proteins FASTA Graphics

☐ 2. RecName: Full=Zinc finger and SCAN domain-containing protein 32; AltName: Full=Human cervical **cancer suppressor gene** 5 protein; Short=HCCS-5; AltName: Full=Zinc finger protein 434
697 aa protein

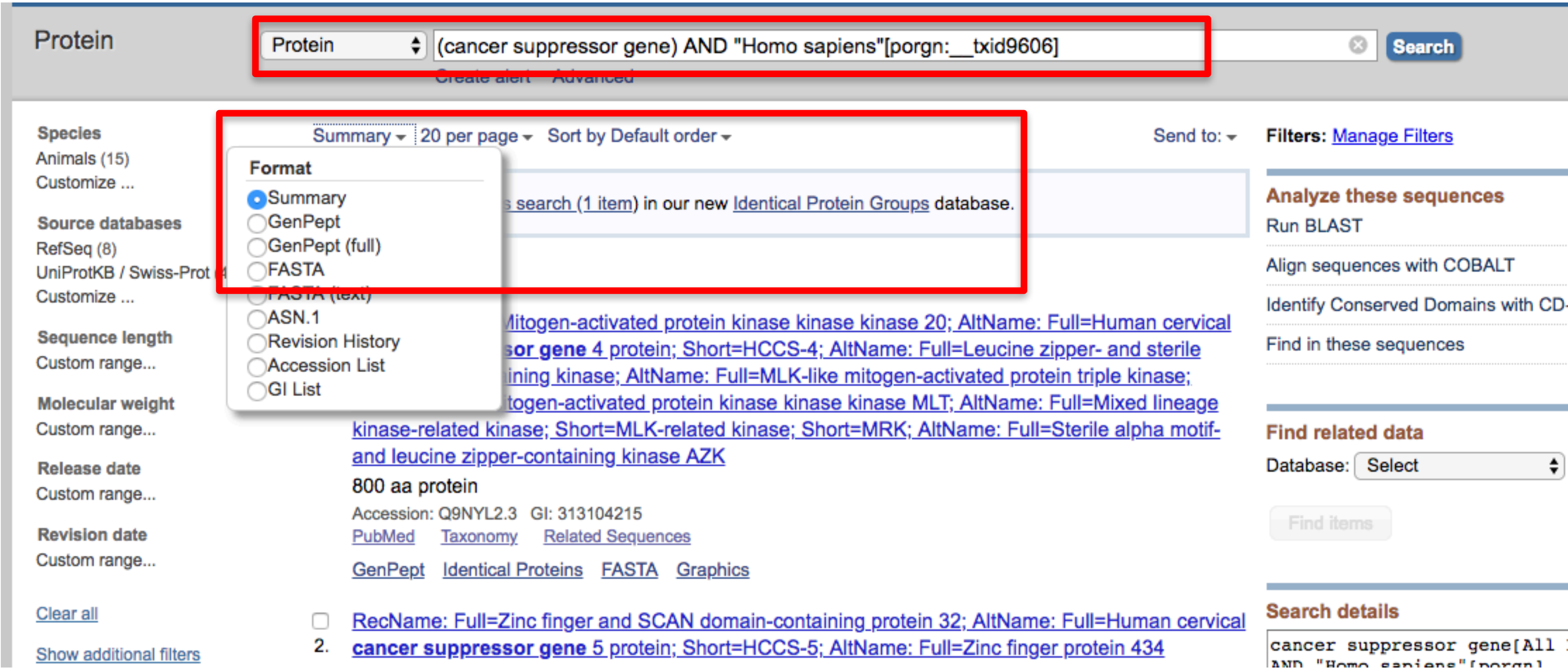
Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Homo sapiens *(15)*
Macaca mulatta *(1)*
Macaca fascicularis *(1)*
Phodopus sungorus *(1)*
Phodopus campbelli *(1)*
All other taxa *(3)*
More...

Find related data
Database: Select
Find items

Search details
```
cancer suppressor gene[All Fie
```



Protein

Protein (cancer suppressor gene) AND "Homo sapiens"[porgn:__txid9606]

Summary 20 per page Sort by Default order

Format

- ● Summary
- ○ GenPept
- ○ GenPept (full)
- ○ FASTA
- ○ FASTA (text)
- ○ ASN.1
- ○ Revision History
- ○ Accession List
- ○ GI List

s search (1 item) in our new Identical Protein Groups database.

Mitogen-activated protein kinase kinase kinase 20; AltName: Full=Human cervical sor gene 4 protein; Short=HCCS-4; AltName: Full=Leucine zipper- and sterile ining kinase; AltName: Full=MLK-like mitogen-activated protein triple kinase; togen-activated protein kinase kinase kinase MLT; AltName: Full=Mixed lineage kinase-related kinase; Short=MLK-related kinase; Short=MRK; AltName: Full=Sterile alpha motif- and leucine zipper-containing kinase AZK

800 aa protein

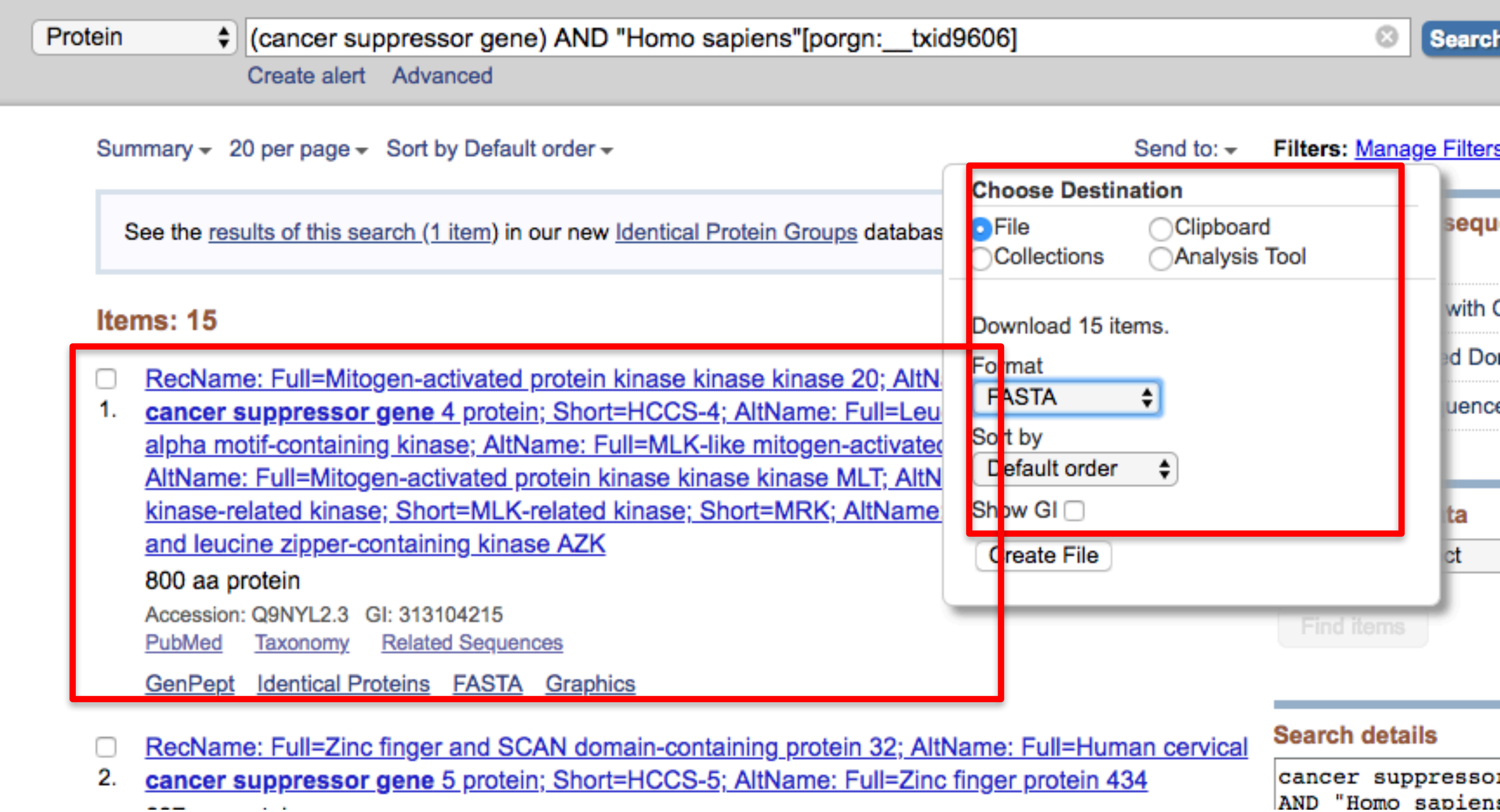
Accession: Q9NYL2.3 GI: 313104215

☐ 2. RecName: Full=Zinc finger and SCAN domain-containing protein 32; AltName: Full=Human cervical **cancer suppressor gene** 5 protein; Short=HCCS-5; AltName: Full=Zinc finger protein 434

Search details

```
cancer suppressor gene[All
AND "Homo sapiens"[porgn]
```

Database: Select



(cancer suppressor gene) AND "Homo sapiens"[porgn:__txid9606]

See the results of this search (1 item) in our new Identical Protein Groups databas

Items: 15

☐ 1. RecName: Full=Mitogen-activated protein kinase kinase kinase 20; AltN **cancer suppressor gene** 4 protein; Short=HCCS-4; AltName: Full=Leu alpha motif-containing kinase; AltName: Full=MLK-like mitogen-activated AltName: Full=Mitogen-activated protein kinase kinase kinase MLT; AltN kinase-related kinase; Short=MLK-related kinase; Short=MRK; AltName and leucine zipper-containing kinase AZK

800 aa protein

Accession: Q9NYL2.3 GI: 313104215

PubMed Taxonomy Related Sequences

GenPept Identical Proteins FASTA Graphics

☐ 2. RecName: Full=Zinc finger and SCAN domain-containing protein 32; AltName: Full=Human cervical **cancer suppressor gene** 5 protein; Short=HCCS-5; AltName: Full=Zinc finger protein 434

Choose Destination

● File ○ Clipboard

○ Collections ○ Analysis Tool

Download 15 items.

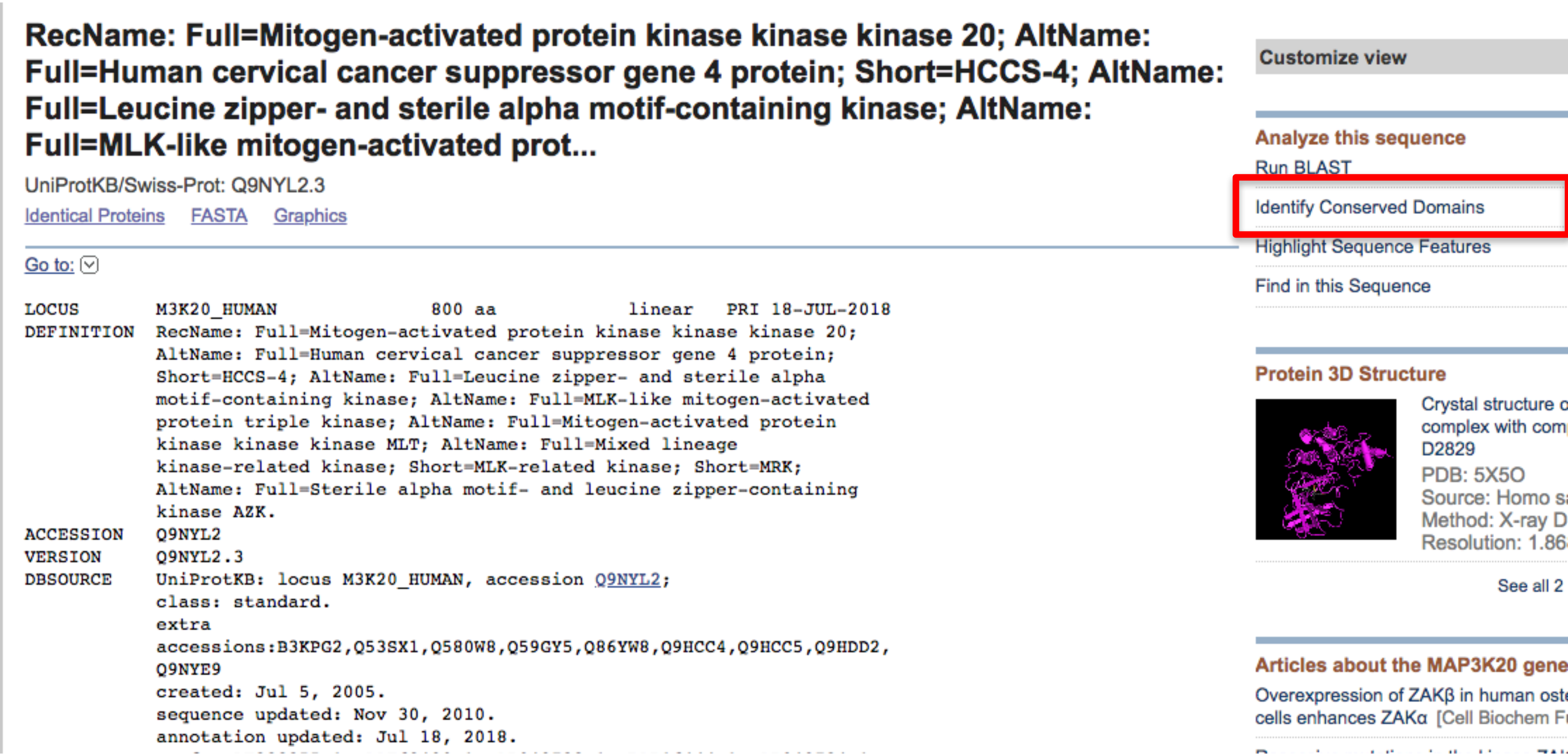
Format: FASTA

Sort by: Default order

Show GI ☐

```
cancer suppresso
AND "Homo sapien
```

UniProtKB/Swiss-Prot: Q9NYL2.3



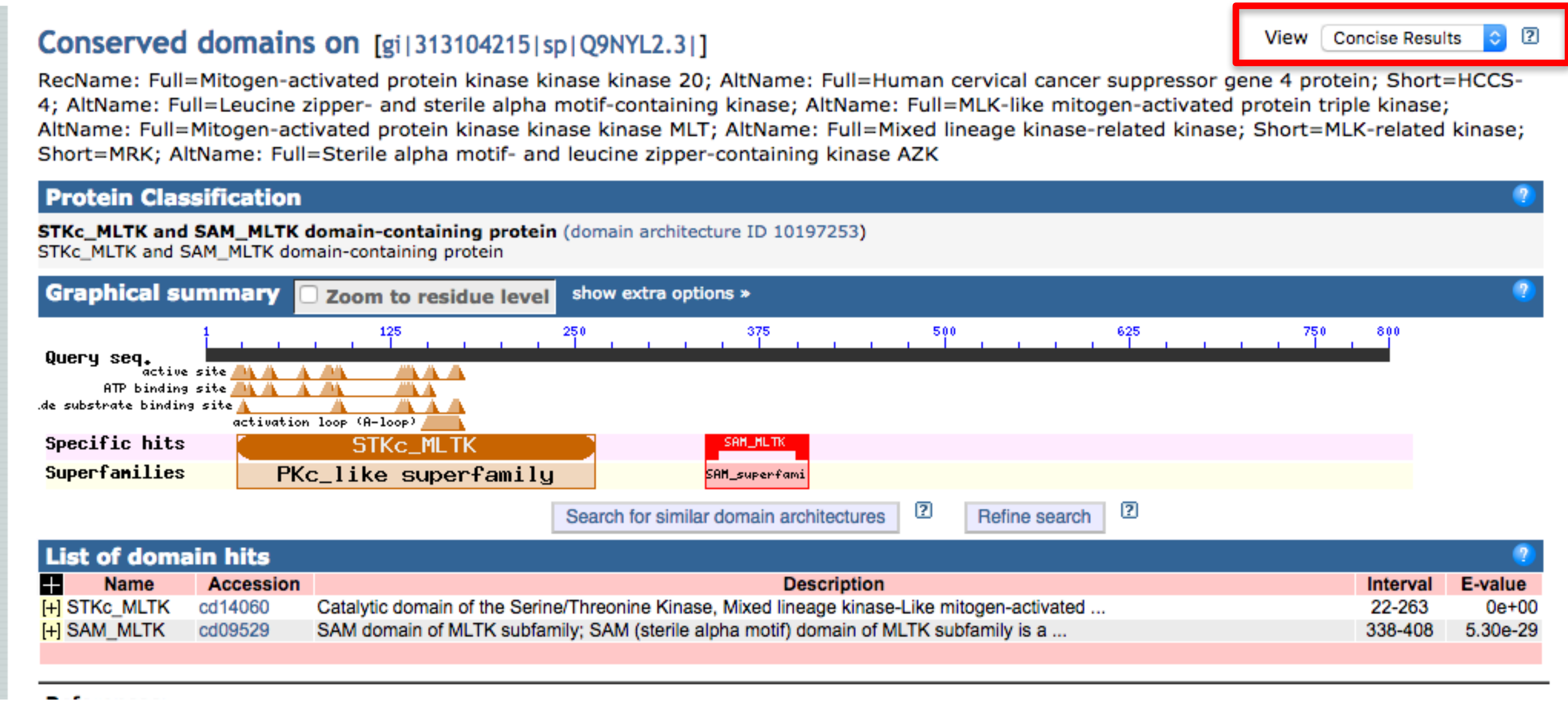
RecName: Full=Mitogen-activated protein kinase kinase kinase 20; AltName: Full=Human cervical cancer suppressor gene 4 protein; Short=HCCS-4; AltName: Full=Leucine zipper- and sterile alpha motif-containing kinase; AltName: Full=MLK-like mitogen-activated prot...

UniProtKB/Swiss-Prot: Q9NYL2.3

```
LOCUS       M3K20_HUMAN              800 aa            linear   PRI 18-JUL-2018
DEFINITION  RecName: Full=Mitogen-activated protein kinase kinase kinase 20;
            AltName: Full=Human cervical cancer suppressor gene 4 protein;
            Short=HCCS-4; AltName: Full=Leucine zipper- and sterile alpha
            motif-containing kinase; AltName: Full=MLK-like mitogen-activated
            protein triple kinase; AltName: Full=Mitogen-activated protein
            kinase kinase kinase MLT; AltName: Full=Mixed lineage
            kinase-related kinase; Short=MLK-related kinase; Short=MRK;
            AltName: Full=Sterile alpha motif- and leucine zipper-containing
            kinase AZK.
ACCESSION   Q9NYL2
VERSION     Q9NYL2.3
DBSOURCE    UniProtKB: locus M3K20_HUMAN, accession Q9NYL2;
            class: standard.
            extra
            accessions:B3KPG2,Q53SX1,Q580W8,Q59GY5,Q86YW8,Q9HCC4,Q9HCC5,Q9HDD2,
            Q9NYE9
            created: Jul 5, 2005.
            sequence updated: Nov 30, 2010.
            annotation updated: Jul 18, 2018.
```

Crystal structure o
complex with com
D2829
PDB: 5X5O
Source: Homo sa
Method: X-ray D
Resolution: 1.86

Overexpression of ZAKβ in human oste
cells enhances ZAKα [Cell Biochem F



Conserved domains on [gi|313104215|sp|Q9NYL2.3|]

RecName: Full=Mitogen-activated protein kinase kinase kinase 20; AltName: Full=Human cervical cancer suppressor gene 4 protein; Short=HCCS-4; AltName: Full=Leucine zipper- and sterile alpha motif-containing kinase; AltName: Full=MLK-like mitogen-activated protein triple kinase; AltName: Full=Mitogen-activated protein kinase kinase kinase MLT; AltName: Full=Mixed lineage kinase-related kinase; Short=MLK-related kinase; Short=MRK; AltName: Full=Sterile alpha motif- and leucine zipper-containing kinase AZK

Protein Classification

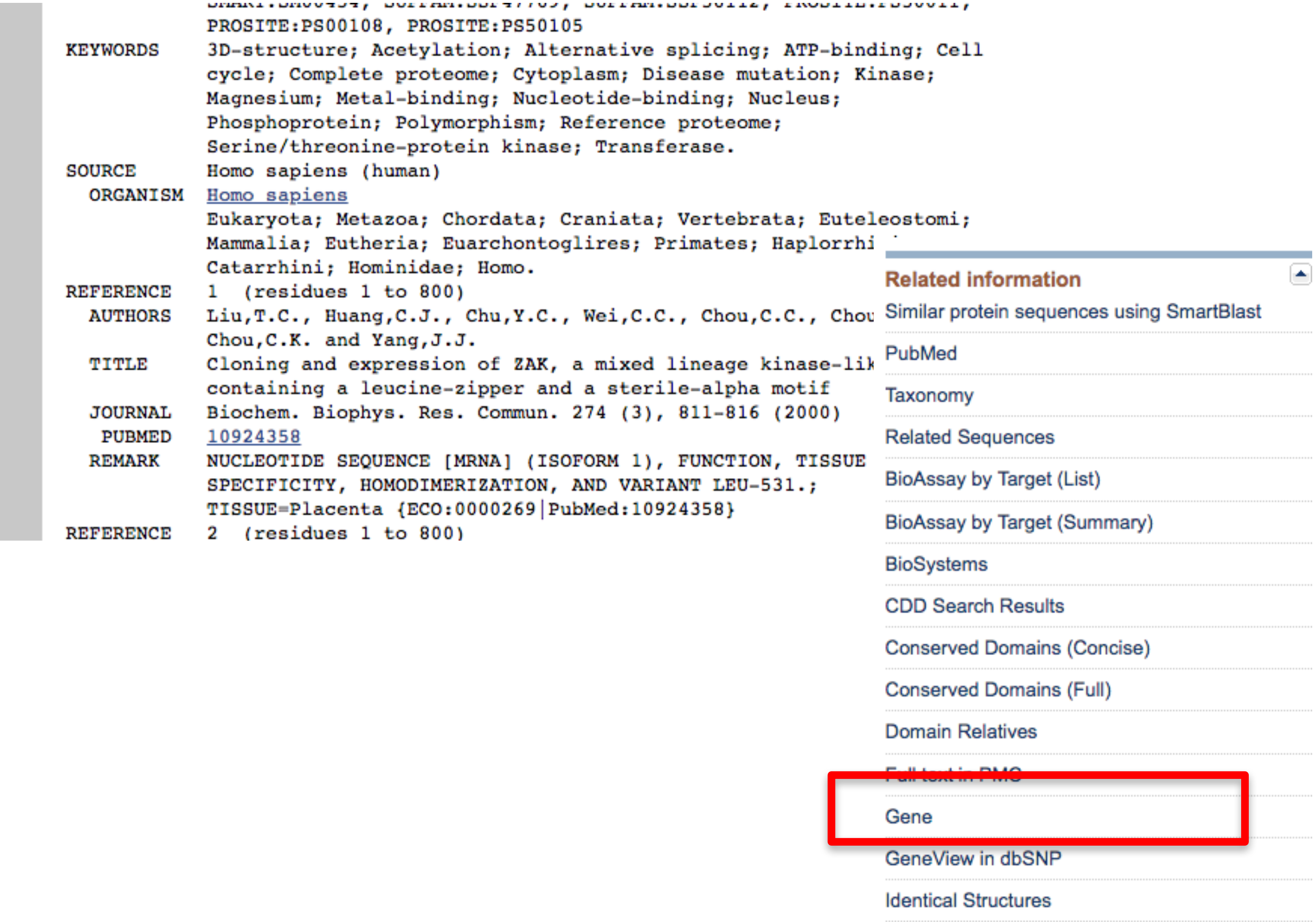
STKc_MLTK and SAM_MLTK domain-containing protein (domain architecture ID 10197253)
STKc_MLTK and SAM_MLTK domain-containing protein

Graphical summary

List of domain hits

	Name	Accession	Description	Interval	E-value
[+]	STKc_MLTK	cd14060	Catalytic domain of the Serine/Threonine Kinase, Mixed lineage kinase-Like mitogen-activated ...	22-263	0e+00
[+]	SAM_MLTK	cd09529	SAM domain of MLTK subfamily; SAM (sterile alpha motif) domain of MLTK subfamily is a ...	338-408	5.30e-29

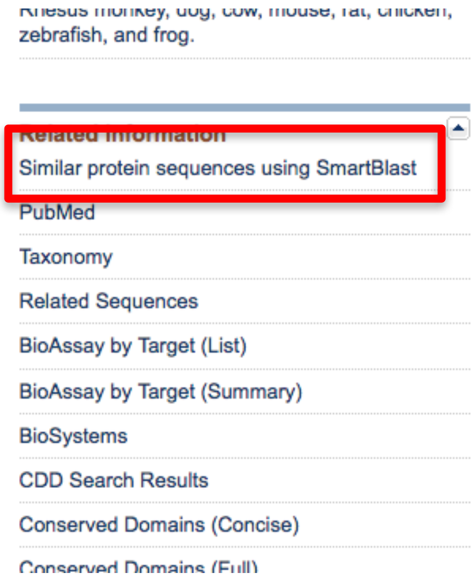
UniProtKB/Swiss-Prot: Q9NYL2.3



```
           PROSITE:PS00108, PROSITE:PS50105
KEYWORDS   3D-structure; Acetylation; Alternative splicing; ATP-binding; Cell
           cycle; Complete proteome; Cytoplasm; Disease mutation; Kinase;
           Magnesium; Metal-binding; Nucleotide-binding; Nucleus;
           Phosphoprotein; Polymorphism; Reference proteome;
           Serine/threonine-protein kinase; Transferase.
SOURCE     Homo sapiens (human)
  ORGANISM Homo sapiens
           Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
           Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhi
           Catarrhini; Hominidae; Homo.
REFERENCE  1  (residues 1 to 800)
  AUTHORS  Liu,T.C., Huang,C.J., Chu,Y.C., Wei,C.C., Chou,C.C., Chou
           Chou,C.K. and Yang,J.J.
  TITLE    Cloning and expression of ZAK, a mixed lineage kinase-lik
           containing a leucine-zipper and a sterile-alpha motif
  JOURNAL  Biochem. Biophys. Res. Commun. 274 (3), 811-816 (2000)
   PUBMED  10924358
  REMARK   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1), FUNCTION, TISSUE
           SPECIFICITY, HOMODIMERIZATION, AND VARIANT LEU-531.;
           TISSUE=Placenta {ECO:0000269|PubMed:10924358}
REFERENCE  2  (residues 1 to 800)
```

Related information

Similar protein sequences using SmartBlast
PubMed
Taxonomy
Related Sequences
BioAssay by Target (List)
BioAssay by Target (Summary)
BioSystems
CDD Search Results
Conserved Domains (Concise)
Conserved Domains (Full)
Domain Relatives
Full text in PMC
Gene
GeneView in dbSNP
Identical Structures

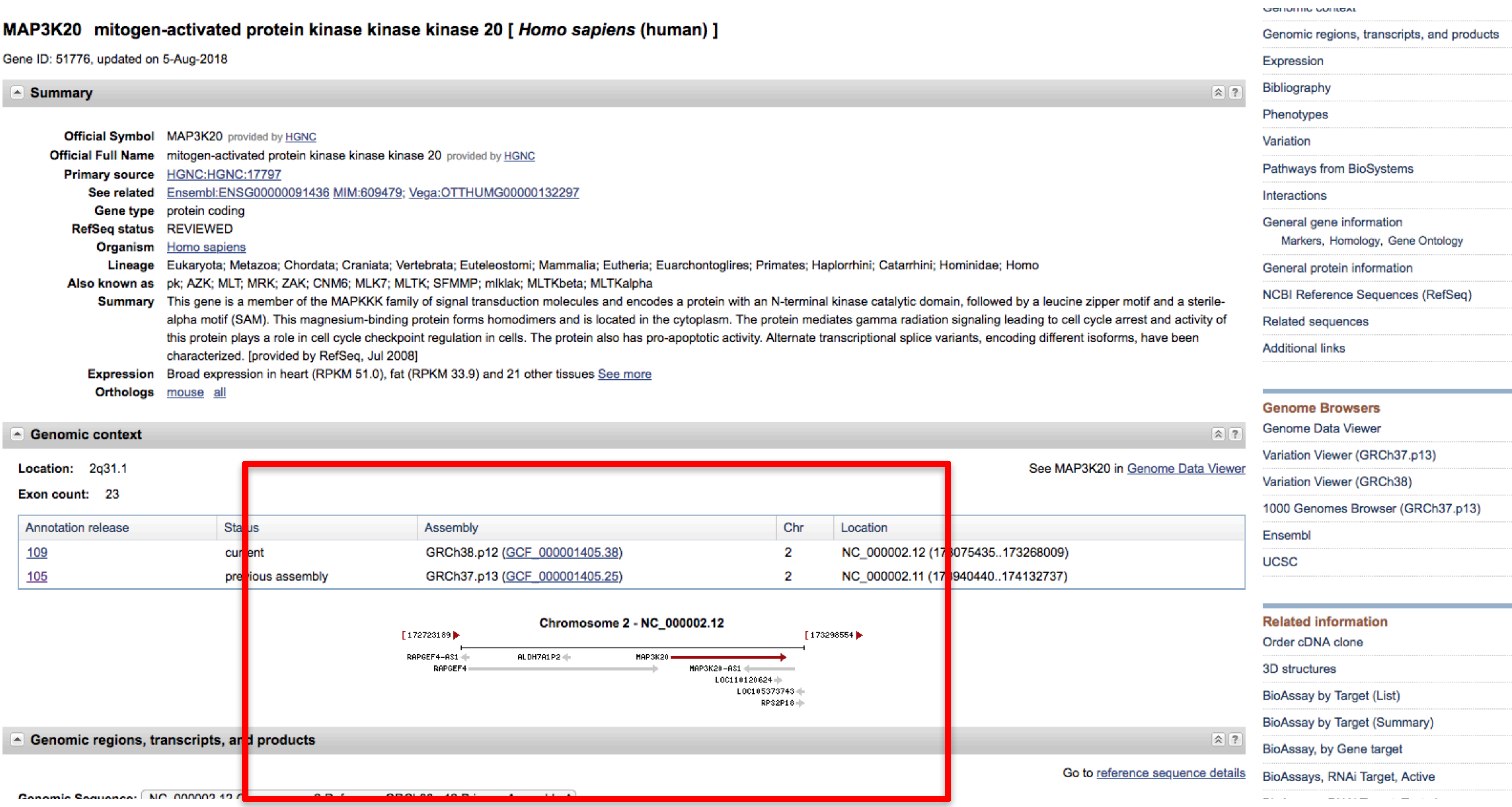


Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish, and frog.

Related information

Similar protein sequences using SmartBlast
PubMed
Taxonomy
Related Sequences
BioAssay by Target (List)
BioAssay by Target (Summary)
BioSystems
CDD Search Results
Conserved Domains (Concise)
Conserved Domains (Full)

Gene page of UniProtKB/Swiss-Prot: Q9NYL2.3



MAP3K20 mitogen-activated protein kinase kinase kinase 20 [*Homo sapiens* (human)]

Gene ID: 51776, updated on 5-Aug-2018

Summary

Official Symbol	MAP3K20 provided by HGNC
Official Full Name	mitogen-activated protein kinase kinase kinase 20 provided by HGNC
Primary source	HGNC:HGNC:17797
See related	Ensembl:ENSG00000091436 MIM:609479; Vega:OTTHUMG00000132297
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	pk; AZK; MLT; MRK; ZAK; CNM6; MLK7; MLTK; SFMMP; mlklak; MLTKbeta; MLTKalpha
Summary	This gene is a member of the MAPKKK family of signal transduction molecules and encodes a protein with an N-terminal kinase catalytic domain, followed by a leucine zipper motif and a sterile-alpha motif (SAM). This magnesium-binding protein forms homodimers and is located in the cytoplasm. The protein mediates gamma radiation signaling leading to cell cycle arrest and activity of this protein plays a role in cell cycle checkpoint regulation in cells. The protein also has pro-apoptotic activity. Alternate transcriptional splice variants, encoding different isoforms, have been characterized. [provided by RefSeq, Jul 2008]
Expression	Broad expression in heart (RPKM 51.0), fat (RPKM 33.9) and 21 other tissues See more
Orthologs	mouse all

Genomic context

Location: 2q31.1

Exon count: 23

See MAP3K20 in Genome Data Viewer

Annotation release	Sta us	Assembly	Chr	Location
109	cur ent	GRCh38.p12 (GCF_000001405.38)	2	NC_000002.12 (17 075435..173268009)
105	pre ious assembly	GRCh37.p13 (GCF_000001405.25)	2	NC_000002.11 (17 940440..174132737)

Chromosome 2 - NC_000002.12

Genomic regions, transcripts, ar d products

Go to reference sequence details

Genomic regions, transcripts, and products
Expression
Bibliography
Phenotypes
Variation
Pathways from BioSystems
Interactions
General gene information
Markers, Homology, Gene Ontology
General protein information
NCBI Reference Sequences (RefSeq)
Related sequences
Additional links

Genome Browsers
Genome Data Viewer
Variation Viewer (GRCh37.p13)
Variation Viewer (GRCh38)
1000 Genomes Browser (GRCh37.p13)
Ensembl
UCSC

Related information
Order cDNA clone
3D structures
BioAssay by Target (List)
BioAssay by Target (Summary)
BioAssay, by Gene target
BioAssays, RNAi Target, Active

Hands-on exercise 2

Given a list of sequence IDs, get their sequences from NCBI

Suppose

- You read a paper which reported a list of genes (with a table e.g. to show all the IDs)

Or

- You have a collaborator sending you a file with all the IDs

You want to quickly get the sequences of these genes

Download the example id file at http://cys.bios.niu.edu/yyin/teach/PBB/gt8-id.txt

Plain text file!

Open notepad++, copy from the above page, paste into the notepad, and save as gt8-id.txt (make sure you know where you've saved it)

Welcome to NCBI

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

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

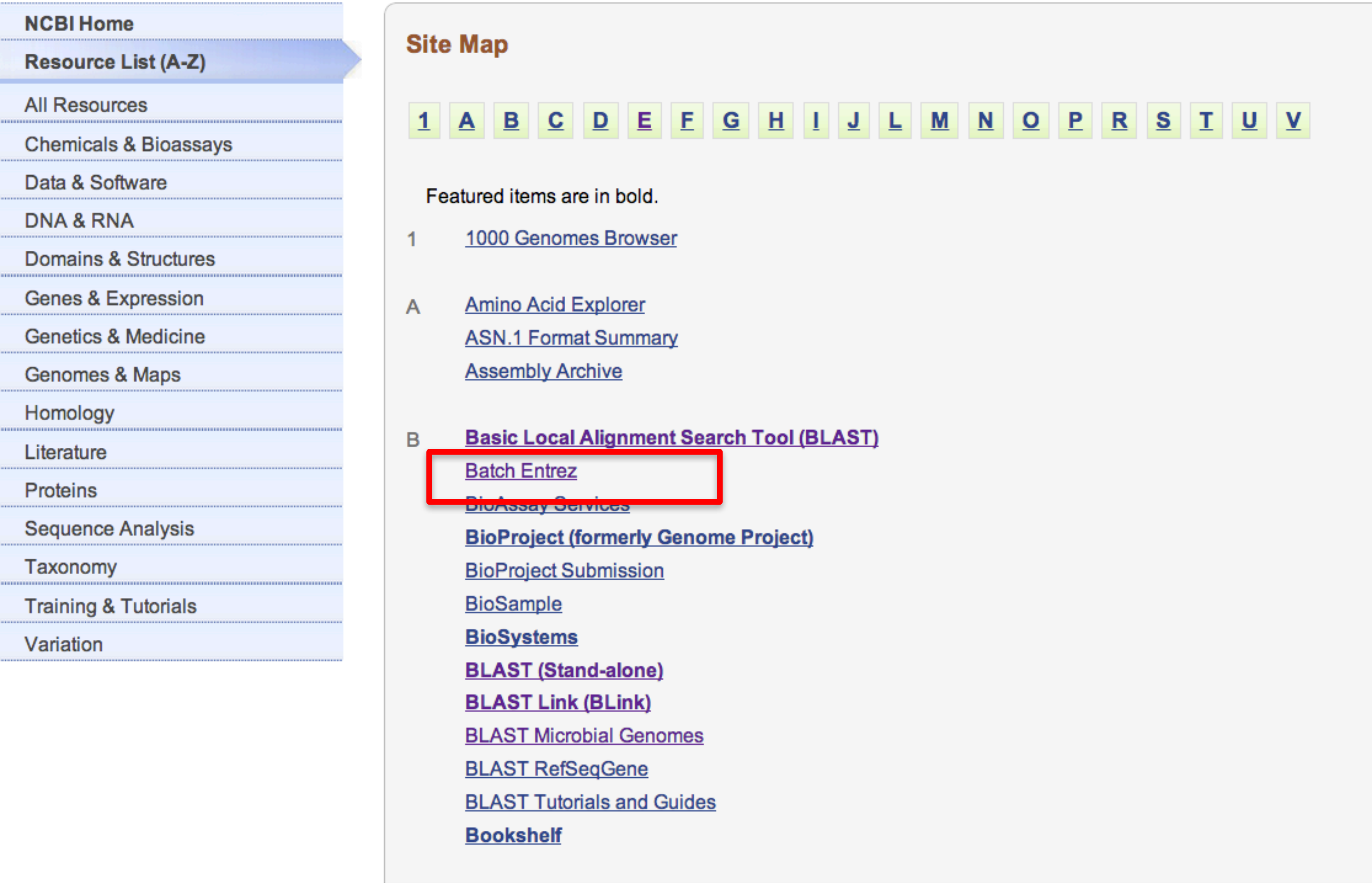
NCBI Facebook page

Find out the latest news about NCBI resources and participate in community discussions.

Come to the NCBI Disco
on February 4&5!

Spaces are still available

New version of Genome
available



NCBI Home

Resource List (A-Z)

- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Site Map

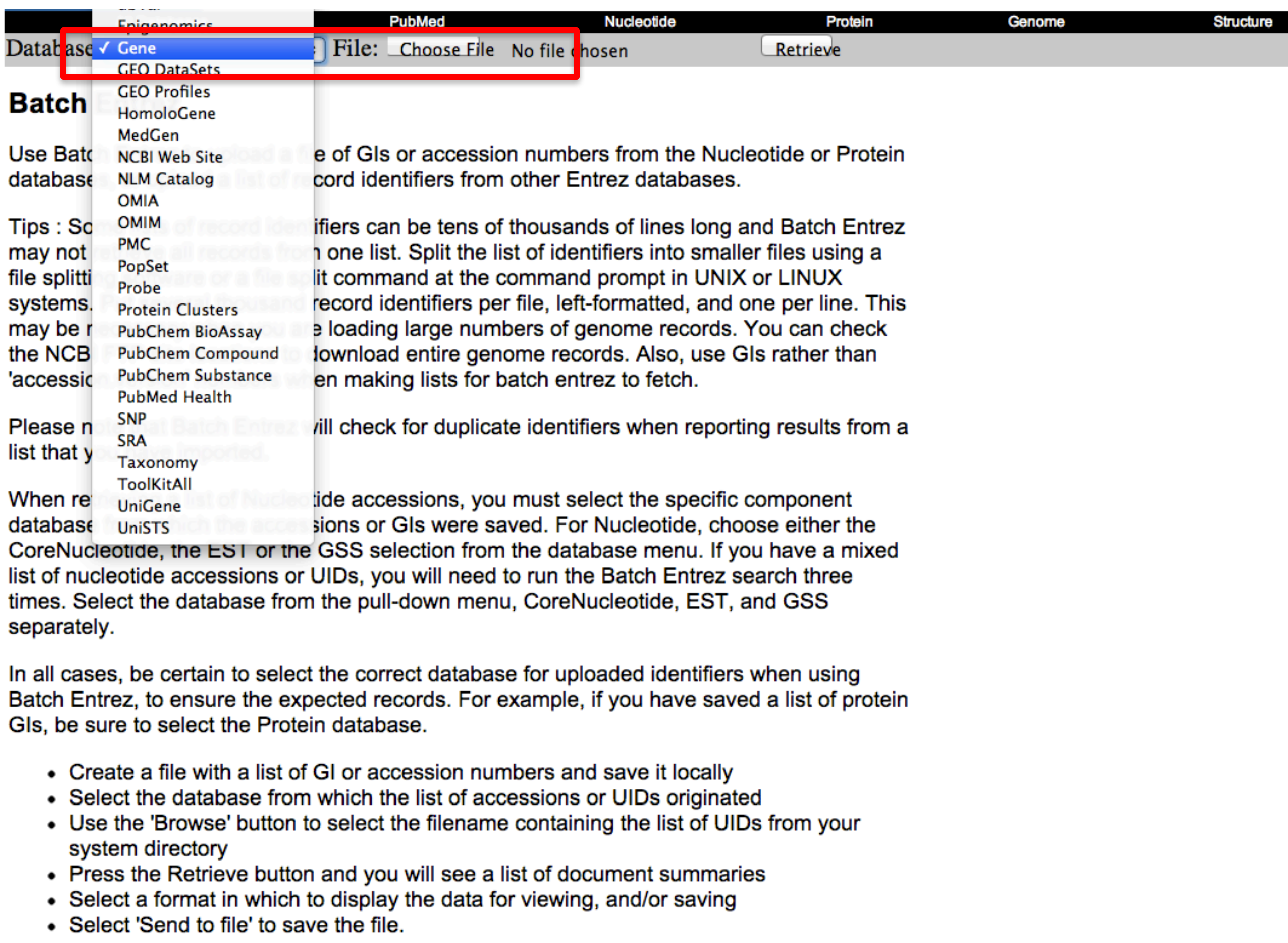
1 A B C D E F G H I J L M N O P R S T U V

Featured items are in bold.

1 1000 Genomes Browser

A Amino Acid Explorer
ASN.1 Format Summary
Assembly Archive

B **Basic Local Alignment Search Tool (BLAST)**
Batch Entrez
BioAssay Services
BioProject (formerly Genome Project)
BioProject Submission
BioSample
BioSystems
BLAST (Stand-alone)
BLAST Link (BLink)
BLAST Microbial Genomes
BLAST RefSeqGene
BLAST Tutorials and Guides
Bookshelf



Batch Entrez

Use Batch Entrez to upload a file of GIs or accession numbers from the Nucleotide or Protein databases, or to upload a list of record identifiers from other Entrez databases.

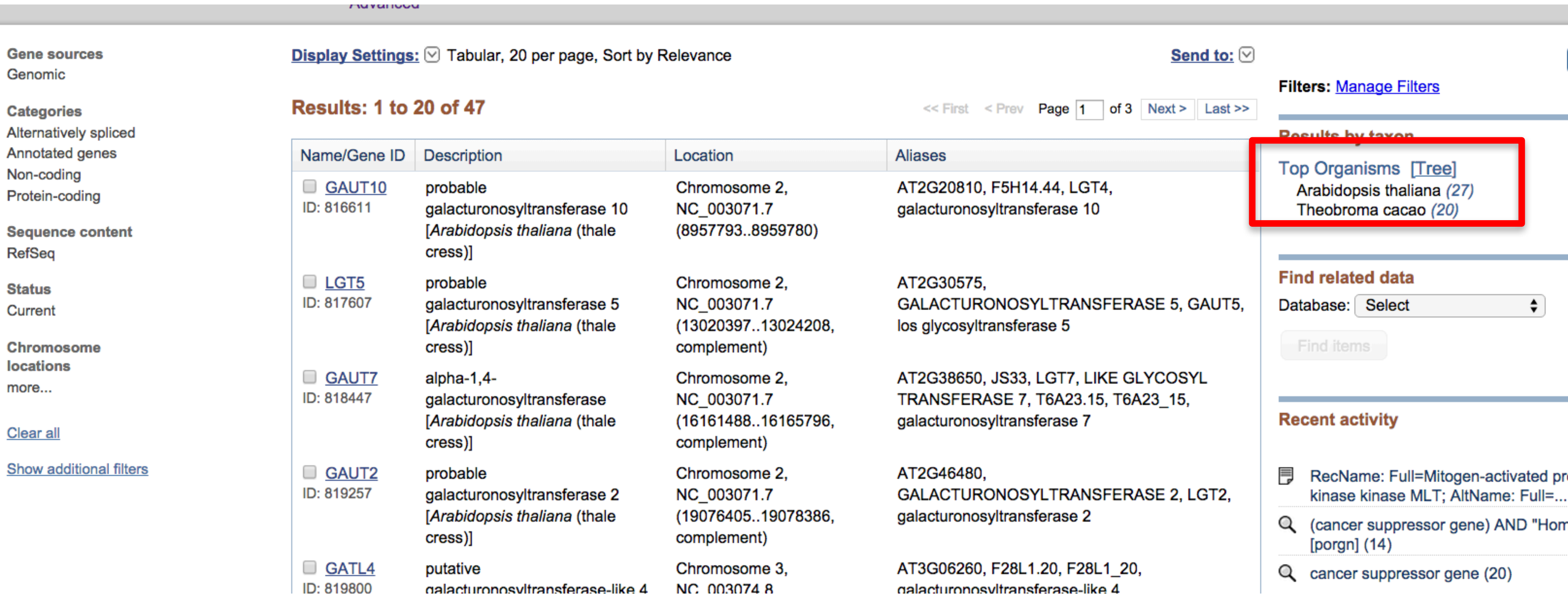
Tips : Some lists of record identifiers can be tens of thousands of lines long and Batch Entrez may not retrieve all records from one list. Split the list of identifiers into smaller files using a file splitting software or a file split command at the command prompt in UNIX or LINUX systems. Put several thousand record identifiers per file, left-formatted, and one per line. This may be necessary when you are loading large numbers of genome records. You can check the NCBI FTP site to download entire genome records. Also, use GIs rather than 'accession.version' identifiers when making lists for batch entrez to fetch.

Please note that Batch Entrez will check for duplicate identifiers when reporting results from a list that you have imported.

When retrieving a list of Nucleotide accessions, you must select the specific component database from which the accessions or GIs were saved. For Nucleotide, choose either the CoreNucleotide, the EST or the GSS selection from the database menu. If you have a mixed list of nucleotide accessions or UIDs, you will need to run the Batch Entrez search three times. Select the database from the pull-down menu, CoreNucleotide, EST, and GSS separately.

In all cases, be certain to select the correct database for uploaded identifiers when using Batch Entrez, to ensure the expected records. For example, if you have saved a list of protein GIs, be sure to select the Protein database.

- Create a file with a list of GI or accession numbers and save it locally
- Select the database from which the list of accessions or UIDs originated
- Use the 'Browse' button to select the filename containing the list of UIDs from your system directory
- Press the Retrieve button and you will see a list of document summaries
- Select a format in which to display the data for viewing, and/or saving
- Select 'Send to file' to save the file.



Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding

Sequence content
RefSeq

Status
Current

Chromosome locations
more...

Clear all

Show additional filters

Display Settings: Tabular, 20 per page, Sort by Relevance

Send to:

Results: 1 to 20 of 47

Name/Gene ID	Description	Location	Aliases
☐ GAUT10 ID: 816611	probable galacturonosyltransferase 10 [*Arabidopsis thaliana* (thale cress)]	Chromosome 2, NC_003071.7 (8957793..8959780)	AT2G20810, F5H14.44, LGT4, galacturonosyltransferase 10
☐ LGT5 ID: 817607	probable galacturonosyltransferase 5 [*Arabidopsis thaliana* (thale cress)]	Chromosome 2, NC_003071.7 (13020397..13024208, complement)	AT2G30575, GALACTURONOSYLTRANSFERASE 5, GAUT5, los glycosyltransferase 5
☐ GAUT7 ID: 818447	alpha-1,4-galacturonosyltransferase [*Arabidopsis thaliana* (thale cress)]	Chromosome 2, NC_003071.7 (16161488..16165796, complement)	AT2G38650, JS33, LGT7, LIKE GLYCOSYL TRANSFERASE 7, T6A23.15, T6A23_15, galacturonosyltransferase 7
☐ GAUT2 ID: 819257	probable galacturonosyltransferase 2 [*Arabidopsis thaliana* (thale cress)]	Chromosome 2, NC_003071.7 (19076405..19078386, complement)	AT2G46480, GALACTURONOSYLTRANSFERASE 2, LGT2, galacturonosyltransferase 2
☐ GATL4 ID: 819800	putative galacturonosyltransferase-like 4	Chromosome 3, NC_003074.8	AT3G06260, F28L1.20, F28L1_20, galacturonosyltransferase-like 4

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Arabidopsis thaliana *(27)*
Theobroma cacao *(20)*

Find related data

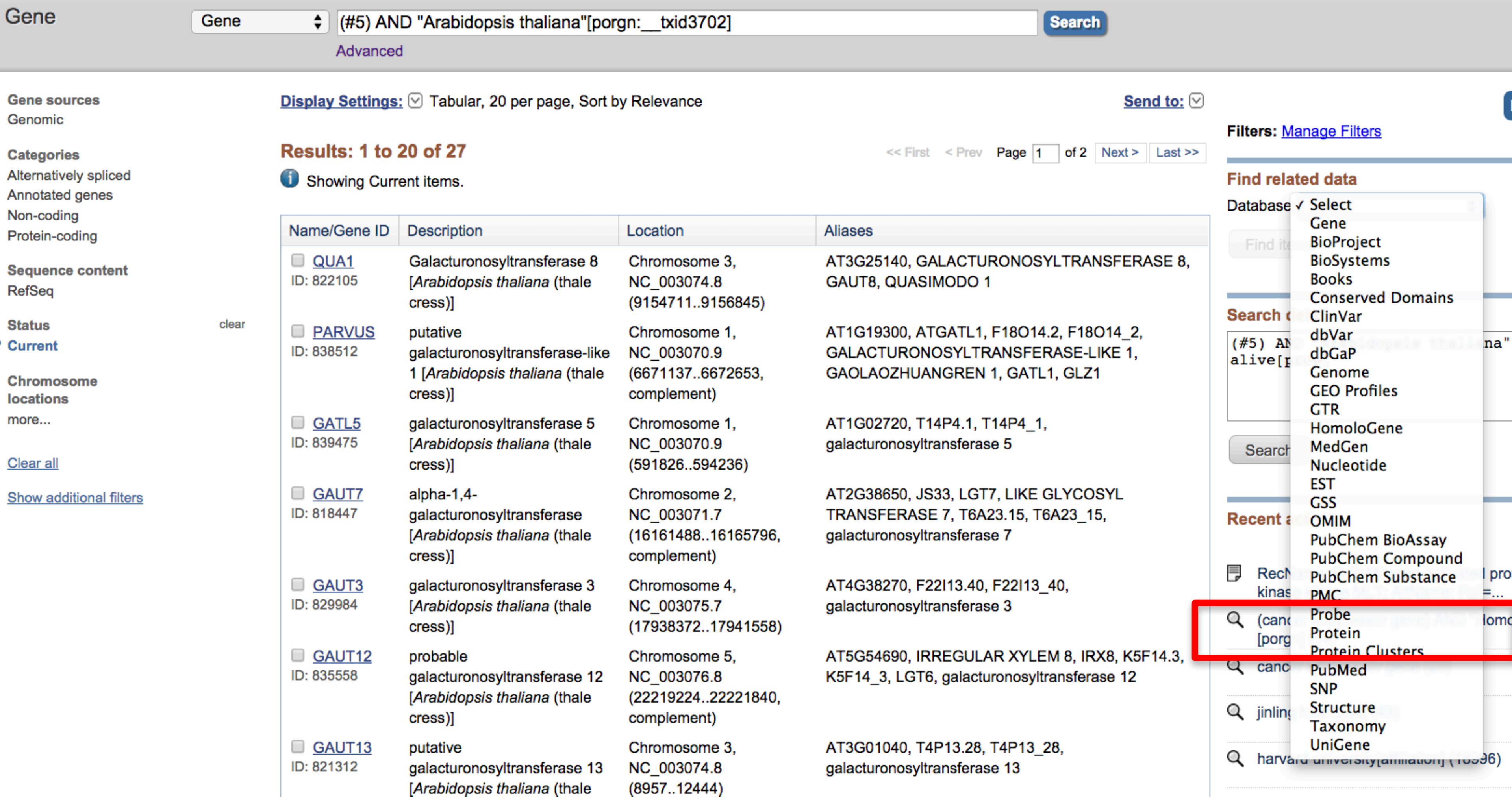
Database: Select

Recent activity

RecName: Full=Mitogen-activated pr
kinase kinase MLT; AltName: Full=...

(cancer suppressor gene) AND "Hom
[porgn] (14)

cancer suppressor gene (20)

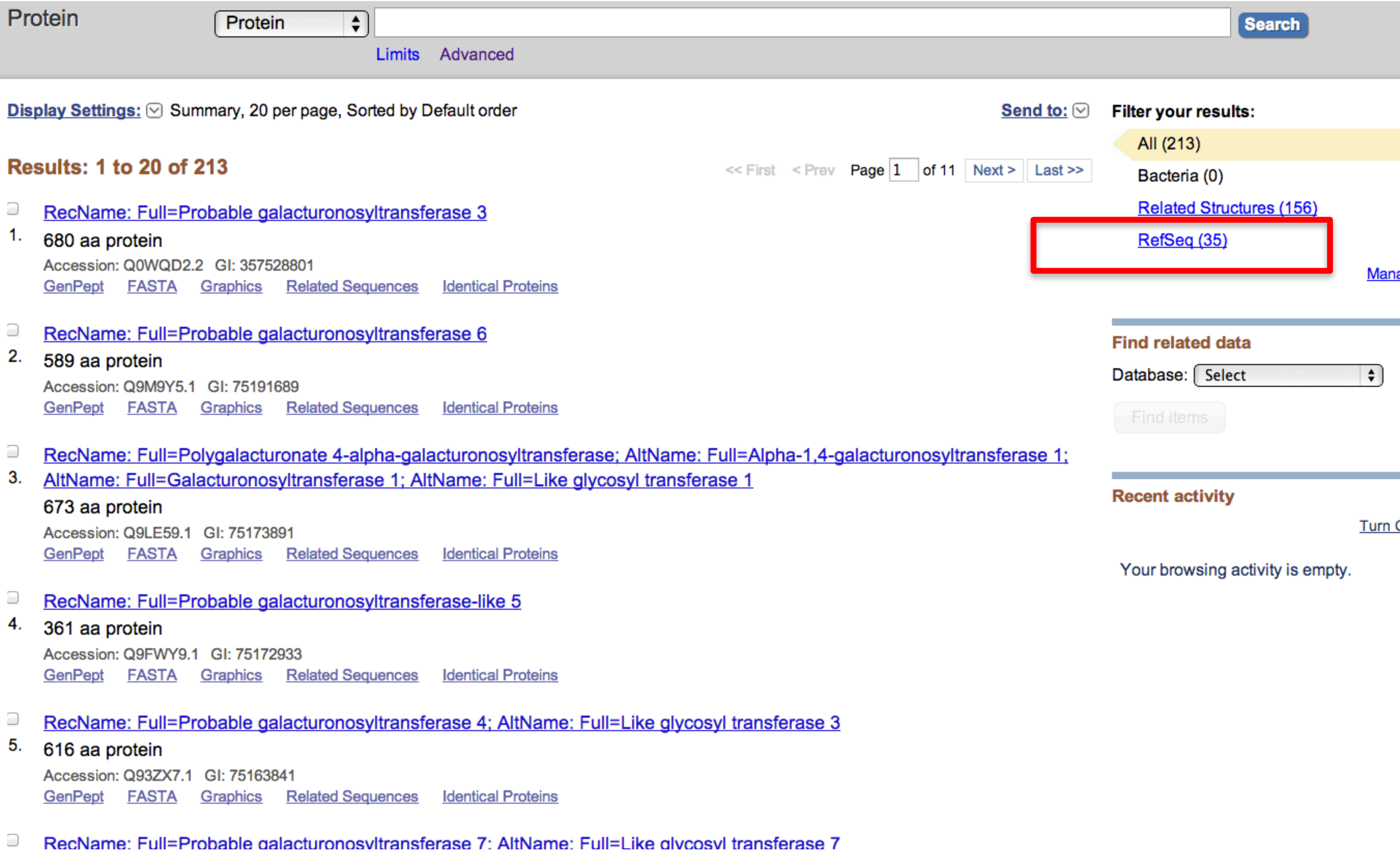


Display Settings: Tabular, 20 per page, Sort by Relevance

Results: 1 to 20 of 27

Showing Current items.

Name/Gene ID	Description	Location	Aliases
☐ QUA1 ID: 822105	Galacturonosyltransferase 8 [*Arabidopsis thaliana* (thale cress)]	Chromosome 3, NC_003074.8 (9154711..9156845)	AT3G25140, GALACTURONOSYLTRANSFERASE 8, GAUT8, QUASIMODO 1
☐ PARVUS ID: 838512	putative galacturonosyltransferase-like 1 [*Arabidopsis thaliana* (thale cress)]	Chromosome 1, NC_003070.9 (6671137..6672653, complement)	AT1G19300, ATGATL1, F18O14.2, F18O14_2, GALACTURONOSYLTRANSFERASE-LIKE 1, GAOLAOZHUANGREN 1, GATL1, GLZ1
☐ GATL5 ID: 839475	galacturonosyltransferase 5 [*Arabidopsis thaliana* (thale cress)]	Chromosome 1, NC_003070.9 (591826..594236)	AT1G02720, T14P4.1, T14P4_1, galacturonosyltransferase 5
☐ GAUT7 ID: 818447	alpha-1,4-galacturonosyltransferase [*Arabidopsis thaliana* (thale cress)]	Chromosome 2, NC_003071.7 (16161488..16165796, complement)	AT2G38650, JS33, LGT7, LIKE GLYCOSYL TRANSFERASE 7, T6A23.15, T6A23_15, galacturonosyltransferase 7
☐ GAUT3 ID: 829984	galacturonosyltransferase 3 [*Arabidopsis thaliana* (thale cress)]	Chromosome 4, NC_003075.7 (17938372..17941558)	AT4G38270, F22I13.40, F22I13_40, galacturonosyltransferase 3
☐ GAUT12 ID: 835558	probable galacturonosyltransferase 12 [*Arabidopsis thaliana* (thale cress)]	Chromosome 5, NC_003076.8 (22219224..22221840, complement)	AT5G54690, IRREGULAR XYLEM 8, IRX8, K5F14.3, K5F14_3, LGT6, galacturonosyltransferase 12
☐ GAUT13 ID: 821312	putative galacturonosyltransferase 13 [*Arabidopsis thaliana* (thale	Chromosome 3, NC_003074.8 (8957..12444)	AT3G01040, T4P13.28, T4P13_28, galacturonosyltransferase 13



Display Settings: Summary, 20 per page, Sorted by Default order

Send to:

Results: 1 to 20 of 213

<< First < Prev Page 1 of 11 Next > Last >>

☐ 1. RecName: Full=Probable galacturonosyltransferase 3
680 aa protein
Accession: Q0WQD2.2 GI: 357528801
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ 2. RecName: Full=Probable galacturonosyltransferase 6
589 aa protein
Accession: Q9M9Y5.1 GI: 75191689
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ 3. RecName: Full=Polygalacturonate 4-alpha-galacturonosyltransferase; AltName: Full=Alpha-1,4-galacturonosyltransferase 1; AltName: Full=Galacturonosyltransferase 1; AltName: Full=Like glycosyl transferase 1
673 aa protein
Accession: Q9LE59.1 GI: 75173891
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ 4. RecName: Full=Probable galacturonosyltransferase-like 5
361 aa protein
Accession: Q9FWY9.1 GI: 75172933
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ 5. RecName: Full=Probable galacturonosyltransferase 4; AltName: Full=Like glycosyl transferase 3
616 aa protein
Accession: Q93ZX7.1 GI: 75163841
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ RecName: Full=Probable galacturonosyltransferase 7; AltName: Full=Like glycosyl transferase 7

Filter your results:

All (213)
Bacteria (0)
Related Structures (156)
RefSeq (35)
Mana

Find related data

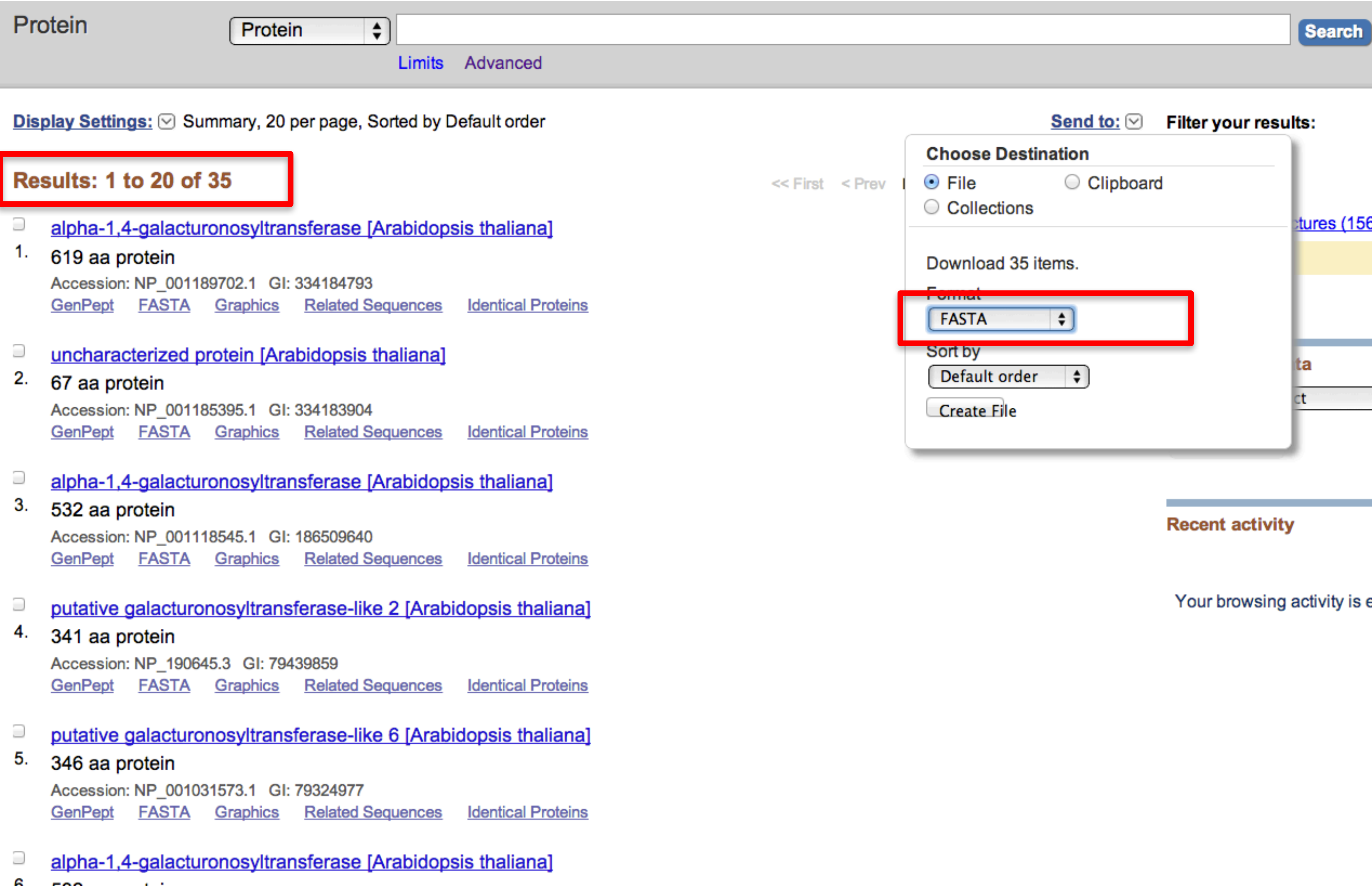
Database: Select

Find items

Recent activity

Turn

Your browsing activity is empty.



Display Settings: Summary, 20 per page, Sorted by Default order

Results: 1 to 20 of 35

☐ alpha-1,4-galacturonosyltransferase [Arabidopsis thaliana]
1. 619 aa protein
Accession: NP_001189702.1 GI: 334184793
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ uncharacterized protein [Arabidopsis thaliana]
2. 67 aa protein
Accession: NP_001185395.1 GI: 334183904
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ alpha-1,4-galacturonosyltransferase [Arabidopsis thaliana]
3. 532 aa protein
Accession: NP_001118545.1 GI: 186509640
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ putative galacturonosyltransferase-like 2 [Arabidopsis thaliana]
4. 341 aa protein
Accession: NP_190645.3 GI: 79439859
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ putative galacturonosyltransferase-like 6 [Arabidopsis thaliana]
5. 346 aa protein
Accession: NP_001031573.1 GI: 79324977
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ alpha-1,4-galacturonosyltransferase [Arabidopsis thaliana]

Choose Destination
● File ○ Clipboard
○ Collections

Download 35 items.

Format: FASTA

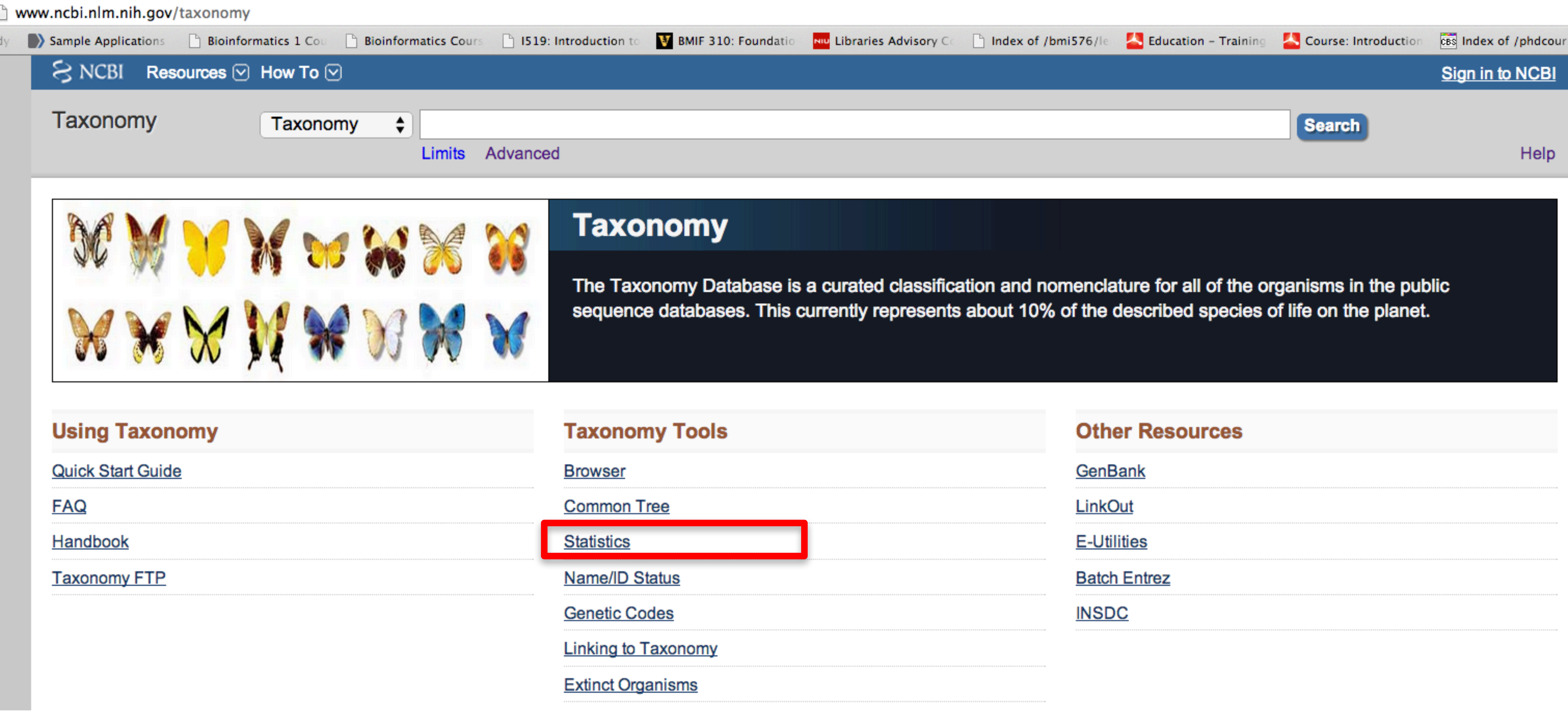
Sort by: Default order

Recent activity

Your browsing activity is e

Hands-on exercise 3

Find sequences through taxonomy database http://www.ncbi.nlm.nih.gov/taxonomy



Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

Using Taxonomy

- Quick Start Guide
- FAQ
- Handbook
- Taxonomy FTP

Taxonomy Tools

- Browser
- Common Tree
- Statistics
- Name/ID Status
- Genetic Codes
- Linking to Taxonomy
- Extinct Organisms

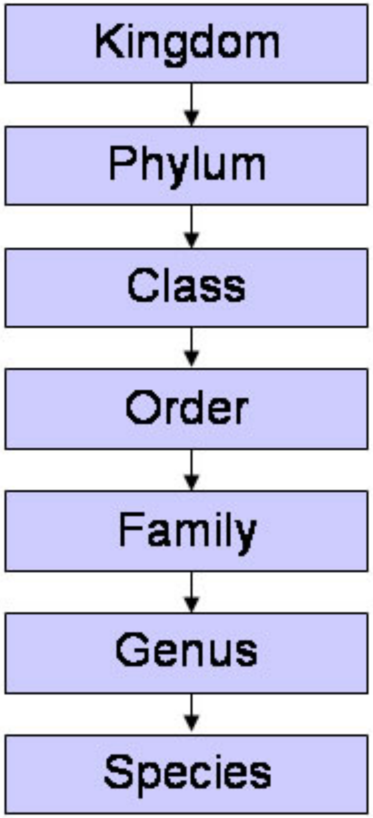
Other Resources

- GenBank
- LinkOut
- E-Utilities
- Batch Entrez
- INSDC

http://nar.oxfordjournals.org/content/early/2011/12/01/nar.gkr1178.full-text-lowres.pdf

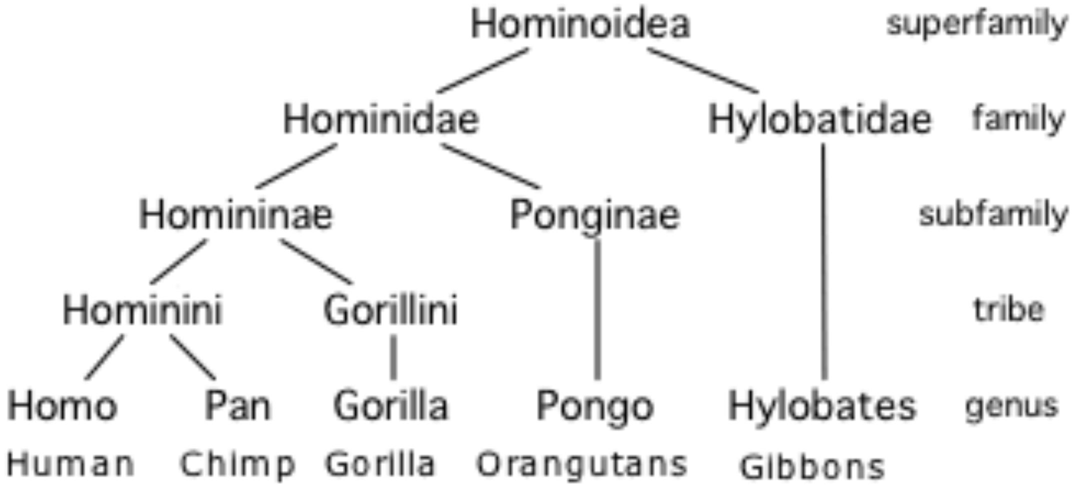
Taxonomy classification of species

Linnaeus's System of Classification



Kingdom
Phylum
Class
Order
Family
Genus
Species

Modern Hominoid Classification



Hominoidea superfamily
Hominidae Hylobatidae family
Homininae Ponginae subfamily
Hominini Gorillini tribe
Homo Pan Gorilla Pongo Hylobates genus
Human Chimp Gorilla Orangutans Gibbons

Every species has a unique taxonomy ID (e.g. human: 9606)

Homo sapiens

Taxonomy ID: 9606

Genbank common name: **human**

Inherited blast name: **primates**

Rank: species

Genetic code: Translation table 1 (Standard)

Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)

Other names:

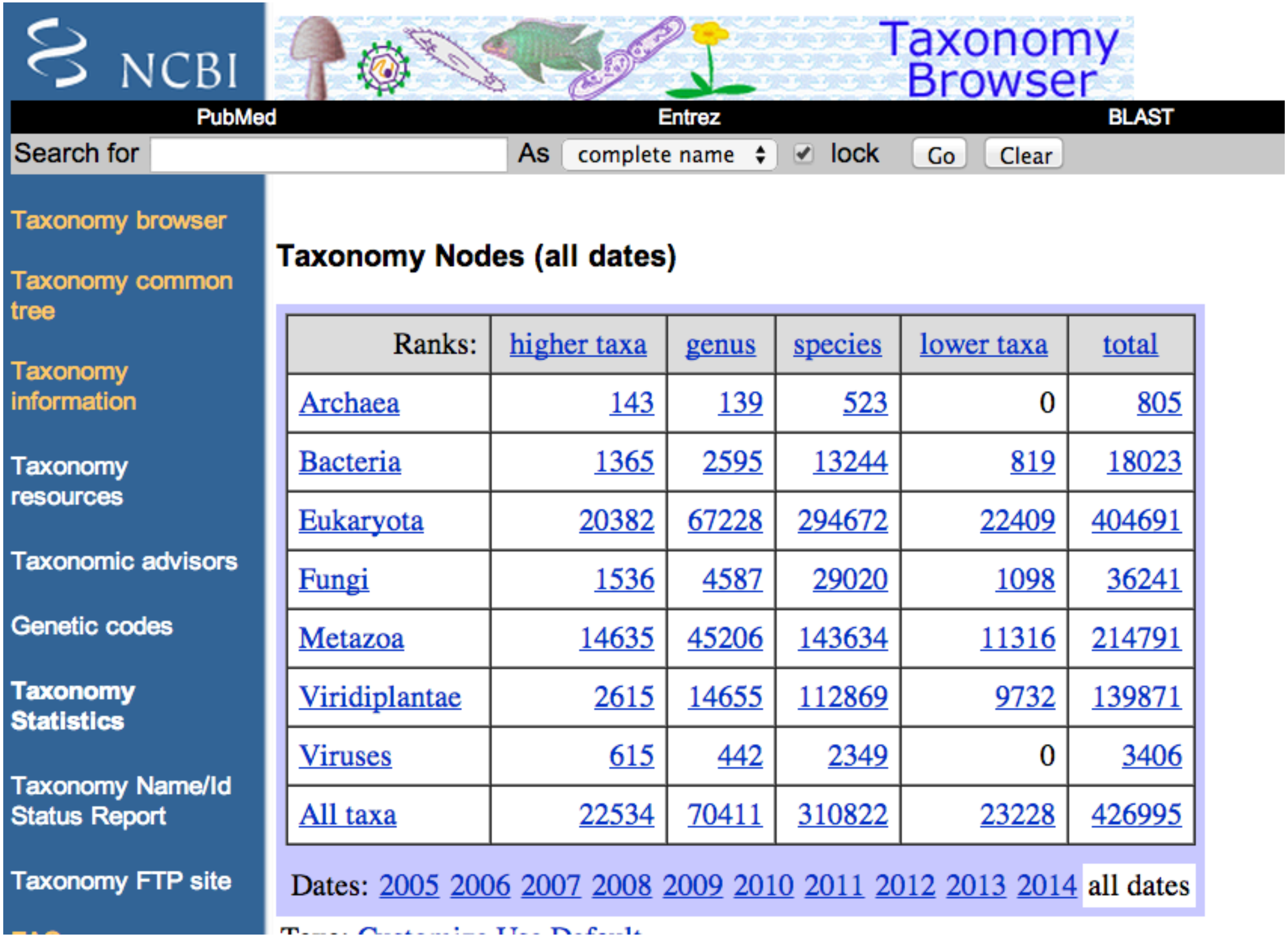
common name: **man**

authority: **Homo sapiens Linnaeus, 1758**

Lineage(full)

cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo

All species of GenBank have a taxonomy ID and lineage info



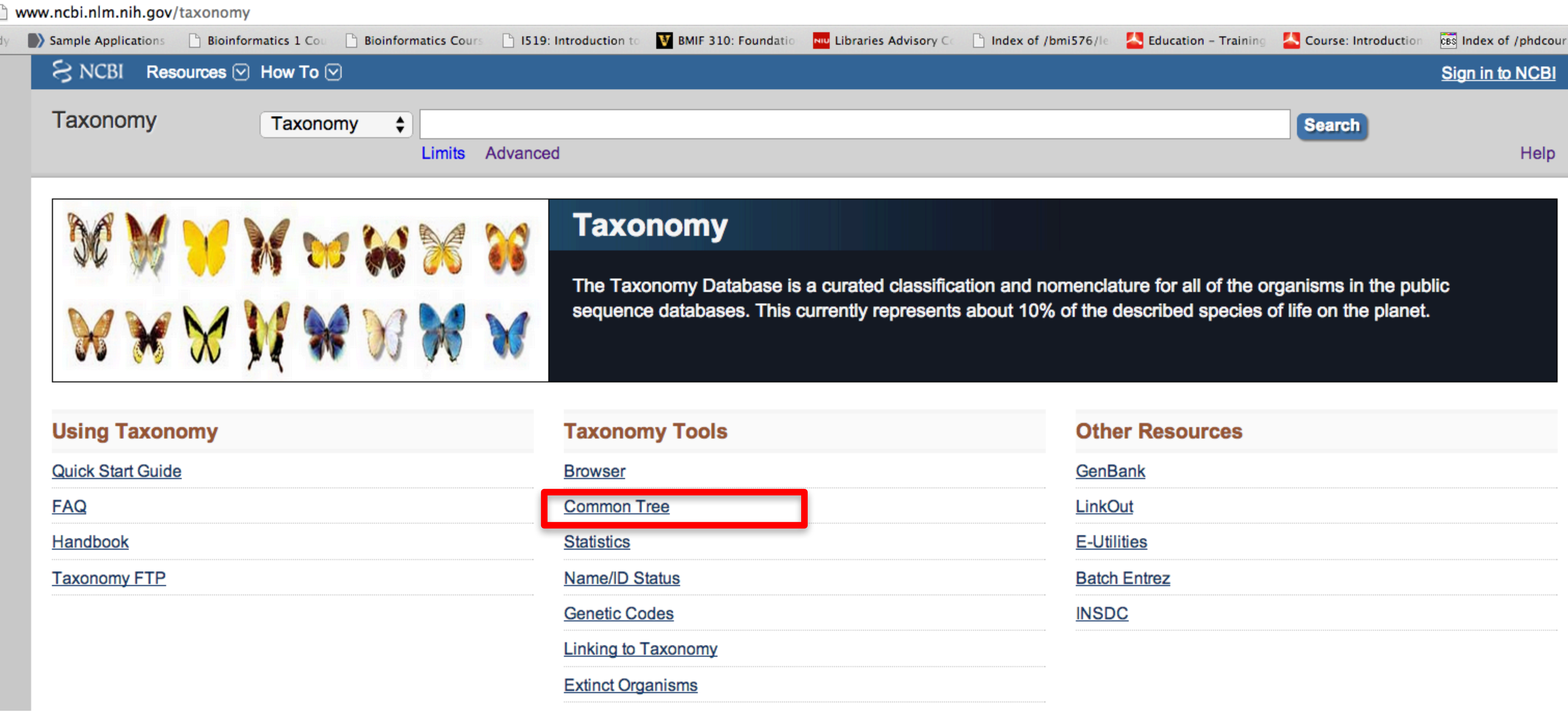
Taxonomy Nodes (all dates)

Ranks:	higher taxa	genus	species	lower taxa	total
Archaea	143	139	523	0	805
Bacteria	1365	2595	13244	819	18023
Eukaryota	20382	67228	294672	22409	404691
Fungi	1536	4587	29020	1098	36241
Metazoa	14635	45206	143634	11316	214791
Viridiplantae	2615	14655	112869	9732	139871
Viruses	615	442	2349	0	3406
All taxa	22534	70411	310822	23228	426995

Dates: 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 all dates

Taxa: Customize Use Default

If you have a list of species names and you want to find out how they are related according to NCBI taxonomy database



Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

Using Taxonomy

- Quick Start Guide
- FAQ
- Handbook
- Taxonomy FTP

Taxonomy Tools

- Browser
- Common Tree
- Statistics
- Name/ID Status
- Genetic Codes
- Linking to Taxonomy
- Extinct Organisms

Other Resources

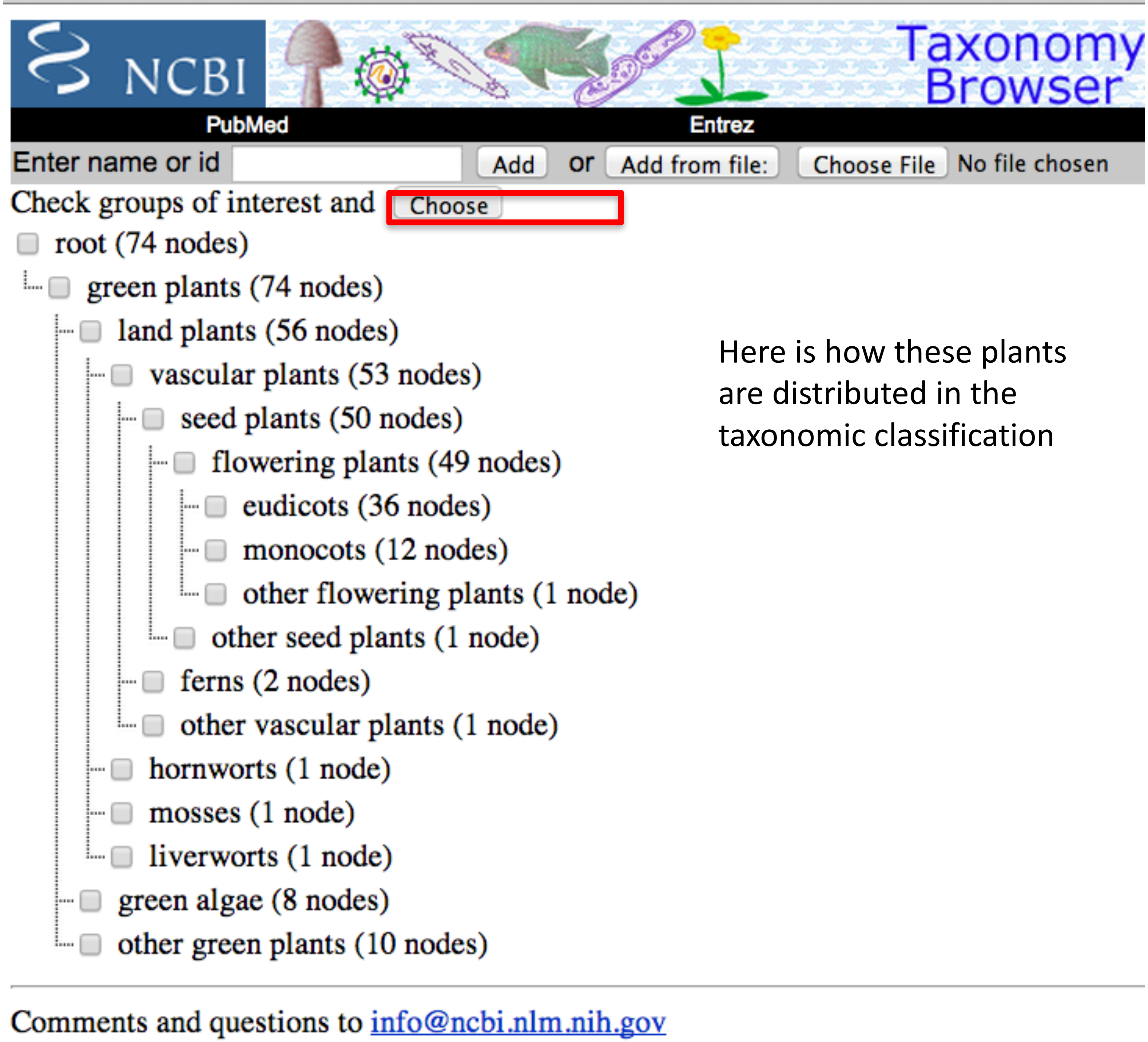
- GenBank
- LinkOut
- E-Utilities
- Batch Entrez
- INSDC

Download a list of plant species from:
http://cys.bios.niu.edu/yyin/teach/PBB/plant-genome.txt

Click this button to add organisms to the tree

Comments and questions to info@ncbi.nlm.nih.gov

Choose file -> Add from file

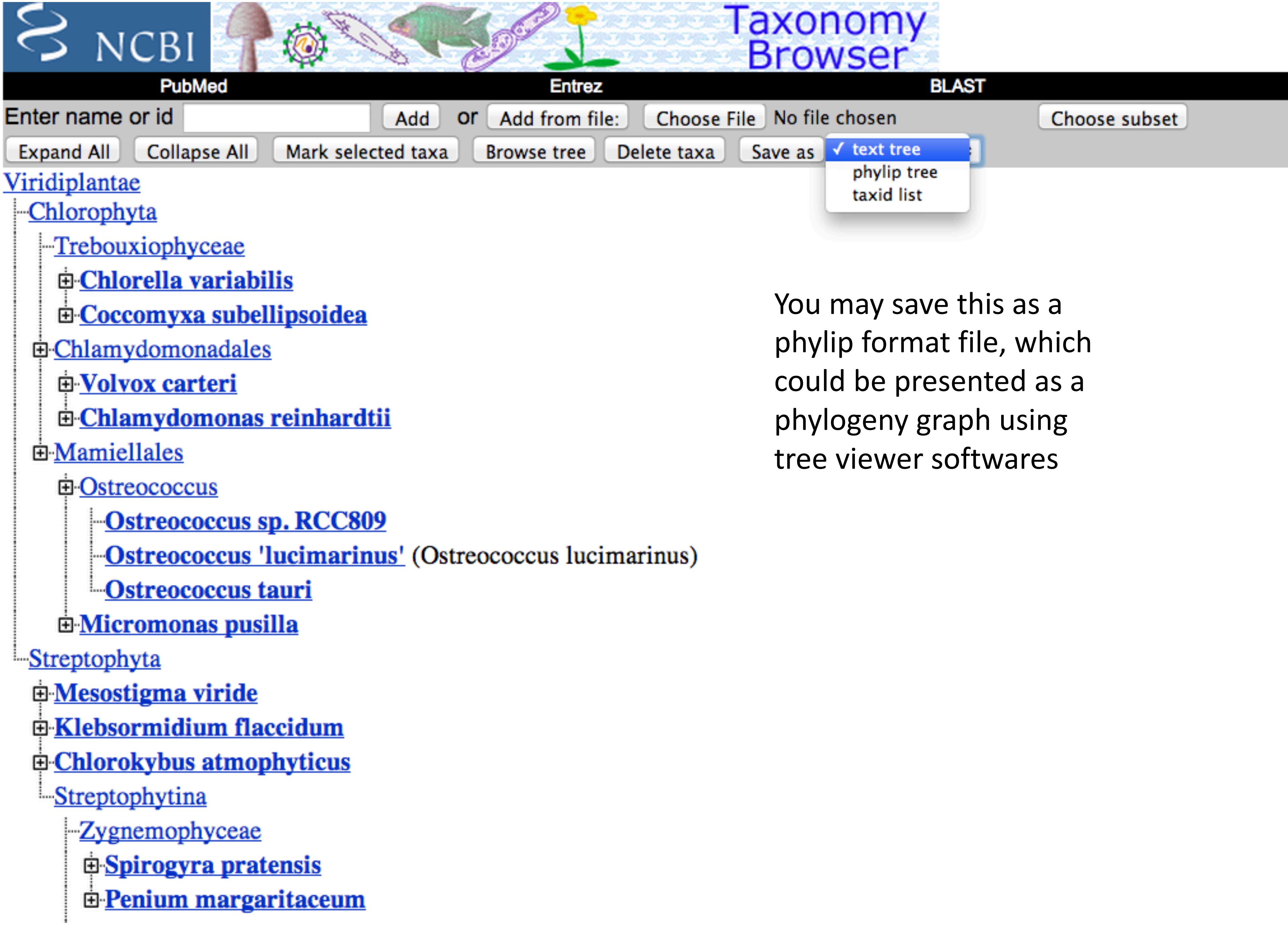


Check groups of interest and [Choose]

- ☐ root (74 nodes)
  - ☐ green plants (74 nodes)
    - ☐ land plants (56 nodes)
      - ☐ vascular plants (53 nodes)
        - ☐ seed plants (50 nodes)
          - ☐ flowering plants (49 nodes)
            - ☐ eudicots (36 nodes)
            - ☐ monocots (12 nodes)
            - ☐ other flowering plants (1 node)
          - ☐ other seed plants (1 node)
        - ☐ ferns (2 nodes)
        - ☐ other vascular plants (1 node)
      - ☐ hornworts (1 node)
      - ☐ mosses (1 node)
      - ☐ liverworts (1 node)
    - ☐ green algae (8 nodes)
    - ☐ other green plants (10 nodes)

Comments and questions to info@ncbi.nlm.nih.gov

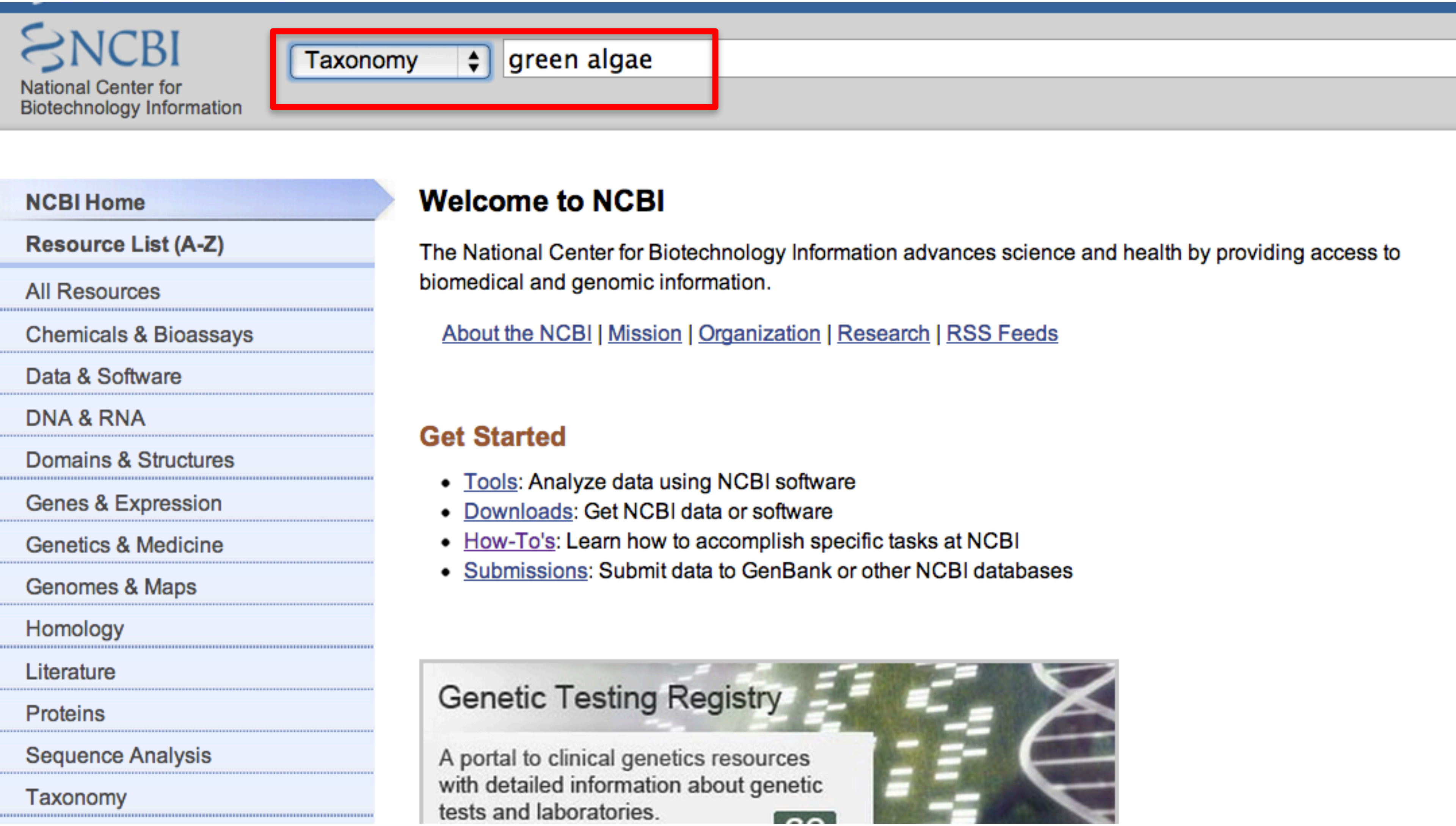
Here is how these plants are distributed in the taxonomic classification



- Viridiplantae
  - Chlorophyta
    - Trebouxiophyceae
      - ⊞ **Chlorella variabilis**
      - ⊞ **Coccomyxa subellipsoidea**
    - ⊞ Chlamydomonadales
      - ⊞ **Volvox carteri**
      - ⊞ **Chlamydomonas reinhardtii**
    - ⊞ Mamiellales
      - ⊞ Ostreococcus
        - **Ostreococcus sp. RCC809**
        - **Ostreococcus 'lucimarinus'** (Ostreococcus lucimarinus)
        - **Ostreococcus tauri**
      - ⊞ **Micromonas pusilla**
  - Streptophyta
    - ⊞ **Mesostigma viride**
    - ⊞ **Klebsormidium flaccidum**
    - ⊞ **Chlorokybus atmophyticus**
    - Streptophytina
      - Zygnemophyceae
        - ⊞ **Spirogyra pratensis**
        - ⊞ **Penium margaritaceum**

You may save this as a phylip format file, which could be presented as a phylogeny graph using tree viewer softwares


Viridiplantae(kingdom)
Chlorophyta(phylum)
Trebouxiophyceae(class)
Chlamydomonadales(order)
Mamiellales(order)
Ostreococcus(genus)
Chlorella variabilis
Coccomyxa subellipsoidea
Volvox carteri
Chlamydomonas reinhardtii
Micromonas pusilla
Ostreococcus sp. RCC809
Ostreococcus lucimarinus
Ostreococcus tauri
Chlorophyta
Green
algae
Streptophyta(phylum)
Mesostigmatophyceae(class)
Klebsormidiophyceae(class)
Chlorokybophyceae(class)
Zygnemophyceae(class)
Charophyceae(class)
Coleochaetales(order)
Coleochaete(genus)
Mesostigma viride
Klebsormidium flaccidum
Chlorokybus atmophyticus
Spirogyra pratensis
Penium margaritaceum
Nitella hyalina
Chara vulgaris
Chaetosphaeridium globosum
Coleochaete sp. CFD
Coleochaete orbicularis
Charophycean
Green
algae
Anthocerotopsida(class)
Bryopsida(class)
Marchantiopsida(class)
Isoetopsida(class)
Polypodiales(order)
Pinidae(subclass)
Amborellales(order)
Anthoceros punctatus
Physcomitrella patens
Marchantia polymorpha
Selaginella moellendorffii
Ceratopteris richardii
Pteridium aquilinum
Picea abies
Amborella trichopoda
Hornwort
Moss
Liverwort
Spike Moss
Fern
Gymnosperm
Liliopsida(class)
Alismatales(order)
Zingiberales(order)
Andropogoneae(tribe)
Panicoideae(subfamily)
commelinids(subclass)
Paniceae(tribe)
Poaceae(family)
Bambusoideae(subfamily)
Ehrhartoideae(subfamily)
Pooideae(subfamily)
Spirodela polyrhiza (duckweed)
Musa acuminata (banana)
Zea mays (maize)
Sorghum bicolor (sorghum)
Setaria italica (foxtail millet)
Panicum hallii (Hall's panicgrass)
Panicum virgatum (switchgrass)
Phyllostachys edulis (moso bamboo)
Oryza sativa (rice)
Brachypodium distachyon
Triticum aestivum (wheat)
Hordeum vulgare (barley)
Monocot
Ranunculales(order)
Caryophyllales(order)
asterids(subclass)
Vitales(order)
Cucurbitales(order)
Rosales(order)
Maloideae(subfamily)
Malpighiales(order)
Euphorbiaceae(family)
Fabales(order)
Phaseoleae(tribe)
Myrtales(order)
Malvales(order)
Sapindales(order)
Caricaceae(family)
Boechereae(tribe)
Brassiceae(tribe)
Brassicales(order)
Eutremeae(tribe)
Brassicaceae(family)
Capsella(genus)
Camelineae(tribe)
Arabidopsis(genus)
rosids(subclass)
Aquilegia coerulea (columbine)
Beta vulgaris (sugar beet)
Erythranthe guttata (monkey flower)
Solanum tuberosum (potato)
Solanum lycopersicum (tomato)
Vitis vinifera (grape)
Cucumis sativus (cucumber)
Citrullus lanatus (watermelon)
Fragaria vesca (strawberry)
Prunus persica (peach)
Malus domestica (apple)
Linum usitatissimum (common flax)
Populus trichocarpa (poplar)
Ricinus communis (castor oil)
Manihot esculenta (cassava)
Lotus japonicus (lotus)
Medicago truncatula (barrel medic)
Cicer arietinum (chickpea)
Phaseolus vulgaris (kidney bean)
Glycine max (soybean)
Cajanus cajan (pigeon pea)
Eucalyptus grandis (rose gum)
Gossypium raimondii (cotton)
Theobroma cacao (cocoa tree)
Citrus clementina (clementine orange)
Citrus sinensis (sweet orange)
Carica papaya (papaya)
Boechera stricta (rockcress)
Brassica rapa (napa cabbage)
Eutrema parvulum
Eutrema salsugineum (salt cress)
Capsella grandiflora
Capsella rubella (shepherd's-purse)
Arabidopsis halleri
Arabidopsis lyrata (rock-cress)
Arabidopsis thaliana (mouseear cress)
Dicot



Welcome to NCBI

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

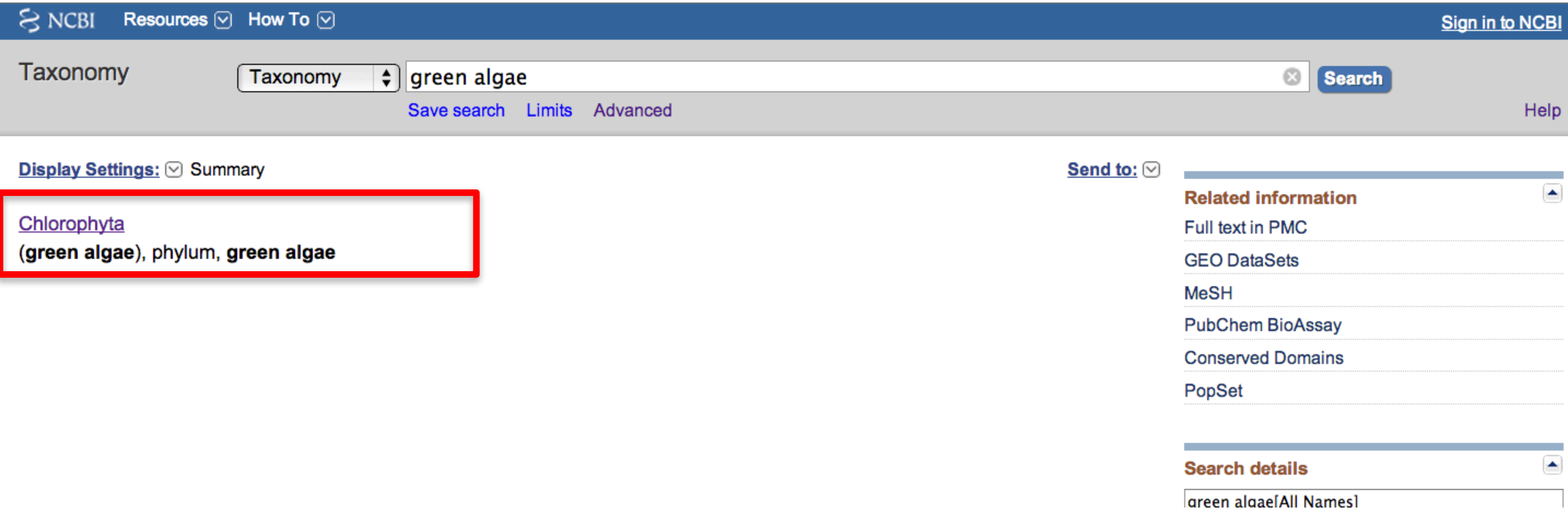
About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Genetic Testing Registry

A portal to clinical genetics resources with detailed information about genetic tests and laboratories.

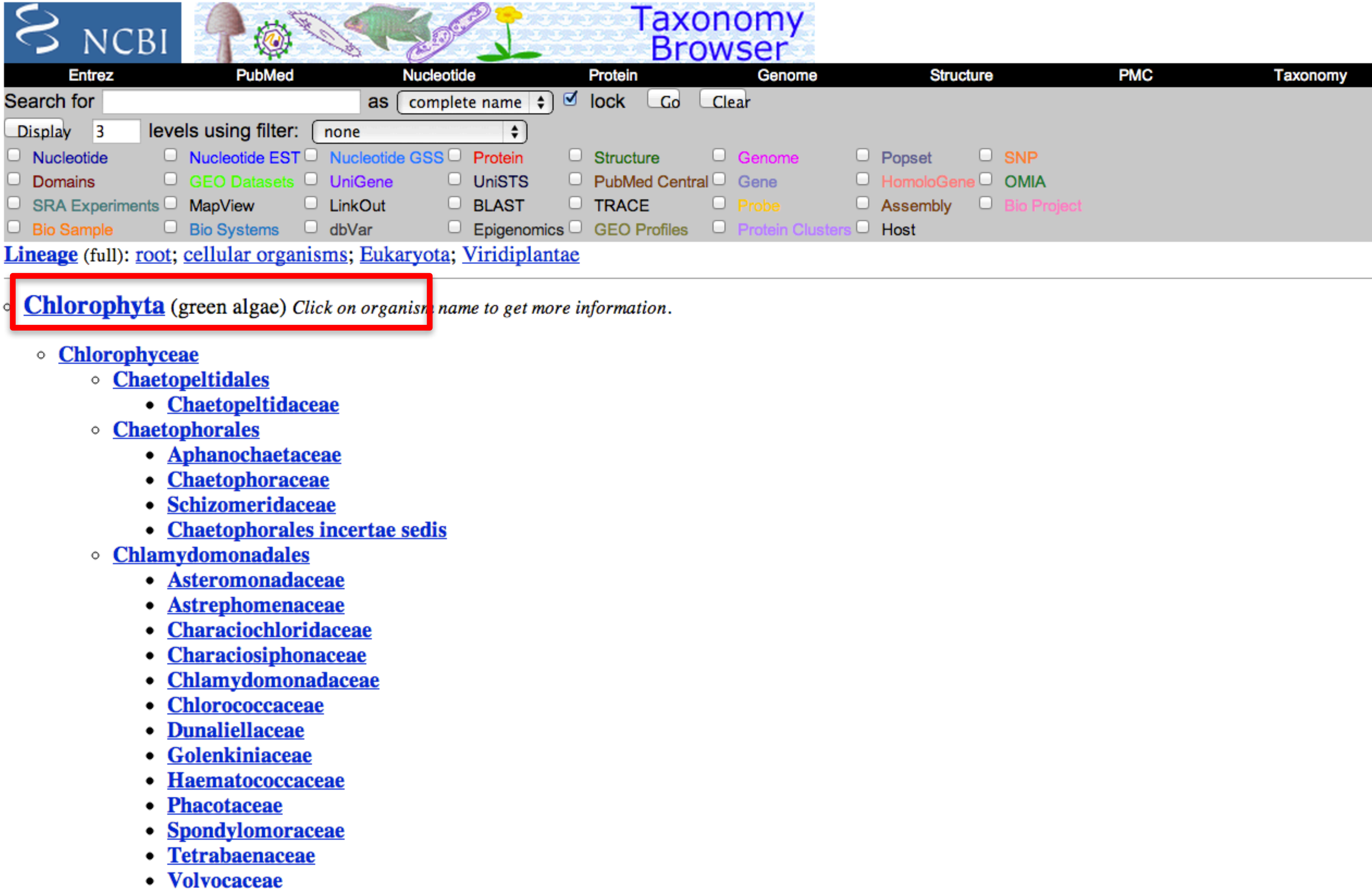


Taxonomy: green algae

Display Settings: Summary

Chlorophyta
(**green algae**), phylum, **green algae**

Search details: green algae[All Names]



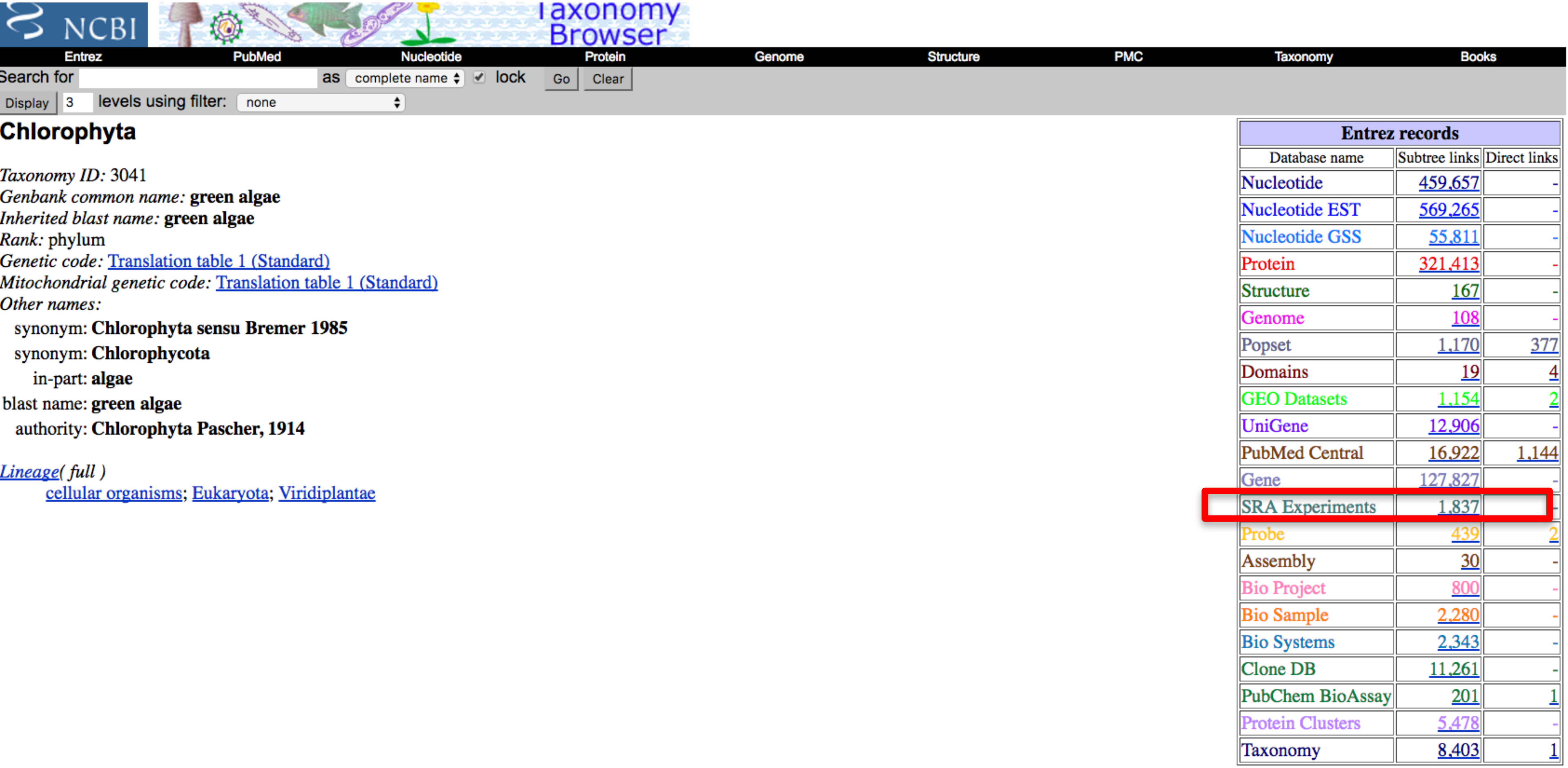
Search for as complete name ☑ lock

Display 3 levels using filter: none

Lineage (full): root; cellular organisms; Eukaryota; Viridiplantae

Chlorophyta (green algae) *Click on organism name to get more information.*

- **Chlorophyceae**
  - **Chaetopeltidales**
    - **Chaetopeltidaceae**
  - **Chaetophorales**
    - **Aphanochaetaceae**
    - **Chaetophoraceae**
    - **Schizomeridaceae**
    - **Chaetophorales incertae sedis**
  - **Chlamydomonadales**
    - **Asteromonadaceae**
    - **Astrephomenaceae**
    - **Characiochloridaceae**
    - **Characiosiphonaceae**
    - **Chlamydomonadaceae**
    - **Chlorococcaceae**
    - **Dunaliellaceae**
    - **Golenkiniaceae**
    - **Haematococcaceae**
    - **Phacotaceae**
    - **Spondylomoraceae**
    - **Tetrabaenaceae**
    - **Volvocaceae**



Search for as complete name ☑ lock

Display 3 levels using filter: none

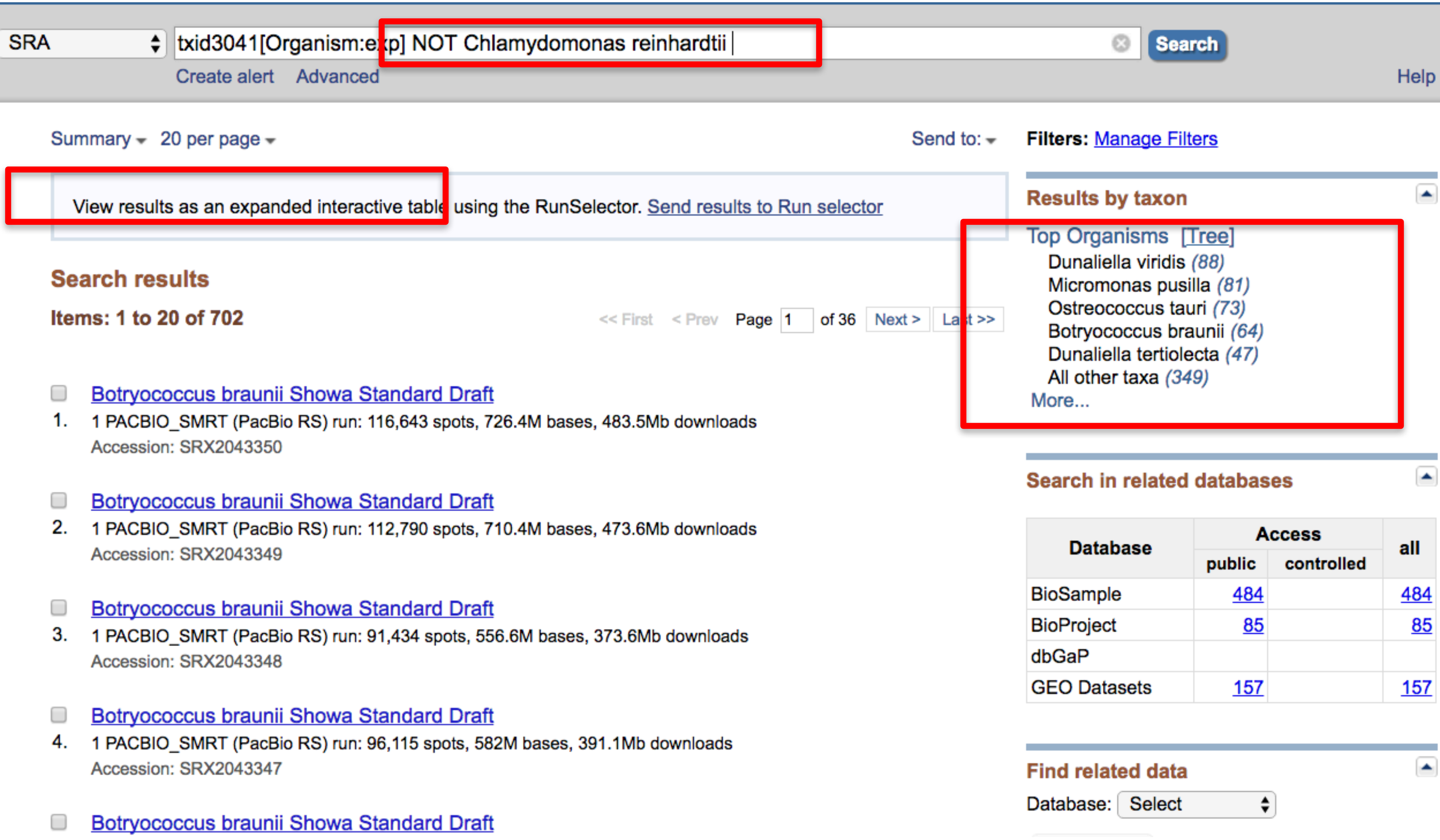
Chlorophyta

Taxonomy ID: 3041
Genbank common name: **green algae**
Inherited blast name: **green algae**
Rank: phylum
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 1 (Standard)
Other names:
synonym: **Chlorophyta sensu Bremer 1985**
synonym: **Chlorophycota**
in-part: **algae**
blast name: **green algae**
authority: **Chlorophyta Pascher, 1914**

Lineage(full)
cellular organisms; Eukaryota; Viridiplantae

Entrez records

Database name	Subtree links	Direct links
Nucleotide	459,657	-
Nucleotide EST	569,265	-
Nucleotide GSS	55,811	-
Protein	321,413	-
Structure	167	-
Genome	108	-
Popset	1,170	377
Domains	19	4
GEO Datasets	1,154	2
UniGene	12,906	-
PubMed Central	16,922	1,144
Gene	127,827	-
SRA Experiments	1,837	-
Probe	439	2
Assembly	30	-
Bio Project	800	-
Bio Sample	2,280	-
Bio Systems	2,343	-
Clone DB	11,261	-
PubChem BioAssay	201	1
Protein Clusters	5,478	-
Taxonomy	8,403	1



txid3041[Organism:exp] NOT Chlamydomonas reinhardtii

View results as an expanded interactive table using the RunSelector. Send results to Run selector

Search results

Items: 1 to 20 of 702

☐ Botryococcus braunii Showa Standard Draft
1. 1 PACBIO_SMRT (PacBio RS) run: 116,643 spots, 726.4M bases, 483.5Mb downloads
Accession: SRX2043350

☐ Botryococcus braunii Showa Standard Draft
2. 1 PACBIO_SMRT (PacBio RS) run: 112,790 spots, 710.4M bases, 473.6Mb downloads
Accession: SRX2043349

☐ Botryococcus braunii Showa Standard Draft
3. 1 PACBIO_SMRT (PacBio RS) run: 91,434 spots, 556.6M bases, 373.6Mb downloads
Accession: SRX2043348

☐ Botryococcus braunii Showa Standard Draft
4. 1 PACBIO_SMRT (PacBio RS) run: 96,115 spots, 582M bases, 391.1Mb downloads
Accession: SRX2043347

☐ Botryococcus braunii Showa Standard Draft

Results by taxon

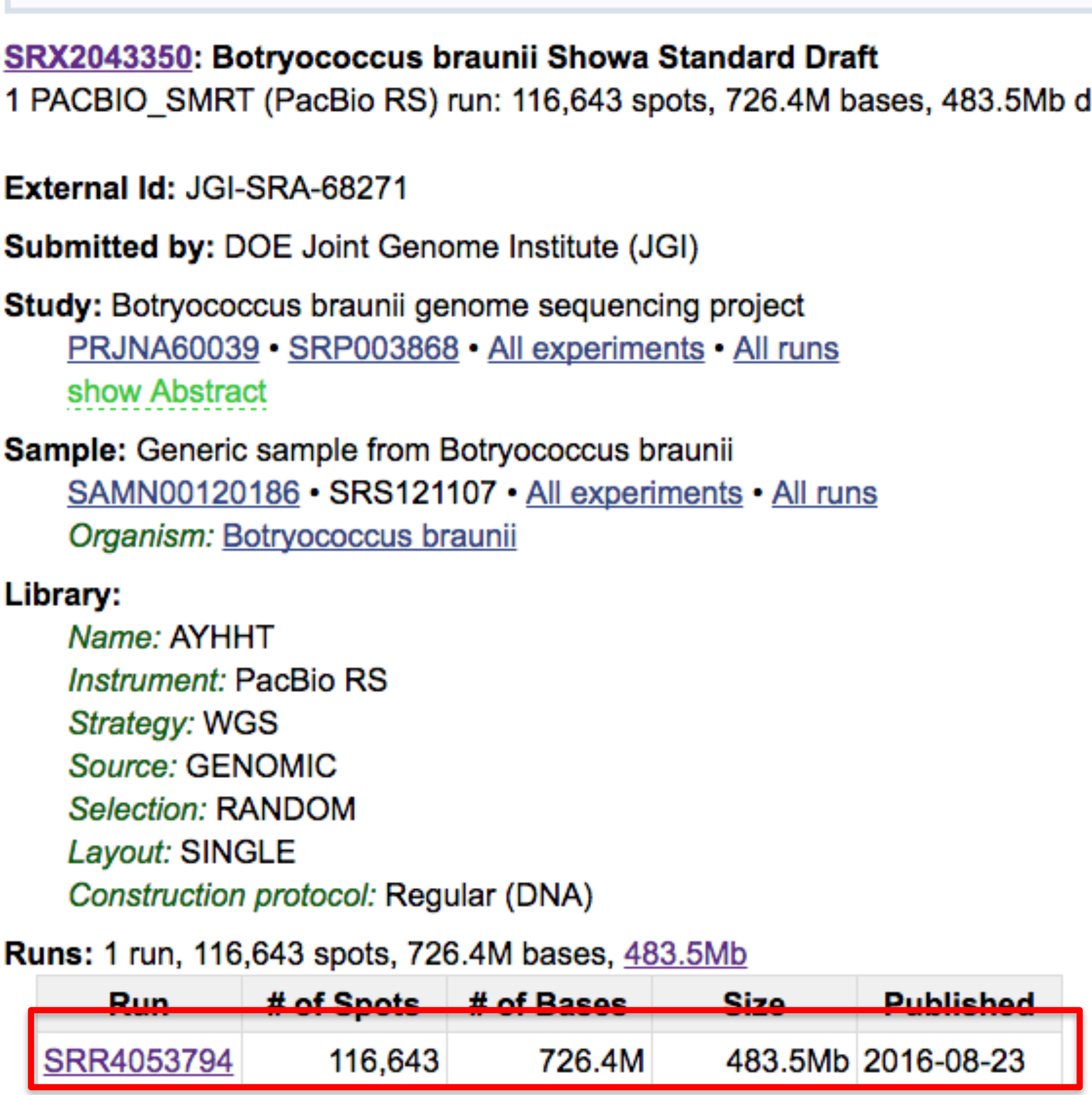
Top Organisms [Tree]

Dunaliella viridis	*(88)*
Micromonas pusilla	*(81)*
Ostreococcus tauri	*(73)*
Botryococcus braunii	*(64)*
Dunaliella tertiolecta	*(47)*
All other taxa	*(349)*

Search in related databases

Database	Access		all
	public	controlled	
BioSample	484		484
BioProject	85		85
dbGaP			
GEO Datasets	157		157

Find related data



SRX2043350: Botryococcus braunii Showa Standard Draft
1 PACBIO_SMRT (PacBio RS) run: 116,643 spots, 726.4M bases, 483.5Mb downloads

External Id: JGI-SRA-68271

Submitted by: DOE Joint Genome Institute (JGI)

Study: Botryococcus braunii genome sequencing project
PRJNA60039 • SRP003868 • All experiments • All runs
show Abstract

Sample: Generic sample from Botryococcus braunii
SAMN00120186 • SRS121107 • All experiments • All runs
Organism: Botryococcus braunii

Library:
Name: AYHHT
Instrument: PacBio RS
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: SINGLE
Construction protocol: Regular (DNA)

Runs: 1 run, 116,643 spots, 726.4M bases, 483.5Mb

Run	# of Spots	# of Bases	Size	Published
SRR4053794	116,643	726.4M	483.5Mb	2016-08-23

Next lecture

NCBI resources II: tools and ftp resources