NCBI web resources I: databases and Entrez

Yanbin Yin

Spring 2013

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

Homework assignment 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

- Using NCBI Entrez to search for fern (Moniliformopses) transcriptome data and relevant literature
- Download the EST data and SRA data to your local computer and transfer it to the Linux server (leu)

- Write a report to include all the operations and screen shots

Due on Feb 5th (send by email or bring printed hard copy to class)

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

Aspects of 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

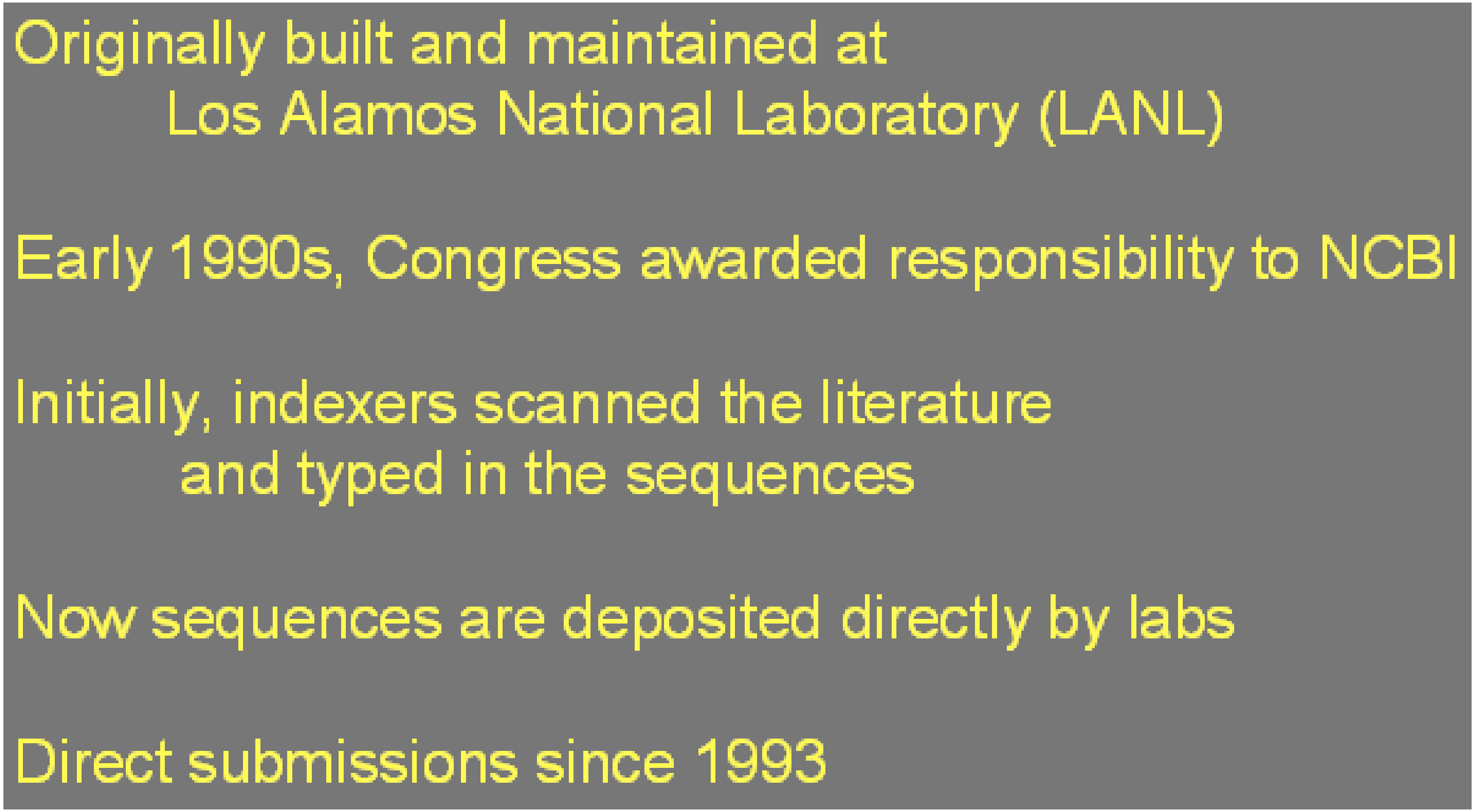
Information Hubs: Analyses

- 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 history



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

Early 1990s, Congress awarded responsibility to NCBI

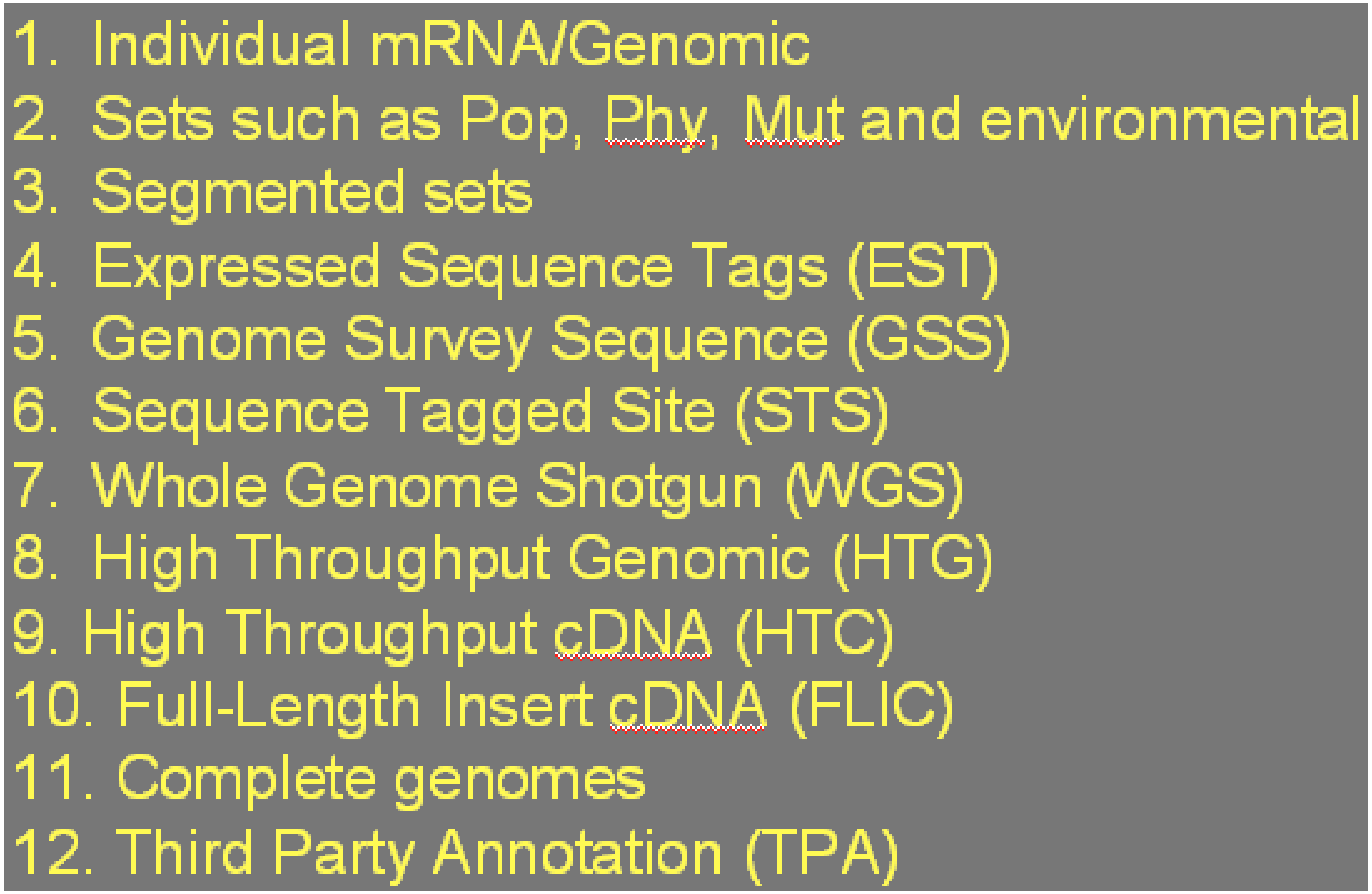
Initially, indexers scanned the literature
and typed in the sequences

Now sequences are deposited directly by labs

Direct submissions since 1993

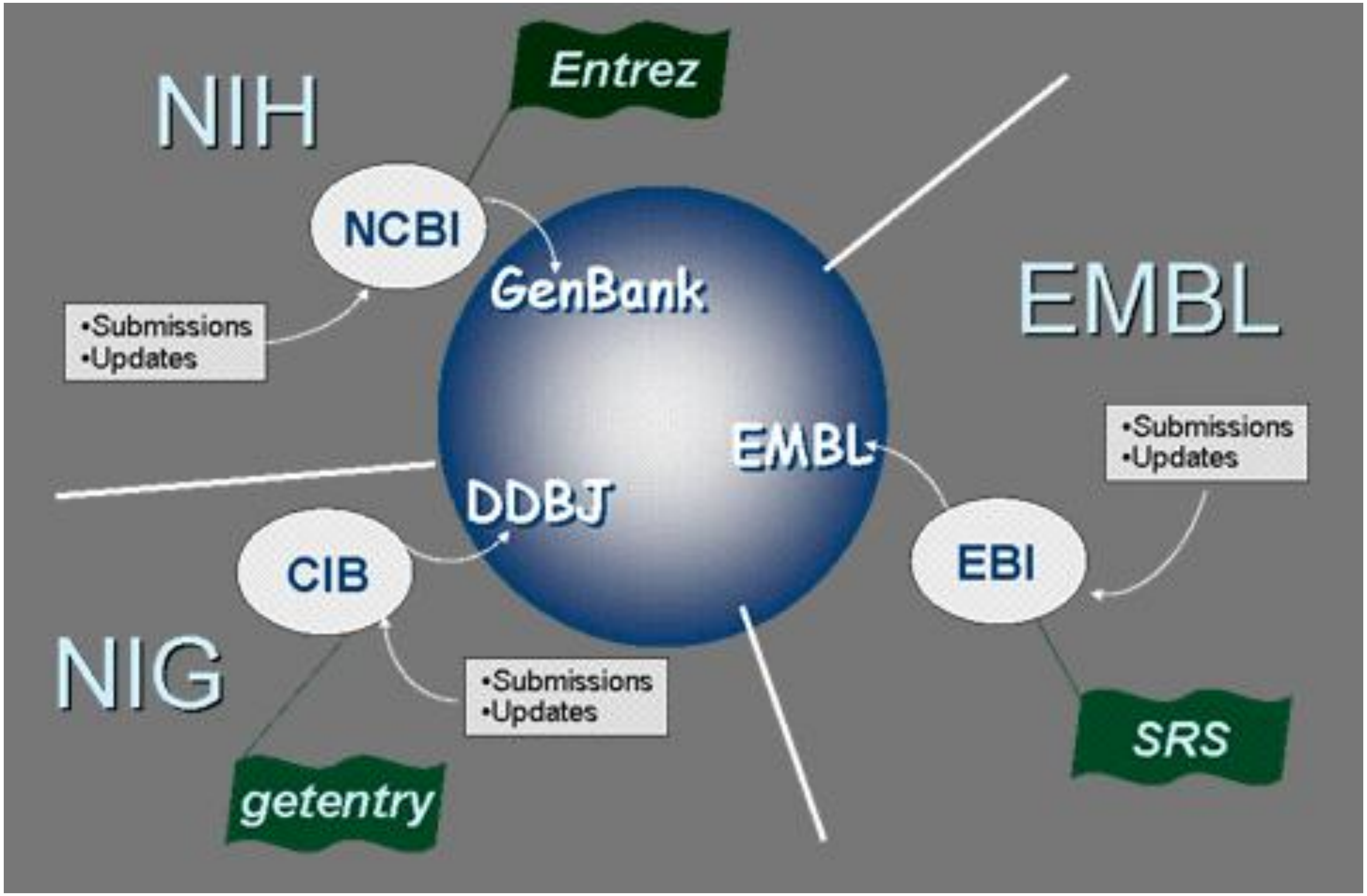
ftp://ftp.ncbi.nih.gov/genbank/

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

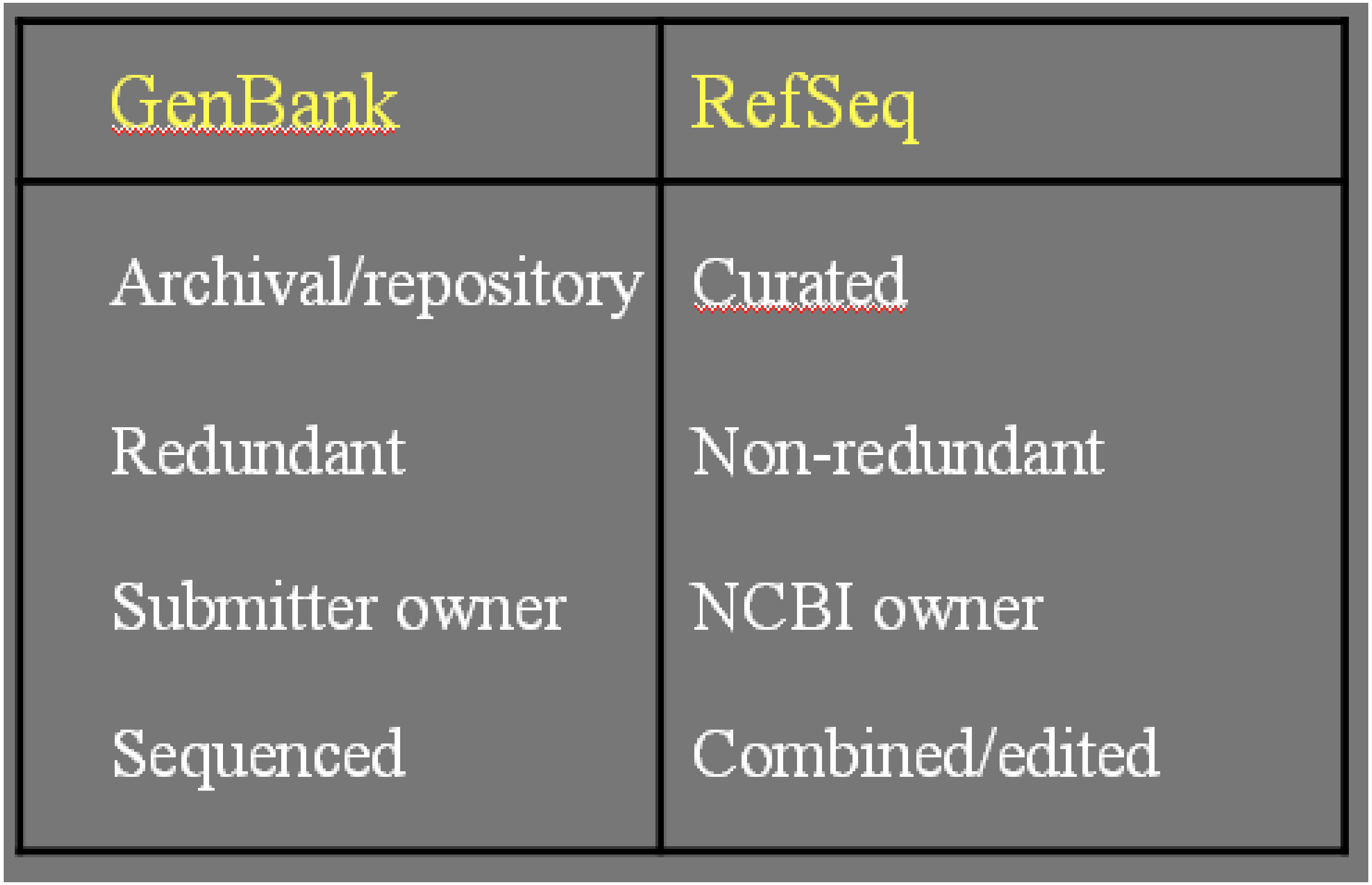


RefSeq: NCBI’s Derivative Sequence Database

- **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

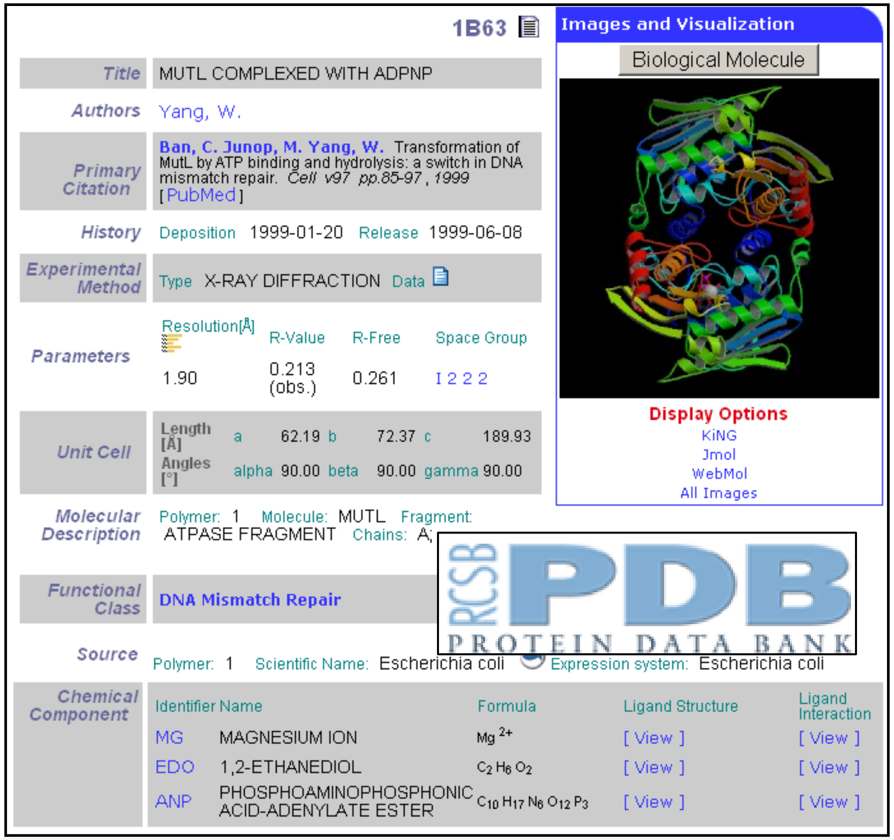
ftp://ftp.ncbi.nih.gov/refseq/release/

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]	R-Value	R-Free	Space Group
1.90	0.213 (obs.)	0.261	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

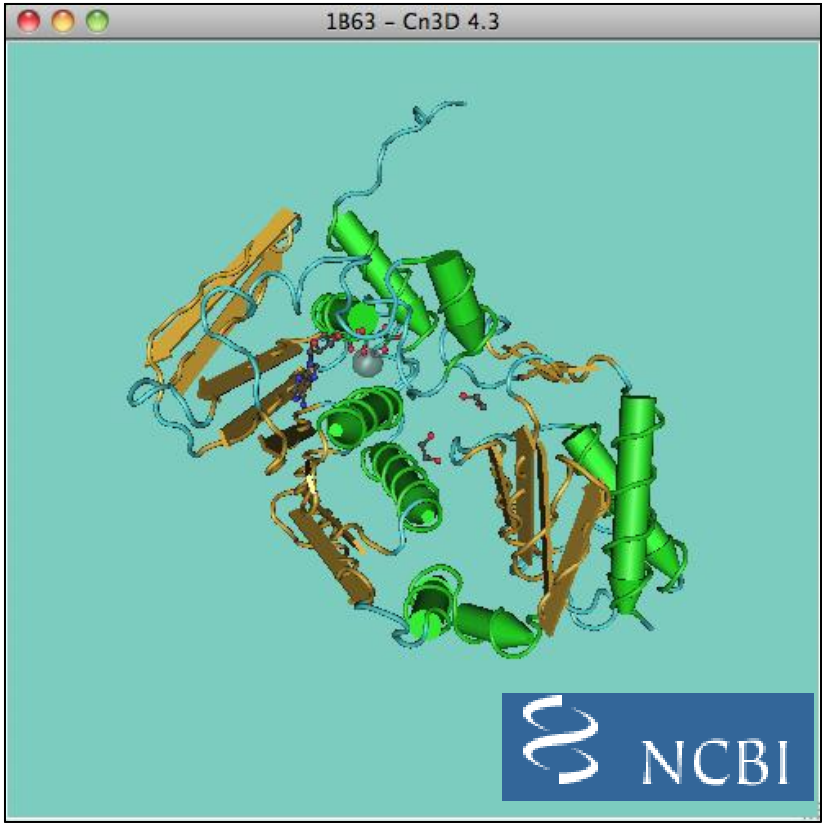
Molecular Description: Polymer: 1 Molecule: MUTL Fragment: ATPASE FRAGMENT Chains: A;

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

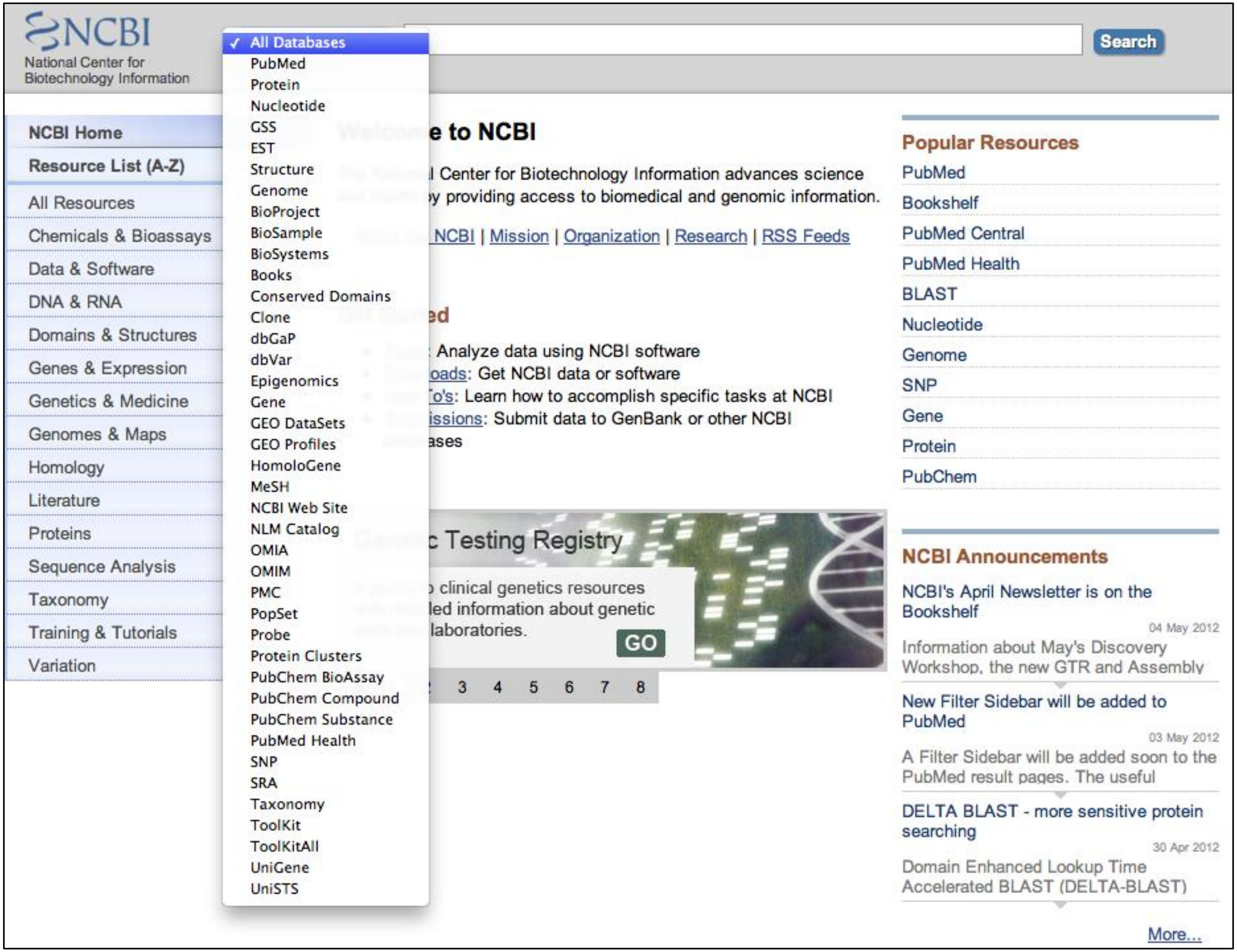
- Searchable with RPS-BLAST
- Sources
  - SMART
  - PFAM
  - COGs
  - NCBI curated domains
    - structure-informed alignments

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 Announcements

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

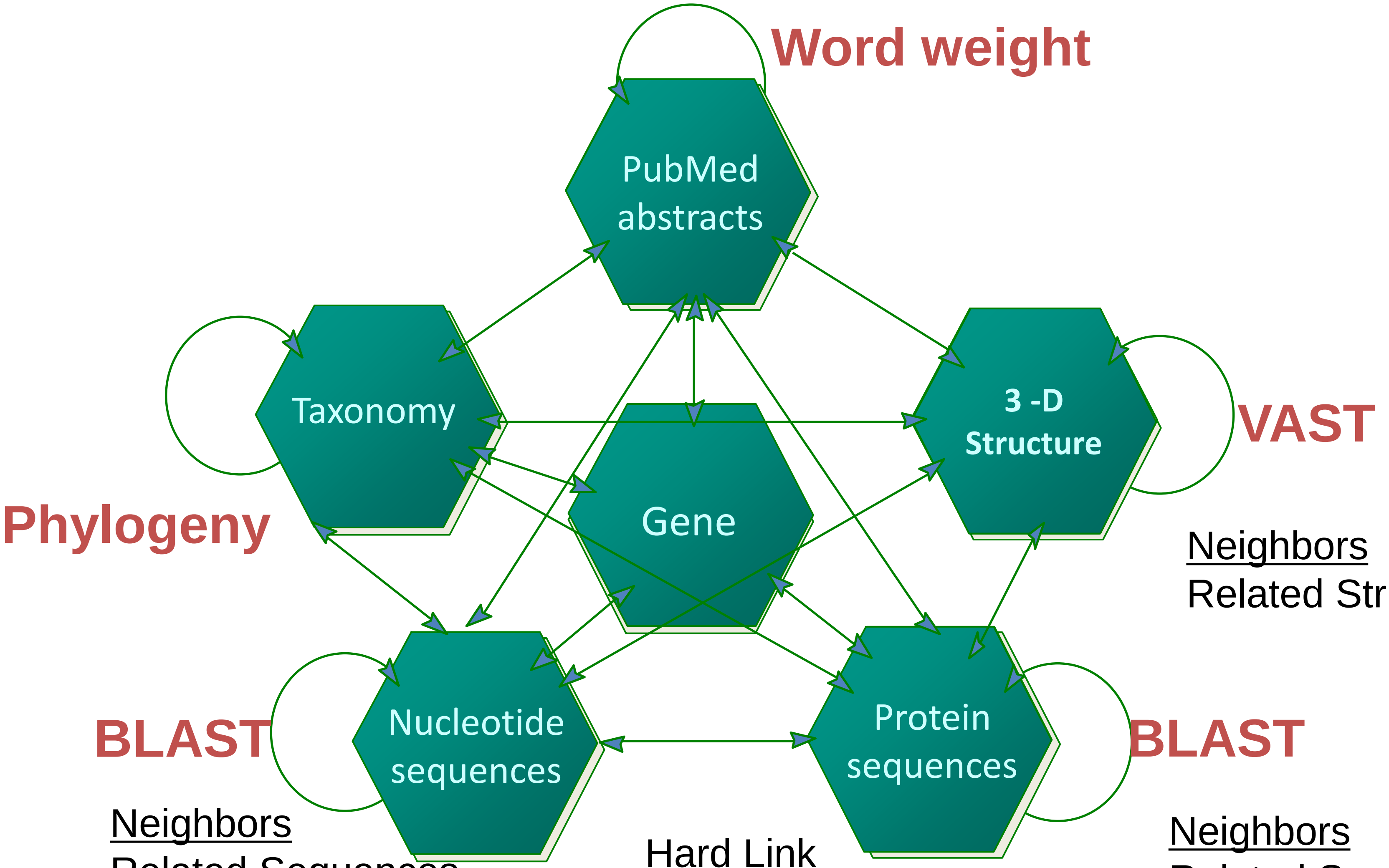
A Filter Sidebar will be added soon to the PubMed result pages. The useful

DELTA BLAST - more sensitive protein searching

30 Apr 2012

Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST)

Entrez: A Discovery System



Word weight

PubMed abstracts

Taxonomy

3 -D Structure

VAST

Gene

Phylogeny

Nucleotide sequences

Protein sequences

BLAST

BLAST

Hard Link

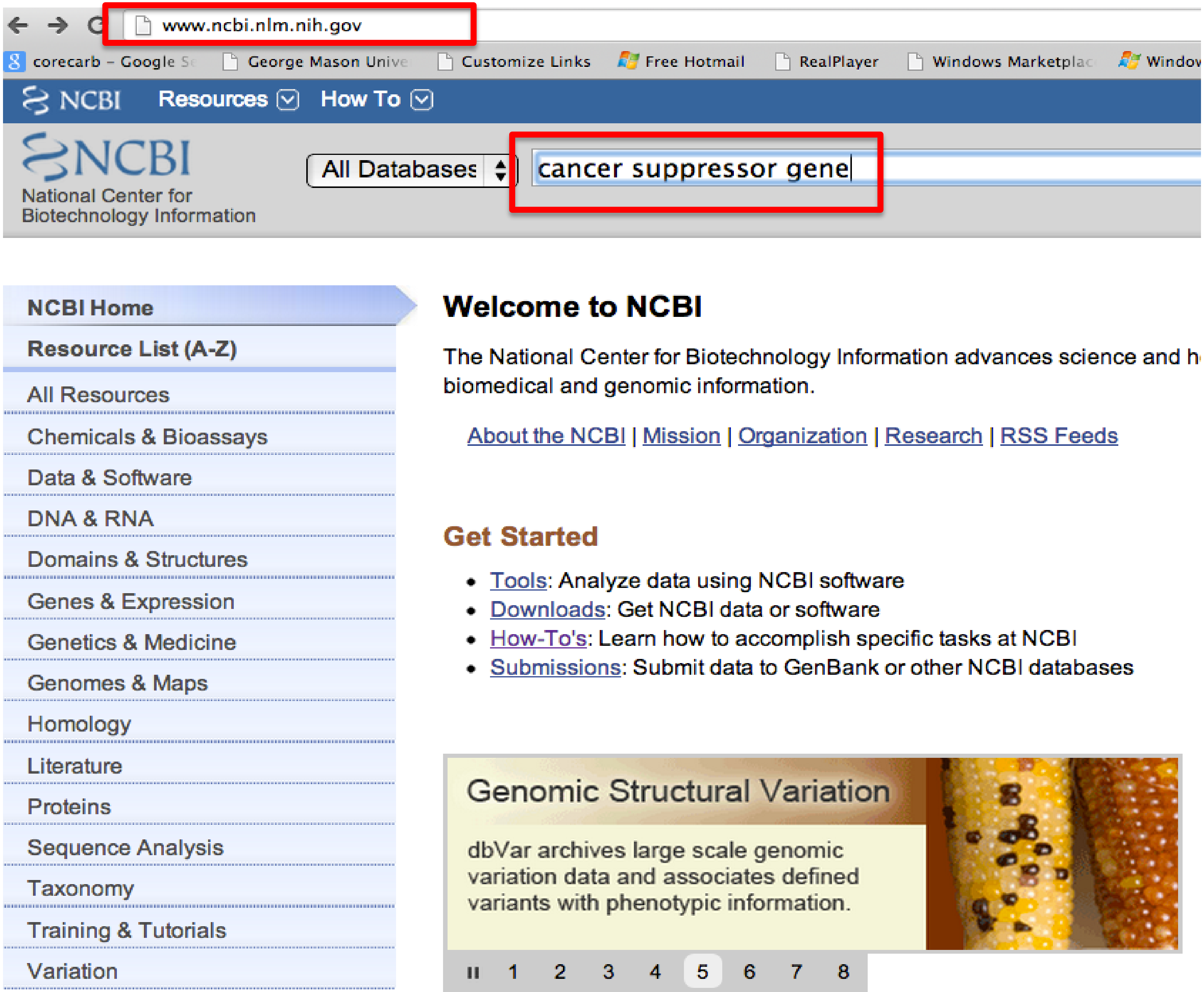
<u>Neighbors</u>
Related Structures

<u>Neighbors</u>
Related Sequences

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

Hands-on exercise 1

Cancer related genes



www.ncbi.nlm.nih.gov

cancer suppressor gene

Welcome to NCBI

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

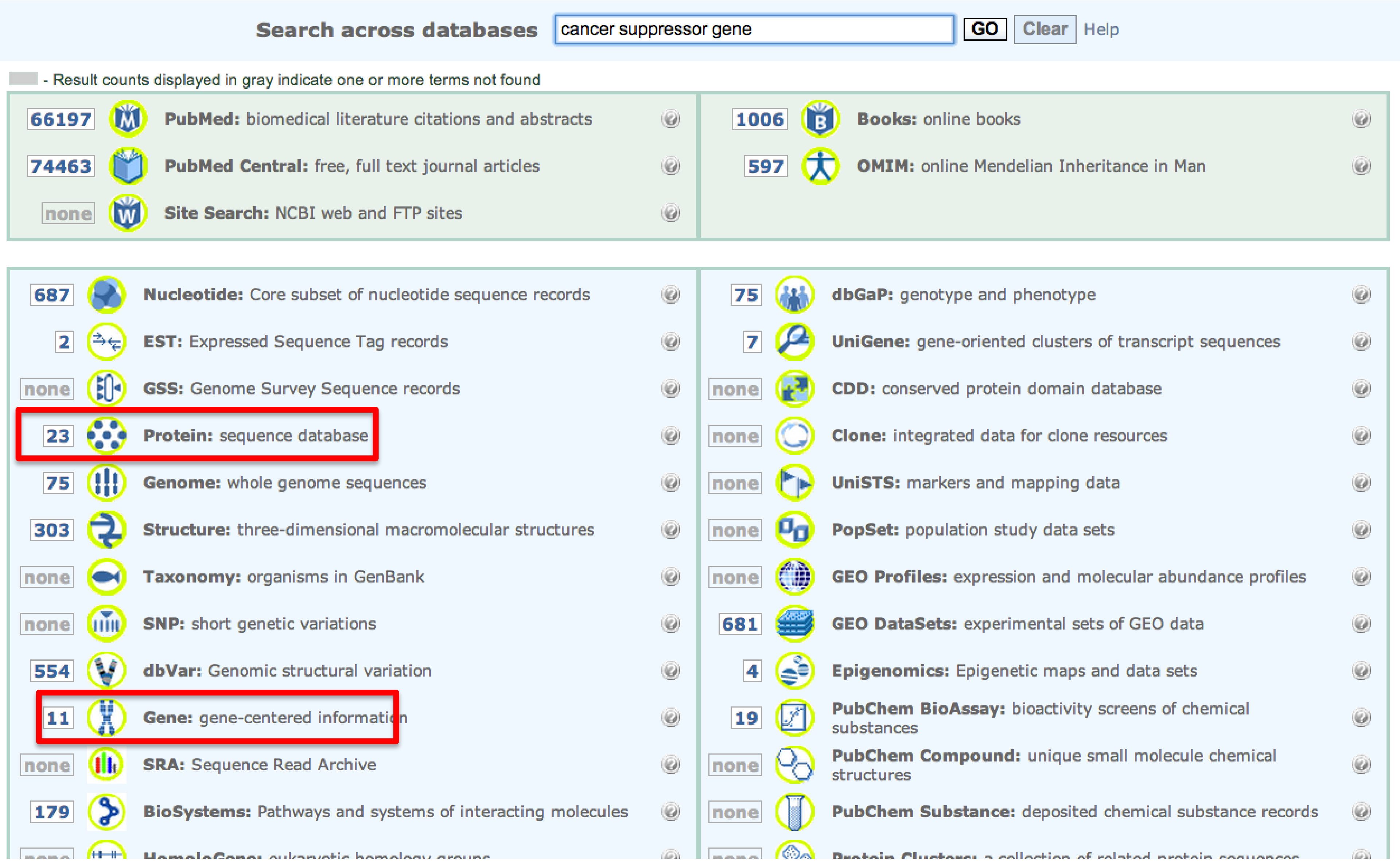
About the NCBI | Mission | Organization | Research | RSS Feeds

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

Genomic Structural Variation

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

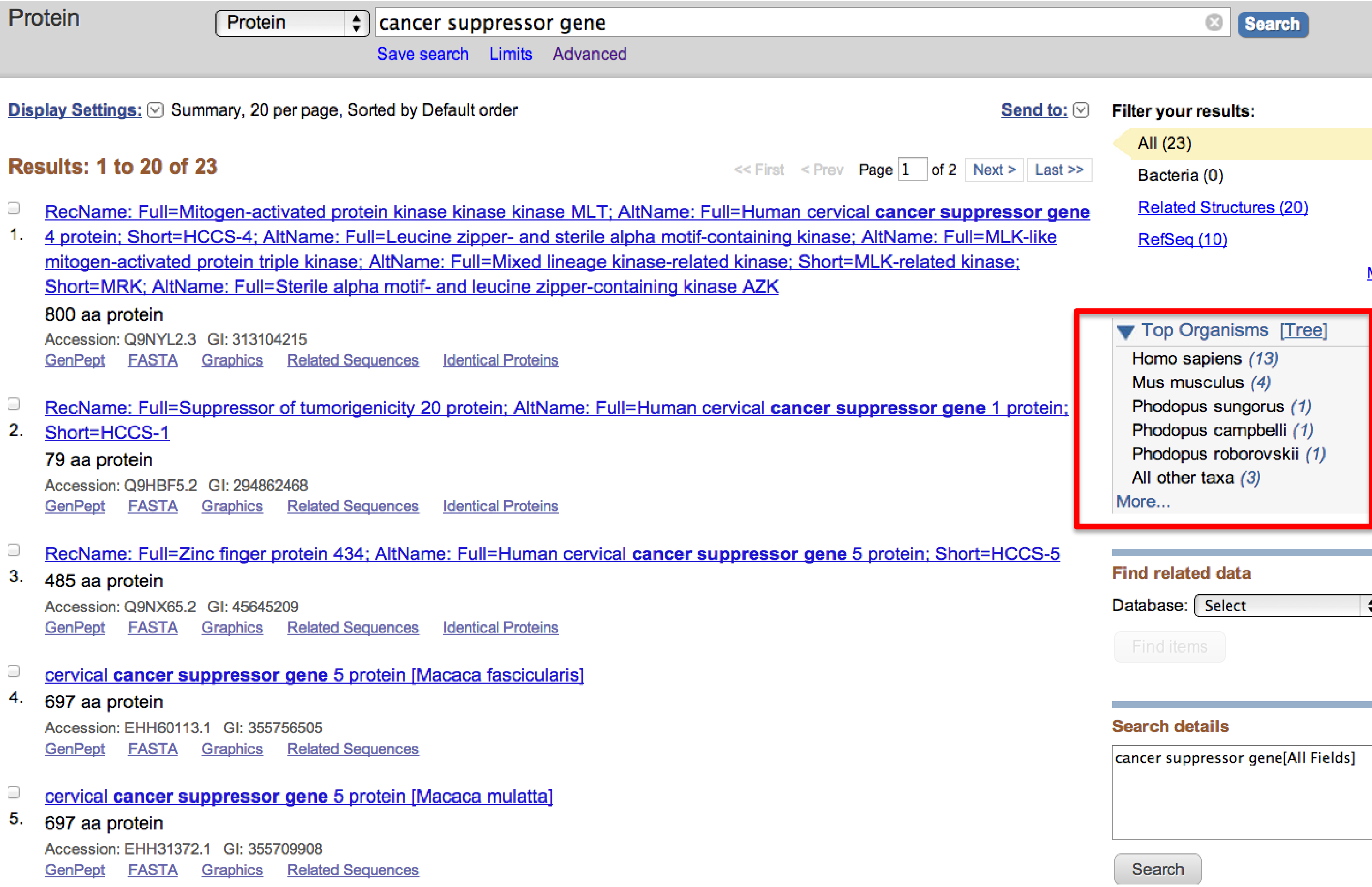


cancer suppressor gene

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

Count	Database
66197	**PubMed:** biomedical literature citations and abstracts
74463	**PubMed Central:** free, full text journal articles
none	**Site Search:** NCBI web and FTP sites
1006	**Books:** online books
597	**OMIM:** online Mendelian Inheritance in Man

Count	Database
687	**Nucleotide:** Core subset of nucleotide sequence records
2	**EST:** Expressed Sequence Tag records
none	**GSS:** Genome Survey Sequence records
23	**Protein:** sequence database
75	**Genome:** whole genome sequences
303	**Structure:** three-dimensional macromolecular structures
none	**Taxonomy:** organisms in GenBank
none	**SNP:** short genetic variations
554	**dbVar:** Genomic structural variation
11	**Gene:** gene-centered information
none	**SRA:** Sequence Read Archive
179	**BioSystems:** Pathways and systems of interacting molecules
none	**HomoloGene:** eukaryotic homology groups
75	**dbGaP:** genotype and phenotype
7	**UniGene:** gene-oriented clusters of transcript sequences
none	**CDD:** conserved protein domain database
none	**Clone:** integrated data for clone resources
none	**UniSTS:** markers and mapping data
none	**PopSet:** population study data sets
none	**GEO Profiles:** expression and molecular abundance profiles
681	**GEO DataSets:** experimental sets of GEO data
4	**Epigenomics:** Epigenetic maps and data sets
19	**PubChem BioAssay:** bioactivity screens of chemical substances
none	**PubChem Compound:** unique small molecule chemical structures
none	**PubChem Substance:** deposited chemical substance records
none	**Protein Clusters:** a collection of related protein sequences



Protein

Protein ⬍ cancer suppressor gene ⊗ **Search**

Save search Limits Advanced

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

Send to:

Results: 1 to 20 of 23

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

☐ 1. RecName: Full=Mitogen-activated protein kinase kinase kinase MLT; 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=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 FASTA Graphics Related Sequences Identical Proteins

☐ 2. RecName: Full=Suppressor of tumorigenicity 20 protein; AltName: Full=Human cervical **cancer suppressor gene** 1 protein; Short=HCCS-1

79 aa protein

Accession: Q9HBF5.2 GI: 294862468

GenPept FASTA Graphics Related Sequences Identical Proteins

☐ 3. RecName: Full=Zinc finger protein 434; AltName: Full=Human cervical **cancer suppressor gene** 5 protein; Short=HCCS-5

485 aa protein

Accession: Q9NX65.2 GI: 45645209

GenPept FASTA Graphics Related Sequences Identical Proteins

☐ 4. cervical **cancer suppressor gene** 5 protein [Macaca fascicularis]

697 aa protein

Accession: EHH60113.1 GI: 355756505

GenPept FASTA Graphics Related Sequences

☐ 5. cervical **cancer suppressor gene** 5 protein [Macaca mulatta]

697 aa protein

Accession: EHH31372.1 GI: 355709908

GenPept FASTA Graphics Related Sequences

Filter your results:

All (23)
Bacteria (0)
Related Structures (20)
RefSeq (10)

▼ Top Organisms [Tree]

Homo sapiens *(13)*
Mus musculus *(4)*
Phodopus sungorus *(1)*
Phodopus campbelli *(1)*
Phodopus roborovskii *(1)*
All other taxa *(3)*
More...

Find related data

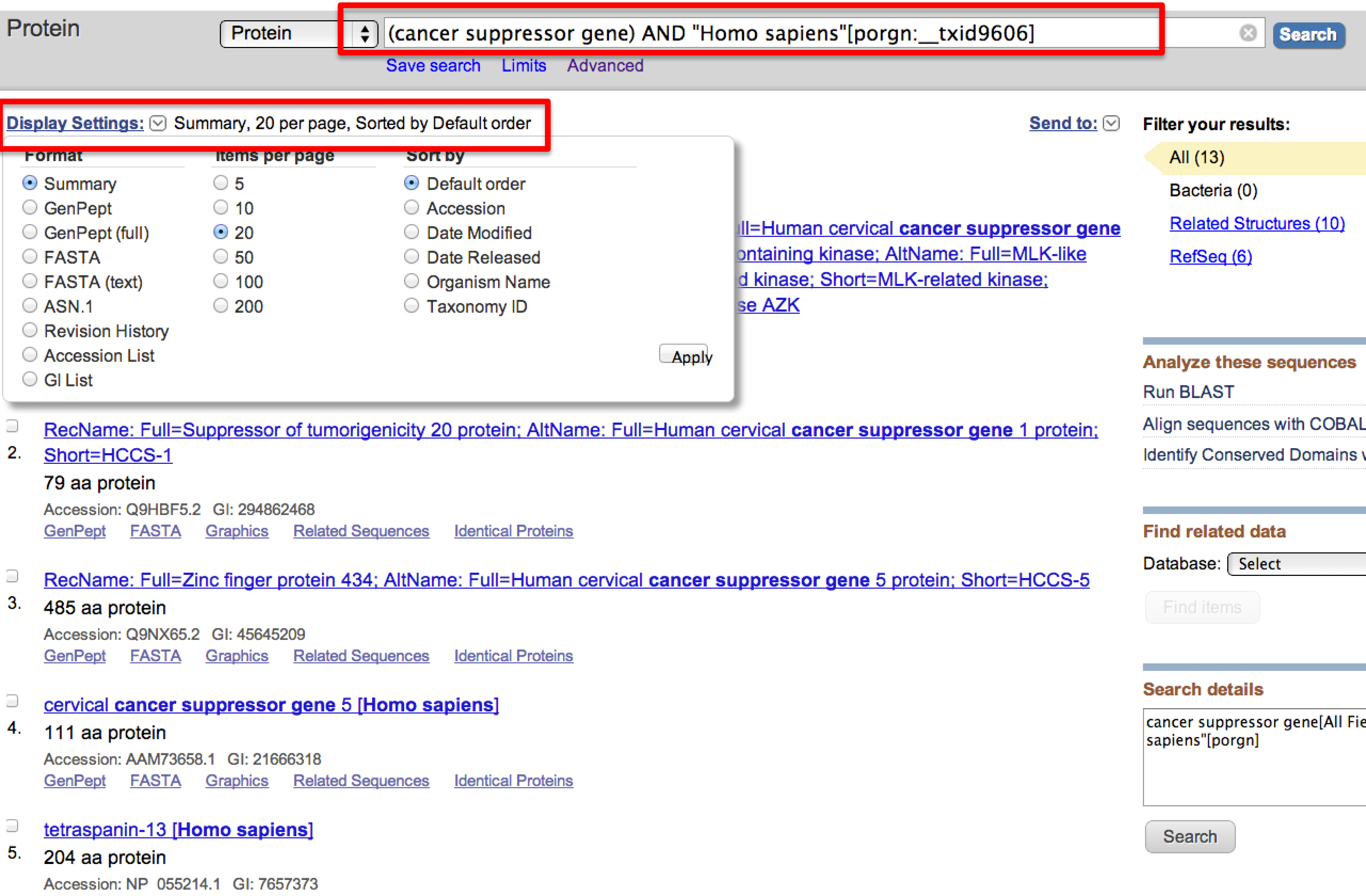
Database: Select

Find items

Search details

cancer suppressor gene[All Fields]

Search



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

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

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

Items per page
- ○ 5
- ○ 10
- ● 20
- ○ 50
- ○ 100
- ○ 200

Sort by
- ● Default order
- ○ Accession
- ○ Date Modified
- ○ Date Released
- ○ Organism Name
- ○ Taxonomy ID

ll=Human cervical **cancer suppressor gene**
ontaining kinase; AltName: Full=MLK-like
d kinase; Short=MLK-related kinase;
se AZK

☐ 2. RecName: Full=Suppressor of tumorigenicity 20 protein; AltName: Full=Human cervical **cancer suppressor gene** 1 protein; Short=HCCS-1
79 aa protein
Accession: Q9HBF5.2 GI: 294862468
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ 3. RecName: Full=Zinc finger protein 434; AltName: Full=Human cervical **cancer suppressor gene** 5 protein; Short=HCCS-5
485 aa protein
Accession: Q9NX65.2 GI: 45645209
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ 4. cervical **cancer suppressor gene** 5 [**Homo sapiens**]
111 aa protein
Accession: AAM73658.1 GI: 21666318
GenPept FASTA Graphics Related Sequences Identical Proteins

☐ 5. tetraspanin-13 [**Homo sapiens**]
204 aa protein
Accession: NP 055214.1 GI: 7657373

Filter your results:
- All (13)
- Bacteria (0)
- Related Structures (10)
- RefSeq (6)

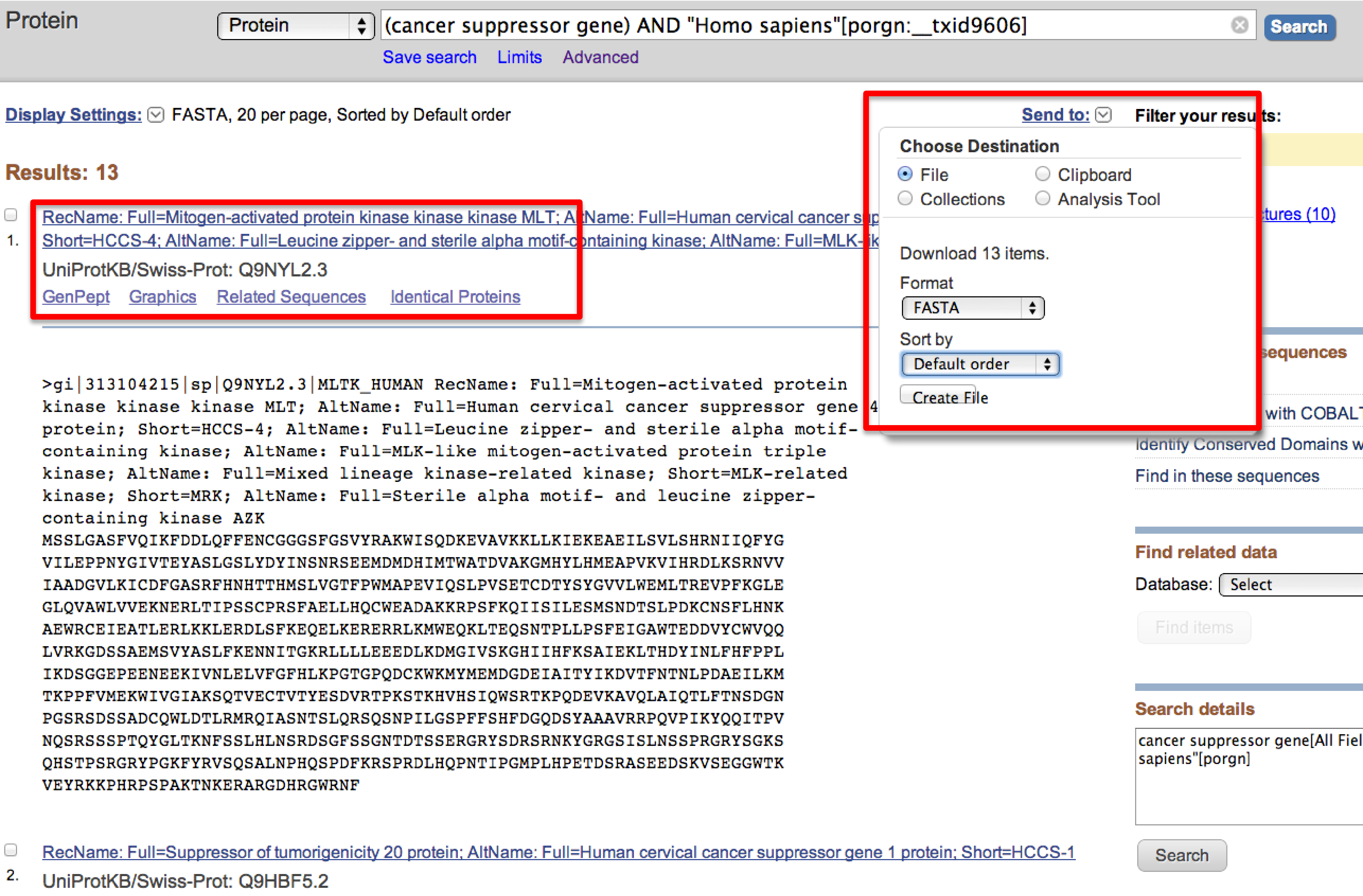
Analyze these sequences
- Run BLAST
- Align sequences with COBAL
- Identify Conserved Domains w

Find related data

Database: Select

Search details

cancer suppressor gene[All Fie
sapiens"[porgn]



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

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

Results: 13

☐ 1. RecName: Full=Mitogen-activated protein kinase kinase kinase MLT; A tName: Full=Human cervical cancer s p Short=HCCS-4; AltName: Full=Leucine zipper- and sterile alpha motif- ntaining kinase; AltName: Full=MLK- ik

UniProtKB/Swiss-Prot: Q9NYL2.3

```
>gi|313104215|sp|Q9NYL2.3|MLTK_HUMAN RecName: Full=Mitogen-activated protein
kinase kinase kinase MLT; 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=Mixed lineage kinase-related kinase; Short=MLK-related
kinase; Short=MRK; AltName: Full=Sterile alpha motif- and leucine zipper-
containing kinase AZK
MSSLGASFVQIKFDDLQFFENCGGGSFGSVYRAKWISQDKEVAVKKLLKIEKEAEILSVLSHRNIIQFYG
VILEPPNYGIVTEYASLGSLYDYINSNRSEEMDMDHIMTWATDVAKGMHYLHMEAPVKVIHRDLKSRNVV
IAADGVLKICDFGASRFHNHTTHMSLVGTFPWMAPEVIQSLPVSETCDTYSYGVVLWEMLTREVPFKGLE
GLQVAWLVVEKNERLTIPSSCPRSFAELLHQCWEADAKKRPSFKQIISILESMSNDTSLPDKCNSFLHNK
AEWRCEIEATLERLKKLERDLSFKEQELKERERRLKMWEQKLTEQSNTPLLPSFEIGAWTEDDVYCWVQQ
LVRKGDSSAEMSVYASLFKENNITGKRLLLLEEEDLKDMGIVSKGHIIHFKSAIEKLTHDYINLFHFPPL
IKDSGGEPEENEEKIVNLELVFGFHLKPGTGPQDCKWKMYMEMDGDEIAITYIKDVTFNTNLPDAEILKM
TKPPFVMEKWIVGIAKSQTVECTVTYESDVRTPKSTKHVHSIQWSRTKPQDEVKAVQLAIQTLFTNSDGN
PGSRSDSSADCQWLDTLRMRQIASNTSLQRSQSNPILGSPFFSHFDGQDSYAAAVRRPQVPIKYQQITPV
NQSRSSSPTQYGLTKNFSSLHLNSRDSGFSSGNTDTSSERGRYSDRSRNKYGRGSISLNSSPRGRYSGKS
QHSTPSRGRYPGKFYRVSQSALNPHQSPDFKRSPRDLHQPNTIPGMPLHPETDSRASEEDSKVSEGGWTK
VEYRKKPHRPSPAKTNKERARGDHRGWRNF
```

☐ 2. RecName: Full=Suppressor of tumorigenicity 20 protein; AltName: Full=Human cervical cancer suppressor gene 1 protein; Short=HCCS-1

UniProtKB/Swiss-Prot: Q9HBF5.2

Choose Destination

● File ○ Clipboard

○ Collections ○ Analysis Tool

Download 13 items.

Format: FASTA

Sort by: Default order

Create File

Find related data

Database: Select

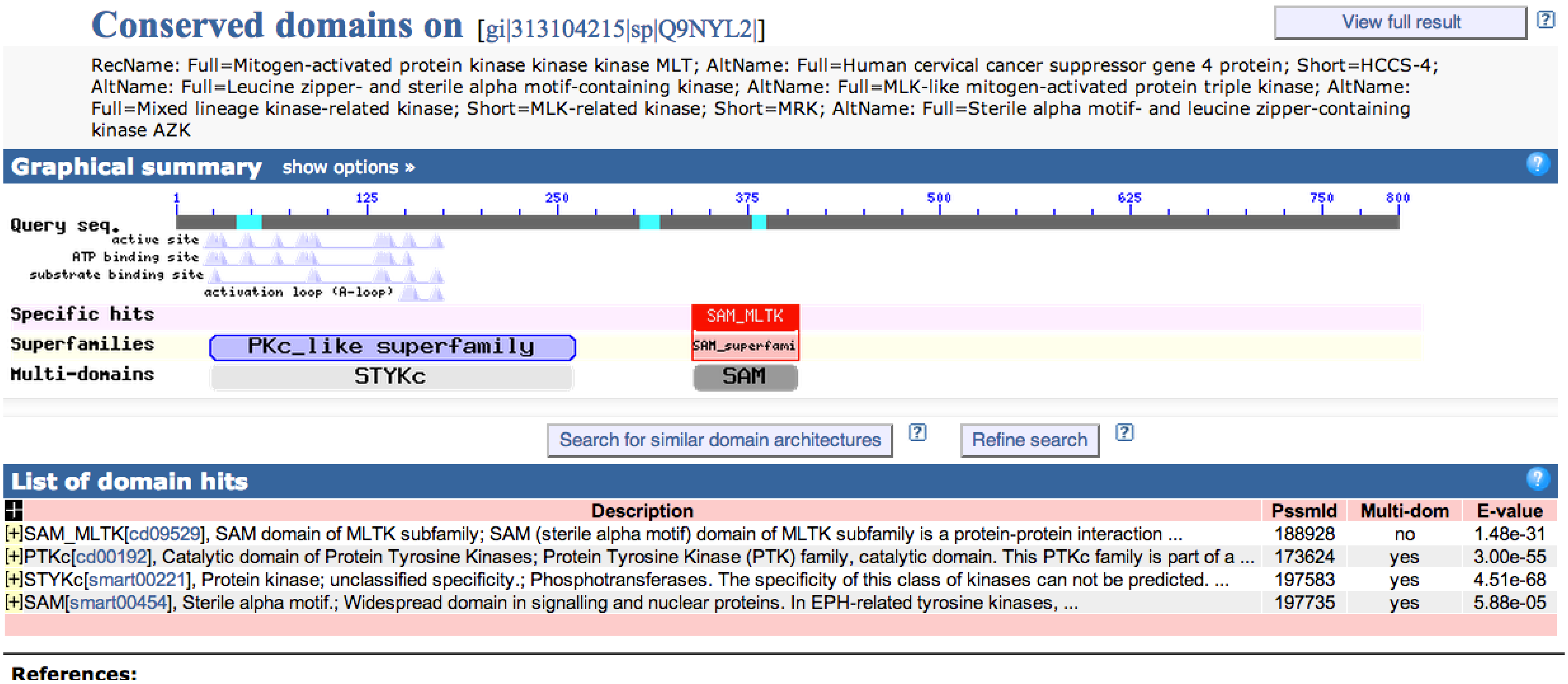
Search details

cancer suppressor gene[All Fiel
sapiens"[porgn]

RecName: Full=Mitogen-activated protein kinase kinase kinase MLT; 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 pro...

UniProtKB/Swiss-Prot: Q9NYL2.3

```
LOCUS       MLTK_HUMAN              800 aa            linear   PRI 28-NOV-2012
DEFINITION  RecName: Full=Mitogen-activated protein kinase kinase kinase MLT;
            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=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  GI:313104215
DBSOURCE    UniProtKB: locus MLTK_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: Nov 28, 2012.
            xrefs: AF238255.1, AAF63490.1, AB049733.1, BAB16444.1, AB049734.1,
            BAB16445.1, AF325454.1, AAK11615.1, AF480461.1, AAL85891.1,
            AF480462.1, AAL85892.1, AB030034.1, BAB12040.1, AF251441.1,
            AAF65822.1, AF465843.1, AA033376.1, AK056310.1, BAG51674.1,
            AB208974.1, BAD92211.1, AC092573.2, AAX82002.1, AC013461.9,
                                               EAX11164.1, BC001401.2,
```

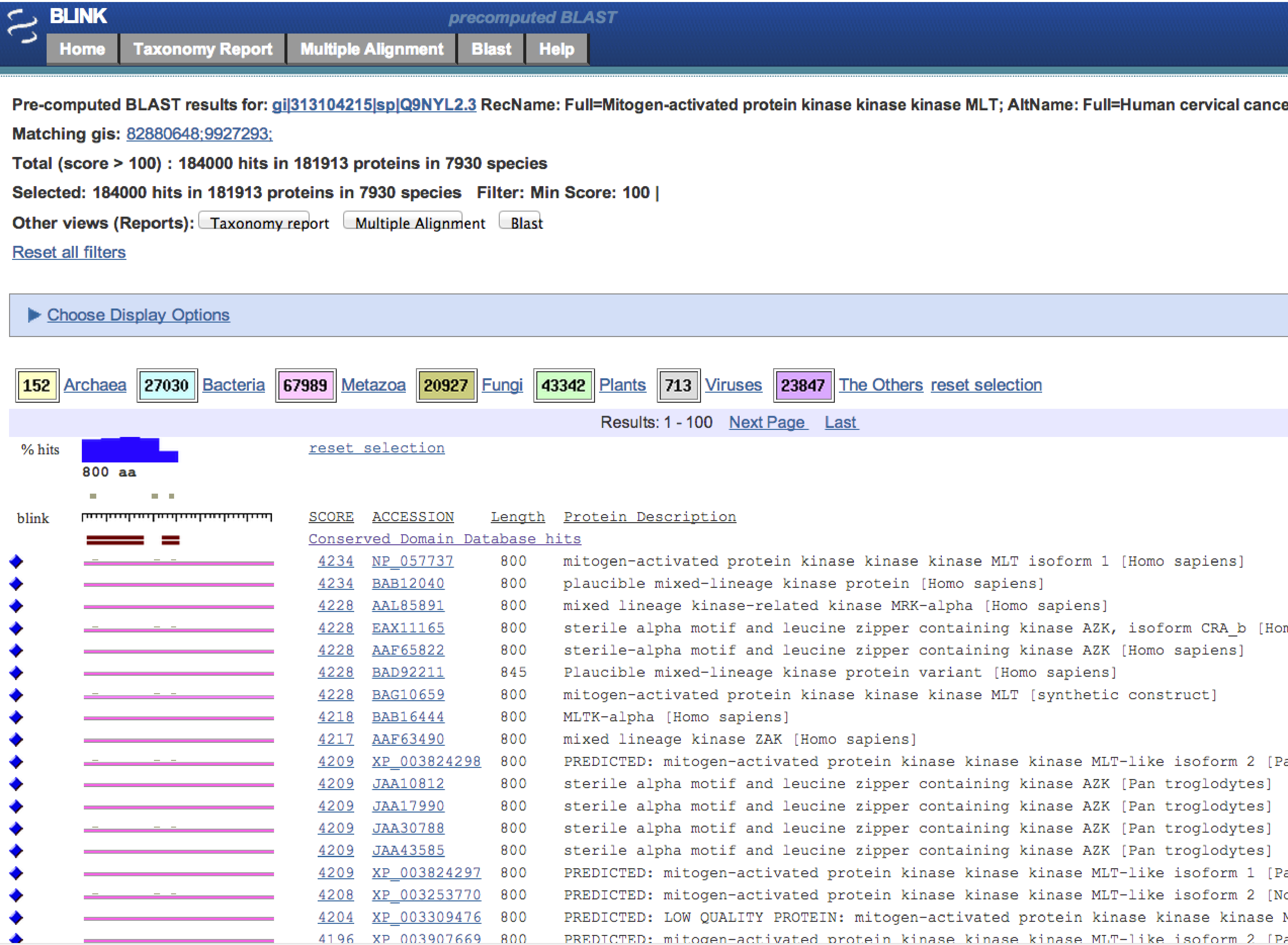


Conserved domains on [gi|313104215|sp|Q9NYL2|]

RecName: Full=Mitogen-activated protein kinase kinase kinase MLT; 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=Mixed lineage kinase-related kinase; Short=MLK-related kinase; Short=MRK; AltName: Full=Sterile alpha motif- and leucine zipper-containing kinase AZK

Description	PssmId	Multi-dom	E-value
[+]SAM_MLTK[cd09529], SAM domain of MLTK subfamily; SAM (sterile alpha motif) domain of MLTK subfamily is a protein-protein interaction ...	188928	no	1.48e-31
[+]PTKc[cd00192], Catalytic domain of Protein Tyrosine Kinases; Protein Tyrosine Kinase (PTK) family, catalytic domain. This PTKc family is part of a ...	173624	yes	3.00e-55
[+]STYKc[smart00221], Protein kinase; unclassified specificity.; Phosphotransferases. The specificity of this class of kinases can not be predicted. ...	197583	yes	4.51e-68
[+]SAM[smart00454], Sterile alpha motif.; Widespread domain in signalling and nuclear proteins. In EPH-related tyrosine kinases, ...	197735	yes	5.88e-05

References:



Pre-computed BLAST results for: gi|313104215|sp|Q9NYL2.3 **RecName: Full=Mitogen-activated protein kinase kinase kinase MLT; AltName: Full=Human cervical cance**

Matching gis: 82880648;9927293;

Total (score > 100) : 184000 hits in 181913 proteins in 7930 species

Selected: 184000 hits in 181913 proteins in 7930 species Filter: Min Score: 100 |

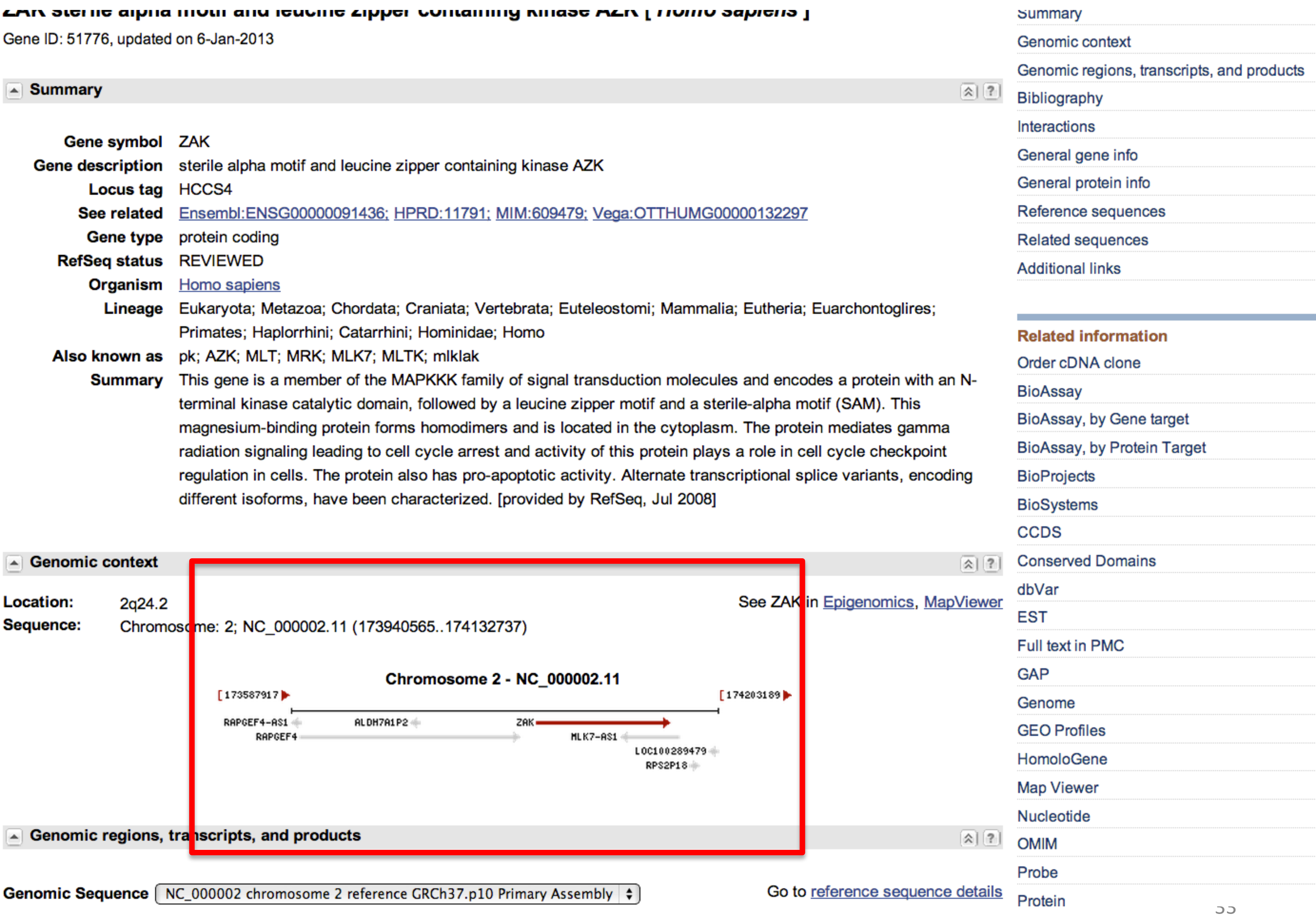
Other views (Reports): Taxonomy report Multiple Alignment Blast

Reset all filters

152 Archaea 27030 Bacteria 67989 Metazoa 20927 Fungi 43342 Plants 713 Viruses 23847 The Others reset selection

Results: 1 - 100 Next Page Last

SCORE	ACCESSION	Length	Protein Description
4234	NP_057737	800	mitogen-activated protein kinase kinase kinase MLT isoform 1 [Homo sapiens]
4234	BAB12040	800	plaucible mixed-lineage kinase protein [Homo sapiens]
4228	AAL85891	800	mixed lineage kinase-related kinase MRK-alpha [Homo sapiens]
4228	EAX11165	800	sterile alpha motif and leucine zipper containing kinase AZK, isoform CRA_b [Hom
4228	AAF65822	800	sterile-alpha motif and leucine zipper containing kinase AZK [Homo sapiens]
4228	BAD92211	845	Plaucible mixed-lineage kinase protein variant [Homo sapiens]
4228	BAG10659	800	mitogen-activated protein kinase kinase kinase MLT [synthetic construct]
4218	BAB16444	800	MLTK-alpha [Homo sapiens]
4217	AAF63490	800	mixed lineage kinase ZAK [Homo sapiens]
4209	XP_003824298	800	PREDICTED: mitogen-activated protein kinase kinase kinase MLT-like isoform 2 [Pa
4209	JAA10812	800	sterile alpha motif and leucine zipper containing kinase AZK [Pan troglodytes]
4209	JAA17990	800	sterile alpha motif and leucine zipper containing kinase AZK [Pan troglodytes]
4209	JAA30788	800	sterile alpha motif and leucine zipper containing kinase AZK [Pan troglodytes]
4209	JAA43585	800	sterile alpha motif and leucine zipper containing kinase AZK [Pan troglodytes]
4209	XP_003824297	800	PREDICTED: mitogen-activated protein kinase kinase kinase MLT-like isoform 1 [Pa
4208	XP_003253770	800	PREDICTED: mitogen-activated protein kinase kinase kinase MLT-like isoform 2 [No
4204	XP_003309476	800	PREDICTED: LOW QUALITY PROTEIN: mitogen-activated protein kinase kinase kinase
4196	XP_003907669	800	PREDICTED: mitogen-activated protein kinase kinase kinase MLT-like isoform 2 [Pa



ZAK sterile alpha motif and leucine zipper containing kinase AZK [*Homo sapiens*]

Gene ID: 51776, updated on 6-Jan-2013

Summary

Gene symbol	ZAK
Gene description	sterile alpha motif and leucine zipper containing kinase AZK
Locus tag	HCCS4
See related	Ensembl:ENSG00000091436; HPRD:11791; 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; MLK7; MLTK; mlklak
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]

Genomic context

Location: 2q24.2

Sequence: Chromosome: 2; NC_000002.11 (173940565..174132737)

See ZAK in Epigenomics, MapViewer

Genomic regions, transcripts, and products

Genomic Sequence NC_000002 chromosome 2 reference GRCh37.p10 Primary Assembly

Go to reference sequence details

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Interactions
- General gene info
- General protein info
- Reference sequences
- Related sequences
- Additional links

Related information

- Order cDNA clone
- BioAssay
- BioAssay, by Gene target
- BioAssay, by Protein Target
- BioProjects
- BioSystems
- CCDS
- Conserved Domains
- dbVar
- EST
- Full text in PMC
- GAP
- Genome
- GEO Profiles
- HomoloGene
- Map Viewer
- Nucleotide
- OMIM
- Probe
- Protein

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)

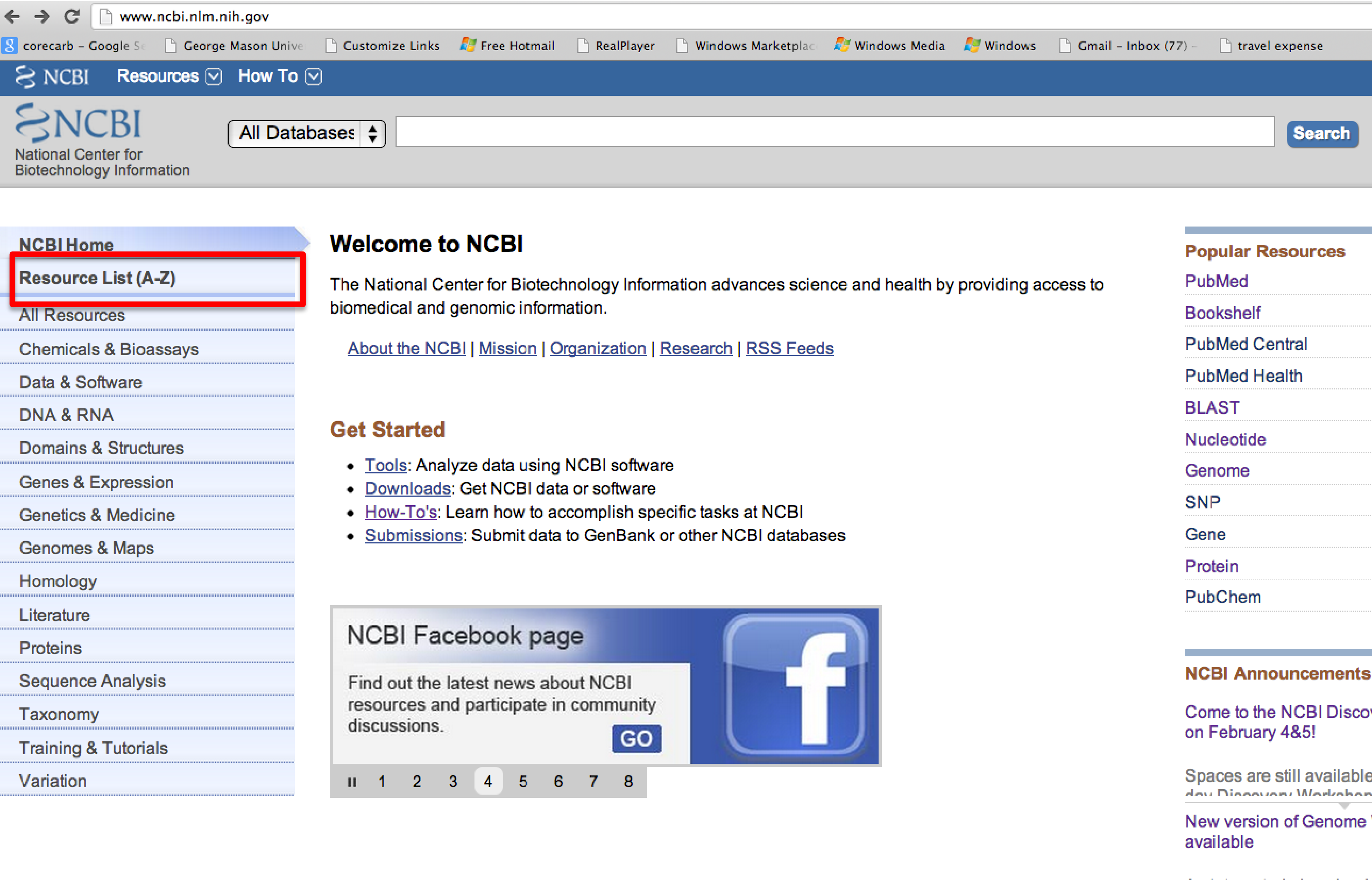
Or

- 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!



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.

Popular Resources

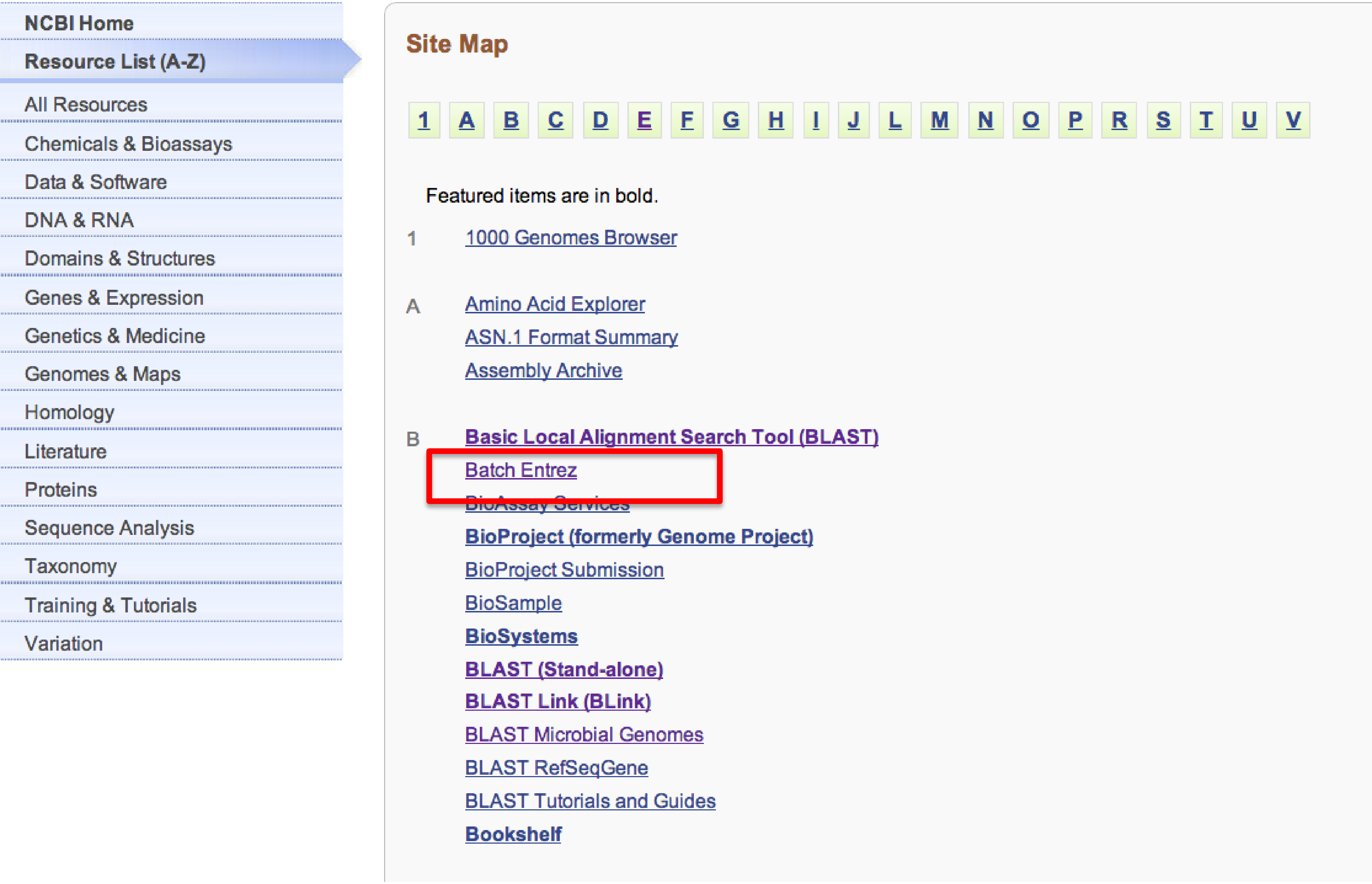
- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

Come to the NCBI Disco
on February 4&5!

Spaces are still available

New version of Genome
available



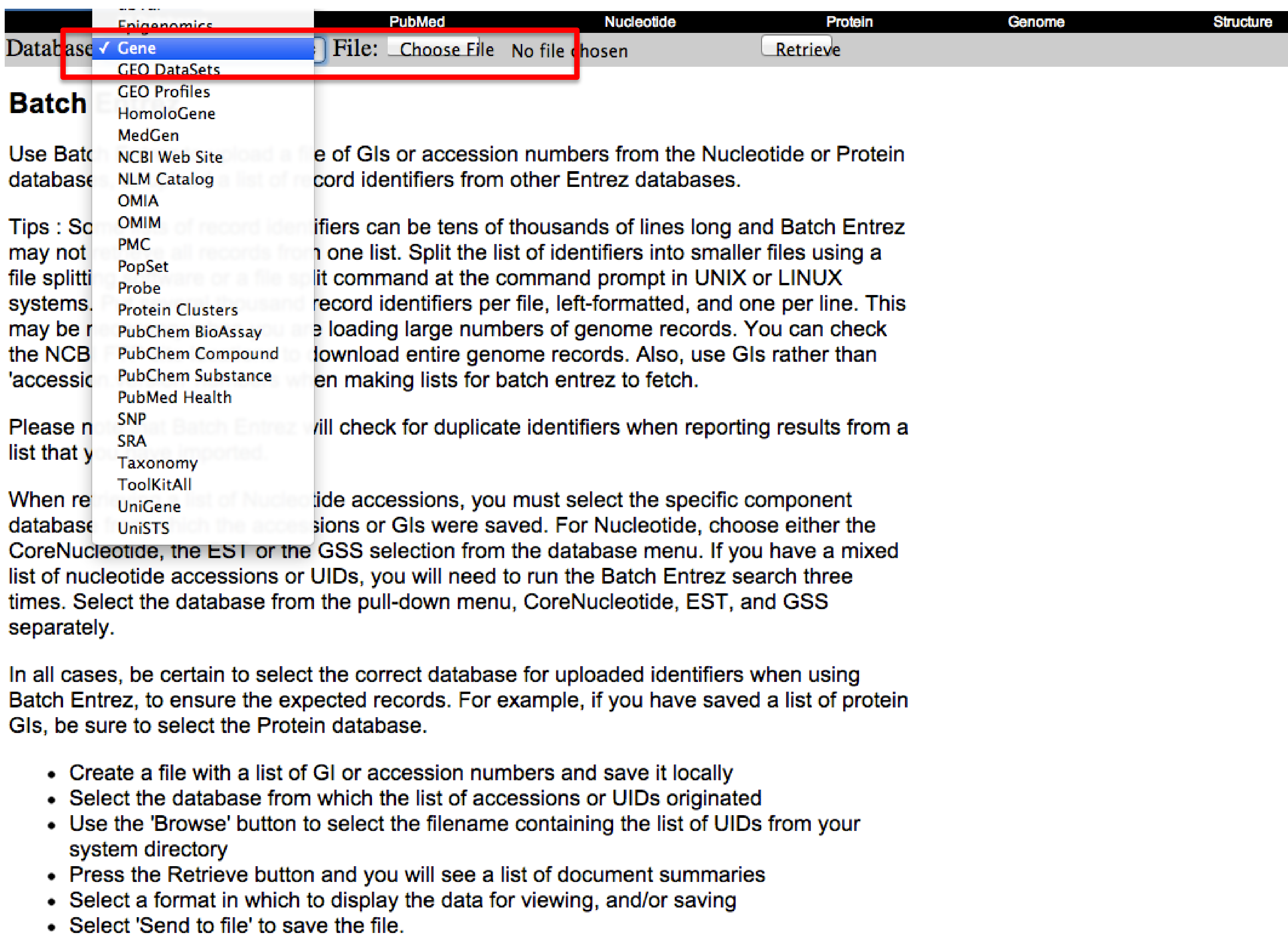
Site Map

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 a list of record identifiers from other Entrez databases.

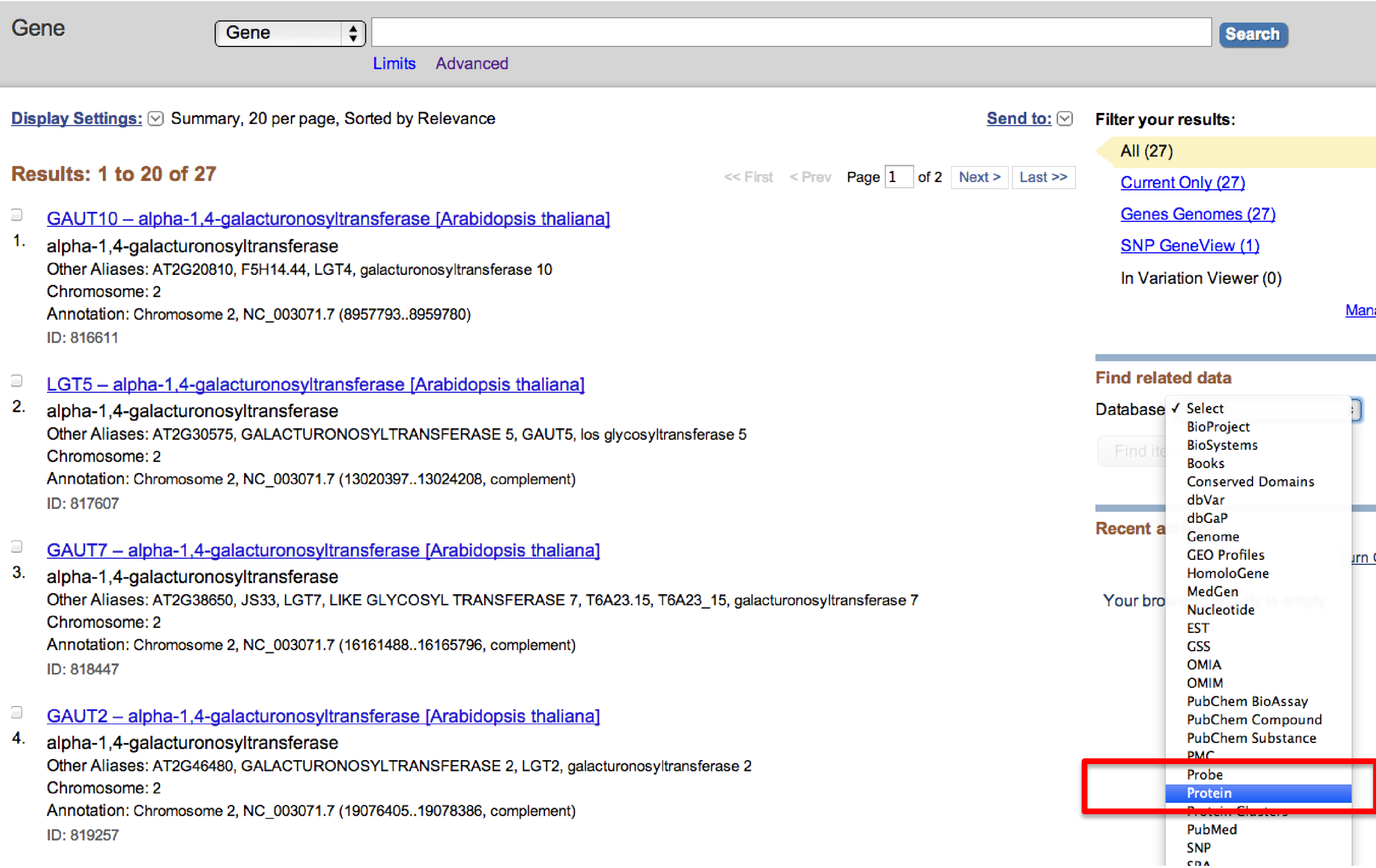
Tips : Some lists of record identifiers can be tens of thousands of lines long and Batch Entrez may not retrieve all records from one list. Split the list of identifiers into smaller files using a file splitting software or a file split command at the command prompt in UNIX or LINUX systems. Put several 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' numbers 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.



Display Settings: Summary, 20 per page, Sorted by Relevance

Results: 1 to 20 of 27

Page 1 of 2

Filter your results:

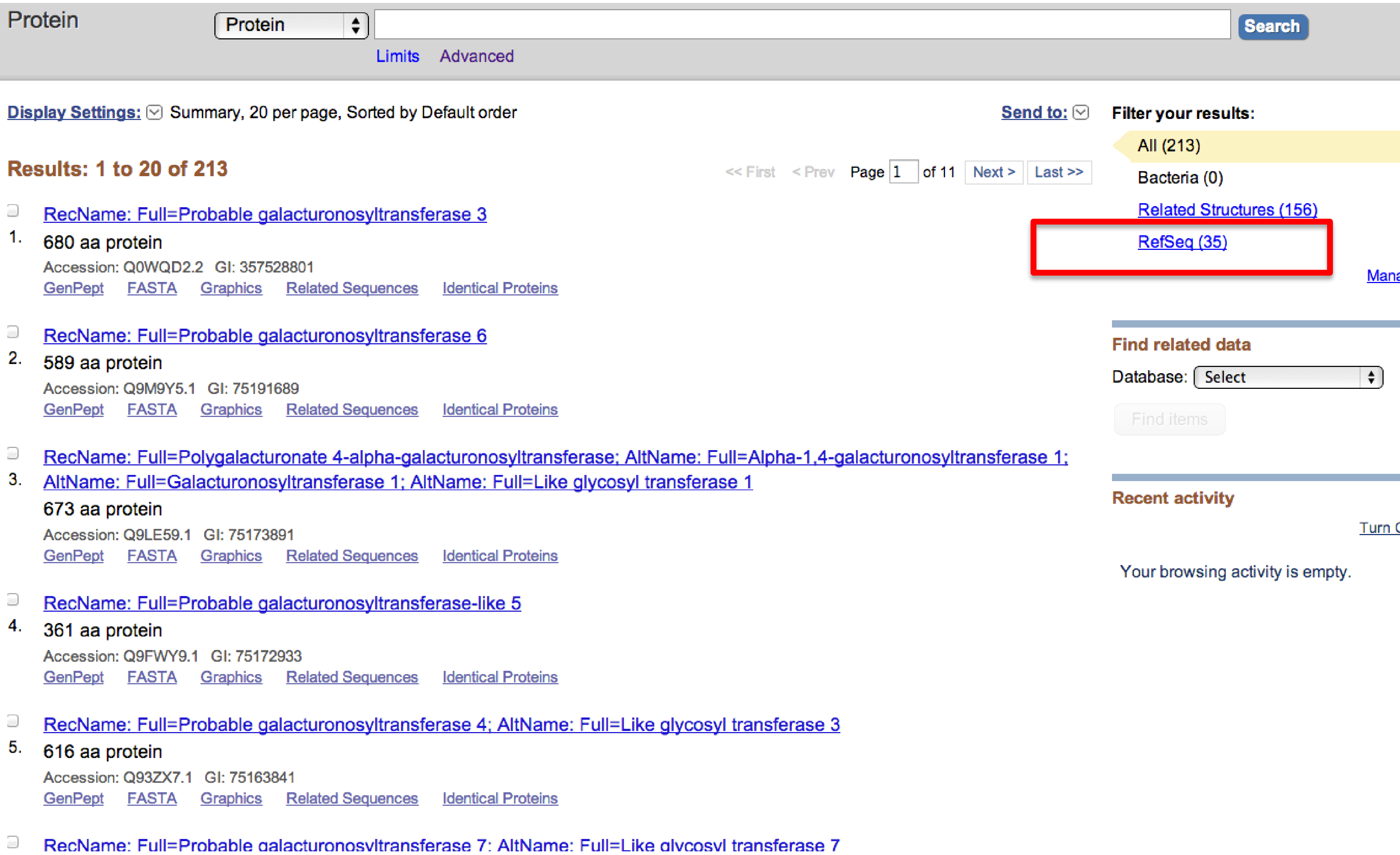
All (27)
Current Only (27)
Genes Genomes (27)
SNP GeneView (1)
In Variation Viewer (0)

1. GAUT10 – alpha-1,4-galacturonosyltransferase [Arabidopsis thaliana]
alpha-1,4-galacturonosyltransferase
Other Aliases: AT2G20810, F5H14.44, LGT4, galacturonosyltransferase 10
Chromosome: 2
Annotation: Chromosome 2, NC_003071.7 (8957793..8959780)
ID: 816611

2. LGT5 – alpha-1,4-galacturonosyltransferase [Arabidopsis thaliana]
alpha-1,4-galacturonosyltransferase
Other Aliases: AT2G30575, GALACTURONOSYLTRANSFERASE 5, GAUT5, los glycosyltransferase 5
Chromosome: 2
Annotation: Chromosome 2, NC_003071.7 (13020397..13024208, complement)
ID: 817607

3. GAUT7 – alpha-1,4-galacturonosyltransferase [Arabidopsis thaliana]
alpha-1,4-galacturonosyltransferase
Other Aliases: AT2G38650, JS33, LGT7, LIKE GLYCOSYL TRANSFERASE 7, T6A23.15, T6A23_15, galacturonosyltransferase 7
Chromosome: 2
Annotation: Chromosome 2, NC_003071.7 (16161488..16165796, complement)
ID: 818447

4. GAUT2 – alpha-1,4-galacturonosyltransferase [Arabidopsis thaliana]
alpha-1,4-galacturonosyltransferase
Other Aliases: AT2G46480, GALACTURONOSYLTRANSFERASE 2, LGT2, galacturonosyltransferase 2
Chromosome: 2
Annotation: Chromosome 2, NC_003071.7 (19076405..19078386, complement)
ID: 819257



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

Results: 1 to 20 of 213

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

Filter your results:

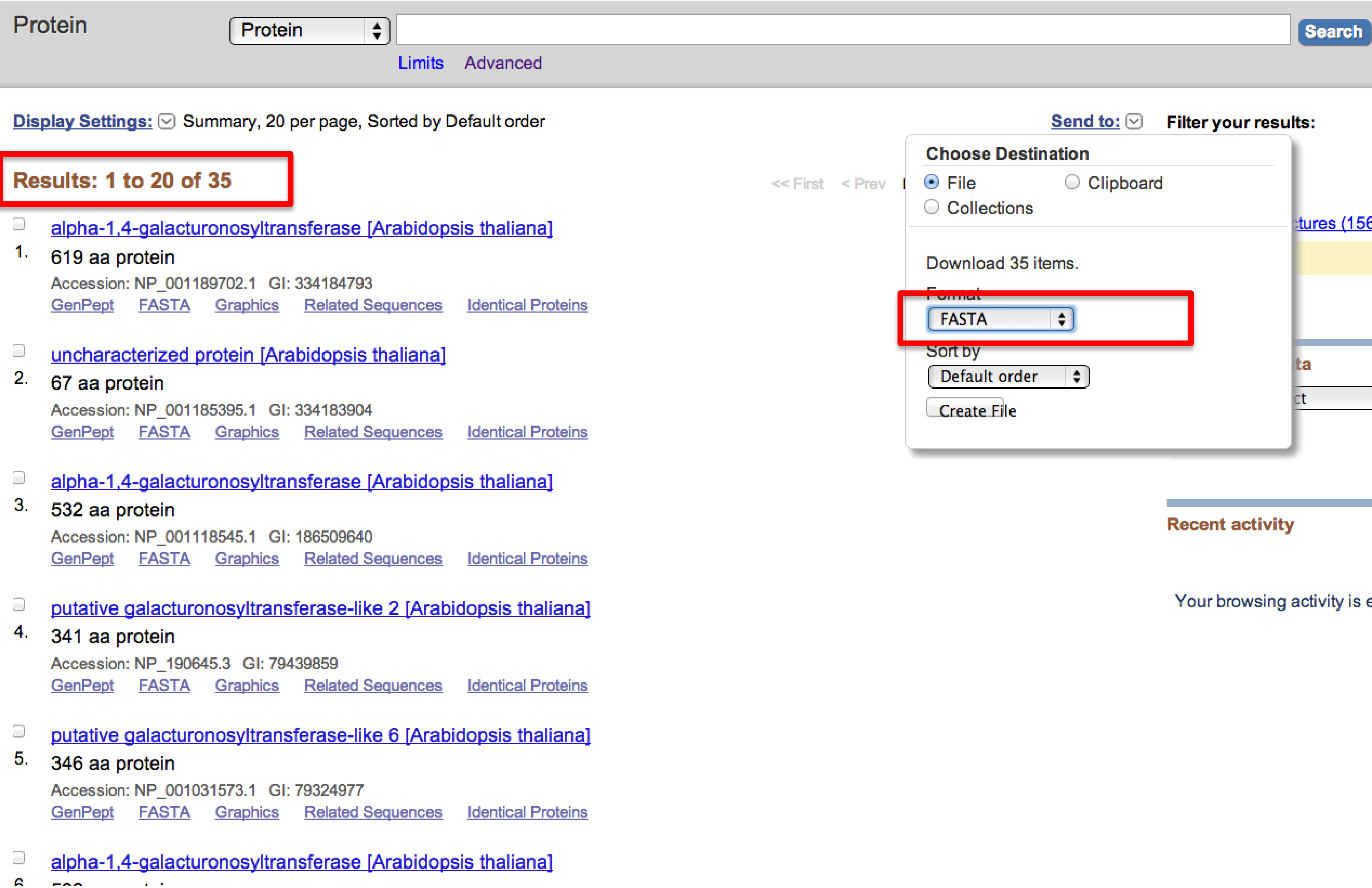
- All (213)
- Bacteria (0)
- Related Structures (156)
- RefSeq (35)

Find related data

Database: Select

Recent activity

Your browsing activity is empty.



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

Send to:

Results: 1 to 20 of 35

Choose Destination
● File ○ Clipboard
○ Collections

Download 35 items.

Format: FASTA

Sort by: Default order

Create File

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

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

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

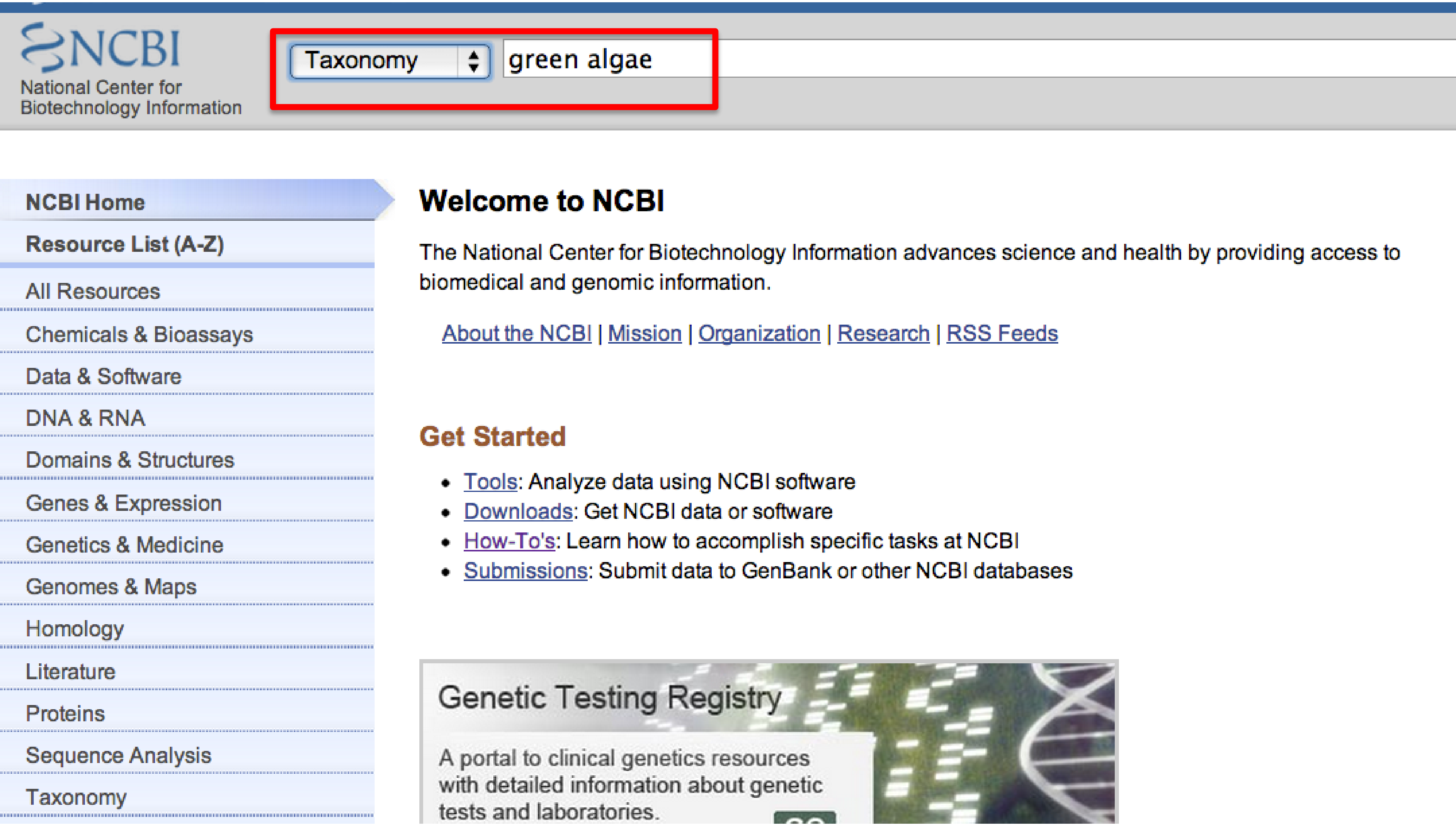
☐ 4. putative galacturonosyltransferase-like 2 [Arabidopsis thaliana]
341 aa protein
Accession: NP_190645.3 GI: 79439859
GenPept FASTA Graphics Related Sequences Identical Proteins

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

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

Hands-on exercise 3

Find sequences through taxonomy database



Welcome to NCBI

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

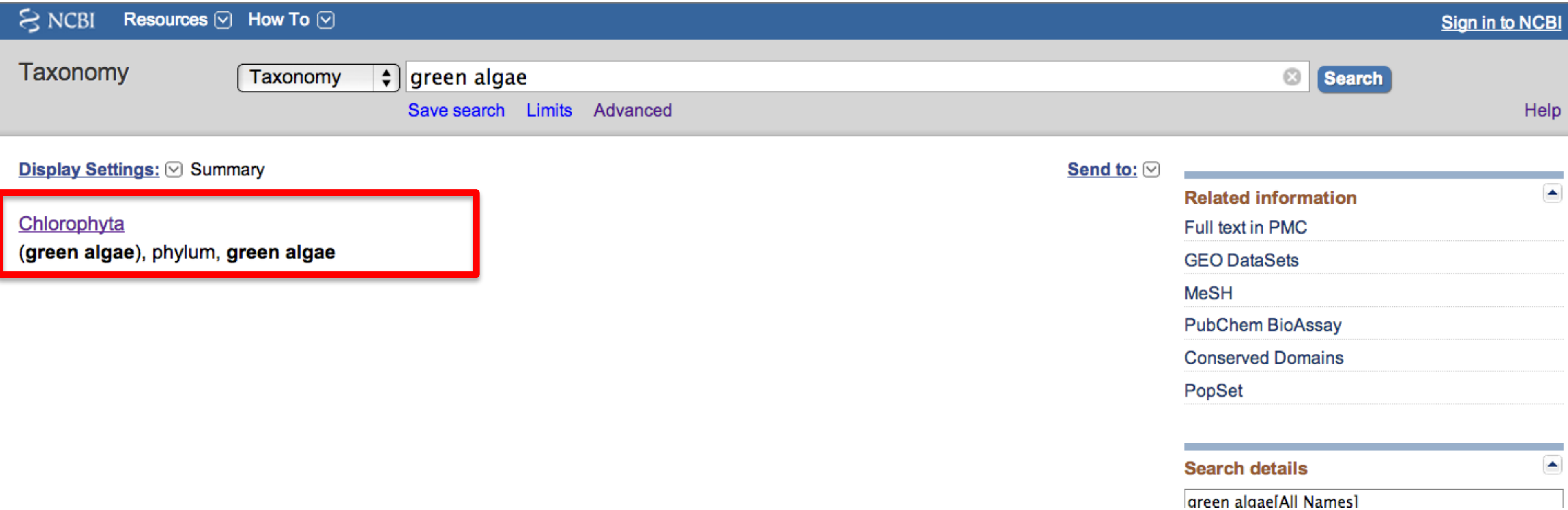
About the NCBI | Mission | Organization | Research | RSS Feeds

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

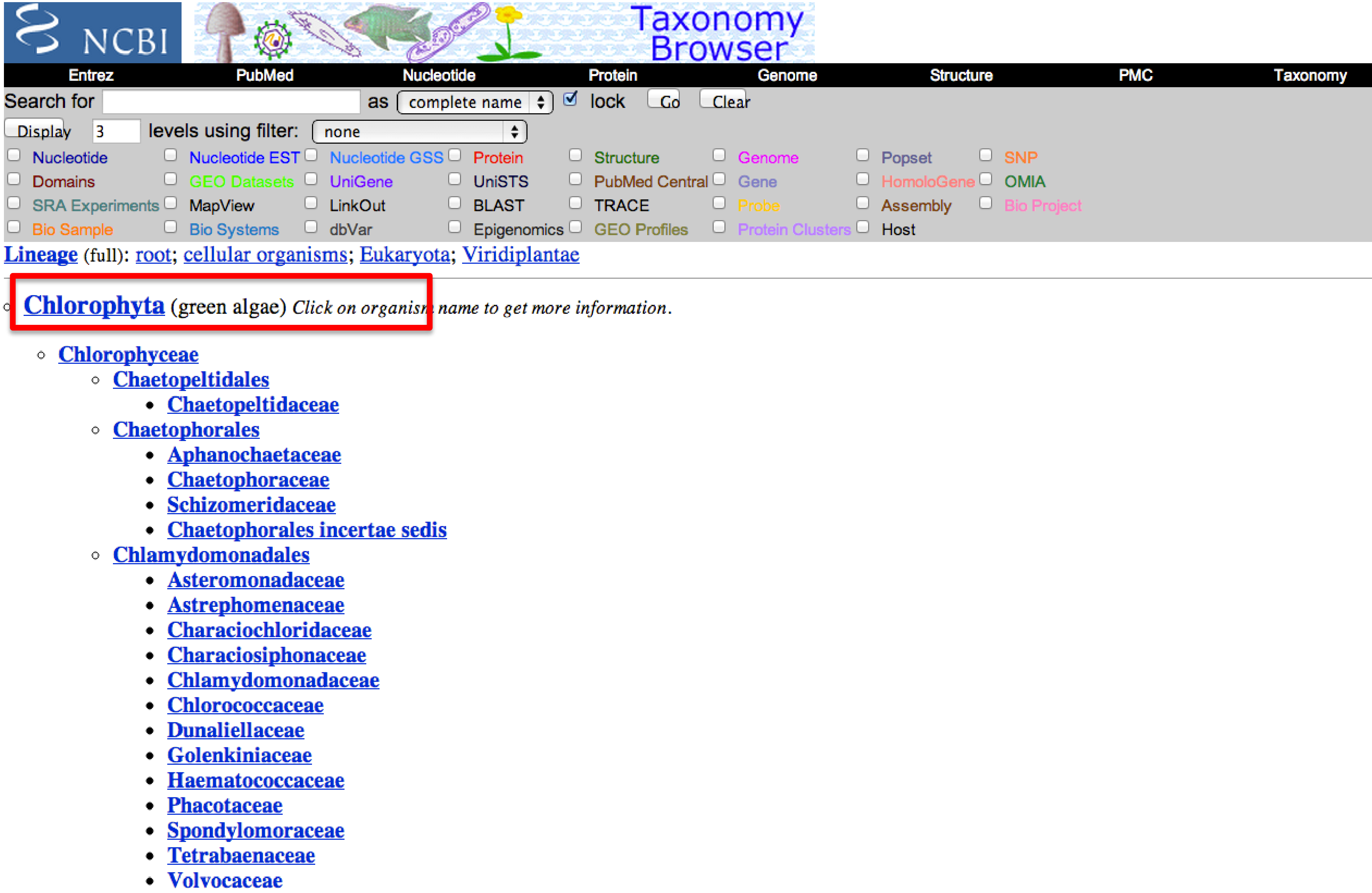
Display Settings: Summary

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

Related information
- Full text in PMC
- GEO DataSets
- MeSH
- PubChem BioAssay
- Conserved Domains
- PopSet

Search details

green algae[All Names]



Search for ☐ as complete name ☑ lock

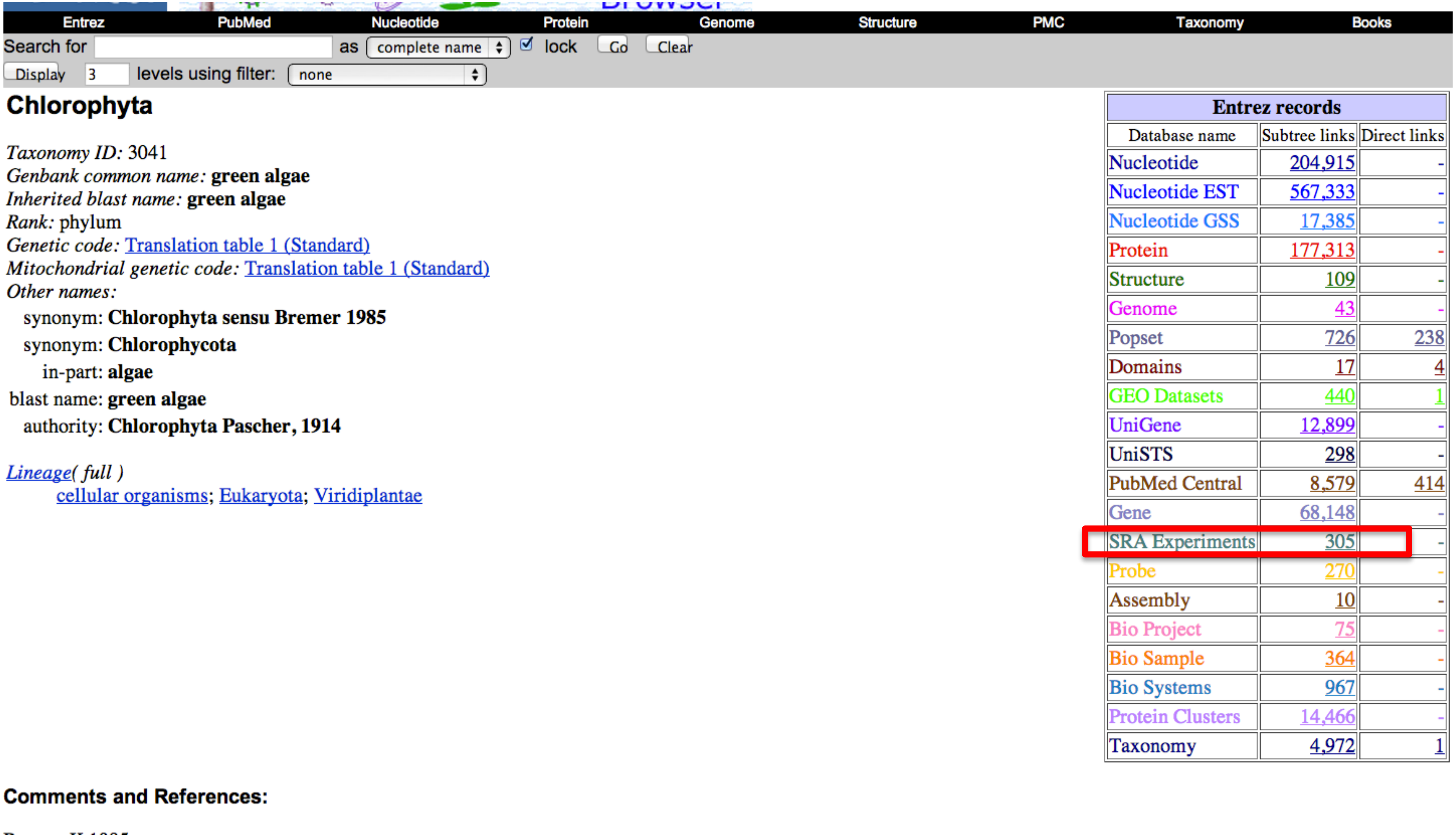
Display 3 levels using filter: none

☐ Nucleotide ☐ Nucleotide EST ☐ Nucleotide GSS ☐ Protein ☐ Structure ☐ Genome ☐ Popset ☐ SNP
☐ Domains ☐ GEO Datasets ☐ UniGene ☐ UniSTS ☐ PubMed Central ☐ Gene ☐ HomoloGene ☐ OMIA
☐ SRA Experiments ☐ MapView ☐ LinkOut ☐ BLAST ☐ TRACE ☐ Probe ☐ Assembly ☐ Bio Project
☐ Bio Sample ☐ Bio Systems ☐ dbVar ☐ Epigenomics ☐ GEO Profiles ☐ Protein Clusters ☐ Host

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

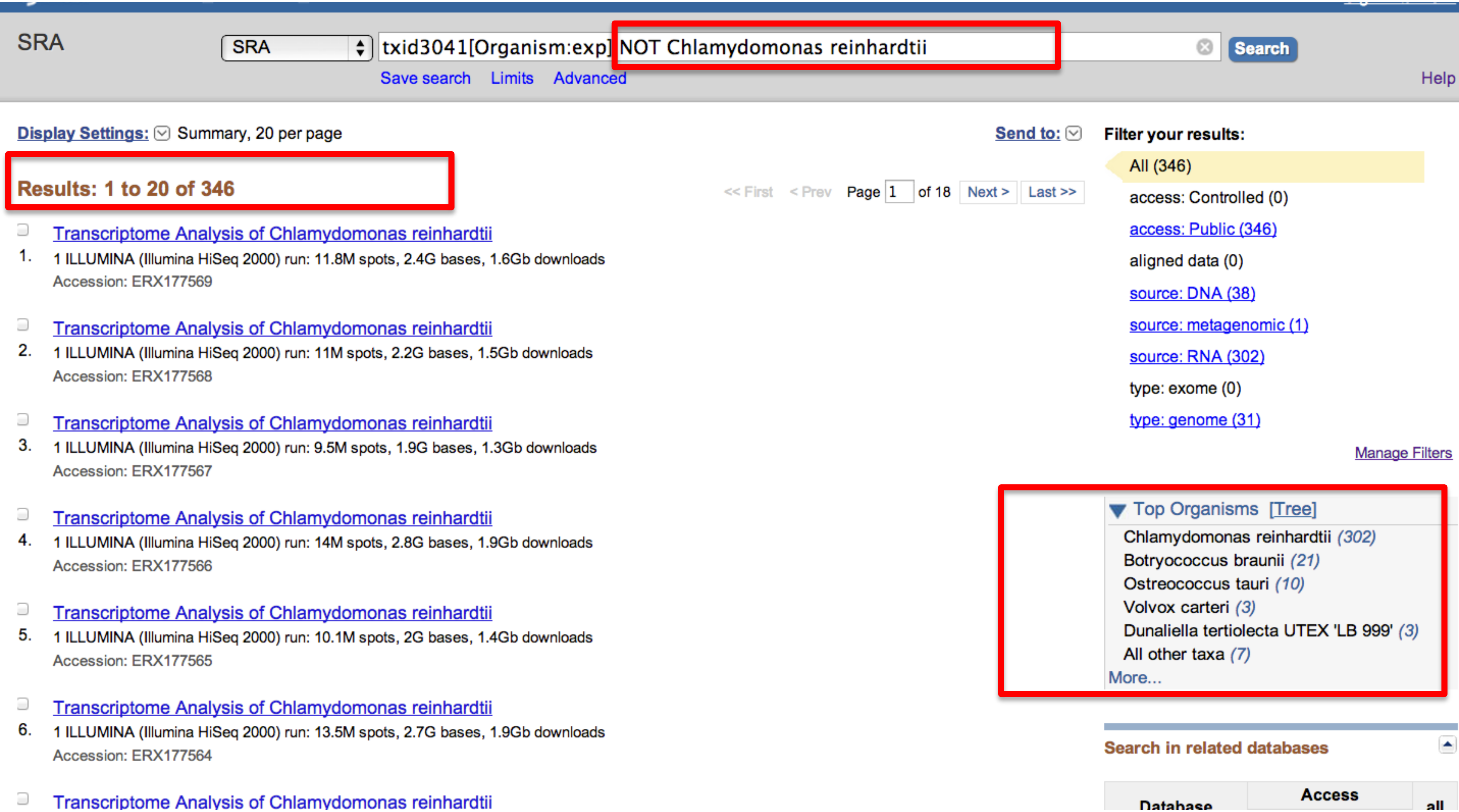
Chlorophyta

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	204,915	-
Nucleotide EST	567,333	-
Nucleotide GSS	17,385	-
Protein	177,313	-
Structure	109	-
Genome	43	-
Popset	726	238
Domains	17	4
GEO Datasets	440	1
UniGene	12,899	-
UniSTS	298	-
PubMed Central	8,579	414
Gene	68,148	-
SRA Experiments	305	-
Probe	270	-
Assembly	10	-
Bio Project	75	-
Bio Sample	364	-
Bio Systems	967	-
Protein Clusters	14,466	-
Taxonomy	4,972	1

Comments and References:



txid3041[Organism:exp] NOT Chlamydomonas reinhardtii

Display Settings: Summary, 20 per page

Results: 1 to 20 of 346

Page 1 of 18

1. ☐ Transcriptome Analysis of Chlamydomonas reinhardtii
1 ILLUMINA (Illumina HiSeq 2000) run: 11.8M spots, 2.4G bases, 1.6Gb downloads
Accession: ERX177569

2. ☐ Transcriptome Analysis of Chlamydomonas reinhardtii
1 ILLUMINA (Illumina HiSeq 2000) run: 11M spots, 2.2G bases, 1.5Gb downloads
Accession: ERX177568

3. ☐ Transcriptome Analysis of Chlamydomonas reinhardtii
1 ILLUMINA (Illumina HiSeq 2000) run: 9.5M spots, 1.9G bases, 1.3Gb downloads
Accession: ERX177567

4. ☐ Transcriptome Analysis of Chlamydomonas reinhardtii
1 ILLUMINA (Illumina HiSeq 2000) run: 14M spots, 2.8G bases, 1.9Gb downloads
Accession: ERX177566

5. ☐ Transcriptome Analysis of Chlamydomonas reinhardtii
1 ILLUMINA (Illumina HiSeq 2000) run: 10.1M spots, 2G bases, 1.4Gb downloads
Accession: ERX177565

6. ☐ Transcriptome Analysis of Chlamydomonas reinhardtii
1 ILLUMINA (Illumina HiSeq 2000) run: 13.5M spots, 2.7G bases, 1.9Gb downloads
Accession: ERX177564

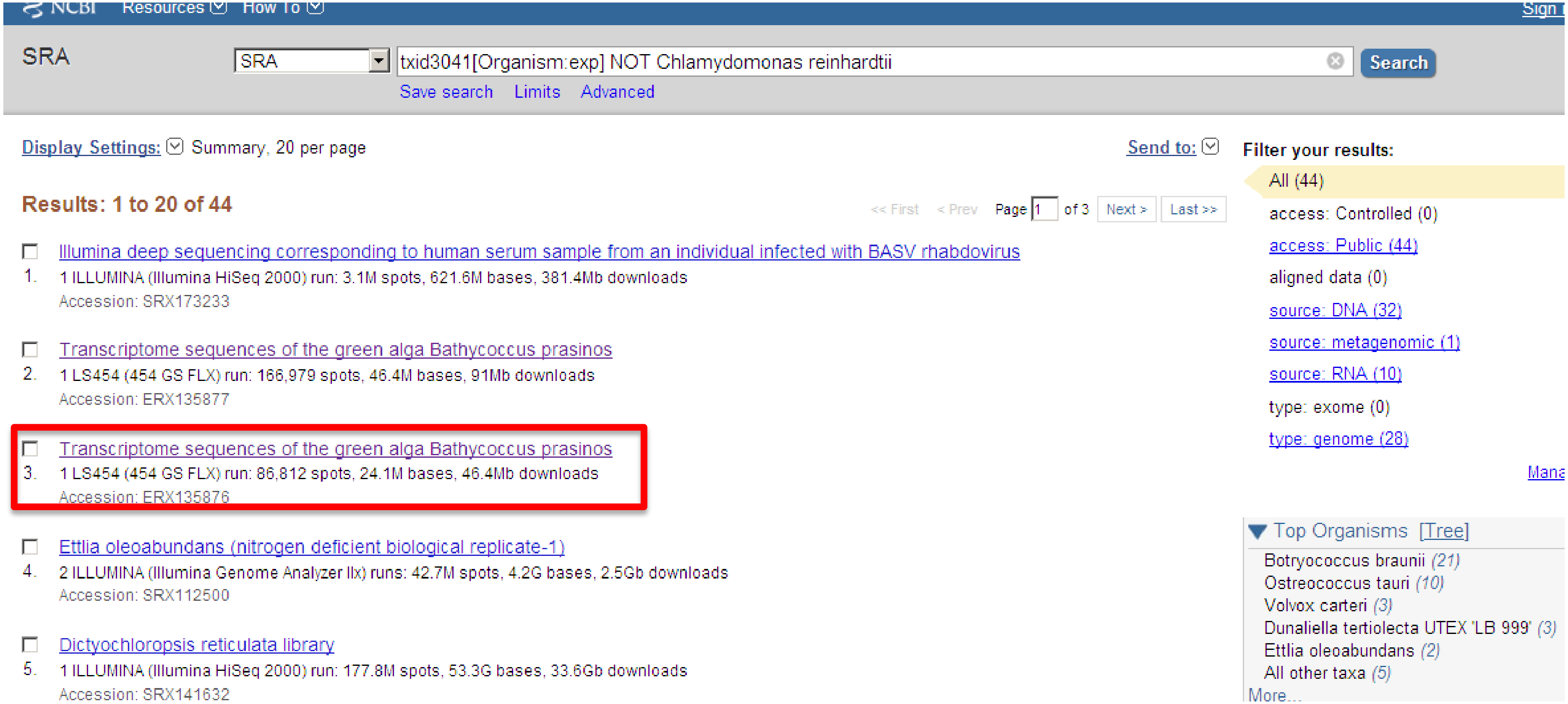
☐ Transcriptome Analysis of Chlamydomonas reinhardtii

Filter your results:

- All (346)
- access: Controlled (0)
- access: Public (346)
- aligned data (0)
- source: DNA (38)
- source: metagenomic (1)
- source: RNA (302)
- type: exome (0)
- type: genome (31)

Top Organisms [Tree]

- Chlamydomonas reinhardtii *(302)*
- Botryococcus braunii *(21)*
- Ostreococcus tauri *(10)*
- Volvox carteri *(3)*
- Dunaliella tertiolecta UTEX 'LB 999' *(3)*
- All other taxa *(7)*



SRA ▾ txid3041[Organism:exp] NOT Chlamydomonas reinhardtii

Display Settings: Summary, 20 per page

Results: 1 to 20 of 44

Page 1 of 3

☐ 1. Illumina deep sequencing corresponding to human serum sample from an individual infected with BASV rhabdovirus
1 ILLUMINA (Illumina HiSeq 2000) run: 3.1M spots, 621.6M bases, 381.4Mb downloads
Accession: SRX173233

☐ 2. Transcriptome sequences of the green alga Bathycoccus prasinos
1 LS454 (454 GS FLX) run: 166,979 spots, 46.4M bases, 91Mb downloads
Accession: ERX135877

☐ 3. Transcriptome sequences of the green alga Bathycoccus prasinos
1 LS454 (454 GS FLX) run: 86,812 spots, 24.1M bases, 46.4Mb downloads
Accession: ERX135876

☐ 4. Ettlia oleoabundans (nitrogen deficient biological replicate-1)
2 ILLUMINA (Illumina Genome Analyzer IIx) runs: 42.7M spots, 4.2G bases, 2.5Gb downloads
Accession: SRX112500

☐ 5. Dictyochloropsis reticulata library
1 ILLUMINA (Illumina HiSeq 2000) run: 177.8M spots, 53.3G bases, 33.6Gb downloads
Accession: SRX141632

Filter your results:

- All (44)
- access: Controlled (0)
- access: Public (44)
- aligned data (0)
- source: DNA (32)
- source: metagenomic (1)
- source: RNA (10)
- type: exome (0)
- type: genome (28)

Top Organisms

- Botryococcus braunii *(21)*
- Ostreococcus tauri *(10)*
- Volvox carteri *(3)*
- Dunaliella tertiolecta UTEX 'LB 999' *(3)*
- Ettlia oleoabundans *(2)*
- All other taxa *(5)*

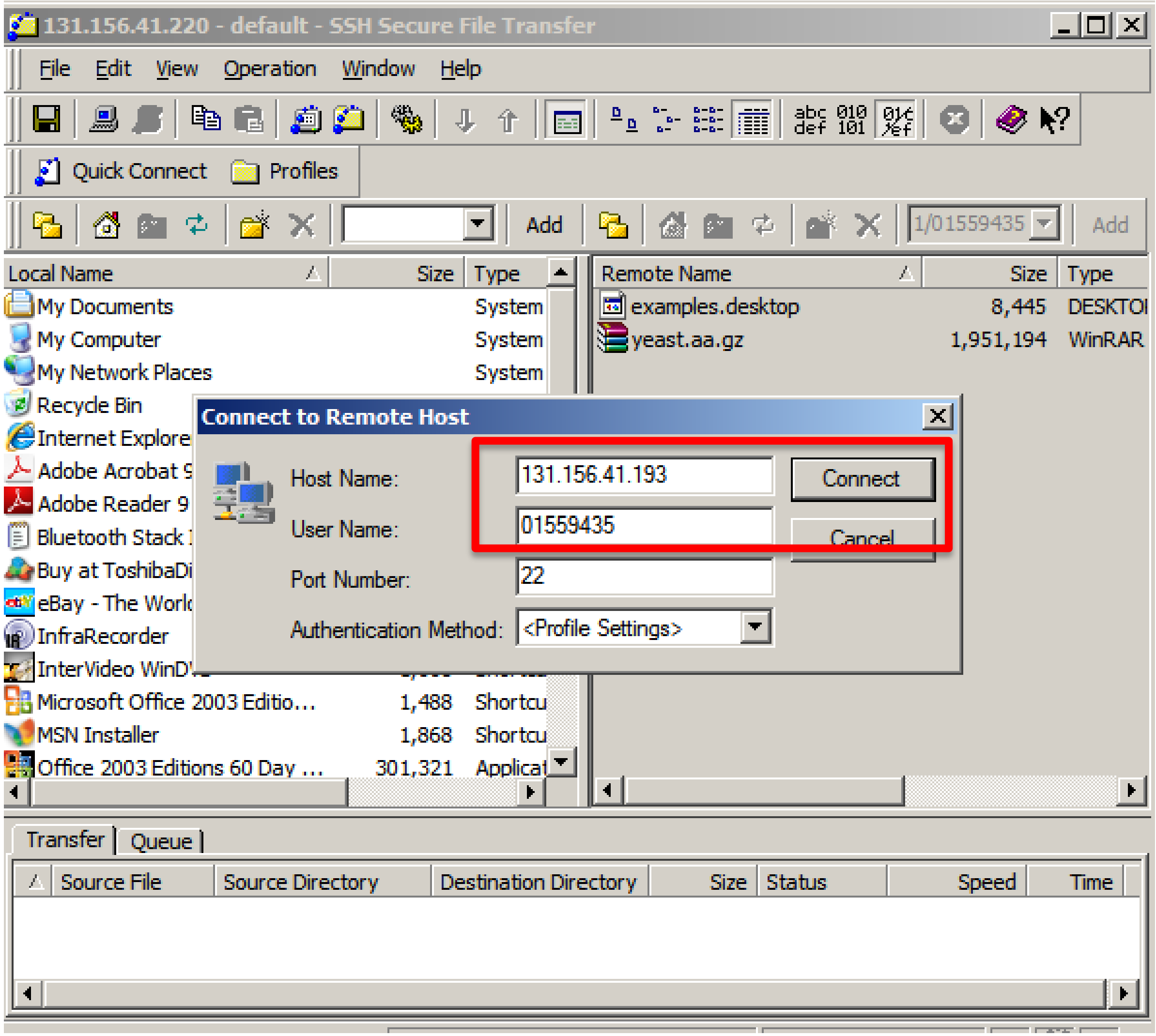
Transcriptome sequences of the green alga Bathycoccus prasinos

Accession: ERX135876
Experiment design: These expression data correspond to a culture in exponential phase of the green alga Bathycoccus prasinos
Submission: ERA148021 by GSC
Study summary: Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage (ERP001629) • Study • All experiments (more...)
Sample: Culture in exponential phase of the green alga Bathycoccus prasinos (ERS158202) (more...)
Library: YKBODS (more...)
Platform: LS454 (more...)
Spot descriptor:

Total: 1 run, 86,812 spots, 24.1M bases, 46.4Mb

#	Run	# of Spots	# of Bases	Size
1	ERR159934	86,812	24.1M	46.4Mb

Transfer files between Windows and remote Linux using SSH Secure File Transfer client

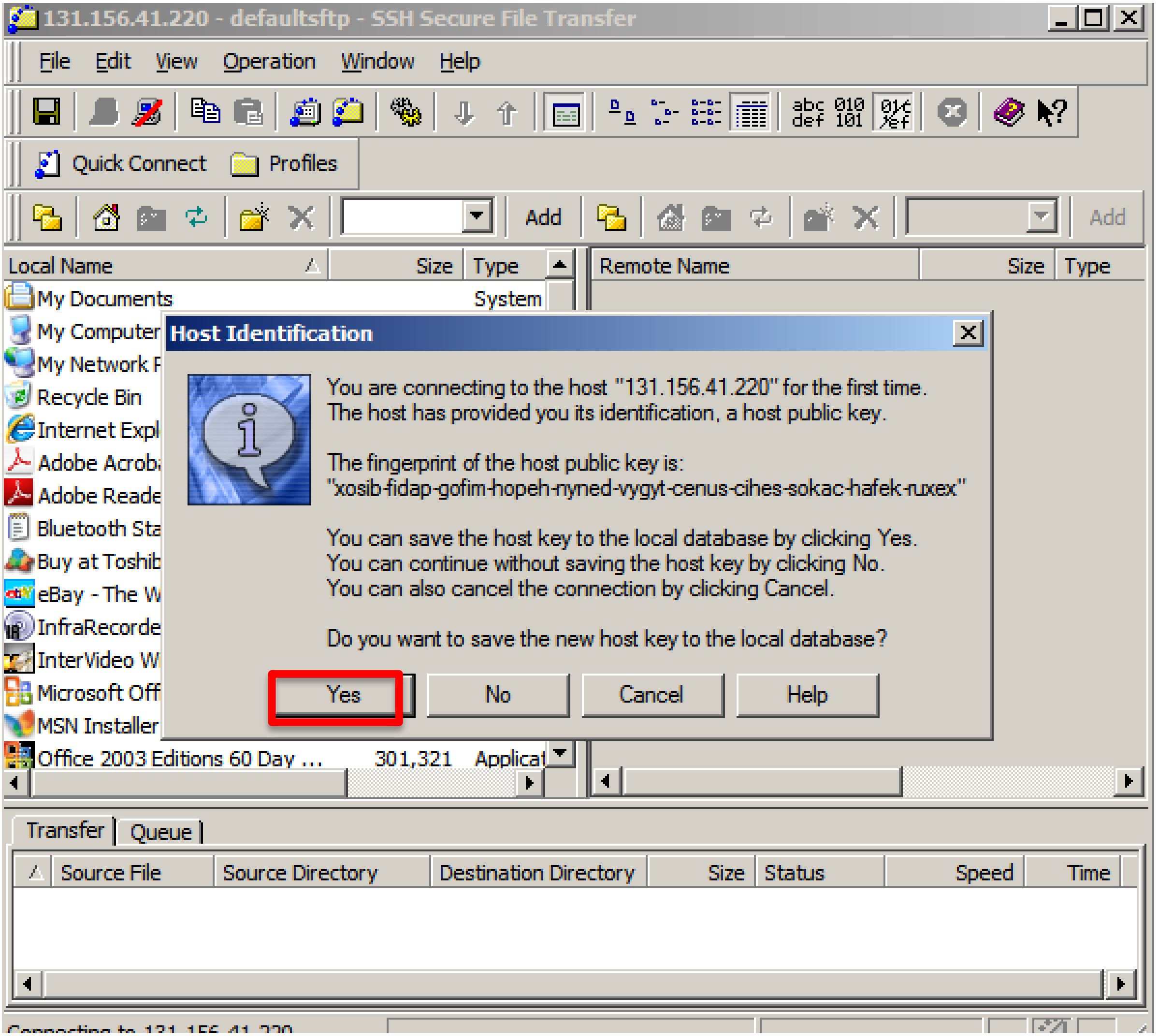


Host Name: 131.156.41.193

User Name: 01559435

Port Number: 22

Authentication Method: <Profile Settings>

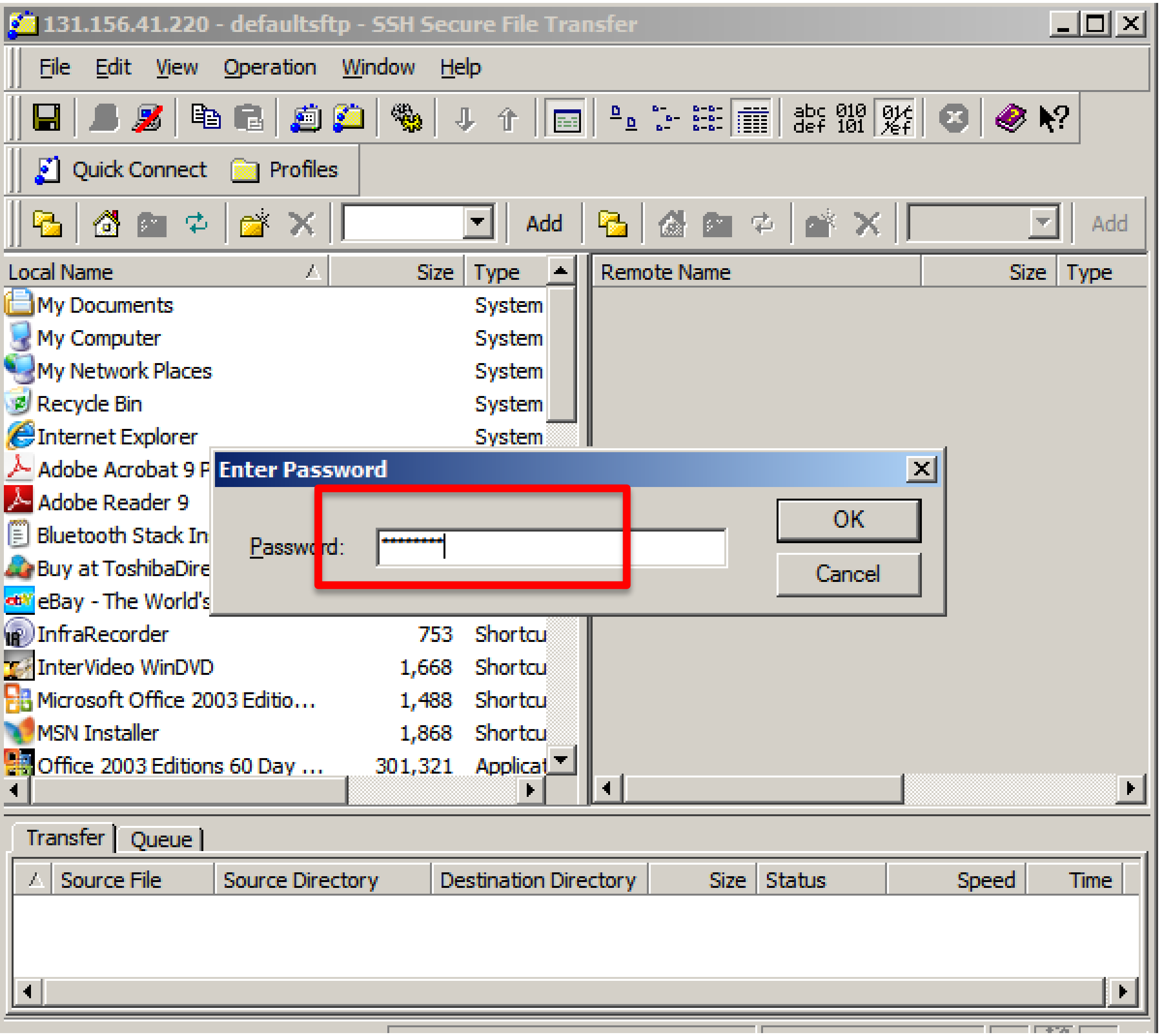


You are connecting to the host "131.156.41.220" for the first time.
The host has provided you its identification, a host public key.

The fingerprint of the host public key is:
"xosib-fidap-gofim-hopeh-nyned-vygyt-cenus-cihes-sokac-hafek-ruxex"

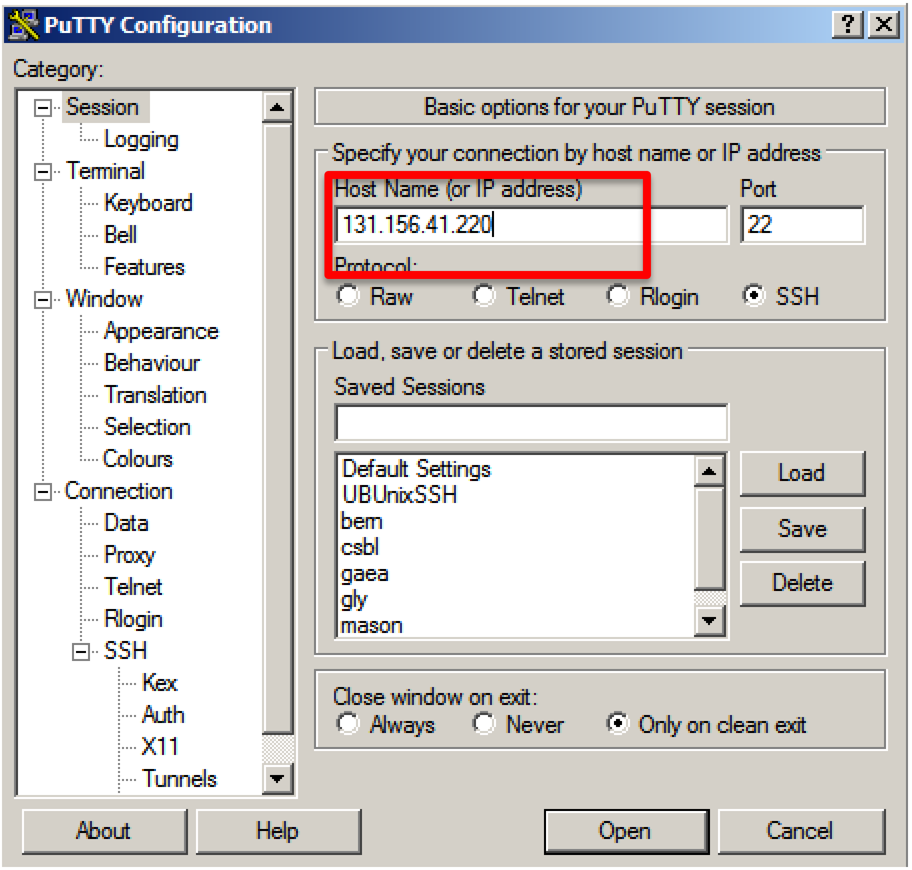
You can save the host key to the local database by clicking Yes.
You can continue without saving the host key by clicking No.
You can also cancel the connection by clicking Cancel.

Do you want to save the new host key to the local database?

131.156.41.220 - defaultsftp - SSH Secure File Transfer
File Edit View Operation Window Help
Quick Connect
Profiles
Add
Add
Local Name
Size
Type
Remote Name
Size
Type
My Documents System
My Computer System
My Network Places System
Recycle Bin System
Internet Explorer System
Adobe Acrobat 9 P
Adobe Reader 9
Bluetooth Stack In
Buy at ToshibaDire
eBay - The World's
InfraRecorder 753 Shortcu
InterVideo WinDVD 1,668 Shortcu
Microsoft Office 2003 Editio… 1,488 Shortcu
MSN Installer 1,868 Shortcu
Office 2003 Editions 60 Day … 301,321 Applicat
Enter Password
OK
Cancel
Transfer
Queue
Source File
Source Directory
Destination Directory
Size
Status
Speed
Time

Password: *********

Connect to a remote Linux server from Windows using SSH client PuTTY



Specify your connection by host name or IP address

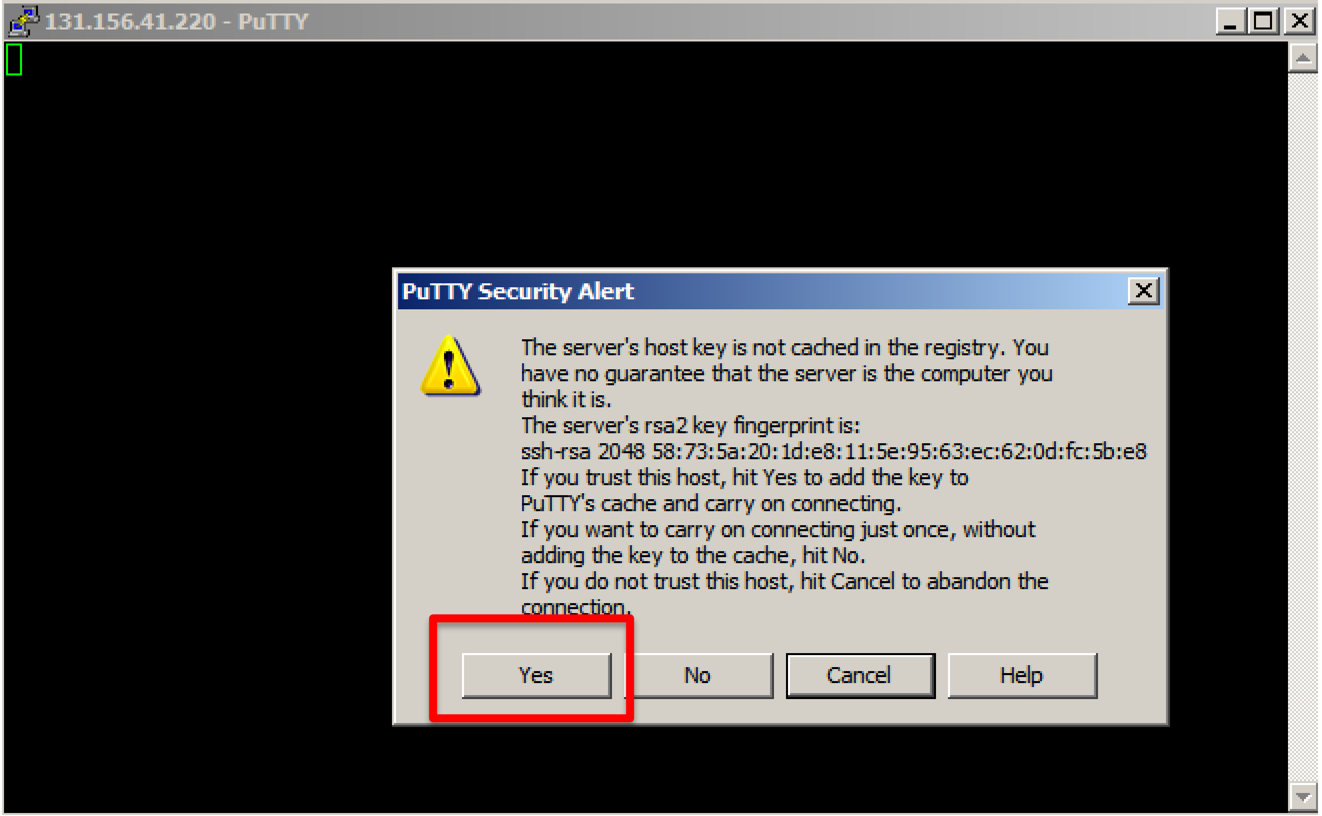
Host Name (or IP address): 131.156.41.220

Port: 22

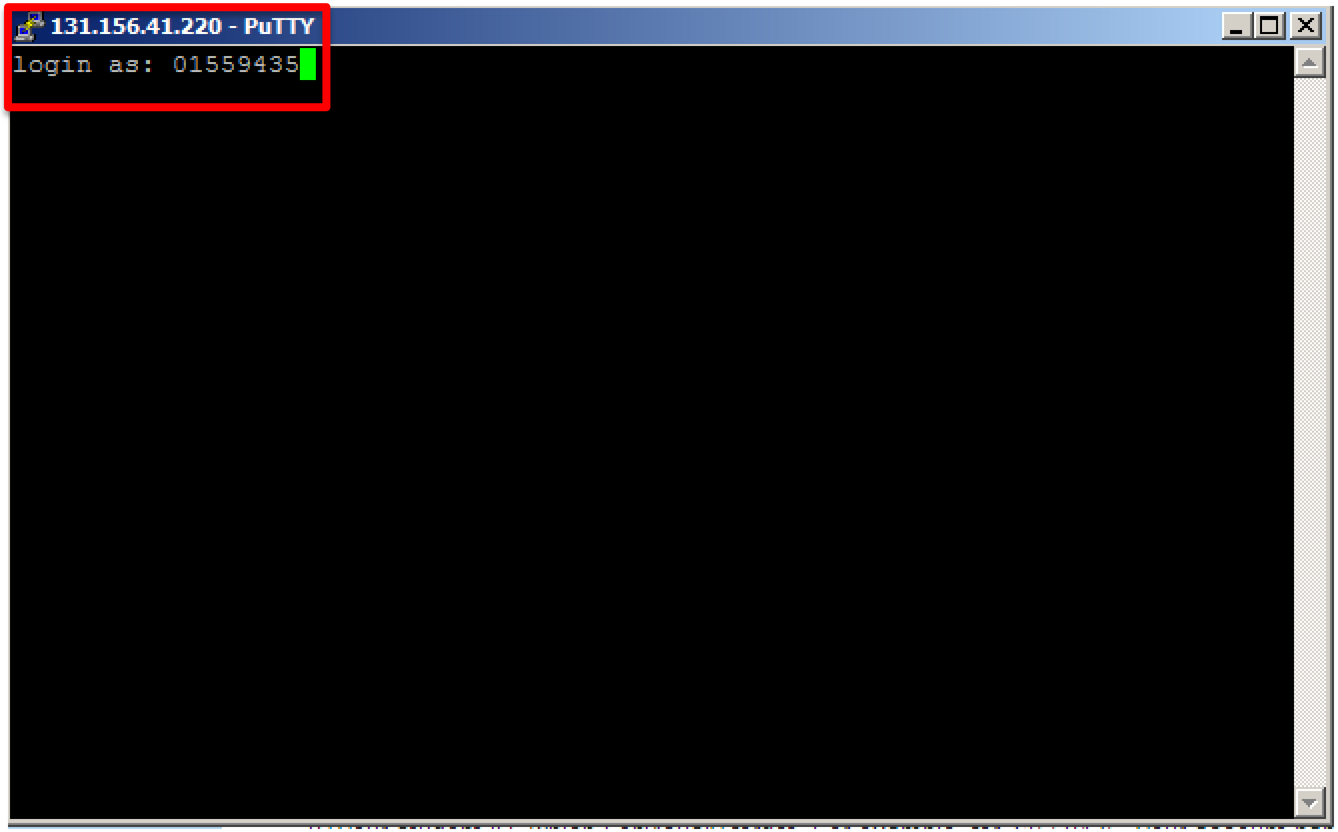
Protocol: ○ Raw ○ Telnet ○ Rlogin ● SSH

Saved Sessions:
Default Settings
UBUnixSSH
bem
csbl
gaea
gly
mason

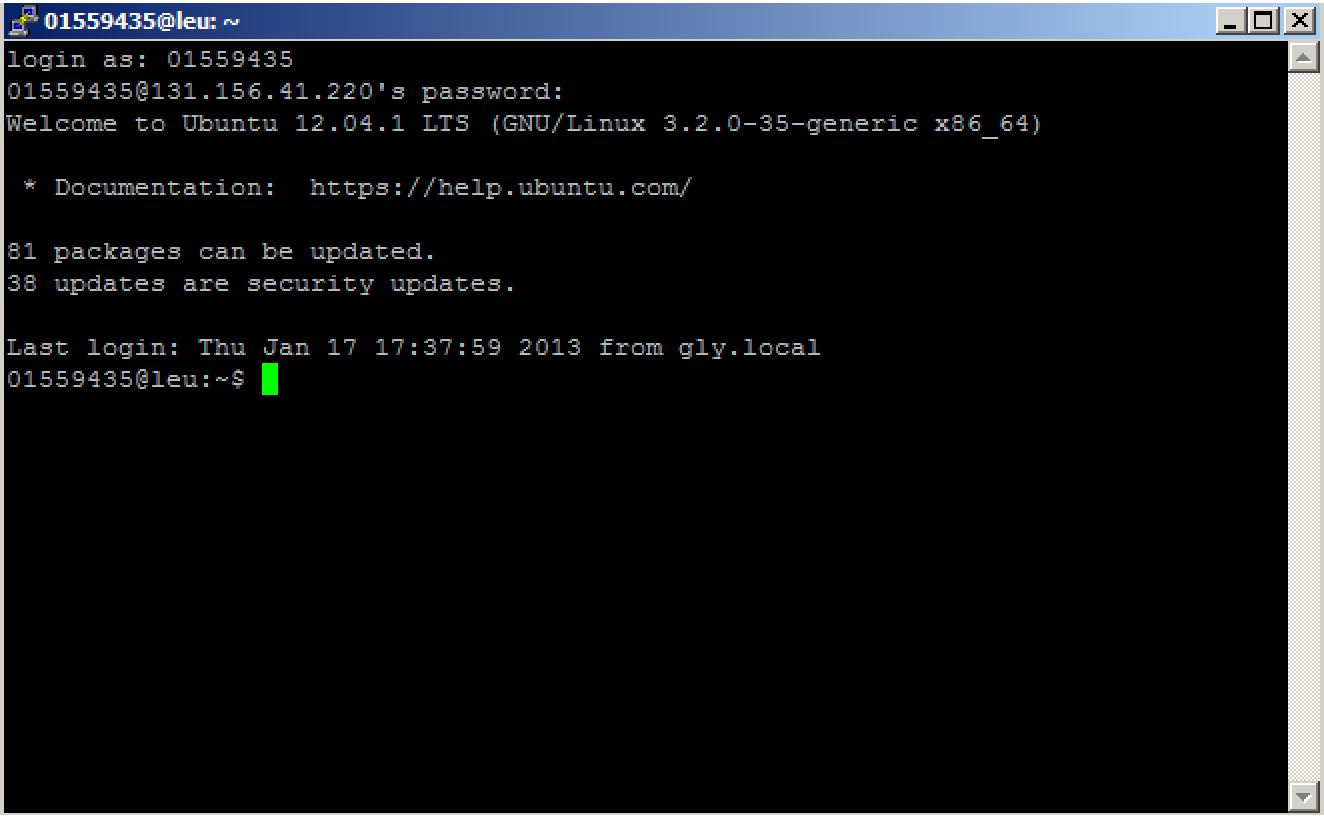
Close window on exit: ○ Always ○ Never ● Only on clean exit



The server's host key is not cached in the registry. You have no guarantee that the server is the computer you think it is.
The server's rsa2 key fingerprint is:
ssh-rsa 2048 58:73:5a:20:1d:e8:11:5e:95:63:ec:62:0d:fc:5b:e8
If you trust this host, hit Yes to add the key to PuTTY's cache and carry on connecting.
If you want to carry on connecting just once, without adding the key to the cache, hit No.
If you do not trust this host, hit Cancel to abandon the connection.



```
login as: 01559435
```



```
login as: 01559435
01559435@131.156.41.220's password:
Welcome to Ubuntu 12.04.1 LTS (GNU/Linux 3.2.0-35-generic x86_64)

 * Documentation:  https://help.ubuntu.com/

81 packages can be updated.
38 updates are security updates.

Last login: Thu Jan 17 17:37:59 2013 from gly.local
01559435@leu:~$
```

Next lecture

NCBI resources II: tools and ftp resources