INTRODUCTION TO BIOINFORMATICS

Please take the initial BIOINF525 questionnaire:

< http://tinyurl.com/bioinf525-questions >

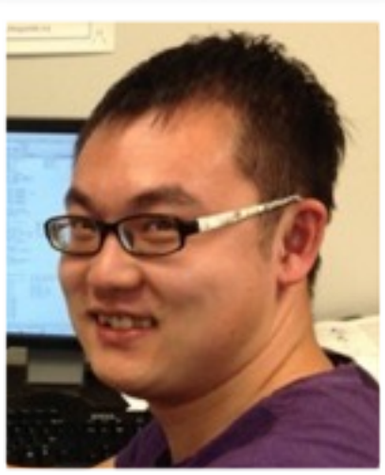
Barry Grant

University of Michigan

www.thegrantlab.org

Barry Grant, Ph.D.
bjgrant@umich.edu

Ryan Mills, Ph.D.
remills@umich.edu



Hongyang Li (GSI)
hyangl@umich.edu

COURSE LOGISTICS

Lectures: Tuesdays 2:30-4:00 PM
Rm. 2062 Palmer Commons

Labs: *Session I:* Thursdays 2:30 - 4:00 PM
Session II: Fridays 10:30 - 12:00 PM
Rm. 2036 Palmer Commons

Website: http://tinyurl.com/bioinf525-w16
Lecture, lab and background reading material
plus homework and course announcements

MODULE OVERVIEW

Objective: Provide an introduction to the practice of bioinformatics as well as a practical guide to using common bioinformatics databases and algorithms

TODAYS MENU

Overview of bioinformatics

- The *what*, *why* and *how* of bioinformatics?
- Major bioinformatics research areas.
- Skepticism and common problems with bioinformatics.

Bioinformatics databases and associated tools

- Primary, secondary and composite databases.
  - Nucleotide sequence databases (GenBank & RefSeq).
  - Protein sequence database (UniProt).
  - Composite databases (PFAM & OMIM).

Database usage vignette

- Searching with ENTREZ and BLAST.
- Reference slides and handout on major databases.

HOMEWORK

- ☑ Complete the **initial course questionnaire**:
  http://tinyurl.com/bioinf525-questions

- ☑ Check out the "**Background Reading**" material on Ctools:
  http://tinyurl.com/bioinf525-w16

- ☑ Complete the **lecture 1.1 homework questions**:
  http://tinyurl.com/bioinf525-quiz1

Q. What is Bioinformatics?

Q. What is Bioinformatics?

"*Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.*"
[After Orengo, 2003]

… Bioinformatics is a hybrid of biology and computer science

… **Bioinformatics is computer aided biology!**

Computer based management and analysis of biological and biomedical data with useful applications in many disciplines, particularly genomics, proteomics, metabolomics, etc...

MORE DEFINITIONS

- “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying **"informatics" techniques** (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**.
  Luscombe NM, *et al*. Methods Inf Med. 2001;40:346.

- “Bioinformatics is research, development, or application of **computational approaches** for expanding the use of **biological**, **medical**, **behavioral** or **health data**, including those to **acquire**, **store**, **organize** and **analyze** such data.”
  National Institutes of Health (NIH) (http://tinyurl.com/l3gxr6b)

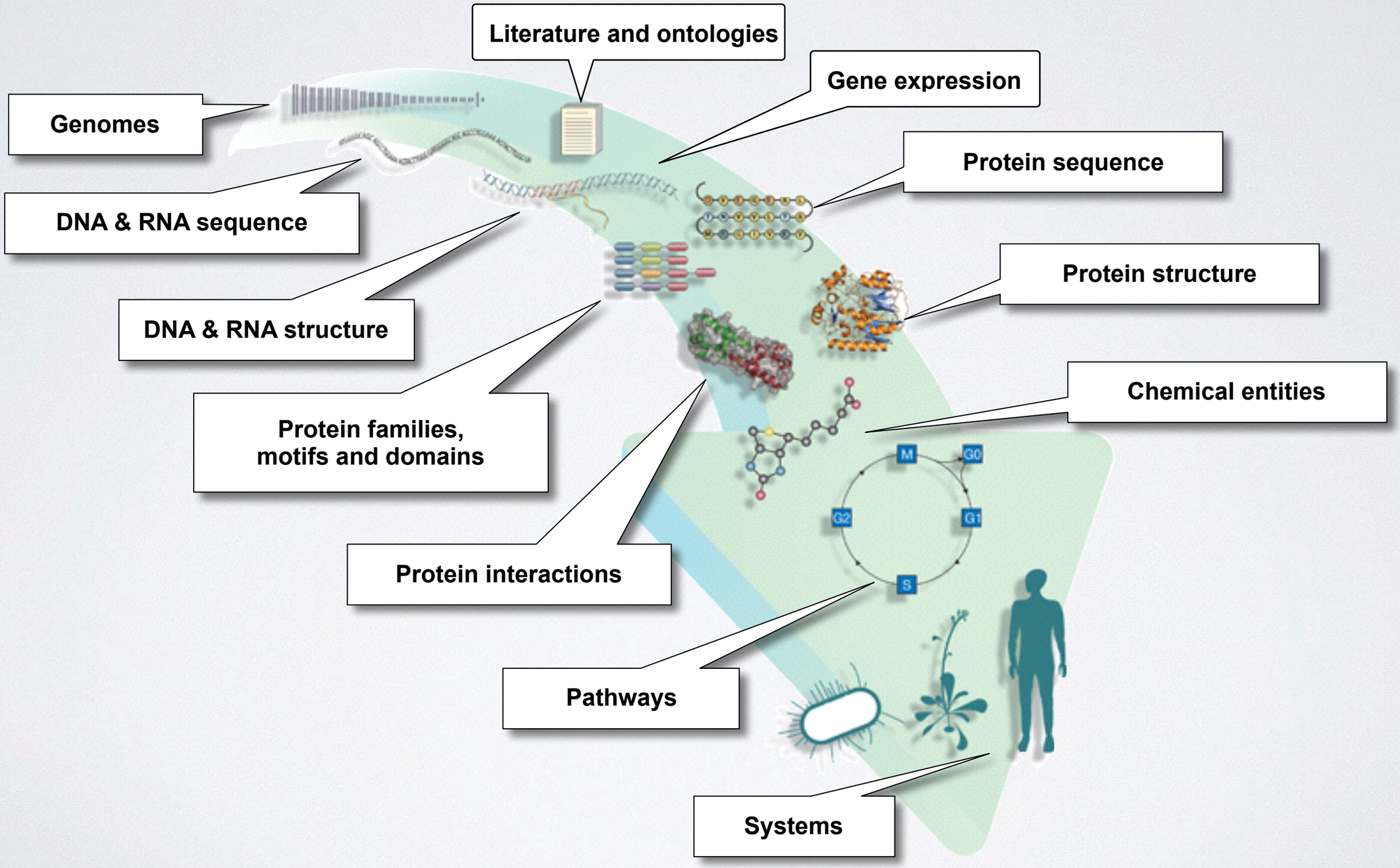
MORE DEFINITIONS

- "Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying **"informatics" techniques** (derived from disciplines such as applied math, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**.
  Luscombe NM, *et al*. Methods Inf Med. 2001;40:346.

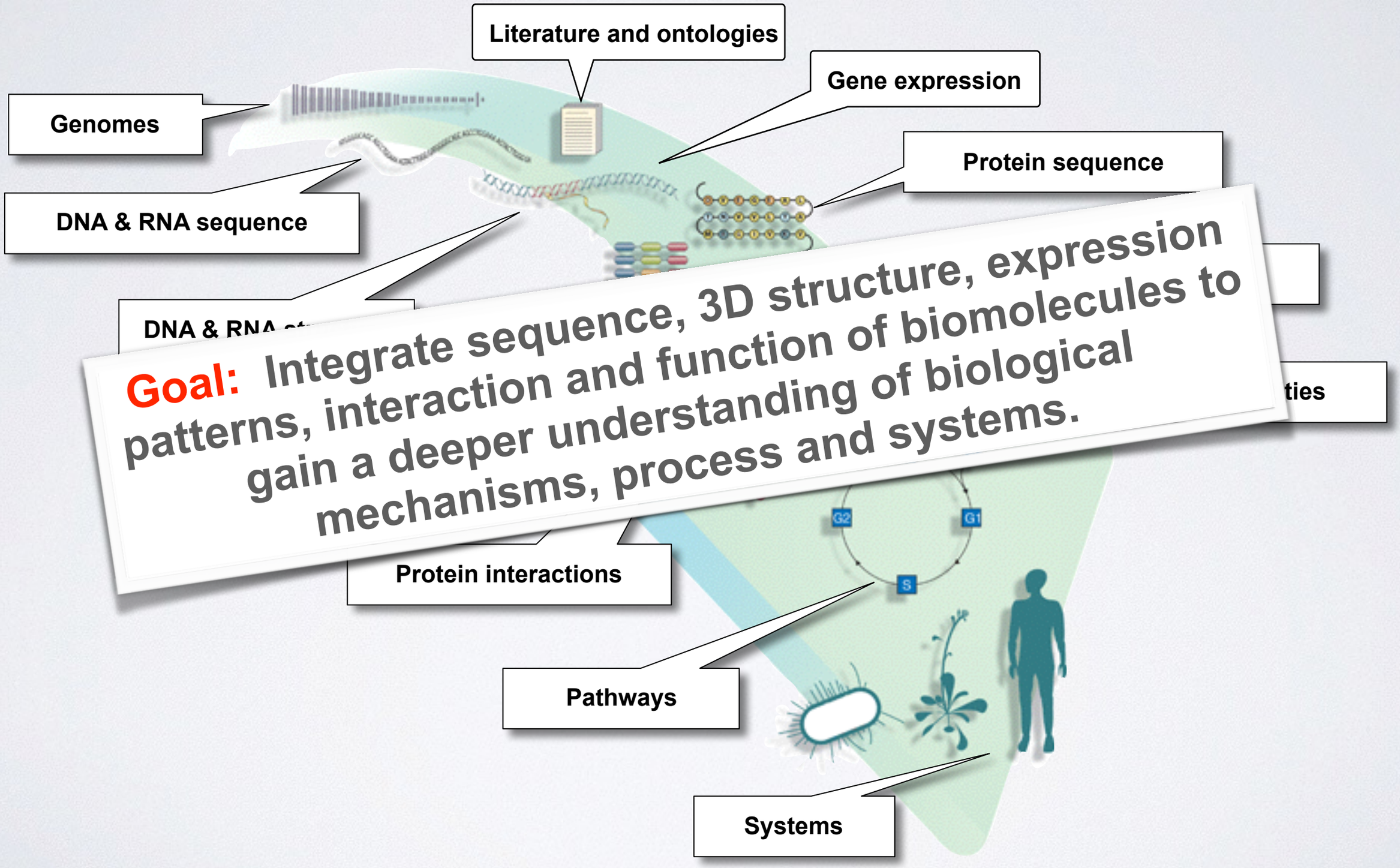
- "Bioinformatics is research, development, or application of **computational approaches** for expanding the use of **biological, medical, behavioral** or **health data**, including those to **acquire**, **store**, **organize** and **analyze** such data."
  National Institutes of Health (NIH) (http://tinyurl.com/l3gxr6b)

Key Point: Bioinformatics is Computer Aided Biology

Major types of Bioinformatics Data

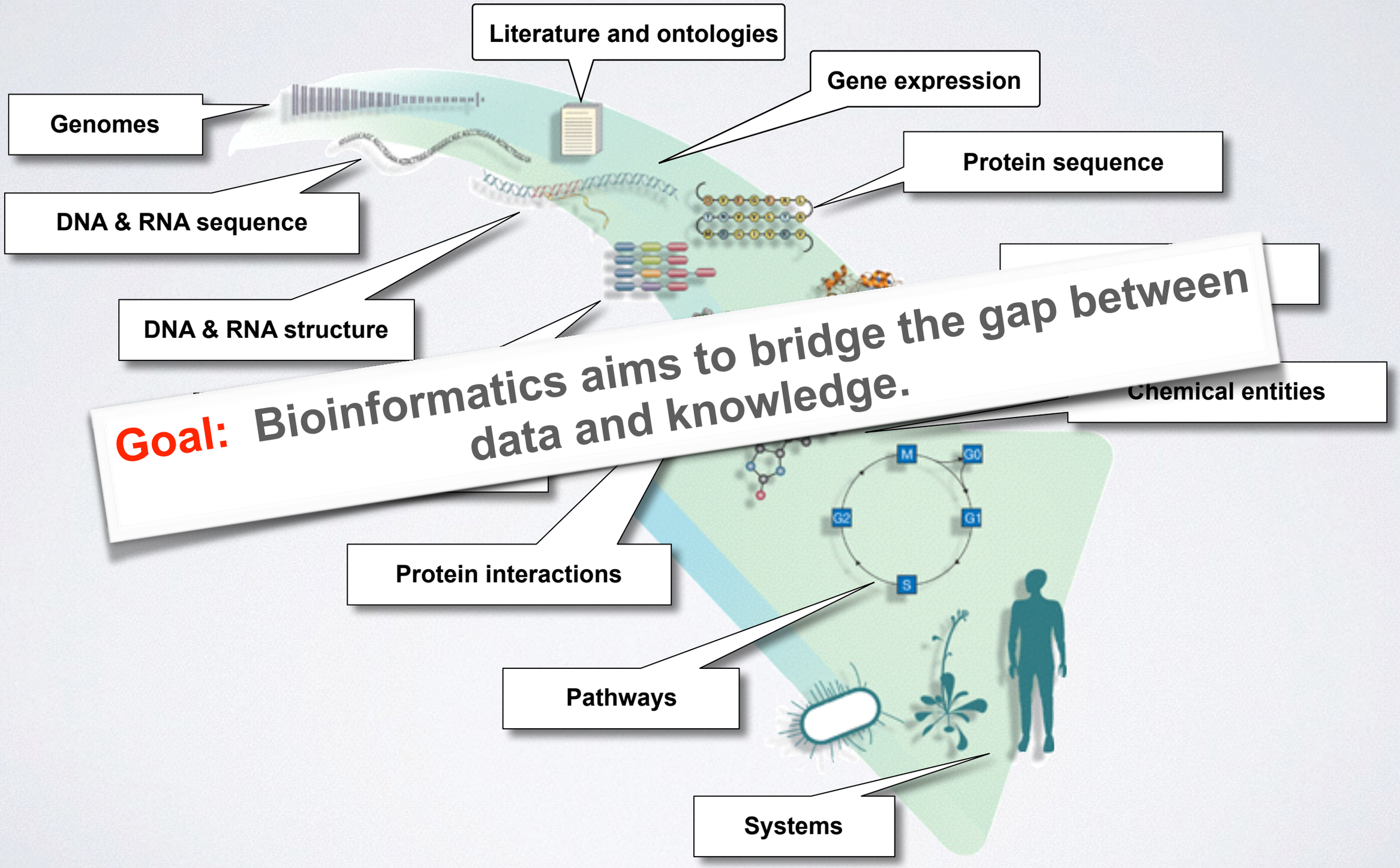


Major types of Bioinformatics Data



Goal: Integrate sequence, 3D structure, expression patterns, interaction and function of biomolecules to gain a deeper understanding of biological mechanisms, process and systems.

Major types of Bioinformatics Data



BIOINFORMATICS RESEARCH AREAS

Include but are not limited to:

- Organization, classification, dissemination and analysis of biological and biomedical data (particularly ‘-omics' data).
- Biological sequence analysis and phylogenetics.
- Genome organization and evolution.
- Regulation of gene expression and epigenetics.
- Biological pathways and networks in healthy & disease states.
- Protein structure prediction from sequence.
- Modeling and prediction of the biophysical properties of biomolecules for binding prediction and drug design.
- Design of biomolecular structure and function.

With applications to Biology, Medicine, Agriculture and Industry

Where did bioinformatics come from?

Bioinformatics arose as molecular biology began to be transformed by the emergence of molecular sequence and structural data

Recap: The key dogmas of molecular biology

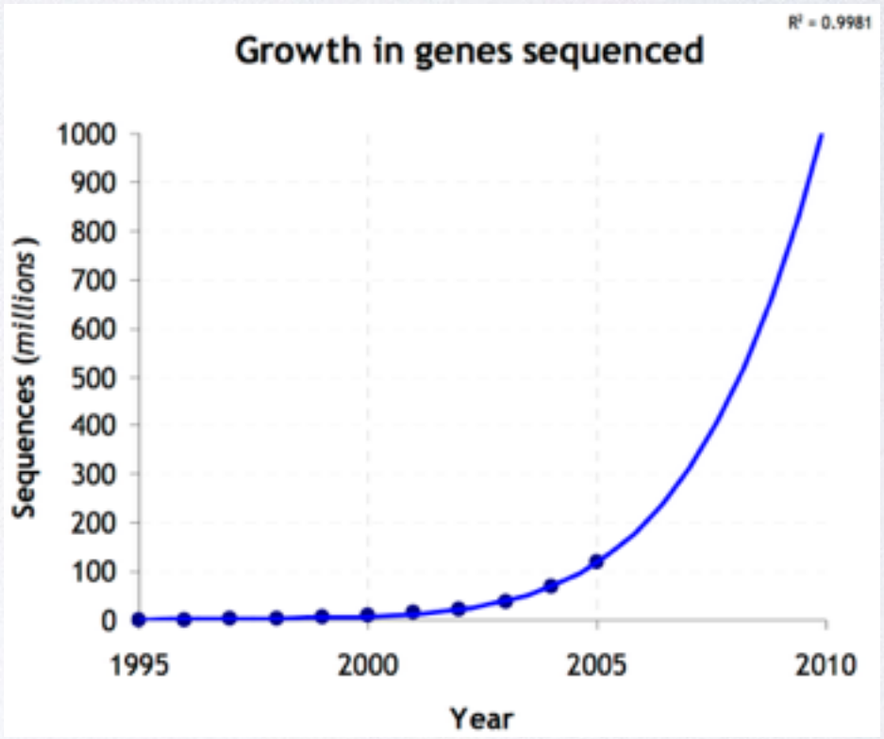
- *DNA sequence* determines *protein sequence.*
- *Protein sequence* determines *protein structure.*
- *Protein structure* determines *protein function.*
- *Regulatory mechanisms* (e.g. gene expression) determine the amount of a particular *function in space and time.*

Bioinformatics is now essential for the archiving, organization and analysis of data related to these processes.

Why do we need Bioinformatics?

Bioinformatics is necessitated by the rapidly expanding quantities and complexity of biomolecular data

- Bioinformatics provides methods for the efficient:
  - **storage**
  - **annotation**
  - **search** and **retrieval**
  - data **integration**
  - data **mining** and **analysis**

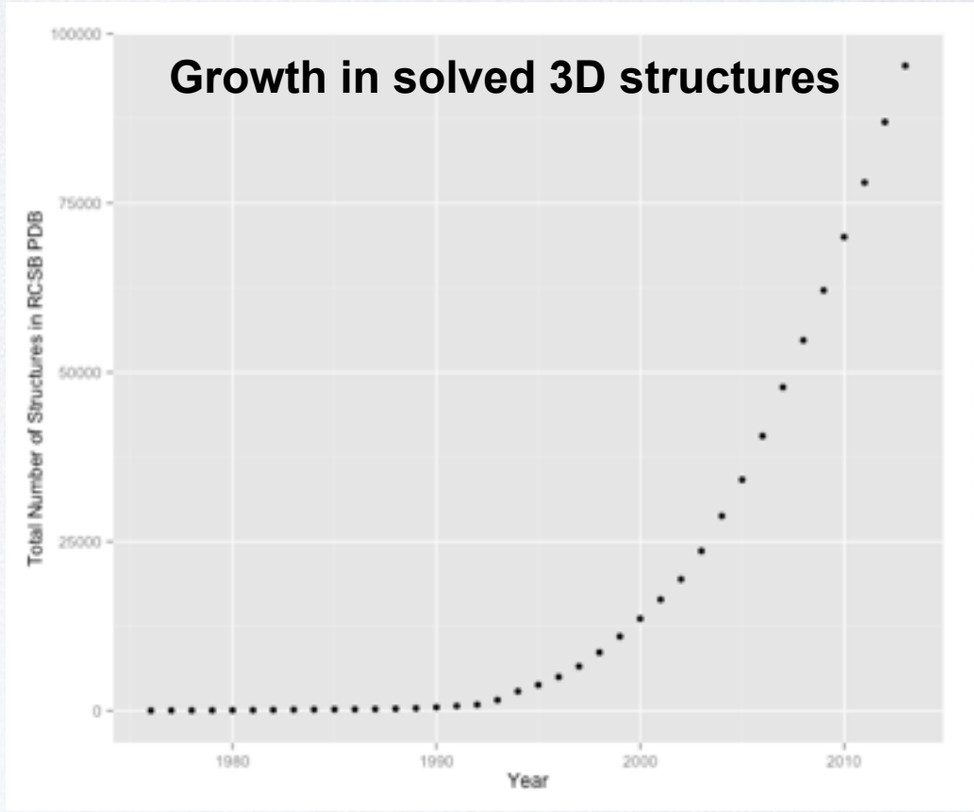


Bioinformatics is essential for the archiving, organization and analysis of data from sequencing, structural genomics, microarrays, proteomics and new high throughput assays.

Why do we need Bioinformatics?

Bioinformatics is necessitated by the rapidly expanding quantities and complexity of biomolecular data

- Bioinformatics provides methods for the efficient:
  - **storage**
  - **annotation**
  - **search** and **retrieval**
  - data **integration**
  - data **mining** and **analysis**



Bioinformatics is essential for the archiving, organization and analysis of data from sequencing, structural genomics, microarrays, proteomics and new high throughput assays.

How do we do Bioinformatics?

- A "*bioinformatics approach*" involves the application of **computer algorithms**, **computer models** and **computer databases** with the broad goal of understanding the action of both individual genes, transcripts, proteins and large collections of these entities.

DNA	→	RNA	→	Protein
Genome	→	**Transcriptome**	→	**Proteome**

How do we actually do Bioinformatics?

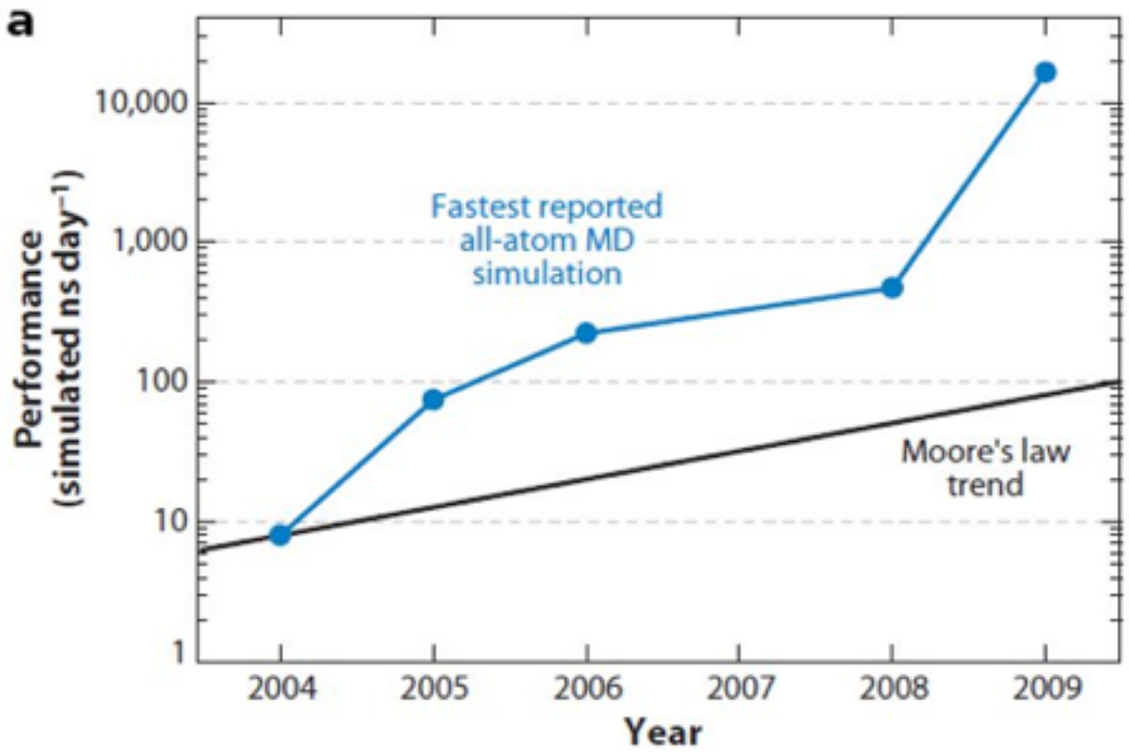
Pre-packaged tools and databases

- Many online
- New tools and time consuming methods frequently require downloading
- Most are free to use

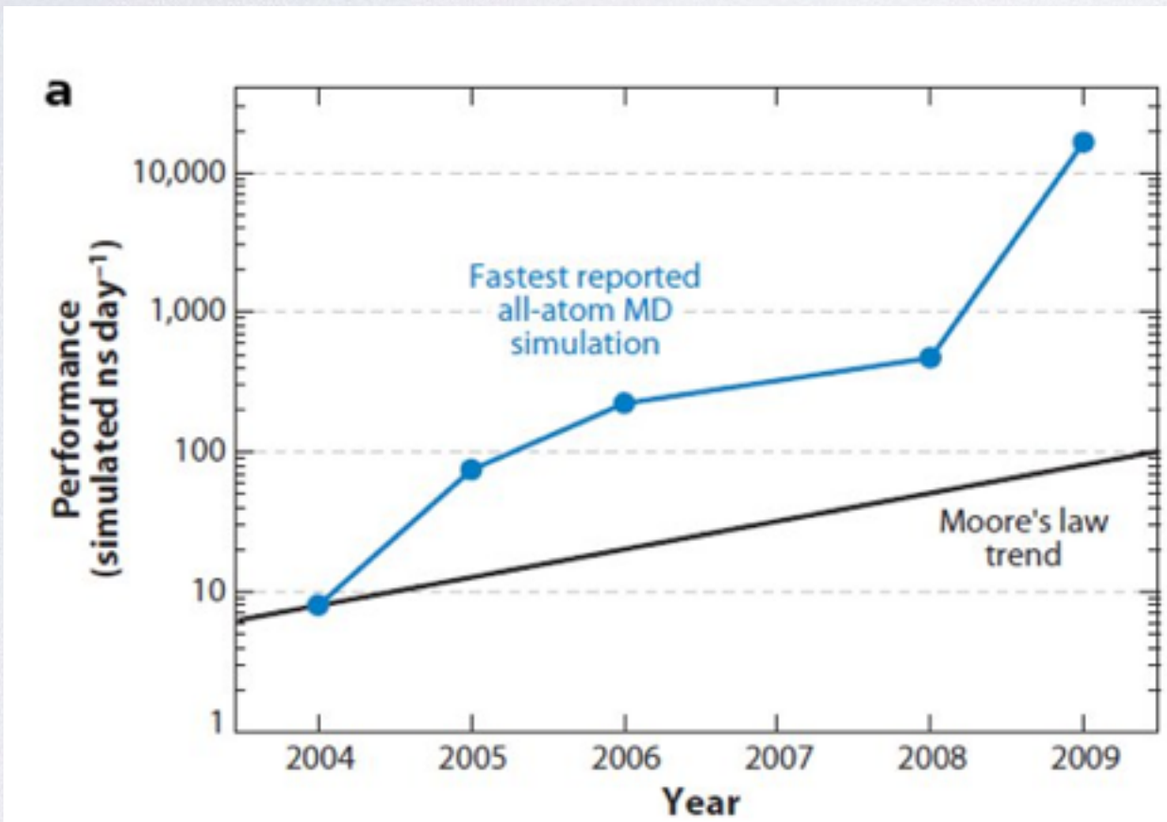
Tool development

- Mostly on a UNIX environment
- Knowledge of programing languages frequently required (Python, Perl, R, C Java, Fortran)
- May require specialized or high performance computing resources…

SIDE-NOTE: SUPERCOMPUTERS AND GPUS



SIDE-NOTE: SUPERCOMPUTERS AND GPUS



HOW COMPUTERS HAVE CHANGED

DATE	COST	SPEED	MEMORY	SIZE
1967	$4OM	0.1 MHz	1 MB	HALL
2013	$4,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

If cars were like computers then a new Volvo would cost $3, would have a top speed of 1,000,000 Km/hr, would carry 50,000 adults and would park in a shoebox

Skepticism & Bioinformatics

We have to approach computational results the same way we do wet-lab results:

- Do they make sense?
- Is it what we expected?
- Do we have adequate controls, and how did they come out?
- Modeling is modeling, but biology is different...

  What does this model actually contribute?
- Avoid the miss-use of ‘black boxes’

Common problems with Bioinformatics

Confusing multitude of tools available

- Each with many options and settable parameters

Most tools and databases are written by and for nerds

- Same is true of documentation - if any exists!

Most are developed independently

Notable exceptions are found at the:

- **EBI** (European Bioinformatics Institute) and
- **NCBI** (National Center for Biotechnology Information)

Protein BLAST: search protein databases using a protein query

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome

General Parameters

Max target sequences: 500
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional sco

Filters and Masking

Filter: ☐ Low complexity regions

Mask: ☐ Mask for lookup table only
☐ Mask lower case letters

PSI/PHI/DELTA BLAST

Upload PSSM Optional: Choose File — no file selected

PSI-BLAST Threshold: 0.005

Pseudocount: 0

Even Blast has many settable parameters

Related tools with different terminology

STEP 3 - Set you

PROGRAM
FASTA

MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
BLOSUM50	-10	-2	2	10	0 (default)

DNA STRAND	HISTOGRAM	FILTER	STATISTICAL ESTIMATES
N/A	no	none	Regress

SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	MULTI HSPs
50	50	START-END	START-END	no

SCORE FORMAT
Default

Key Online Bioinformatics Resources: NCBI & EBI

The NCBI and EBI are invaluable, publicly available resources for biomedical research

Please take the initial BIOINF525 questionnaire:

< tinyurl.com/bioinf525-questions >

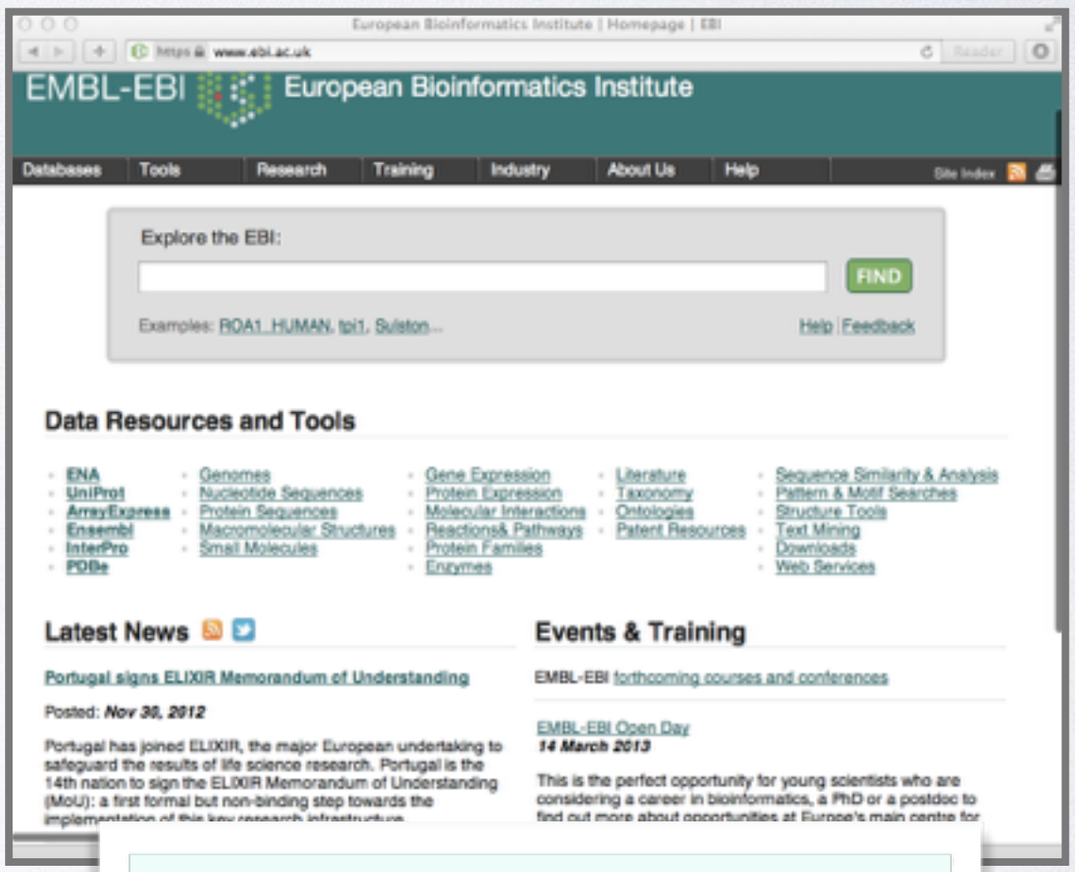
http://www.ncbi.nlm.nih.gov

https://www.ebi.ac.uk

Key Online Bioinformatics Resources: NCBI & EBI

The NCBI and EBI are invaluable, publicly available resources for biomedical research

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



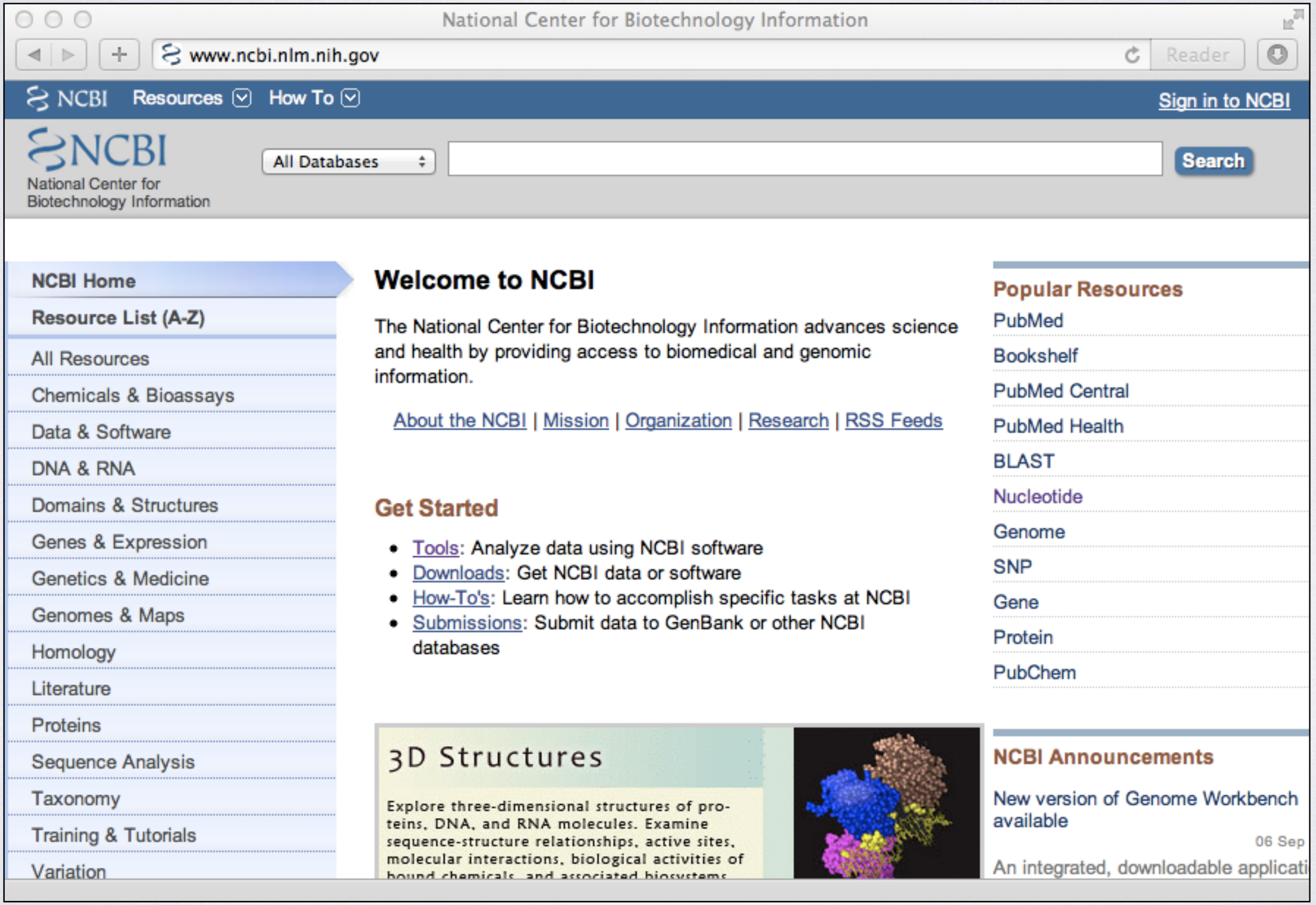
https://www.ebi.ac.uk

National Center for Biotechnology Information (NCBI)

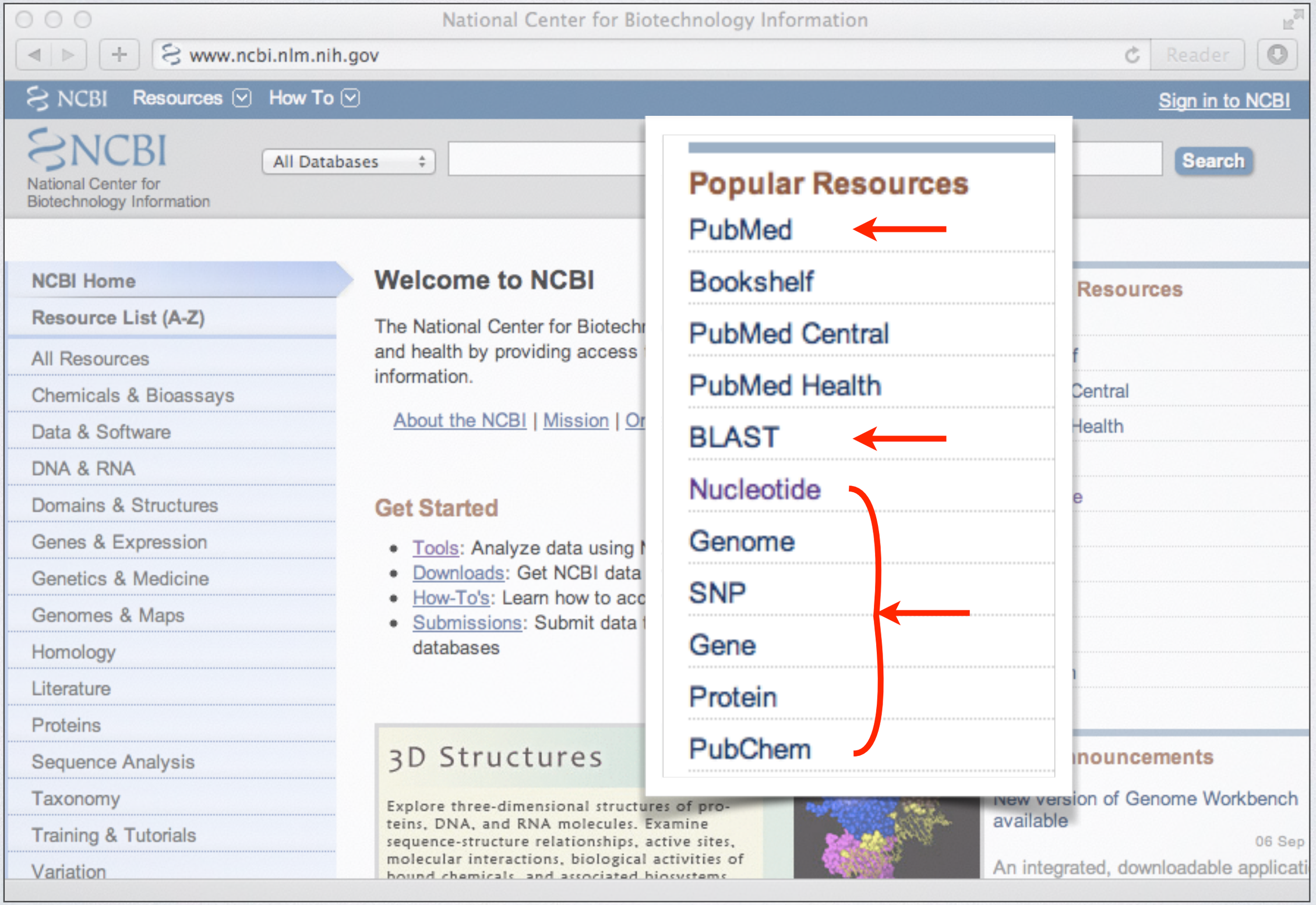
- Created in 1988 as a part of the National Library of Medicine (NLM) at the National Institutes of Health
- NCBI's mission includes:
  - Establish **public databases**
  - Develop **software tools**
  - **Education** on and dissemination of biomedical information

- We will cover a number of core NCBI databases and software tools in the lecture

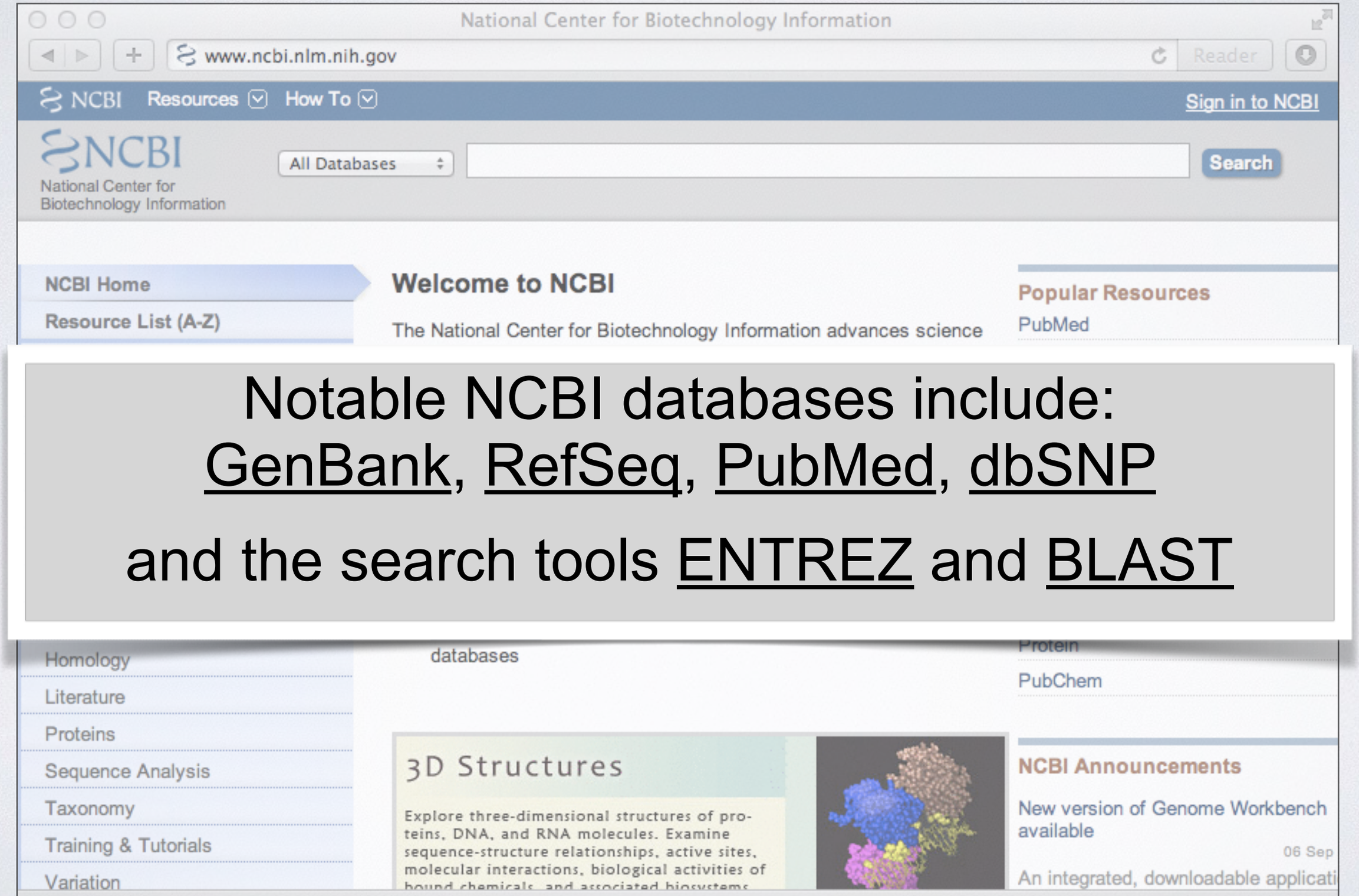
http://www.ncbi.nlm.nih.gov



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



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



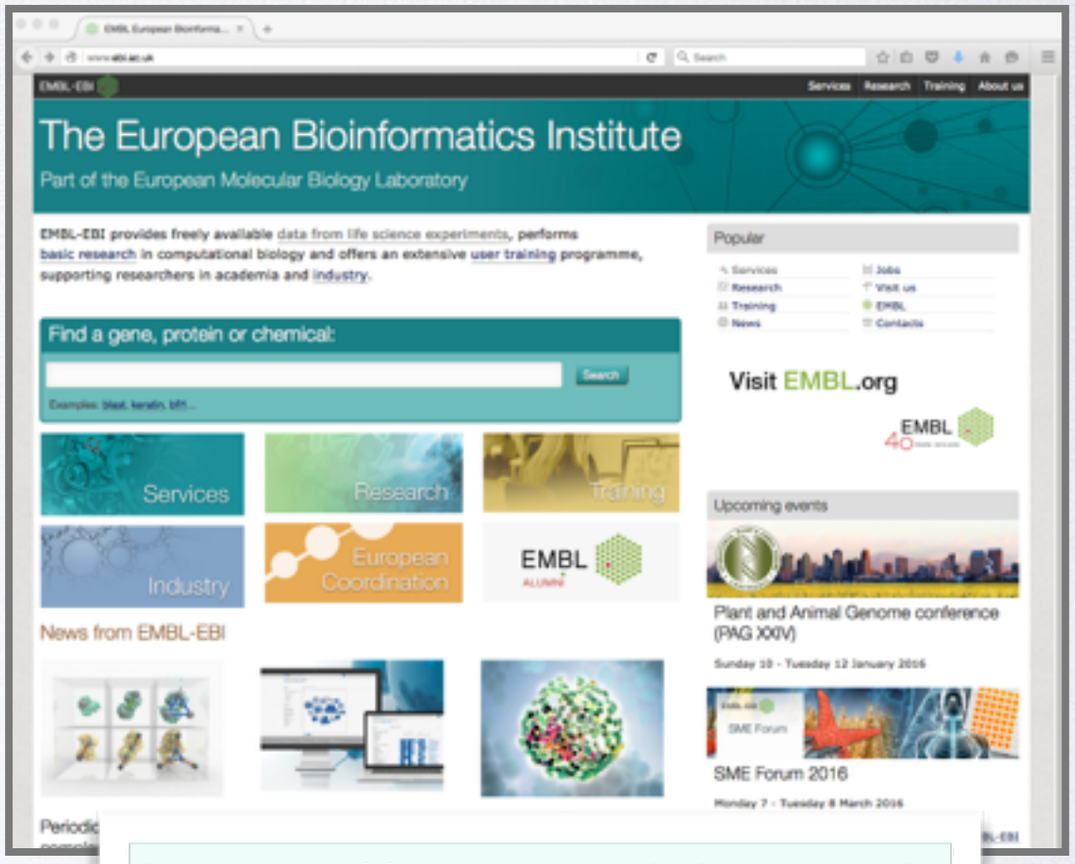
Notable NCBI databases include:
GenBank, RefSeq, PubMed, dbSNP

and the search tools ENTREZ and BLAST

Key Online Bioinformatics Resources: NCBI & EBI

The NCBI and EBI are invaluable, publicly available resources for biomedical research

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



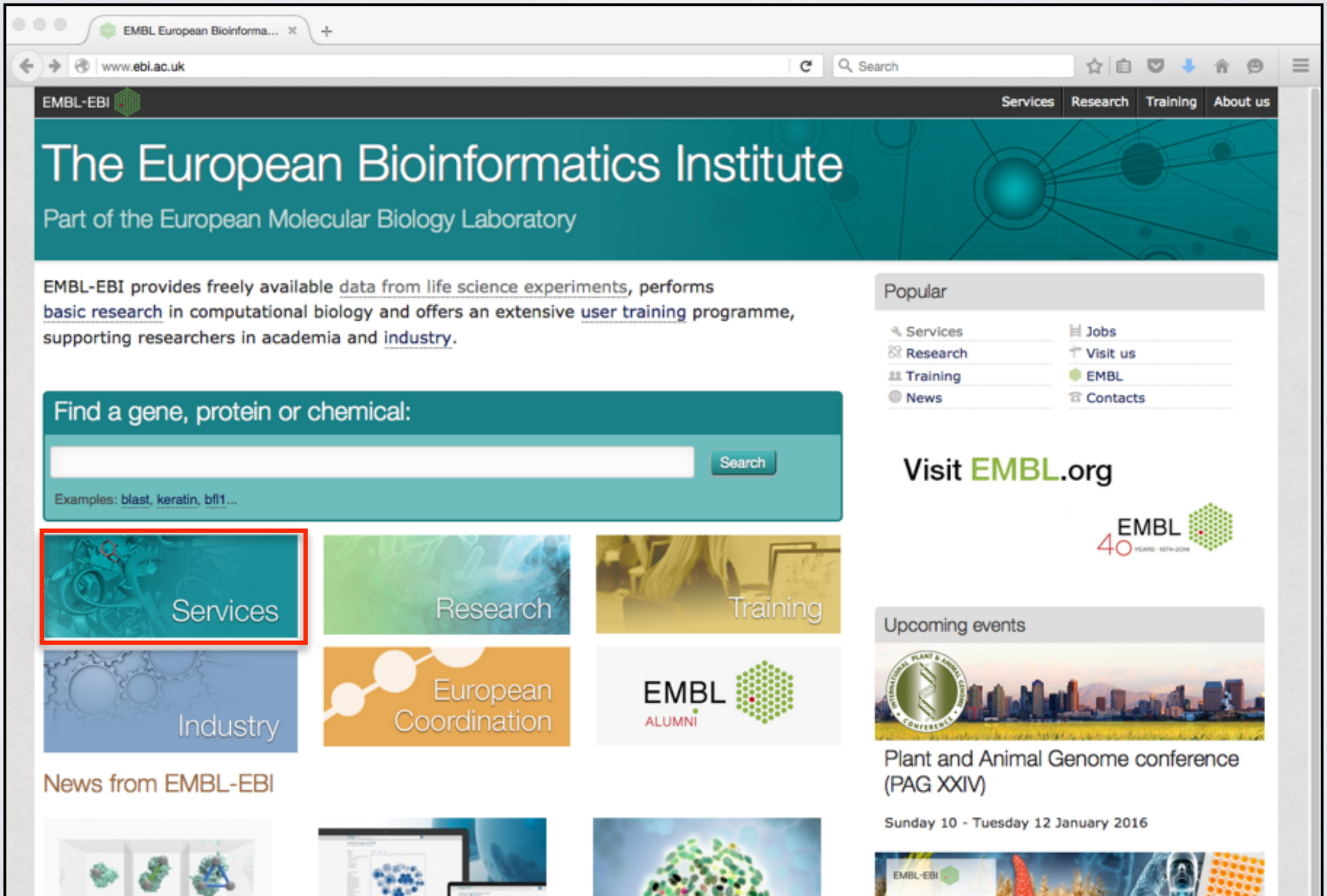
https://www.ebi.ac.uk

European Bioinformatics Institute (EBI)

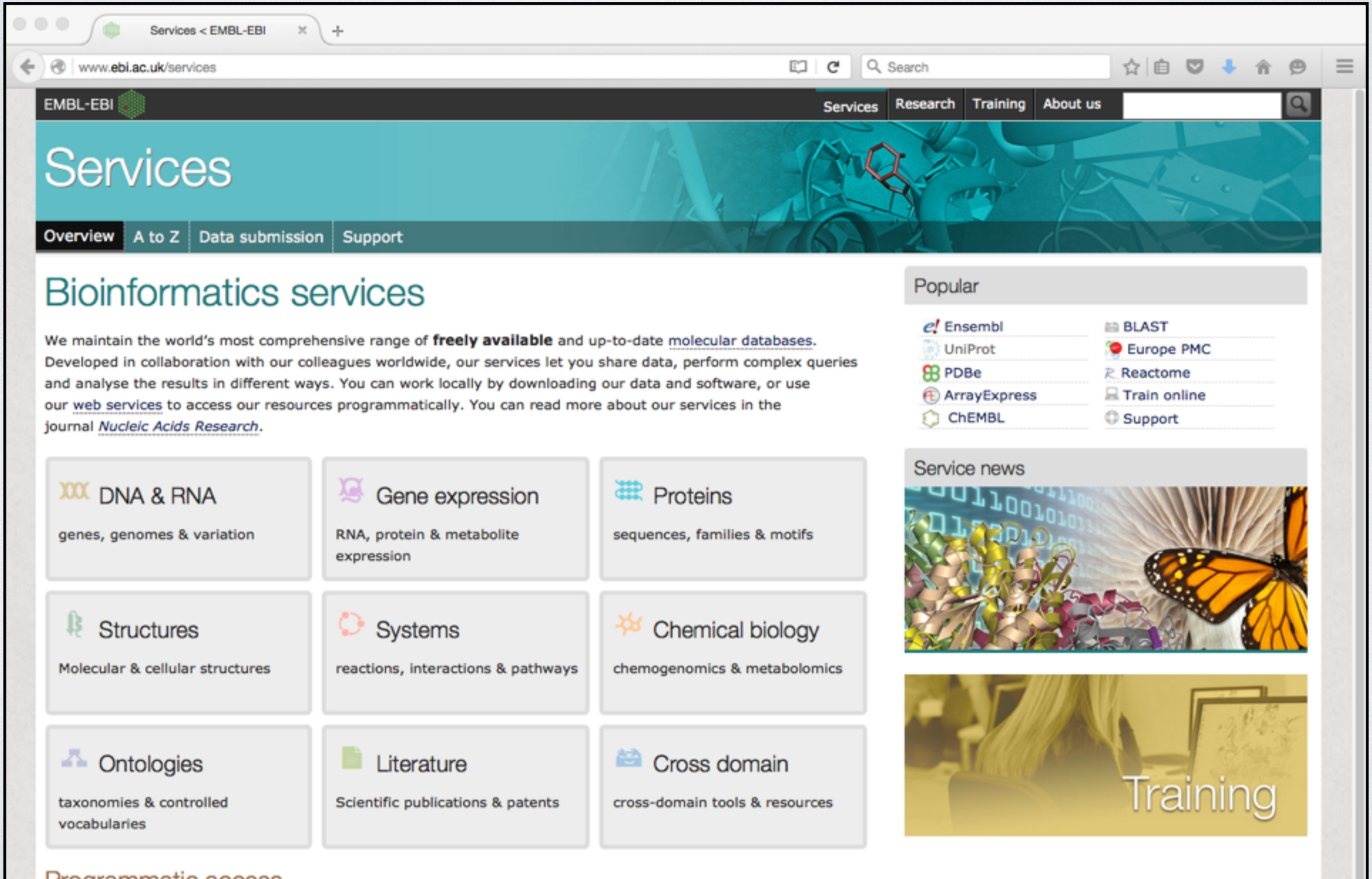
- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
  - providing freely available **data** and **bioinformatics services**
  - and providing advanced **bioinformatics training**

- We will briefly cover several EBI databases and tools that have advantages over those offered at NCBI

The EBI maintains a number of high quality curated **secondary databases** and associated tools



The EBI maintains a number of high quality curated **secondary databases** and associated tools



The EBI maintains a number of high quality curated **secondary databases** and associated tools

Services < EMBL-EBI

www.ebi.ac.uk/services

Search

EMBL-EBI

Services | Research | Training | About us

Services

Overview | A to Z | Data submission | Support

Bioinformatics services

We maintain the world's most comprehensive range of **freely available** and up-to-date molecular databases. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically. You can read more about our services in the journal *Nucleic Acids Research*.

- **DNA & RNA** — genes, genomes & variation
- **Gene expression** — RNA, protein & metabolite expression
- **Proteins** — sequences, families & motifs
- **Structures** — Molecular & cellular structures
- **Systems** — reactions, interactions & pathways
- **Chemical biology** — chemogenomics & metabolomics
- **Ontologies** — taxonomies & controlled vocabularies
- **Literature** — Scientific publications & patents
- **Cross domain** — cross-domain tools & resources

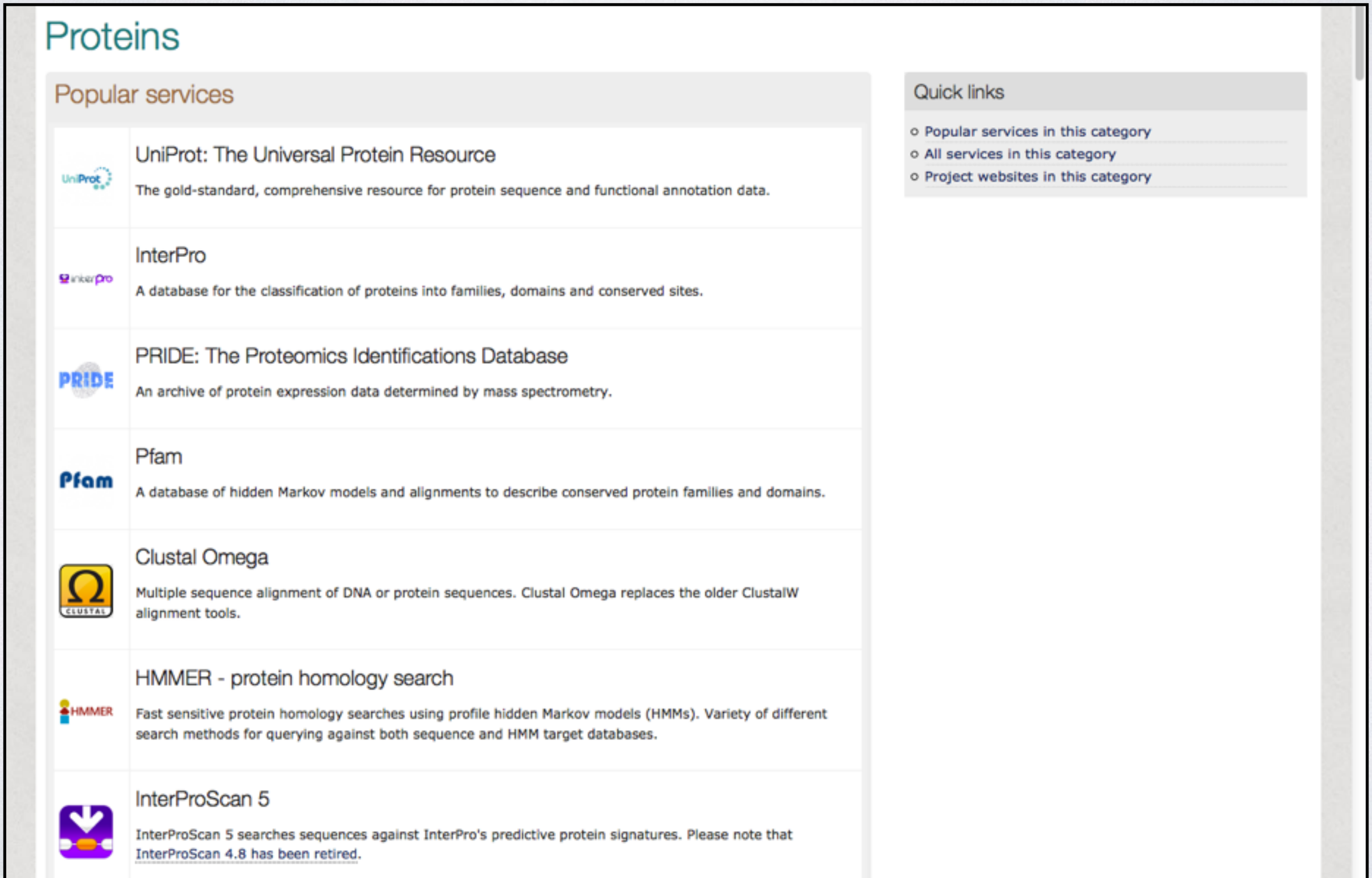
Programmatic access

Popular

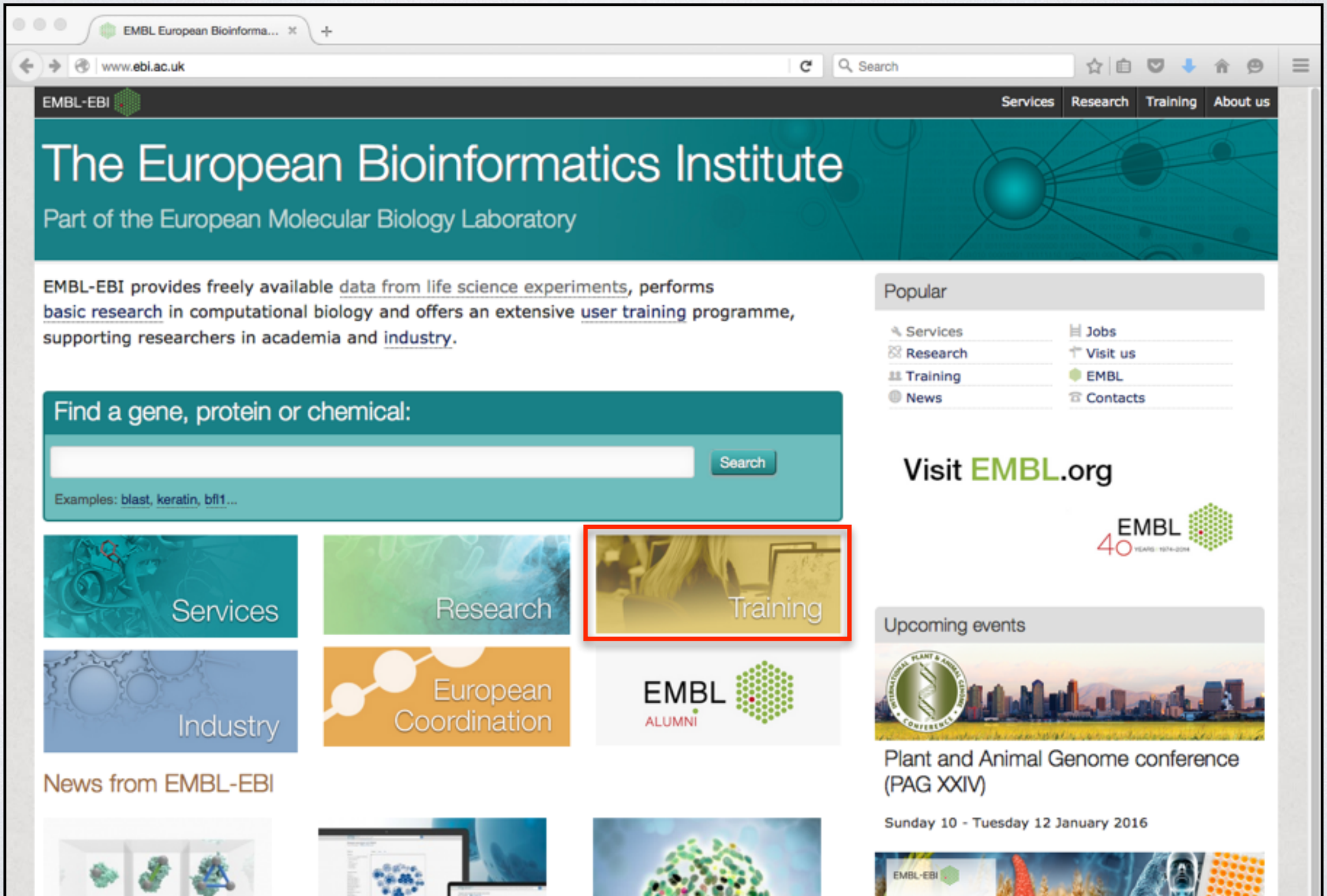
- Ensembl
- UniProt
- PDBe
- ArrayExpress
- ChEMBL

https://www.ebi.ac.uk

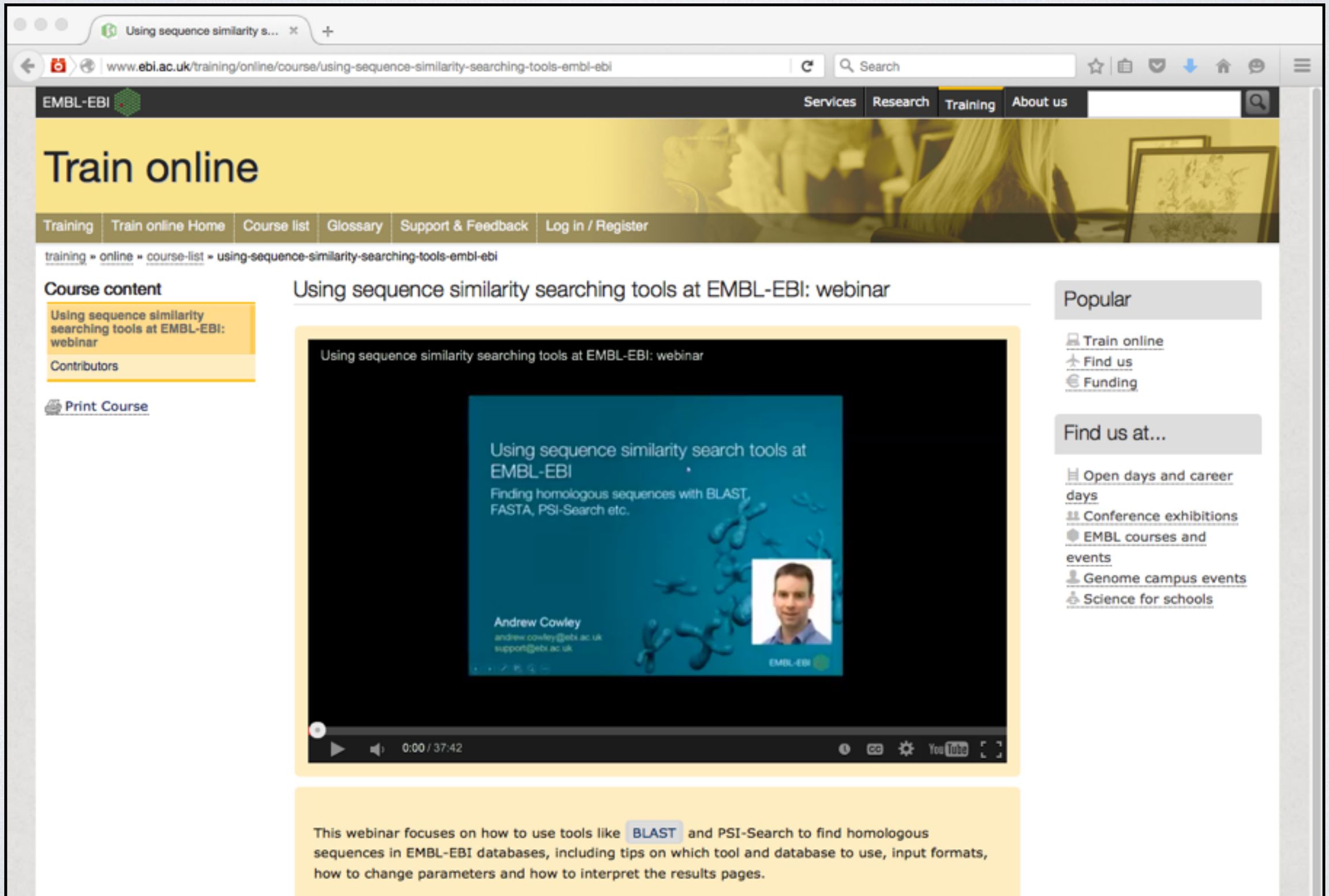
The EBI makes available a wider variety of **online tools** than NCBI



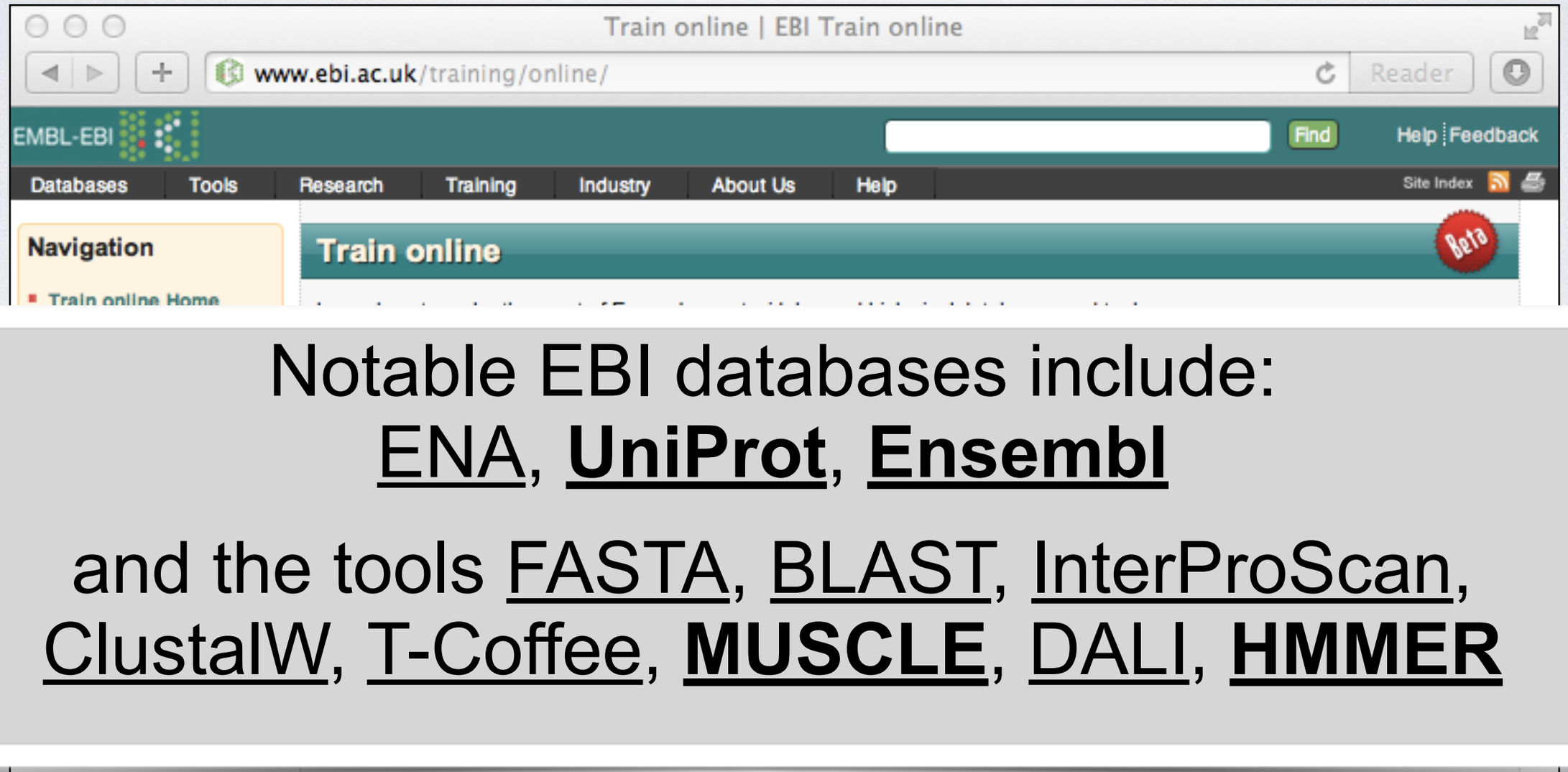
The EBI also provides a growing selection of **online tutorials** on EBI databases and tools



The EBI also provides a growing selection of **online tutorials** on EBI databases and tools



The EBI also provides a growing selection of **online tutorials** on EBI databases and tools



Notable EBI databases include:
ENA, **UniProt**, **Ensembl**

and the tools FASTA, BLAST, InterProScan, ClustalW, T-Coffee, **MUSCLE**, DALI, **HMMER**

BIOINFORMATICS DATABASES AND ASSOCIATED TOOLS

What is a database?

Computerized store of data that is organized to provide efficient retrieval.

- Uses standardized data (record) formats to enable computer handling

Key database features allow for:

- Adding, changing, removing and merging of records
- User-defined queries and extraction of specified records

Desirable features include:

- Contains the data you are interested in
- Allows fast data access
- Provides annotation and curation of entries
- Provides links to additional information (possibly in other databases)
- Allows you to make discoveries

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb,BBDB, BCGD, Beanref, BioImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSub, 0-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene,Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT, TeIDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..... !!!!

Bioinformatics Databases

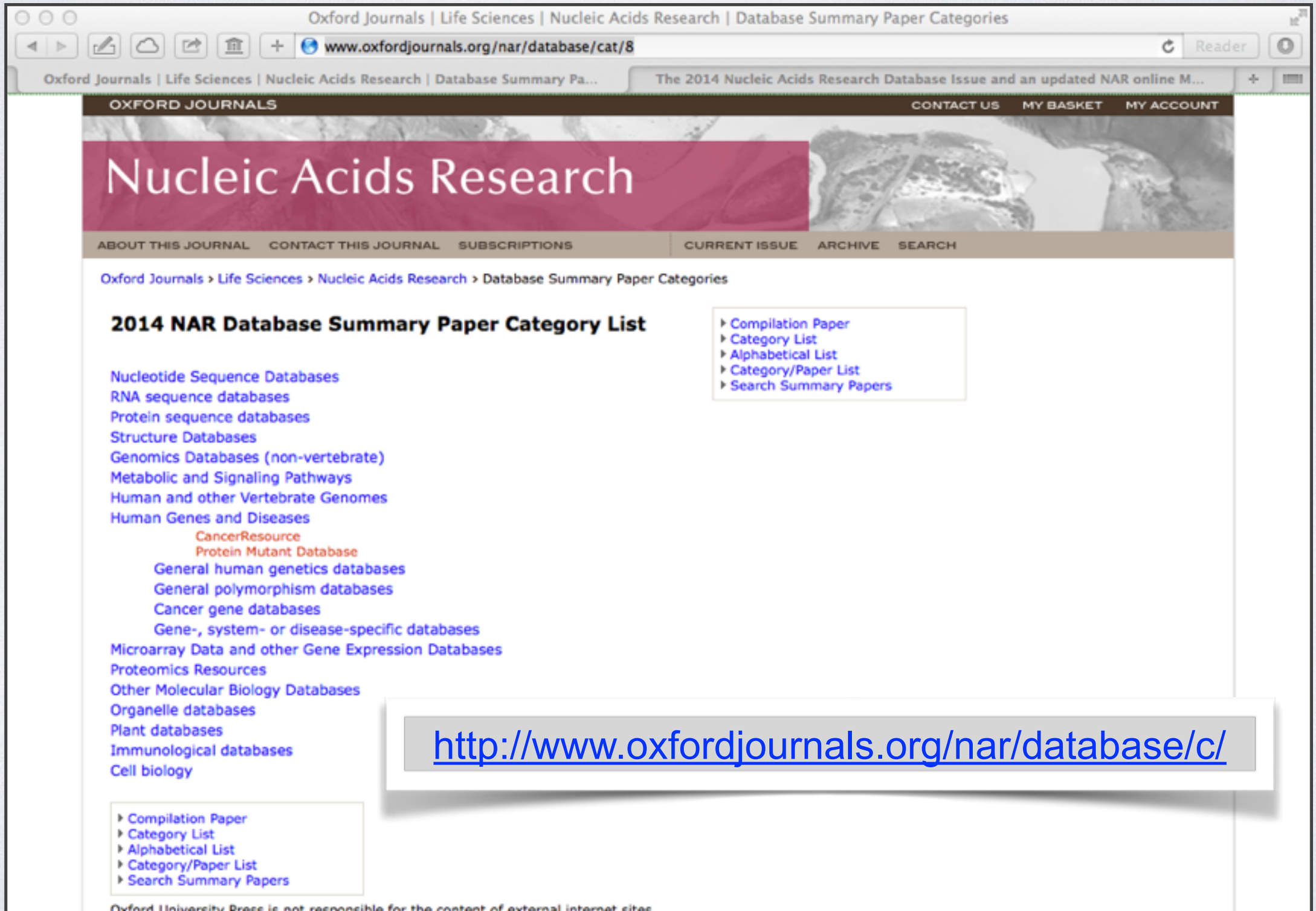
AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb,BBDB, BCGD, Beanref,
BioImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVM… TKbase,
CANSITE, CarbBank, CARBHYD, CATH, CAZY, C… AP,
ChickGBASE, Colibri, COPE, CottonDB, C… bEST,
dbSTS, DDBJ, DGP, DictyDb, Ri… CDC,
ECGC, EC02DBASE, E… THER,
FlyBase, Fl… Link,
G… DB, HAEMB,
H… HIVdb, HotMolecBase,
H… IMGT, Kabat, KDNA,
K… Medline, Mendel, MEROMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase,
MHC… OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase,
Myc… 0-lycBase,
PD… DD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS,
ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP,
SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK,
StyGene,Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL
Repository, SWISS-PROT, TeIDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene,
URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..... !!!!

There are lots of Bioinformatics Databases

For a annotated listing of major bioinformatics databases please see the Ctools handout

< Handout_Major_Databases.pdf >

Finding Bioinformatics Databases



Major Molecular Databases

The most popular bioinformatics databases focus on:

- Biomolecular sequence (e.g. GenBank, UniProt)
- Biomolecular structure (e.g. PDB)
- Vertebrate genomes (e.g. Ensemble)
- Small molecules (e.g. PubChem)
- Biomedical literature (e.g. PubMed)

The are also many popular "*boutique*" databases for:

- Classifying protein families, domains and motifs (e.g. PFAM, PROSITE)
- Specific organisms (e.g. WormBase, FlyBase)
- Specific proteins of biomedical importance (e.g. KinaseDB, GPCRDB)
- Specific diseases, mutations (e.g. OMIM, HGMD)
- Specific fields or methods of study (e.g. GOA, IEDB)

Major Molecular Databases

The most popular bioinformatics databases focus on:

- Biomolecular sequence (e.g. GenBank, UniProt)
- Biomolecular structure (e.g. PDB)
- Vertebrate genomes (e.g. Ensemble)
- Small molecules (e.g. PubChem)
- Biomedical literature (e.g. PubMed)

The are also many r… *…tique*" databases for:

- Classifyi… …ilies, domains and motifs (e.g. PFAM, PROSITE)
- Spe… …isms (e.g. WormBase, FlyBase)
- Specific proteins of biomedical importance (e.g. KinaseDB, GPCRDB)
- Specific diseases, mutations (e.g. OMIM, HGMD)
- Specific fields or methods of study (e.g. GOA, IEDB)

See Online: 'Handout_Major_Databases.pdf'

Primary, secondary & composite databases

Bioinformatics databases can be usefully classified into *primary*, *secondary* and *composite* according to their data source.

- **Primary databases** (or *archival databases*) consist of data derived experimentally.
  - **GenBank**: NCBI's primary nucleotide sequence database.
  - **PDB:** Protein X-ray crystal and NMR structures.

- **Secondary databases** (or *derived databases*) contain information derived from a primary database.
  - **RefSeq**: non redundant set of curated reference sequences primarily from GenBank
  - **PFAM**: protein sequence families primarily from UniProt and PDB

- **Composite databases** (or *metadatabases*) join a variety of different primary and secondary database sources.
  - **OMIM**: catalog of human genes, genetic disorders and related literature
  - **GENE**: molecular data and literature related to genes with extensive links to other databases.

GENBANK & REFSEQ:

NCBI'S NUCLEOTIDE SEQUENCE DATABASES

What is GenBank?

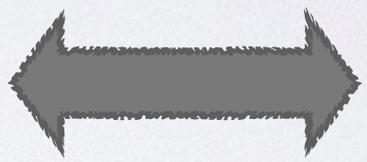
- GenBank is NCBI's primary **nucleotide only** sequence database
  - Archival in nature - reflects the state of knowledge at time of submission
  - Subjective - reflects the submitter point of view
  - Redundant - can have many copies of the same nucleotide sequence

- GenBank is actually three collaborating international databases from the US, Japan and Europe
  - GenBank (US)
  - DNA Database of Japan (DDBJ)
  - European Nucleotide Archive (ENA)

GenBank, ENA and DDBJ Share and synchronize data

ENA
Housed at EBI
European Bioinformatics Institute

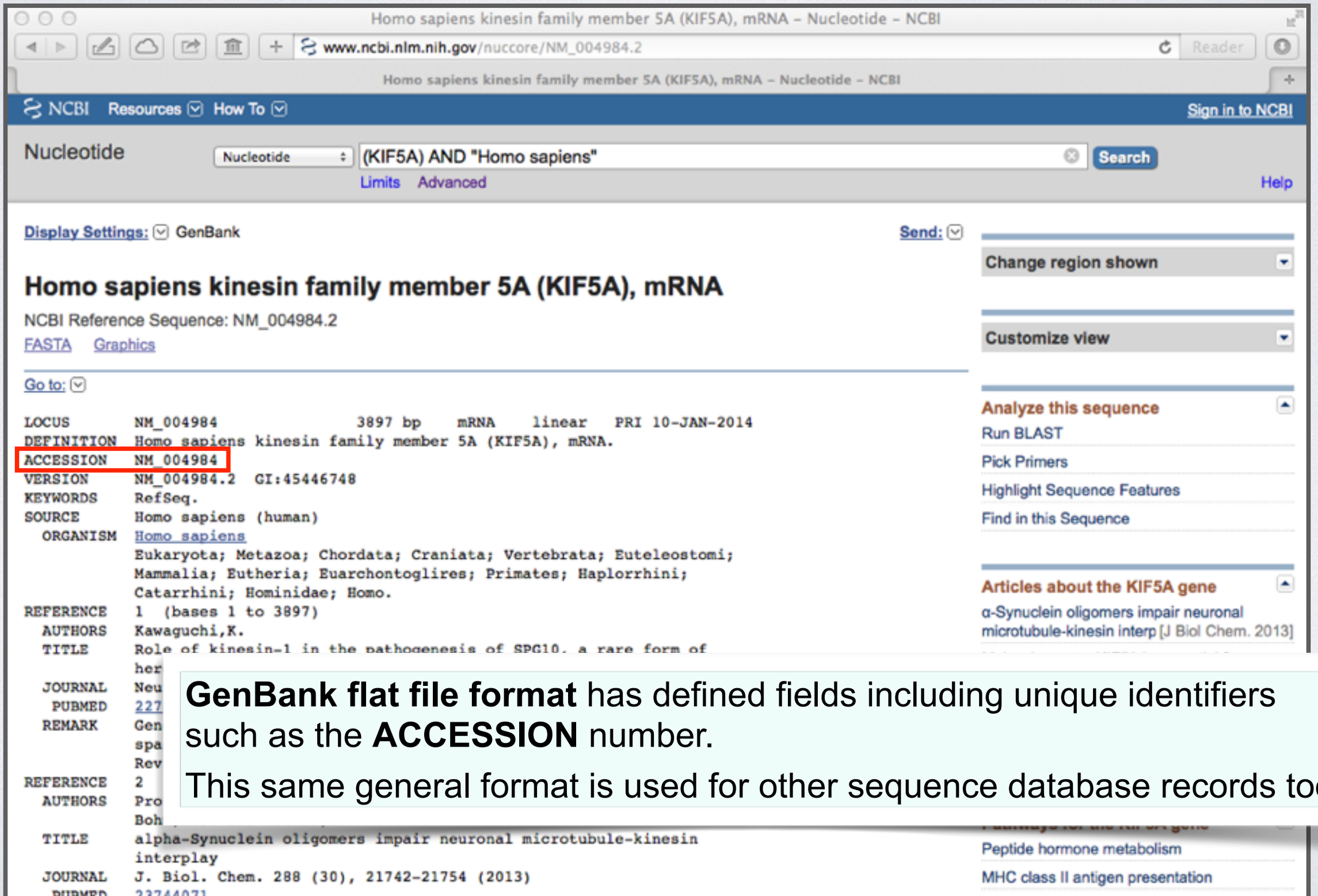
GenBank
Housed at NCBI
National Center for Biotechnology Information



DDBJ
Housed in Japan

- The underlying raw DNA sequences are identical
  - The different sites provide different views and ways to navigate through the data
- Access to GenBank (and other NCBI databases including RefSeq) is typically through **Entrez**, (the Google of NCBI) - more on this later

GenBank sequence record



GenBank flat file format has defined fields including unique identifiers such as the **ACCESSION** number.

This same general format is used for other sequence database records too.

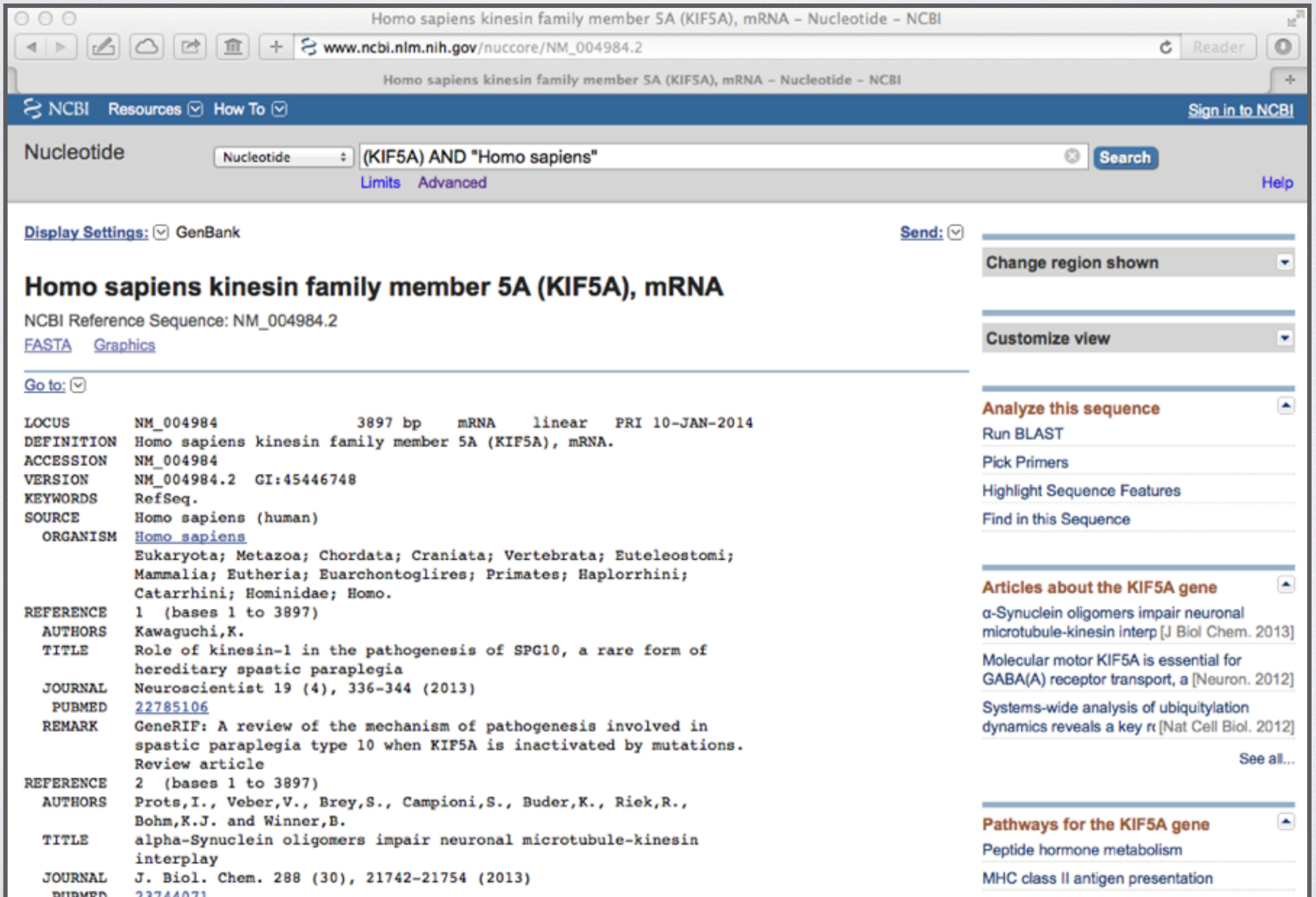
Side node: Database accession numbers

Database **accession numbers** are strings of letters and numbers used as **identifying labels** for sequences and other data within databases

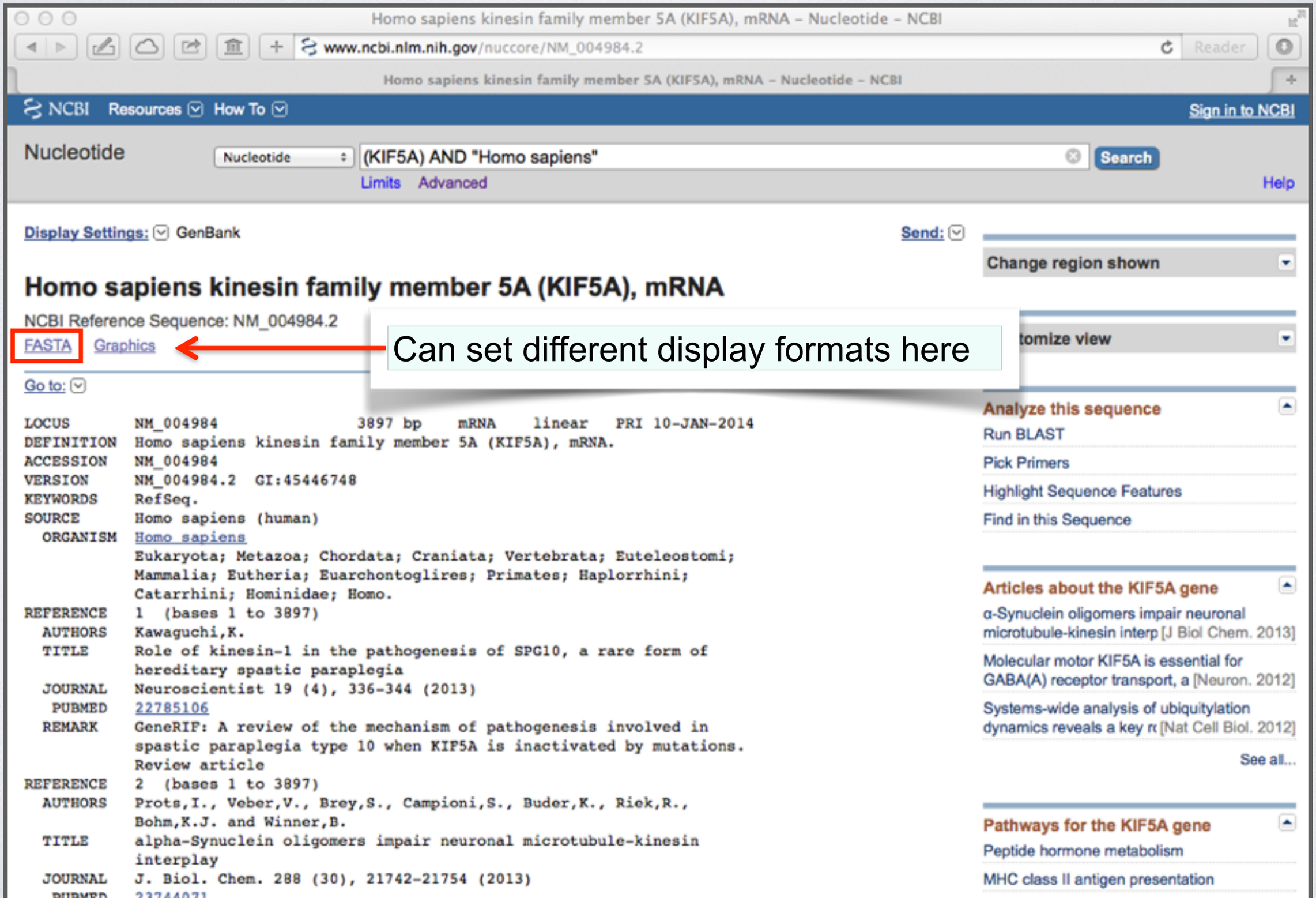
- Examples (all for retinol-binding protein, RBP4):

X02775	GenBank genomic DNA sequence	DNA
NT_030059	Genomic contig	
N91759.1	An expressed sequence tag (1 of 170)	RNA
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	Protein
AAC02945	GenBank protein	
Q28369	UniProtKB/SwissProt protein	
1KT7	Protein Data Bank structure record	
PMID: 12205585	PubMed IDs identify articles at NCBI/NIH	Literature

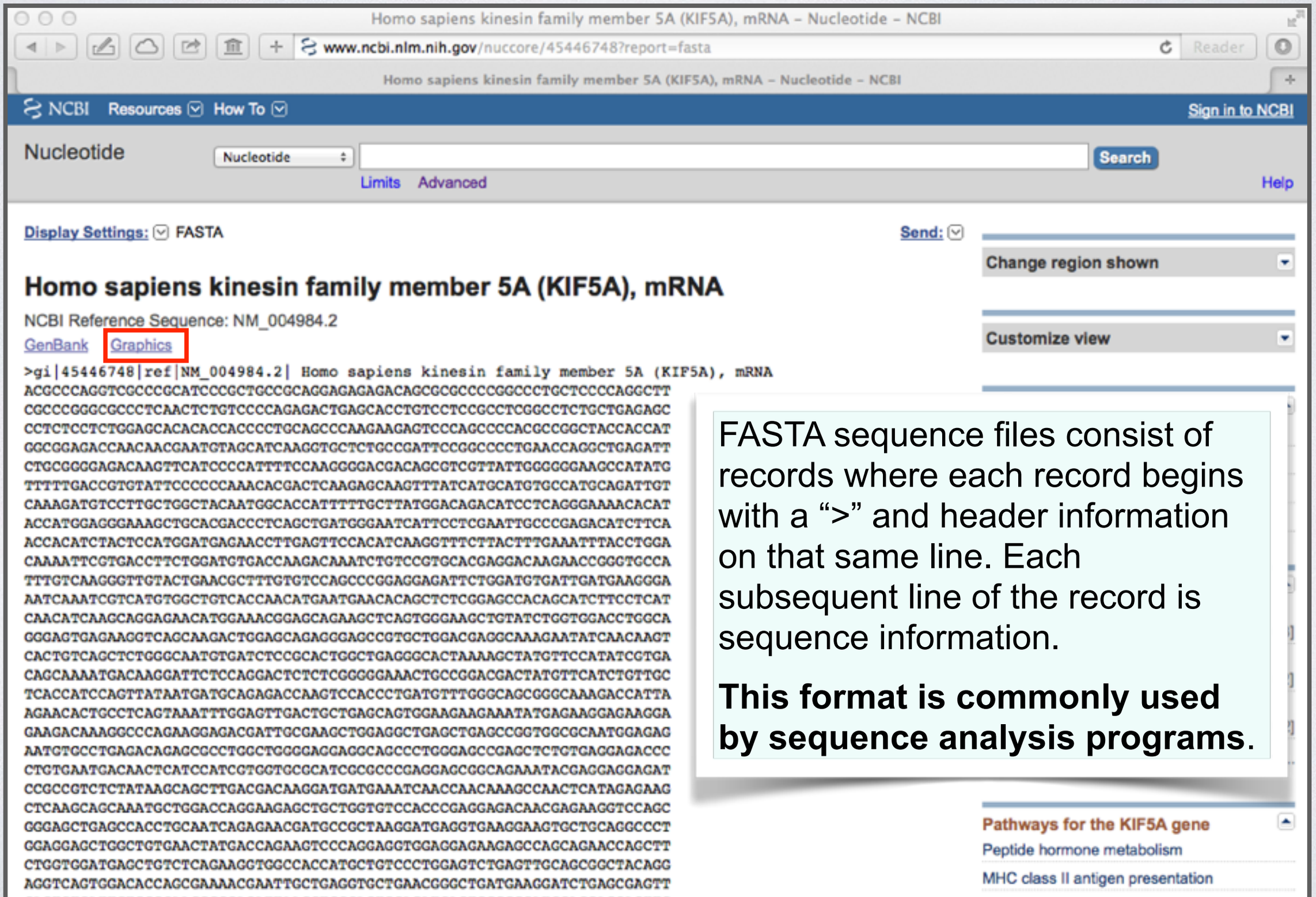
GenBank sequence record



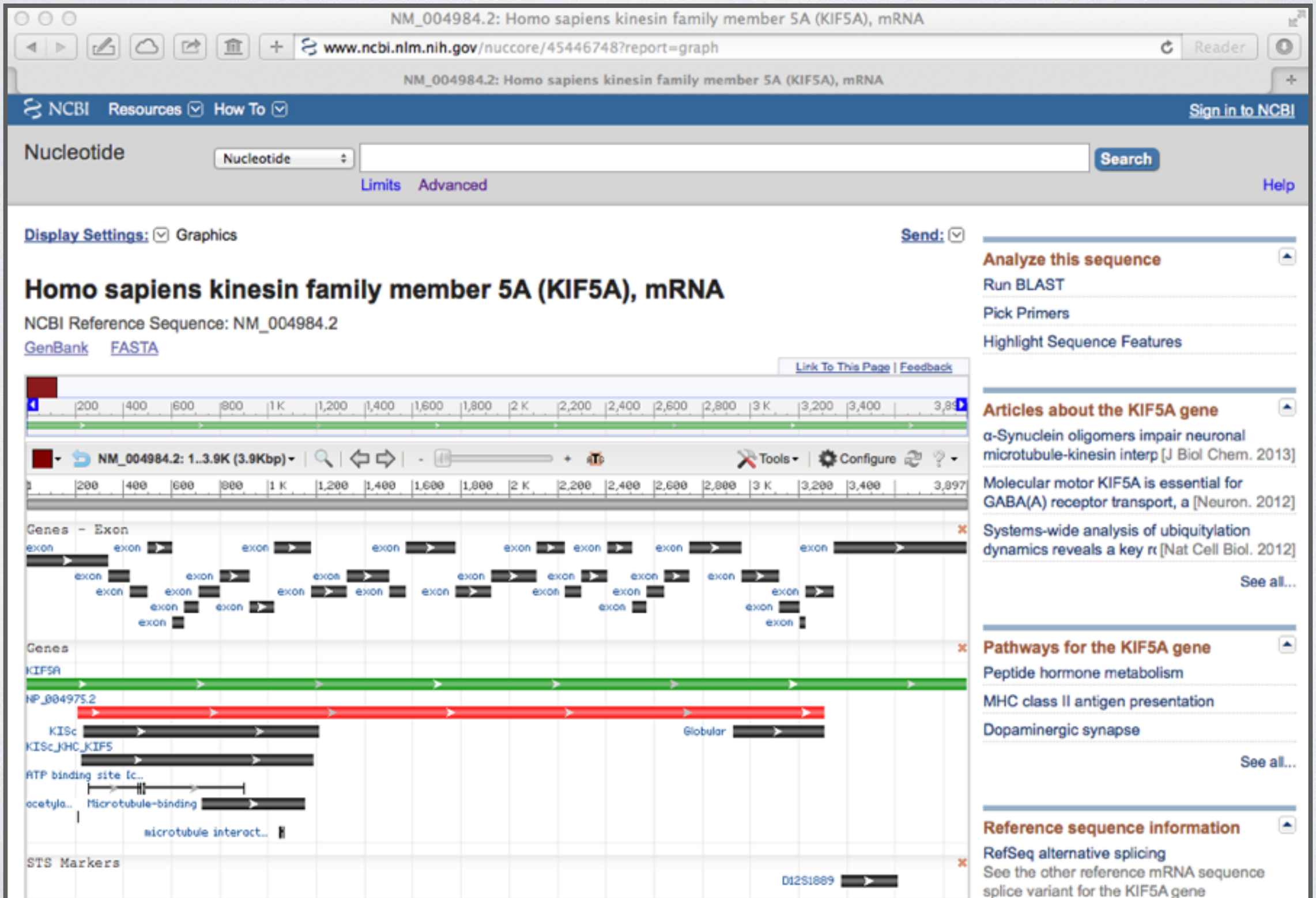
GenBank sequence record



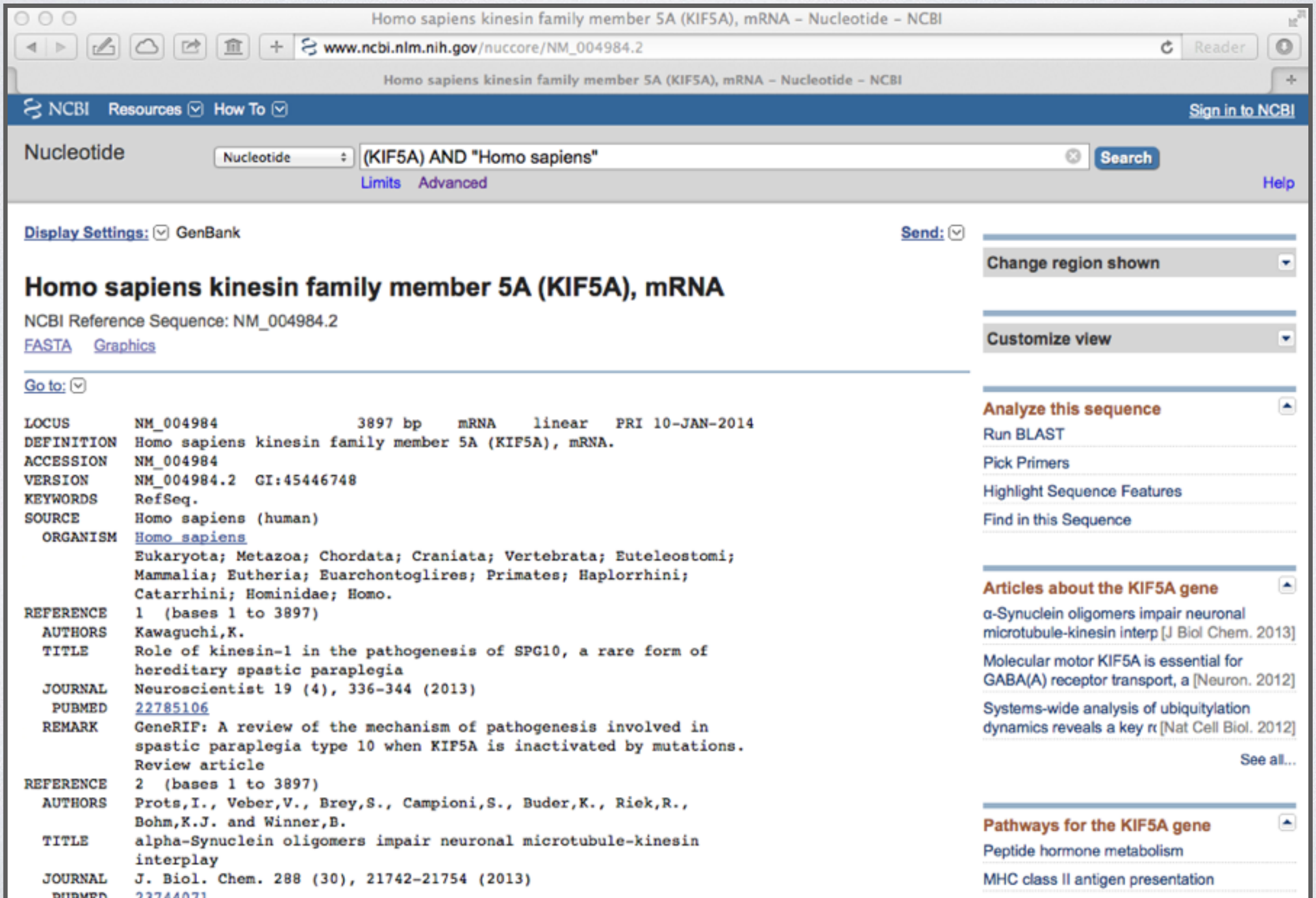
FASTA sequence record



GenBank ‘graphics’ sequence record



GenBank sequence record, cont.



GenBank sequence record, cont.

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nuccore/45446748?report=genbank&to=3897#feature_45446748

Reader

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

```
FEATURES             Location/Qualifiers
     source          1..3897
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="12"
                     /map="12q13.13"
     gene            1..3897
                     /gene="KIF5A"
                     /gene_synonym="D12S1889; MY050; NKHC; SPG10"
                     /note="kinesin family member 5A"
                     /db_xref="GeneID:3798"
                     /db_xref="HGNC:6323"
                     /db_xref="HPRD:09108"
                     /db_xref="MIM:602821"
     exon            1..337
                     /gene="KIF5A"
                     /gene_synonym="D12S1889; MY050; NKHC; SPG10"
                     /inference="alignment:Splign:1.39.8"
     misc_feature    134..136
                     /gene="KIF5A"
                     /gene_synonym="D12S1889; MY050; NKHC; SPG10"
                     /note="upstream in-frame stop codon"
     CDS             209..3307
                     /gene="KIF5A"
                     /gene_synonym="D12S1889; MY050; NKHC; SPG10"
                     /note="kinesin, heavy chain, neuron-specific; KIF5A
                     variant protein; neuronal kinesin heavy chain; kinesin
                     heavy chain neuron-specific 1"
                     /codon_start=1
                     /product="kinesin heavy chain isoform 5A"
                     /protein_id="NP_004975.2"
                     /db_xref="GI:45446749"
                     /db_xref="CCDS:CCDS8945.1"
                     /db_xref="GeneID:3798"
                     /db_xref="HGNC:6323"
                     /db_xref="HPRD:09108"
                     /db_xref="MIM:602821"
                     /translation="MAETNNECSIKVLCRFRPLNQAEILRGDKFIPIFQGDDSVVIGG
                     KPYVFDRVFPPNTTQEQVYHACAMQIVKDVLAGYNGTIFAYGQTSSGKTHTMEGKLHD
                     PQLMGIIPRIARDIFNHIYSMDENLEFHIKVSYFEIYLDKIRDLLDVTKTNLSVHEDK
                     NRVPFVKGCTERFVSSPEEILDVIDEGKSNRHVAVTNMNEHSSRSHSIFLINIKQENM
                     ETEQKLSGKLYLVDLAGSEKVSKTGAEGAVLDEAKNINKSLSALGNVISALAEGTKSY
                     VPYRDSKMTRILQDSLGGNCRTTMFICCSPSSYNDAETKSTLMFGQRAKTIKNTASVN
```

OMIM
Probe
Protein
PubMed
PubMed (RefSeq)

The **FEATURES** section contains annotations including a conceptual translation of the nucleotide sequence.

Recent activity

Turn Off Clear

Homo sapiens kinesin family member 5A (KIF5A), mRNA — Nucleotide

(kinesin) AND "Homo sapiens"[porgn] (1351) — Nucleotide

kinesin (37064) — Nucleotide

See more...

GenBank sequence record, cont.

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nuccore/45446748?report=genbank&to=3897#sequence_45446748

Reader

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

```
                     /gene_synonym="D12S1889; MY050; NKHC; SPG10"
                     /standard_name="D12S1889"
                     /db_xref="UniSTS:48006"
ORIGIN
        1 acgcccaggt cgcccgcatc ccgctgccgc aggagagaga cagcgcgccc cggccctgct
       61 ccccaggctt cgcccgggcg ccctcaactc tgtccccaga gactgagcac ctgtcctccg
      121 cctcggcctc tgctgagagc cctctcctct ggagcacaca ccacccctgc agcccaagaa
      181 gagtcccagc cccacgccgg ctaccaccat ggcggagacc aacaacgaat gtagcatcaa
      241 ggtgctctgc cgattccggc ccctgaacca ggctgagatt ctgcggggag acaagttcat
      301 ccccattttc caaggggacg acagcgtcgt tattgggggg aagccatatg tttttgaccg
      361 tgtattcccc ccaaacacga ctcaagagca agtttatcat gcatgtgcca tgcagattgt
      421 caaagatgtc cttgctggct acaatggcac cattttgct tatggacaga catcctcagg
      481 gaaaacacat accatggagg gaaagctgca cgaccctcag ctgatgggaa tcattcctcg
      541 aattgcccga gacatcttca accacatcta ctccatggat gagaaccttg agttccacat
      601 caaggtttct tactttgaaa tttacctgga caaaattcgt gaccttctgg atgtgaccaa
      661 gacaaatctg tccgtgcacg aggacaagaa ccgggtgcca tttgtcaagg gttgtactga
      721 acgctttgtg tccagcccgg aggagattct ggatgtgatt gatgaaggga aatcaaatcg
      781 tcatgtggct gtcaccaaca tgaatgaaca cagctctcgg agccacagca tcttcctcat
      841 caacatcaag caggagaaca tggaaacgga gcagaagctc agtgggaagc tgtatctggt
      901 ggacctggca gggagtgaga aggtcagcaa gactggagca gagggagccg tgctggacga
      961 ggcaaagaat atcaacaagt cactgtcagc tctgggcaat gtgatctccg cactggctga
     1021 gggcactaaa agctatgttc catatcgtga cagcaaaatg acaaggattc tccaggactc
     1081 tctcggggga aactgccgga cgactatgtt catctgttgc tcaccatcca gttataatga
     1141 tgcagagacc aagtccaccc tgatgtttgg gcagcgggca aagaccatta agaacactgc
     1201 ctcagtaaat ttggagttga ctgctgagca gtggaagaag aaatatgaga aggagaagga
     1261 gaagacaaag gcccagaagg agacgattgc gaagctggag gctgagctga gccggtggcg
     1321 caatggagag aatgtgcctg agacagagcg cctggctggg gaggaggcag ccctgggagc
     1381 cgagctctgt gaggagaccc ctgtgaatga caactcatcc atcgtggtgc gcatcgcgcc
     1441 cgaggagcgg cagaaatacg aggaggagat ccgccgtctc tataagcagc ttgacgacaa
     1501 ggatgatgaa atcaaccaac aaagccaact catagagaag ctcaagcagc aaatgctgga
     1561 ccaggaagag ctgctggtgt ccacccgagg agacaacgag aaggtccagc gggagctgag
     1621 ccacctgcaa tcagagaacg atgccgctaa ggatgaggtg aaggaagtgc tgcaggccct
     1681 ggaggagctg gctgtgaact atgaccagaa gtcccaggag gtggaggaga agagccagca
     1741 gaaccagctt ctggtggatg agctgtctca gaaggtggcc accatgctgt ccctggagtc
     1801 tgagttgcag cggctacagg aggtcagtgg acaccagcga aaacgaattg ctgaggtgct
     1861 gaacgggctg atgaaggatc tgagcgagtt cagtgtcatt gtgggcaacg gggagattaa
     1921 gctgccagtg gagatcagtg gggccatcga ggaggagttc actgtggccc gactctacat
     1981 cagcaaaatc aaatcagaag tcaagtctgt ggtcaagcgg tgccggcagc tggagaacct
     2041 ccaggtggag tgtcaccgca agatggaagt gaccgggcgg gagctctcat cctgccagct
     2101 cctcatctct cagcatgagg ccaagatccg ctcgcttacg gaatacatgc agagcgtgga
     2161 gctaaagaag cggcacctgg aagagtccta tgactccttg agcgatgagc tggccaagct
     2221 ccaggcccag gaaactgtgc atgaagtggc cctgaaggac aaggagcctg acactcagga
     2281 tgcagatgaa gtgaagaagg ctctggagct gcagatggag agtcaccggg aggcccatca
     2341 ccggcagctg gcccggctcc gggacgagat caacgagaag cagaagacca ttgatgagct
```

The actual sequence entry starts after the word **ORIGIN**

RefSeq: NCBI's Derivative Sequence Database

- RefSeq entries are hand curated best representation of a transcript or protein (in their judgement)
- Non-redundant for a given species although alternate transcript forms will be included if there is good evidence

- Experimentally verified transcripts and proteins
  accession numbers begin with "NM_" or "NP_"
- Model transcripts and proteins based on bioinformatics predictions with little experimental support
  accession numbers begin with "XM_" or "XP_"
- RefSeq also contains contigs and chromosome records

UNIPROT:

THE PREMIER PROTEIN SEQUENCE DATABASE

UniProt: Protein sequence database

UniProt is a comprehensive, high-quality resource of protein sequence and functional information

- UniProt comprises four databases:

1. UniProtKB (Knowledgebase)

Containing **Swiss-Prot** and **TrEMBL** components (these correspond to hand curated and automatically annotated entries respectively)

2. UniRef (Reference Clusters)

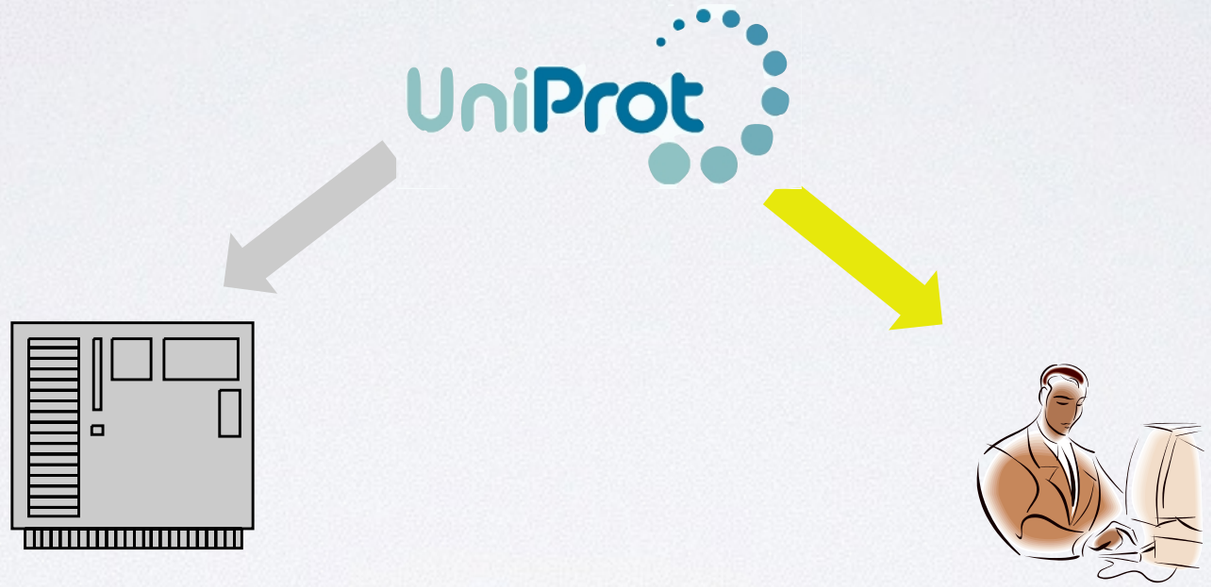
Filtered version of UniProtKB at various levels of sequence identity

e.g. UniRef90 contains sequences with a maximum of 90% sequence identity to each other

3. UniParc (Archive) with database cross-references to source.

4. UniMES (Metagenomic and Environmental Sequences)

The two sides of UniProtKB



UniProtKB/**TrEMBL**

Redundant, automatically annotated - unreviewed

UniProtKB/**Swiss-Prot**

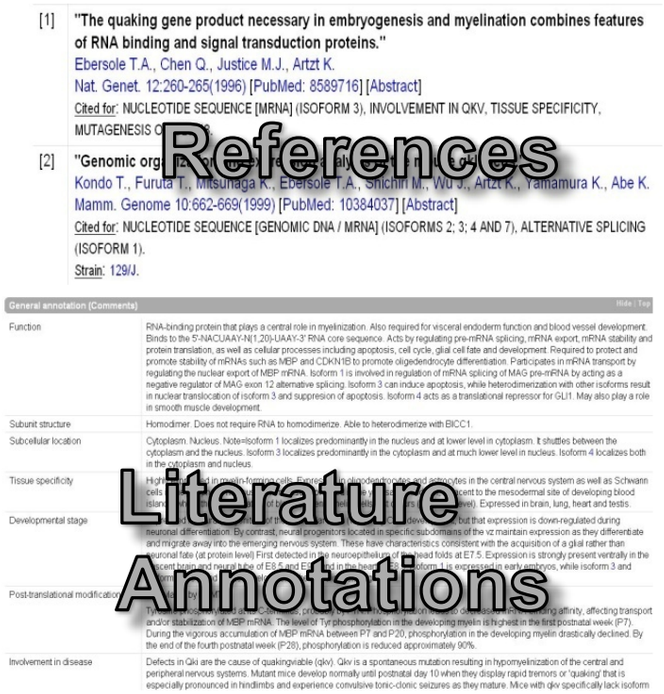
Non-redundant, high-quality manual annotation - reviewed

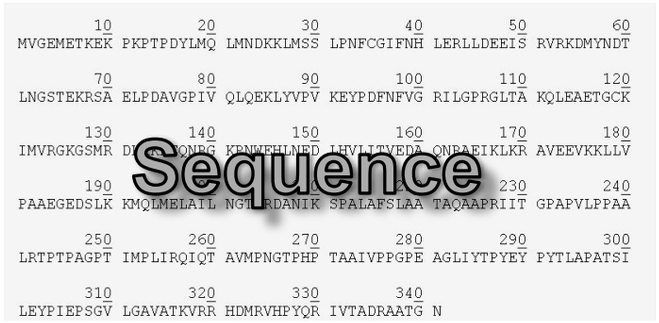
☆ Unreviewed, UniProtKB/TrEMBL **Q9N0H9** (Q9N0H9_EQUAS)

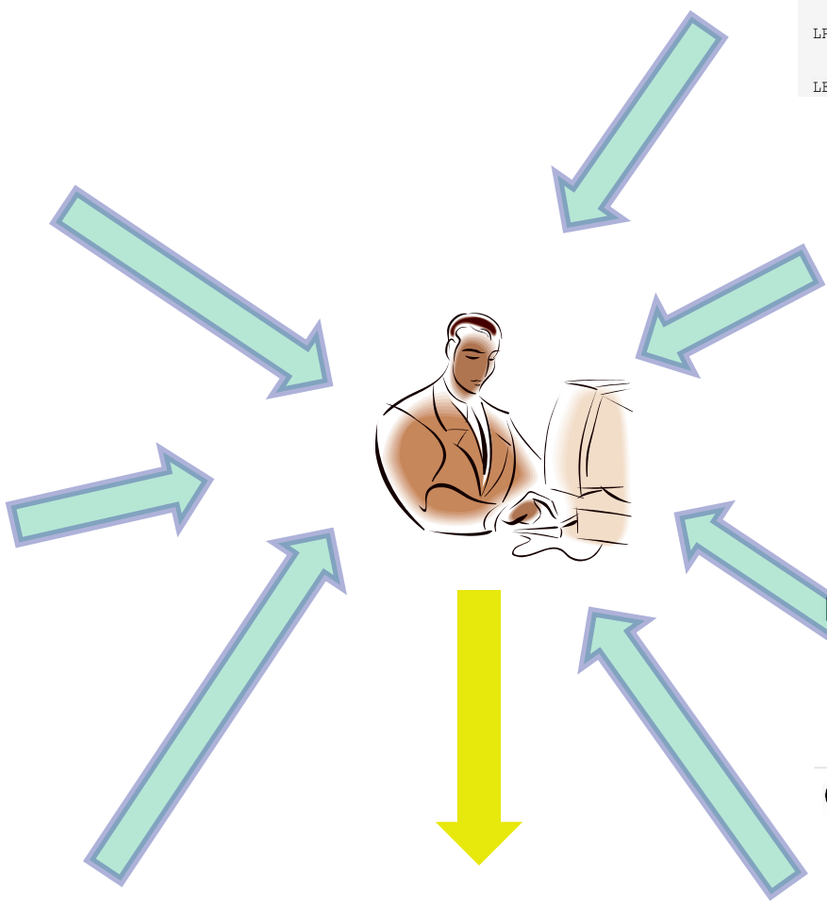
★ Reviewed, UniProtKB/Swiss-Prot **P38398** (BRCA1_HUMAN)

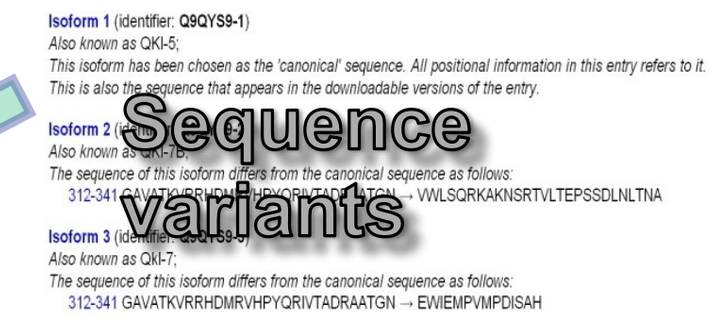
Indicators of which part of UniProt an entry belongs to include the color of the stars and the ID

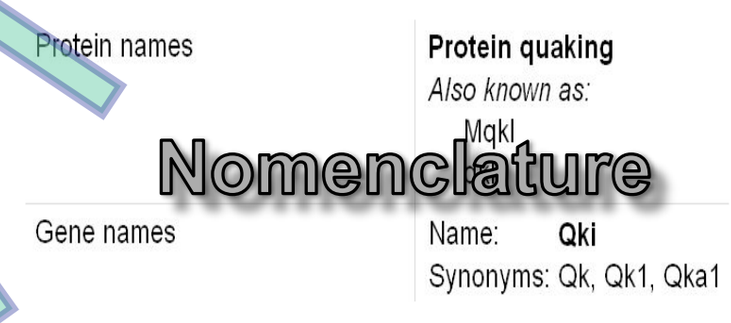
The main information added to a UniProt/Swiss-Prot entry

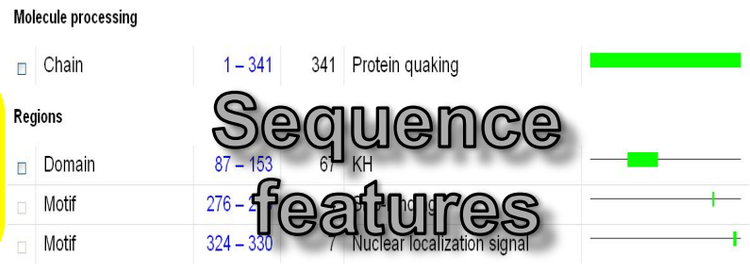












Organism-specific databases

AGD
BuruList
CGD
CYGD
dictyBase
EchoBASE
EcoGene
euHCVdb
FlyBase
GeneDB_Spombe
GeneFarm
GeneLynx
Gramene
H-InvDB
HGNC
HPA
LegioList
Leproma
ListiList
MaizeGDB
MGI
MIM
MypuList
Orphanet
PharmGKB
PseudoCAP
PhotoList
RGD
SagaList
SGD
StyGene
SubtiList
TAIR
TubercuList
WormBase
WormPep
ZFIN

Sequence databases

EMBL
PIR
RefSeq
UniGene

Enzyme and pathway databases

BioCyc
Reactome

Family and domain databases

Gene3D
HAMAP
InterPro
PANTHER
PIRSF
Pfam
PRINTS
ProDom
PROSITE
SMART
TIGRFAMs

2D-gel databases

2DBase-Ecoli
ANU-2DPAGE
Aarhus/Ghent-2DPAGE
COMPLUYEAST-2DPAGE
Cornea-2DPAGE
DOSAC-COBS-2DPAGE
ECO2DBASE
HSC-2DPAGE
OGP
PHCI-2DPAGE
PMMA-2DPAGE
Rat-heart-2DPAGE
REPRODUCTION-2DPAGE
Siena-2DPAGE
SWISS-2DPAGE
World-2DPAGE

UniProt

UniProt provides cross-references to a large number other resources and can serve as a useful “portal” when you first begin to investigate a particular protein

Genome annotation databases

Ensembl
GenomeReviews
GeneID
KEGG
NMPDR
TIGR
Vectorbase

PTM databases

GlycoSuiteDB
PhosSite
PhosphoSite

3D structure databases

DisProt
HSSP
PDB
PDBsum
SMR

Protein family/group databases

GermOnline
MEROPS
PeroxiBase
PptaseDB
REBASE

Miscellaneous

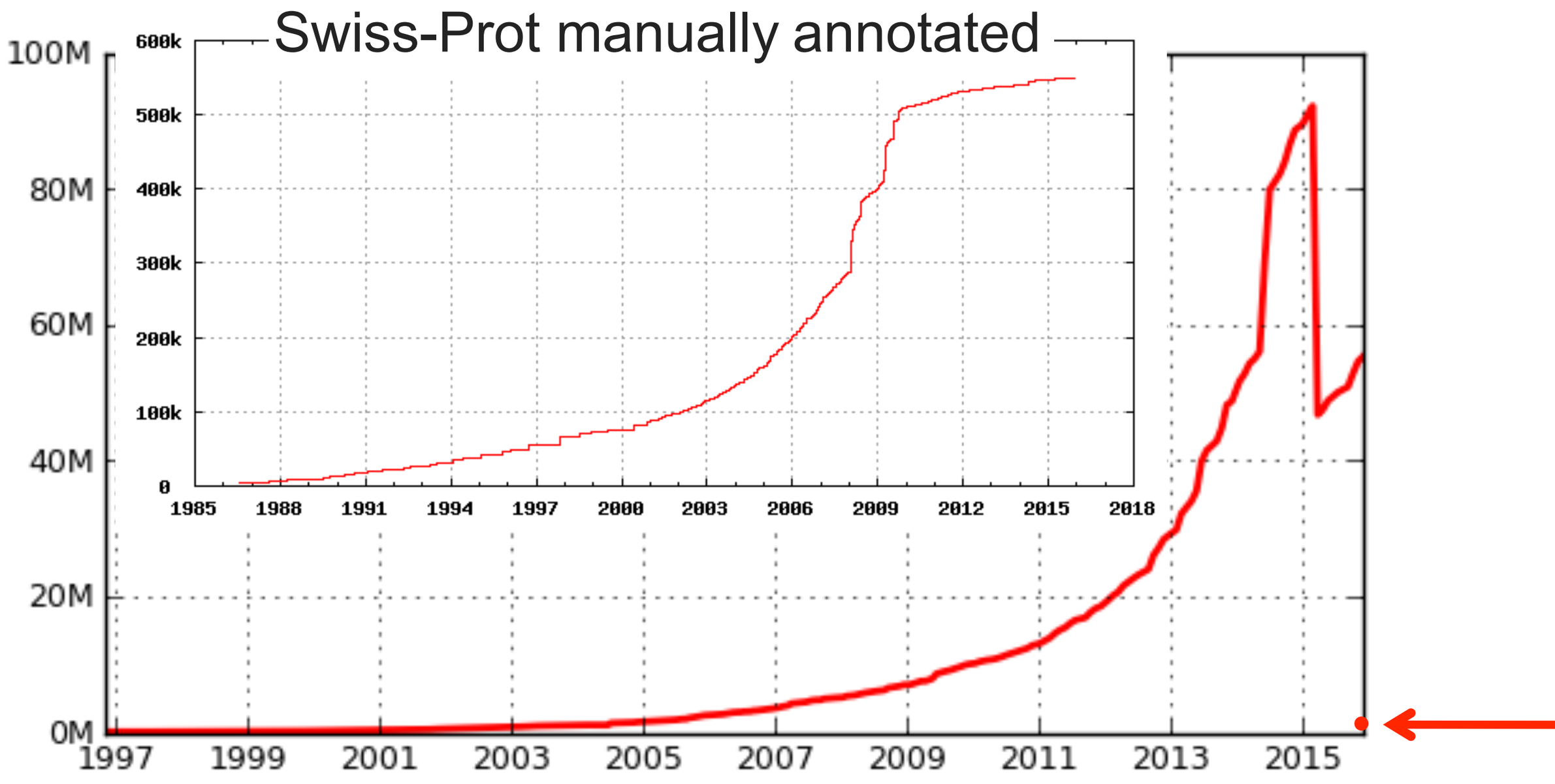
ArrayExpress
BindingDB
CleanEx
dbSNP
DIP
DrugBank
GO
HOGENOM
HOVERGEN
IntAct
LinkHub
PeptideAtlas
ProMEX

UniProt/Swiss-Prot *vs* UniProt/TrEMBL

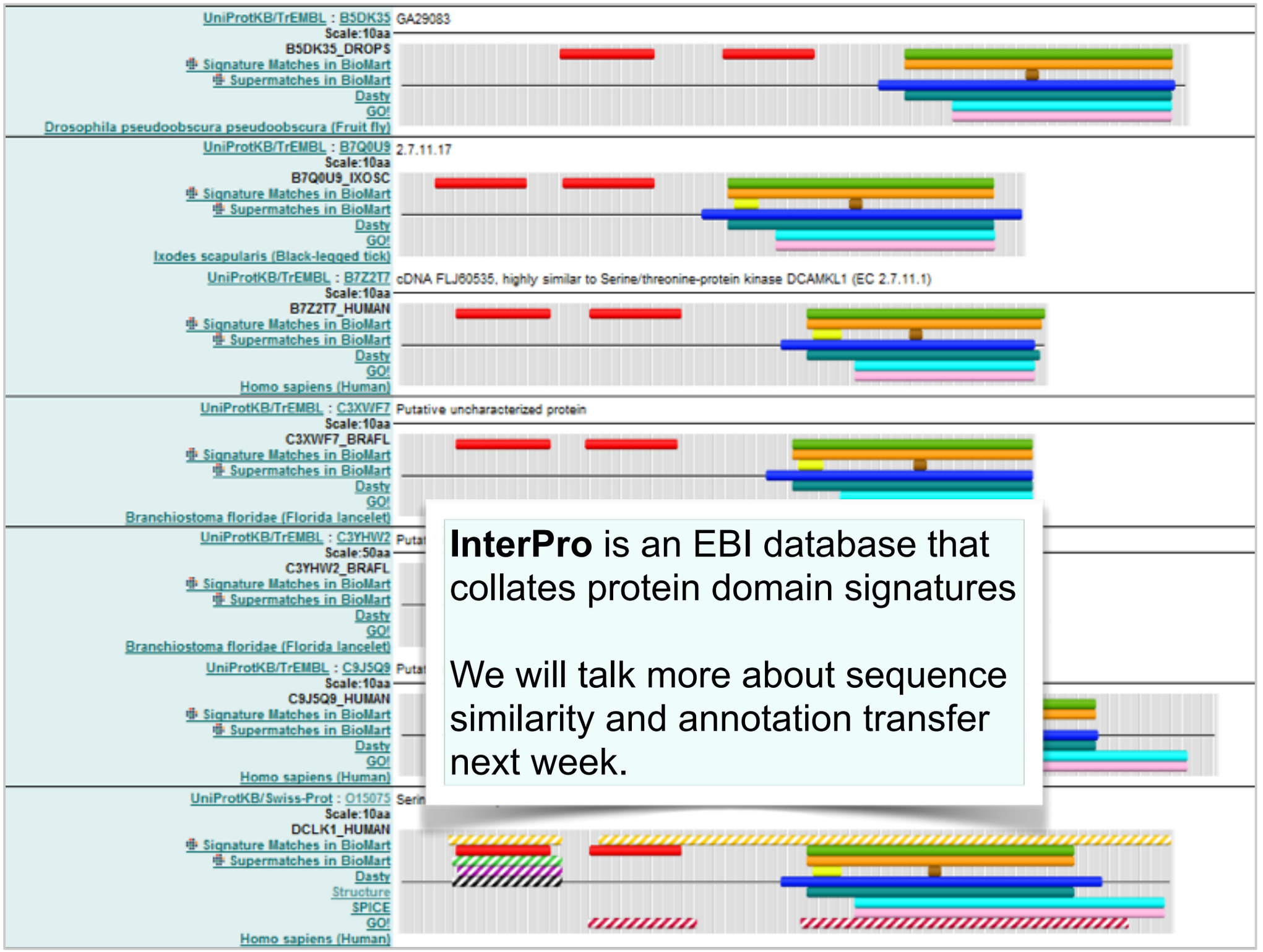
- *UniProtKB/Swiss-Prot* is a **non-redundant** database with one entry per protein

- *UniProtKB/TrEMBL* is a **redundant** database with one entry per translated ENA entry (ENA is the EBI's equivalent of GenBank)
  - Therefore TrEMBL can contain multiple entries for the same protein
  - Multiple UniProtKB/TrEMBL entries for the same protein can arise due to:
    - Erroneous gene model predictions
    - Sequence errors (Frame shifts)
    - Polymorphisms
    - Alternative start sites
    - Isoforms
    - OR because the same sequence was submitted by different people

Side note: Automatic Annotation (sharing the wealth)



Same domain composition
= same function = annotation transfer



InterPro is an EBI database that collates protein domain signatures

We will talk more about sequence similarity and annotation transfer next week.

DATABASE VIGNETTE

You have just come out a seminar about gastric cancer and one of your co-workers asks:

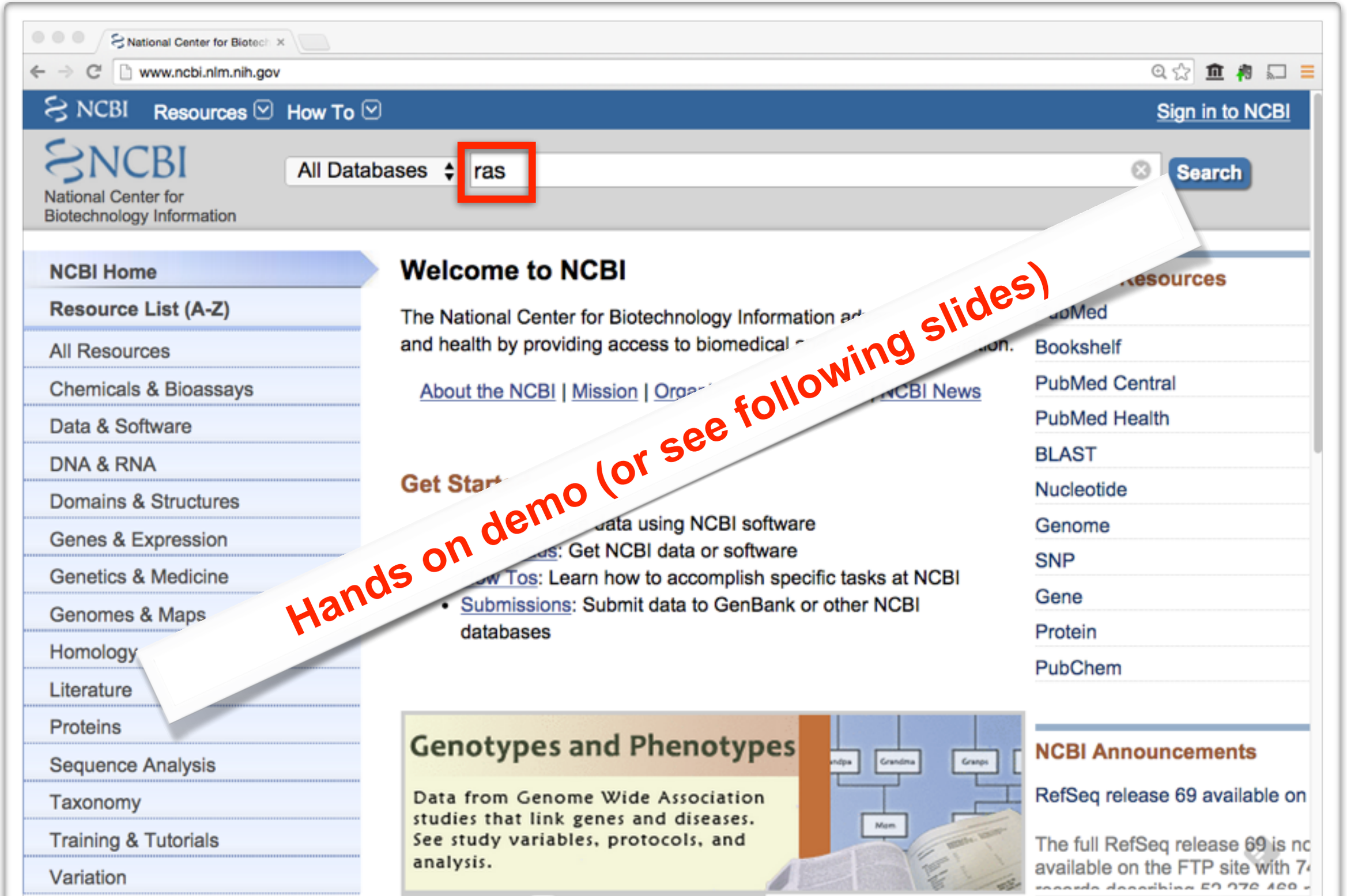
> *"What do you know about that 'Kras' gene the speaker kept taking about?*"

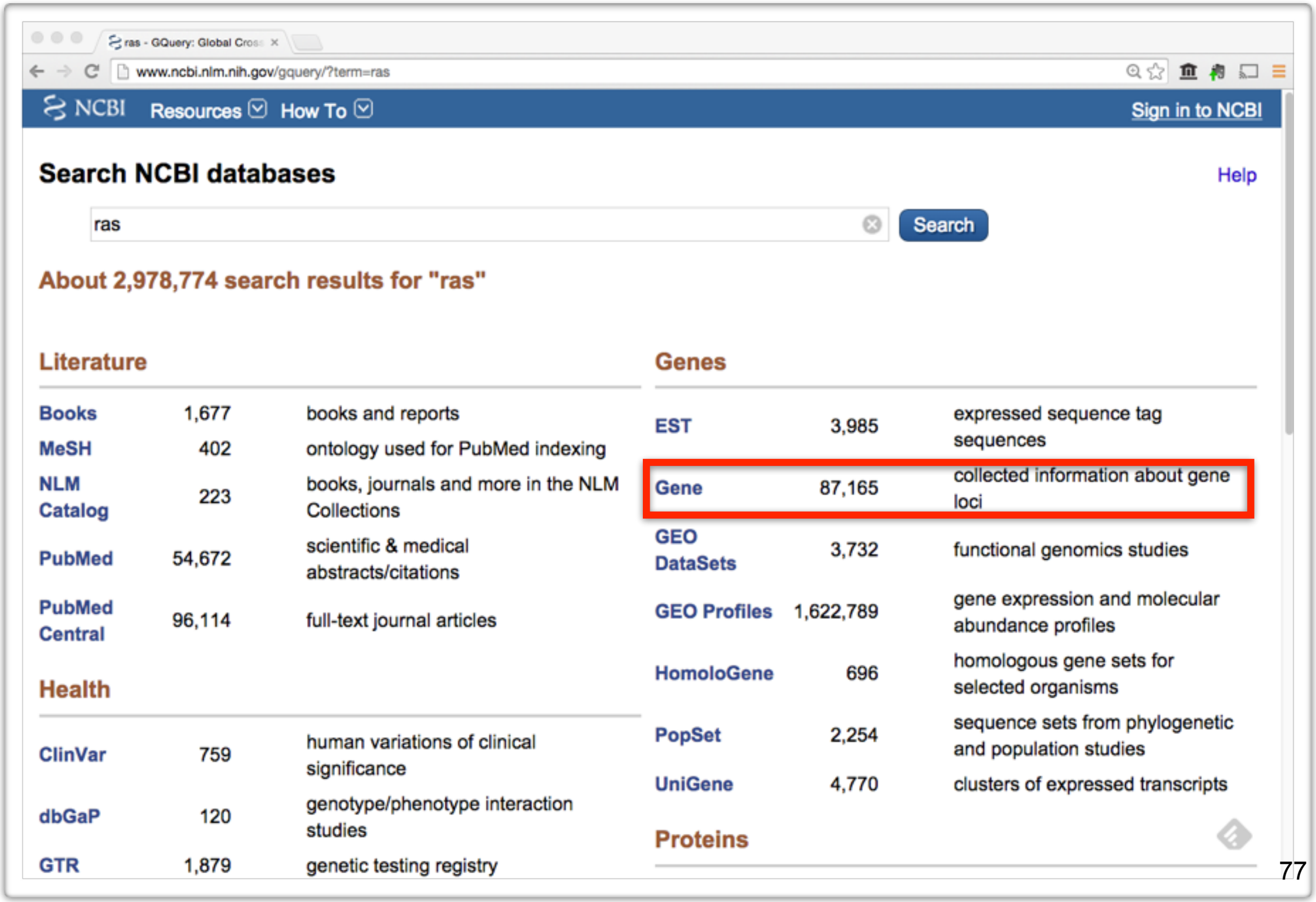
You have some recollection about hearing of 'Ras' before. How would you find out more?

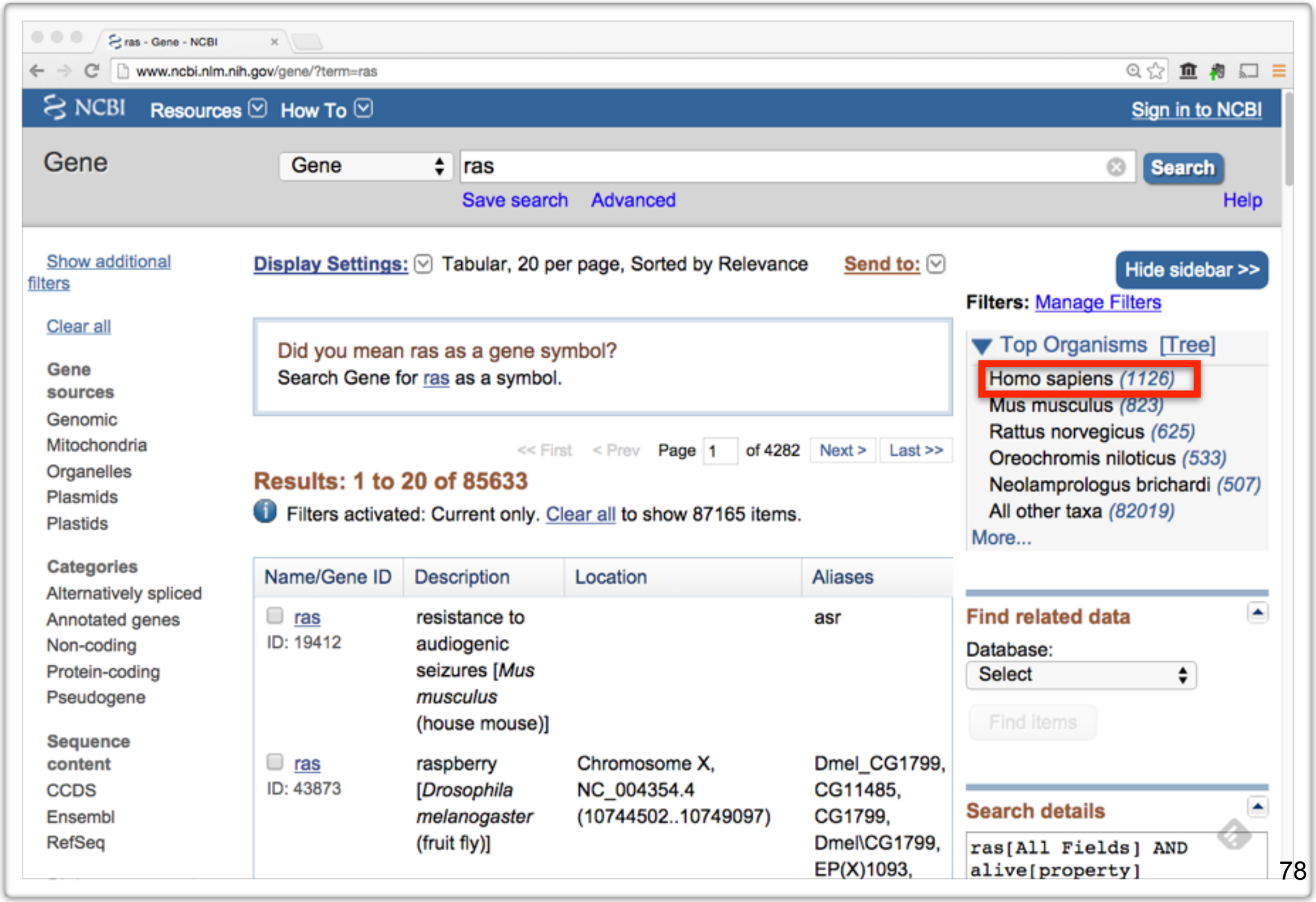
- Google?
- Library?
- **Bioinformatics databases at NCBI and EBI!**

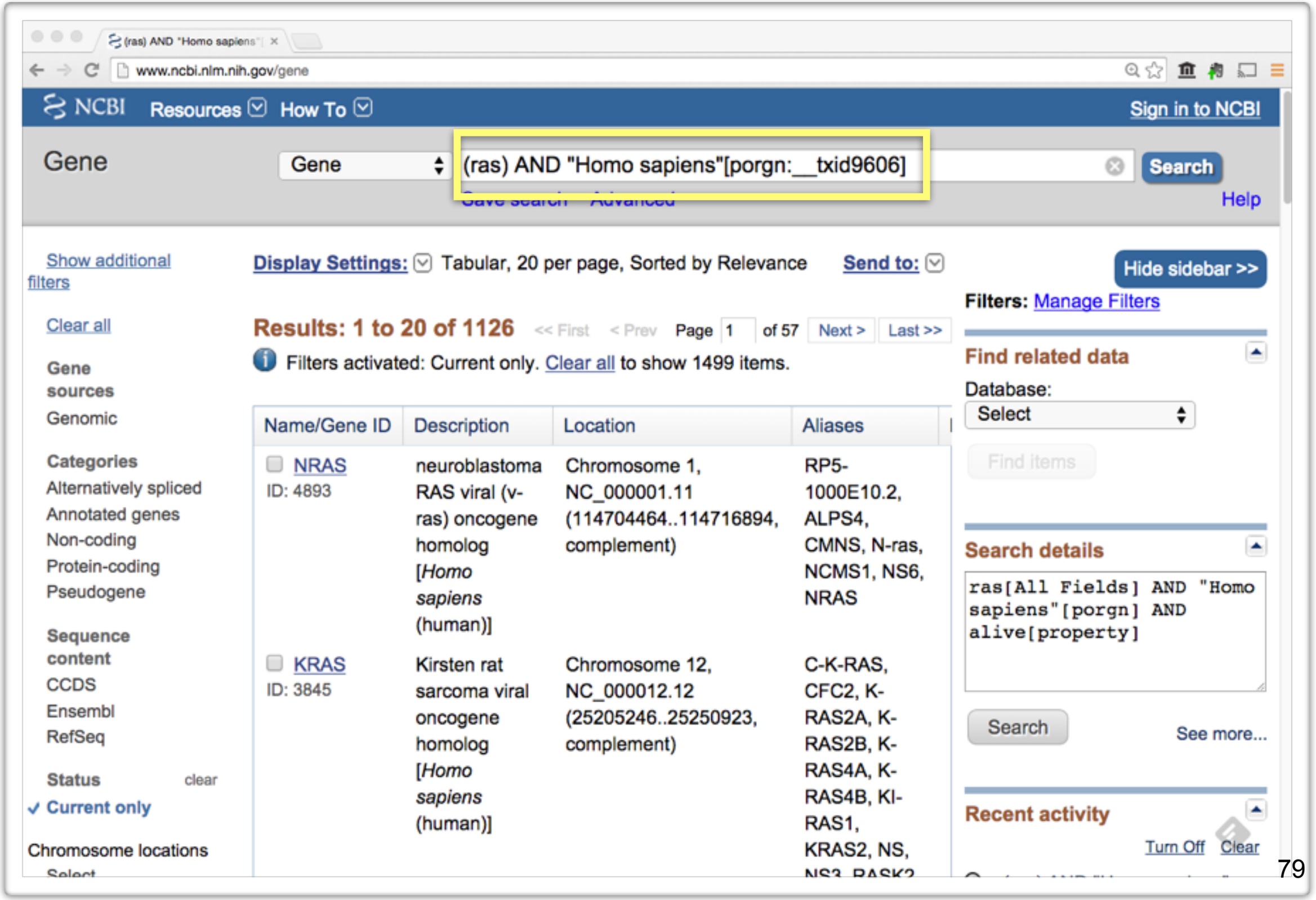
http://www.ncbi.nlm.nih.gov/

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

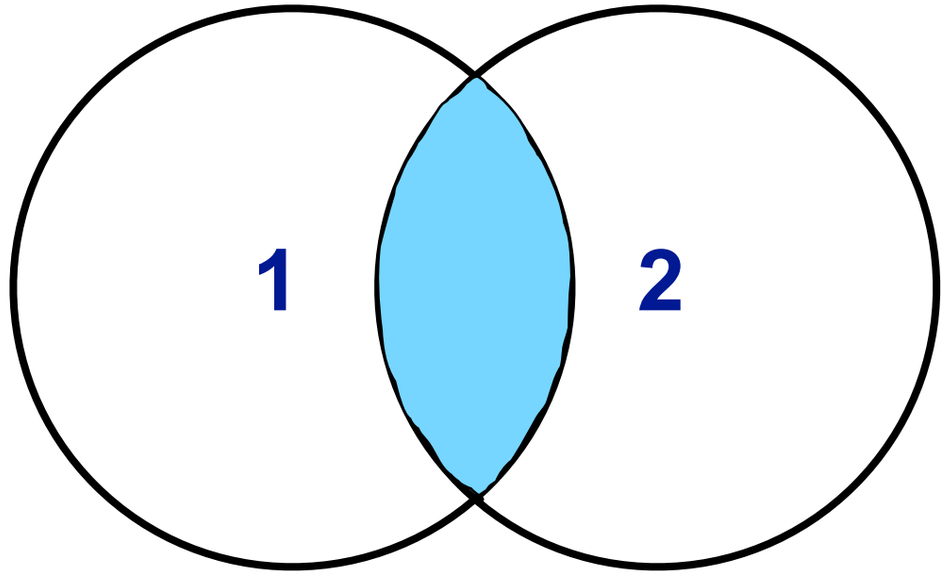



ras - GQuery: Global Cross
www.ncbi.nlm.nih.gov/gquery/?term=ras
NCBI
Resources
How To
Sign in to NCBI
Search NCBI databases
Help
ras
Search
About 2,978,774 search results for "ras"
Literature
Books 1,677 books and reports
MeSH 402 ontology used for PubMed indexing
NLM Catalog 223 books, journals and more in the NLM Collections
PubMed 54,672 scientific & medical abstracts/citations
PubMed Central 96,114 full-text journal articles
Health
ClinVar 759 human variations of clinical significance
dbGaP 120 genotype/phenotype interaction studies
GTR 1,879 genetic testing registry
Genes
EST 3,985 expressed sequence tag sequences
Gene 87,165 collected information about gene loci
GEO DataSets 3,732 functional genomics studies
GEO Profiles 1,622,789 gene expression and molecular abundance profiles
HomoloGene 696 homologous gene sets for selected organisms
PopSet 2,254 sequence sets from phylogenetic and population studies
UniGene 4,770 clusters of expressed transcripts
Proteins


ras - Gene - NCBI
www.ncbi.nlm.nih.gov/gene/?term=ras
NCBI
Resources
How To
Sign in to NCBI
Gene
Gene
ras
Search
Save search
Advanced
Help
Show additional filters
Display Settings: Tabular, 20 per page, Sorted by Relevance
Send to:
Hide sidebar >>
Filters: Manage Filters
Clear all
Did you mean ras as a gene symbol?
Search Gene for ras as a symbol.
Top Organisms [Tree]
Homo sapiens (1126)
Mus musculus (823)
Rattus norvegicus (625)
Oreochromis niloticus (533)
Neolamprologus brichardi (507)
All other taxa (82019)
More...
Gene sources
Genomic
Mitochondria
Organelles
Plasmids
Plastids
Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene
Sequence content
CCDS
Ensembl
RefSeq
<< First
< Prev
Page 1 of 4282
Next >
Last >>
Results: 1 to 20 of 85633
Filters activated: Current only. Clear all to show 87165 items.
Name/Gene ID
Description
Location
Aliases
ras
ID: 19412
resistance to audiogenic seizures [Mus musculus (house mouse)]
asr
ras
ID: 43873
raspberry [Drosophila melanogaster (fruit fly)]
Chromosome X, NC_004354.4 (10744502..10749097)
Dmel_CG1799, CG11485, CG1799, Dmel\CG1799, EP(X)1093,
Find related data
Database:
Select
Find items
Search details
ras[All Fields] AND alive[property]

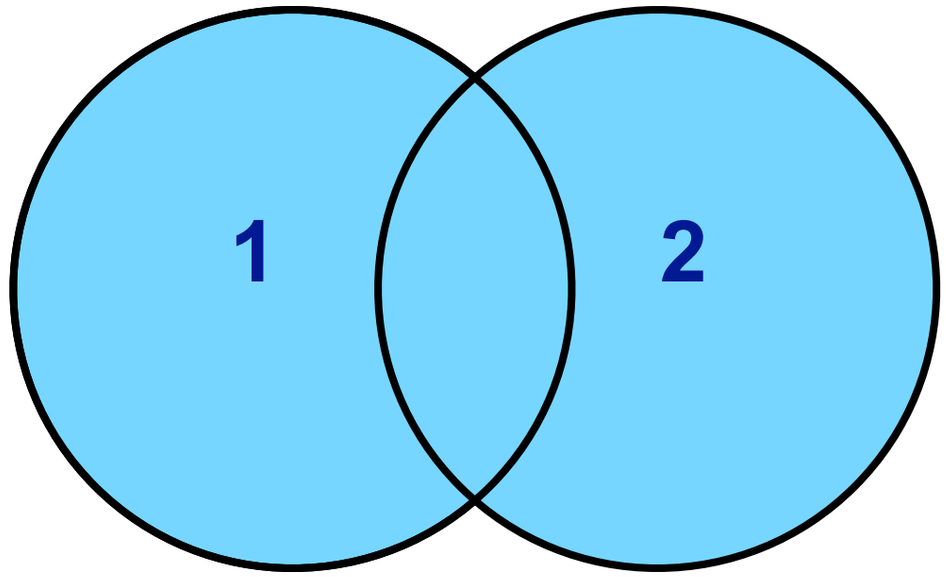

(ras) AND "Homo sapiens"
www.ncbi.nlm.nih.gov/gene
NCBI
Resources
How To
Sign in to NCBI
Gene
Gene
(ras) AND "Homo sapiens"[porgn:__txid9606]
Search
Help
Show additional filters
Display Settings: Tabular, 20 per page, Sorted by Relevance
Send to:
Hide sidebar >>
Clear all
Results: 1 to 20 of 1126
<< First
< Prev
Page 1
of 57
Next >
Last >>
Filters: Manage Filters
Filters activated: Current only. Clear all to show 1499 items.
Gene sources
Genomic
Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene
Sequence content
CCDS
Ensembl
RefSeq
Status
clear
Current only
Chromosome locations
Select
Name/Gene ID
Description
Location
Aliases
NRAS
ID: 4893
neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]
Chromosome 1, NC_000001.11 (114704464..114716894, complement)
RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS
ID: 3845
Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)]
Chromosome 12, NC_000012.12 (25205246..25250923, complement)
C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS1, KRAS2, NS,
Find related data
Database:
Select
Find items
Search details
ras[All Fields] AND "Homo sapiens"[porgn] AND alive[property]
Search
See more...
Recent activity
Turn Off
Clear

1 AND 2



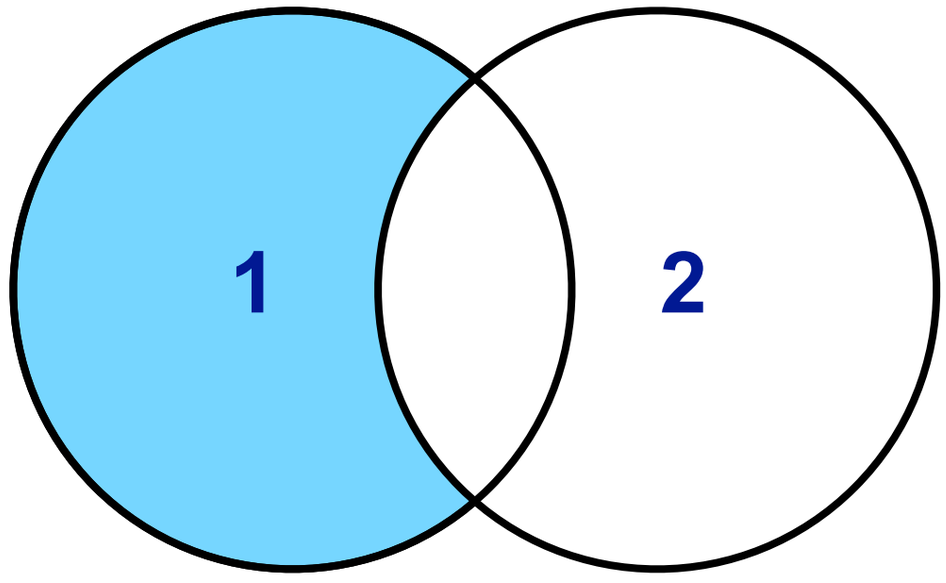
ras AND disease
(1185 results)

1 OR 2



ras OR disease
(134,872 results)

1 NOT 2



ras NOT disease
(84,448 results)

(ras) AND "Homo sapiens"

www.ncbi.nlm.nih.gov/gene

NCBI Resources How To Sign in to NCBI

Gene

Gene (ras) AND "Homo sapiens"[porgn:__txid9606] Search

Save search Advanced Help

Show additional filters

Display Settings: Tabular, 20 per page, Sorted by Relevance Send to:

Hide sidebar >>

Clear all

Results: 1 to 20 of 1126 << First < Prev Page 1 of 57 Next > Last >>

Filters activated: Current only. Clear all to show 1499 items.

Filters: Manage Filters

Gene sources

Genomic

Categories

Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene

Sequence content

CCDS
Ensembl
RefSeq

Status clear

✓ Current only

Chromosome locations

Select

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [*Homo sapiens* (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [*Homo sapiens* (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS1, KRAS2, NS,

Find related data

Database:

Select

Find items

Search details

```
ras[All Fields] AND "Homo
sapiens"[porgn] AND
alive[property]
```

Search See more...

Recent activity

Turn Off Clear

KRAS Kirsten rat sarcoma

www.ncbi.nlm.nih.gov/gene/3845

NCBI Resources How To Sign in to NCBI

Gene Gene Search

Advanced Help

Display Settings: Full Report Send to: Hide sidebar >>

KRAS Kirsten rat sarcoma viral oncogene homolog [*Homo sapiens* (human)]

Gene ID: 3845, updated on 4-Jan-2015

Summary

Official Symbol	KRAS provided by HGNC
Official Full Name	Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source	HGNC:HGNC:6407
See related	Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
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	NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Table of contents

KRAS Kirsten rat sarcoma

www.ncbi.nlm.nih.gov/gene/3845

NCBI Resources How To Sign in to NCBI

Gene Search Help

Display S Hide sidebar >>

Example Questions:

What chromosome location and what genes are in the vicinity?

KRAS

(human)

Gene ID: 3845, updated on 4-Jan-2015

Summary

Official Symbol	KRAS provided by HGNC
Official Full Name	Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source	HGNC:HGNC:6407
See related	Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
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	NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Table of contents

KRAS Kirsten rat sarcoma

www.ncbi.nlm.nih.gov/gene/3845#genomic-context

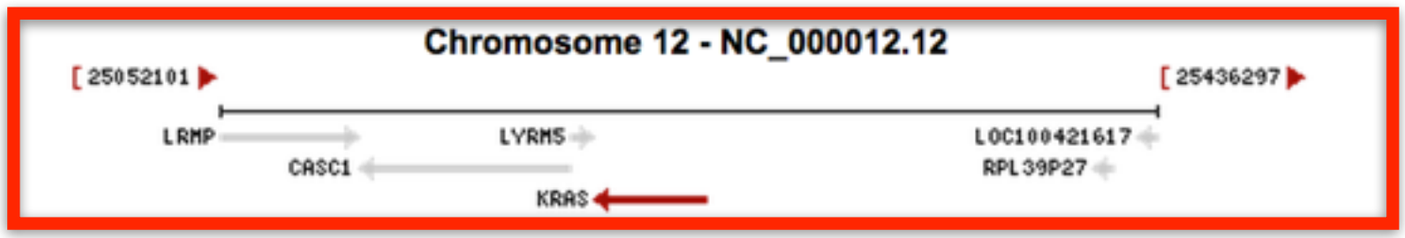
Genomic context

Location: 12p12.1

Exon count: 6

See KRAS in Epigenomics, MapViewer

Annotation release	Status	Assembly	Chr	Location
106	current	GRCh38 (GCF_000001405.26)	12	NC_000012.12 (25205246..25250923, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	12	NC_000012.11 (25358180..25403870, complement)



Genomic regions, transcripts, and products

Go to reference sequence details

Genomic Sequence: NC_000012.12 chromosome 12 reference GRCh38 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank

BioAssay by Target (List)
BioAssay by Target (Summary)
BioAssay, by Gene target
BioAssays, RNAi Target, Active
BioAssays, RNAi Target, Tested
BioProjects
BioSystems
Books
CCDS
ClinVar
Conserved Domains
dbVar
EST
Full text in PMC
Full text in PMC_nucleotide
Gene neighbors
Genome
GEO Profiles
GTR
HomoloGene
Map Viewer
MedGen
Nucleotide

KRAS Kirsten rat sarcoma

www.ncbi.nlm.nih.gov/gene/3845

NCBI Resources How To Sign in to NCBI

Gene Search Help

Display Settin

KRAS Ki
(human)]

Gene ID: 3845

Summa

Example Questions:

What 'molecular functions', 'biological processes', and 'cellular component' information is available?

Official Symbol	KRAS provided by HGNC
Official Full Name	Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source	HGNC:HGNC:6407
See related	Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
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	NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Hide sidebar >>

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Variation
- HIV-1 interactions
- Pathways from BioSystems
- Interactions
- General gene information
  - Markers, Related pseudogene(s), Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)

KRAS Kirsten rat sarcoma

www.ncbi.nlm.nih.gov/gene/3845#gene-ontology

Gene Ontology Provided by GOA

Function	Evidence Code	Pubs
GDP binding	IEA	
GMP binding	IEA	
GTP binding	IEA	
LRR domain binding	IEA	
protein binding	IPI	PubMed
protein complex binding	IDA	PubMed

Items 1 - 25 of 33 < Prev Page 1 of 2 Next >

Process	Evidence Code	Pubs
Fc-epsilon receptor signaling pathway	TAS	
GTP catabolic process	IEA	
MAPK cascade	TAS	
Ras protein signal transduction	TAS	
actin cytoskeleton organization	IEA	
activation of MAPKK activity	TAS	
axon guidance	TAS	
blood coagulation	TAS	

GO: Gene Ontology

GO provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data

KRAS Kirsten rat sarcoma | UniProt-GOA < EMBL-EBI

www.ebi.ac.uk/GOA

EMBL-EBI

Services | Research | Training | About us

UniProt-GOA

Search

Examples: GO:0006915, tropomyosin, P06727

Overview | New to UniProt-GOA | FAQ | Contact Us

Gene Ontology Annotation (UniProt-GOA) Database

The UniProt GO annotation program aims to provide high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB). The assignment of GO terms to UniProt records is an integral part of UniProt biocuration . UniProt manual and electronic GO annotations are supplemented with manual annotations supplied by external collaborating GO Consortium groups, to ensure a comprehensive GO annotation dataset is supplied to users .

UniProt is a member of the GO Consortium .

Menu

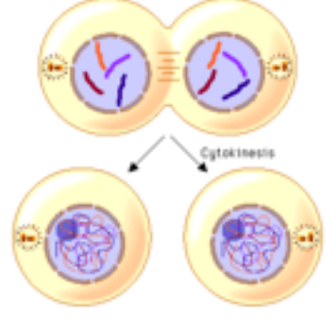
- Downloads
- Searching UniProt-GOA
- Annotation Methods
  - Annotation Tutorial
  - Manual Annotation Efforts
    - Reference Genome Annotation Initiative
    - Cardiovascular Gene Ontology Annotation Initiative
    - Renal Gene Ontology Annotation Initiative
    - Exosome Gene

Why do we need Ontologies?

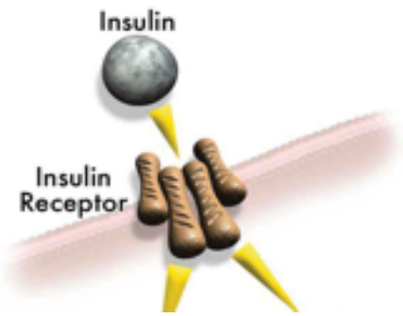
- Annotation is essential for capturing the understanding and knowledge associated with a sequence or other molecular entity
- Annotation is traditionally recorded as “free text”, which is easy to read by humans, but has a number of disadvantages, including:
  - Difficult for computers to parse
  - Quality varies from database to database
  - Terminology used varies from annotator to annotator
- Ontologies are annotations using standard vocabularies that try to address these issues
- GO is integrated with UniProt and many other databases including a number at NCBI

GO Ontologies

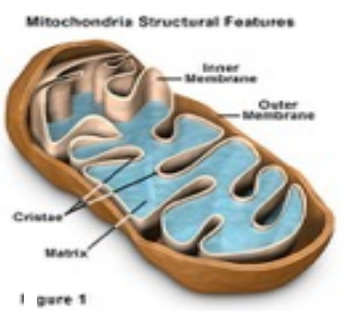
- There are three ontologies in GO:
  - **Biological Process**
    A commonly recognized series of events
    e.g. cell division, mitosis,



  - **Molecular Function**
    An elemental activity, task or job
    e.g. kinase activity, insulin binding



  - **Cellular Component**
    Where a gene product is located
    e.g. mitochondrion, mitochondrial membrane



KRAS Kirsten rat sarcoma

www.ncbi.nlm.nih.gov/gene/3845#gene-ontology

Gene Ontology Provided by GOA

Function	Evidence Code	Pubs
GDP binding		
GMP binding		
GTP binding		
LRR domain binding		
protein binding		
protein complex binding		

Process	Evidence Code	Pubs
Fc-epsilon receptor signaling pathway	TAS	
GTP catabolic process	IEA	
MAPK cascade	TAS	
Ras protein signal transduction	TAS	
actin cytoskeleton organization	IEA	
activation of MAPKK activity	TAS	
axon guidance	TAS	
blood coagulation	TAS	

The 'Gene Ontology' or GO is actually maintained by the EBI so lets switch or link over to UniProt also from the EBI.

Scroll down to **UniProt** link

UniProt will detail much more information for protein coding genes such as this one

KRAS Kirsten rat sarcoma

www.ncbi.nlm.nih.gov/gene/3845#gene-ontology

genomic	X01669.1	CAA25828.1

Items 1 - 25 of 43 < Prev Page 1 of 2 Next >

Protein Accession	Links	
	GenPept Link	UniProtKB Link
P01116.1	GenPept	UniProtKB/Swiss-Prot:P01116

Additional links

Scroll down to **UniProt** link

You are here: NCBI > Genes & Expression > Gene

Write to the Help Desk

GETTING STARTED
NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials

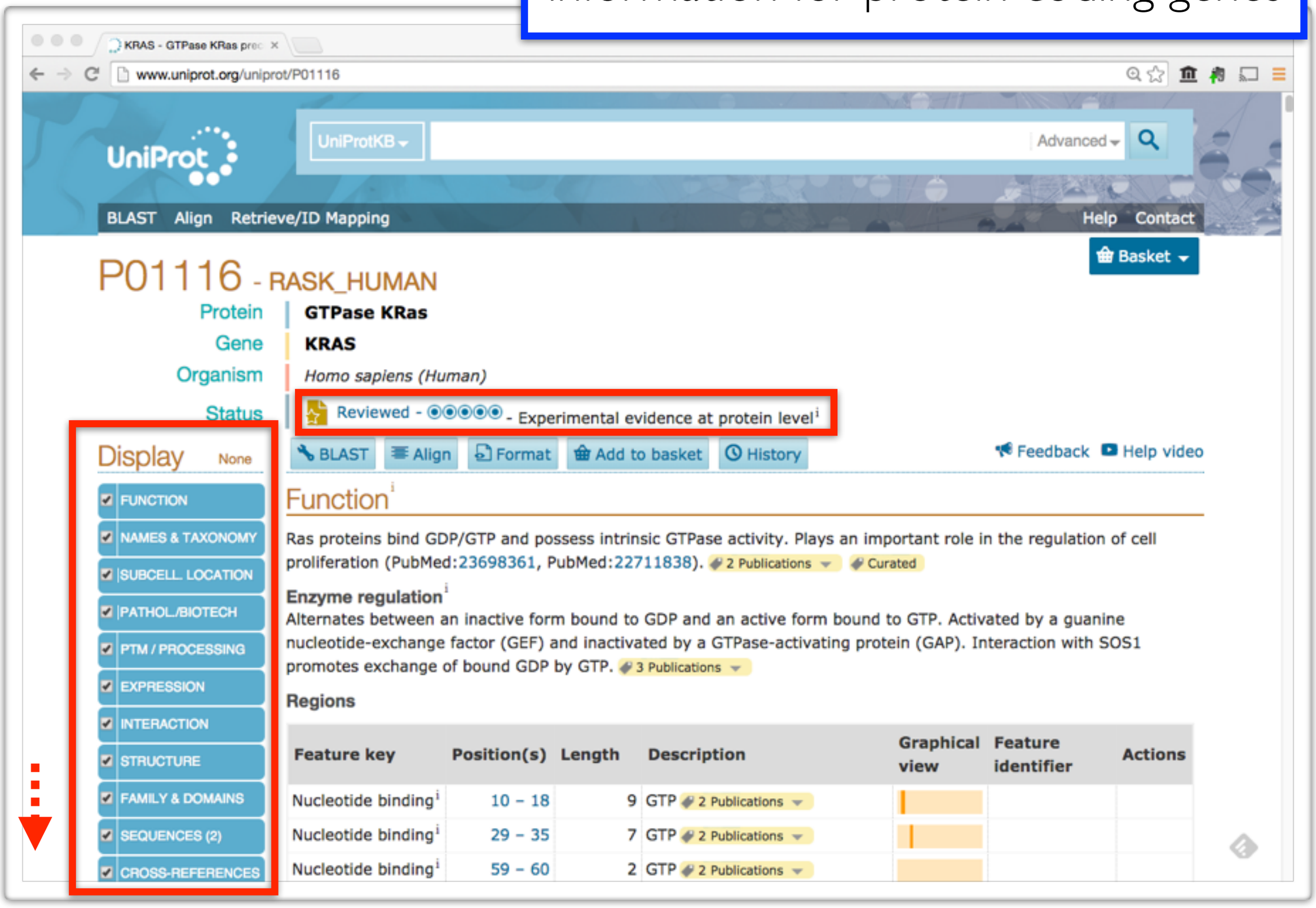
RESOURCES
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy

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

FEATURED
Genetic Testing Registry
PubMed Health
GenBank
Reference Sequences
Gene Expression Omnibus
Map Viewer
Human Genome
Mouse Genome
Influenza Virus
Primer-BLAST
Sequence Read Archive

NCBI INFORMATION
About NCBI
Research at NCBI
NCBI News
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter
NCBI on YouTube

UniProt will detail much more information for protein coding genes



KRAS - GTPase KRas proc

www.uniprot.org/uniprot/P01116

UniProt

UniProtKB | Advanced

BLAST | Align | Retrieve/ID Mapping | Help | Contact

Basket

P01116 - RASK_HUMAN

Protein	**GTPase KRas**
Gene	**KRAS**
Organism	*Homo sapiens (Human)*
Status	Reviewed - ◉◉◉◉◉

Example Questions:

What positions in the protein are responsible for GTP binding?

BLAST | Align | Fo

Display None

- FUNCTION
- NAMES & TAXONOMY
- SUBCELL. LOCATION
- PATHOL./BIOTECH
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (2)
- CROSS-REFERENCES

Function

Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23698361, PubMed:22711838). 2 Publications | Curated

Enzyme regulation

Alternates between an inactive form bound to GDP and an active form bound to GTP. Activated by a guanine nucleotide-exchange factor (GEF) and inactivated by a GTPase-activating protein (GAP). Interaction with SOS1 promotes exchange of bound GDP by GTP. 3 Publications

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding	10 – 18	9	GTP 2 Publications			
Nucleotide binding	29 – 35	7	GTP 2 Publications			
Nucleotide binding	59 – 60	2	GTP 2 Publications			

Example Questions:

What variants of this enzyme are involved in gastric cancer and other human diseases?

KRAS - GTPase KRas proc

www.uniprot.org/uniprot/P01116

Display None

- FUNCTION
- NAMES & TAXONOMY
- SUBCELL. LOCATION
- PATHOL./BIOTECH
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (2)
- CROSS-REFERENCES
- PUBLICATIONS
- ENTRY INFORMATION
- MISCELLANEOUS
- SIMILAR PROTEINS

▲Top

Pathology & Biotech

Involvement in disease

LEUKEMIA, ACUTE MYELOGENOUS (AML)

[MIM:601626]: A subtype of acute leukemia, a cancer of the white blood cells. AML is a malignant disease of bone marrow characterized by maturational arrest of hematopoietic precursors at an early stage of development. Clonal expansion of myeloid blasts occurs in bone marrow, blood, and other tissue. Myelogenous leukemias develop from changes in cells that normally produce neutrophils, basophils, eosinophils and monocytes. 1 Publication

Note: The disease is caused by mutations affecting the gene represented in this entry.

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Natural variant	10 – 10	1	G → GG in one individual with AML; expression in 3T3 cell causes cellular transformation; expression in COS cells activates the Ras-MAPK signaling pathway; lower GTPase activity; faster GDP dissociation rate. 1 Publication		VAR_034601	

LEUKEMIA, JUVENILE MYELOMONOCYTIC (JMML)

[MIM:607785]: An aggressive pediatric myelodysplastic syndrome/myeloproliferative disorder characterized by malignant transformation in the hematopoietic stem cell compartment with proliferation of differentiated progeny. Patients have splenomegaly, enlarged lymph nodes, rashes, and hemorrhages.

Note: The disease is caused by mutations affecting the gene represented in this entry.

NOONAN SYNDROME 3 (NS3)

[MIM:609942]: A form of Noonan syndrome, a disease characterized by short stature, facial dysmorphic features such as hypertelorism, a downward eyeslant and low-set posteriorly rotated ears, and a high incidence of congenital heart

Example Questions:

Are high resolution protein structures available to examine the details of these mutations?

KRAS - GTPase KRas proc

www.uniprot.org/uniprot/P01116

Display None

- FUNCTION
- NAMES & TAXONOMY
- SUBCELL. LOCATION
- PATHOL./BIOTECH
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (2)
- CROSS-REFERENCES
- PUBLICATIONS
- ENTRY INFORMATION
- MISCELLANEOUS
- SIMILAR PROTEINS

▲Top

Structure

Secondary structure

1 189

Legend: Helix Turn Beta strand

Show more details

3D structure databases

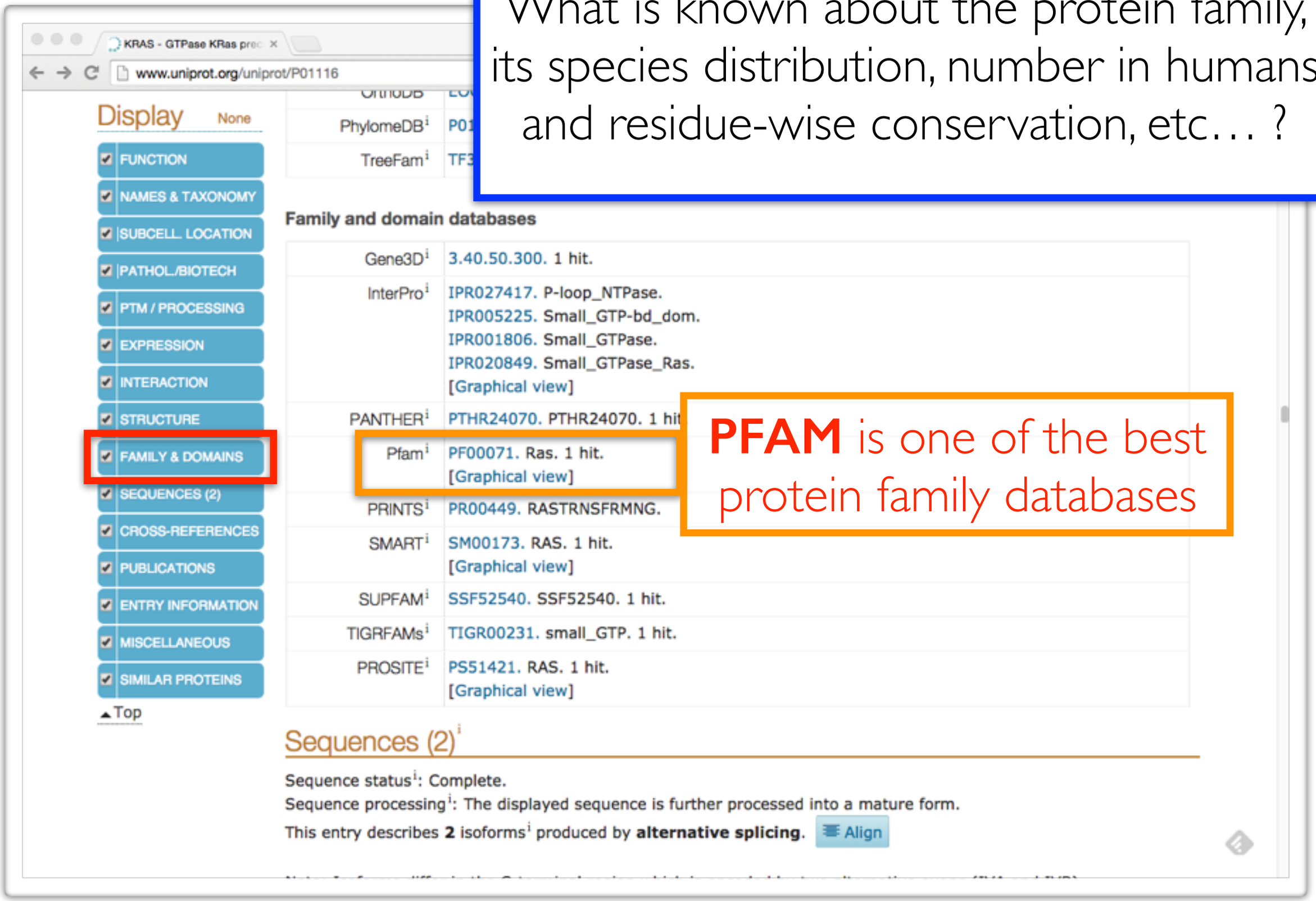
Select the link destinations:

- PDBe
- RCSB PDB
- PDBj

Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
1D8D	X-ray	2.00	P	178-188	[»]
1D8E	X-ray	3.00	P	178-188	[»]
1KZO	X-ray	2.20	C	169-173	[»]
1KZP	X-ray	2.10	C	169-173	[»]
3GFT	X-ray	2.27	A/B/C/D/E/F	1-164	[»]
4DSN	X-ray	2.03	A	2-164	[»]
4DSO	X-ray	1.85	A	2-164	[»]
4EPR	X-ray	2.00	A	1-164	[»]
4EPT	X-ray	2.00	A	1-164	[»]
4EPV	X-ray	1.35	A	1-164	[»]
4EPW	X-ray	1.70	A	1-164	[»]
4EPX	X-ray	1.76	A	1-164	[»]
4EPY	X-ray	1.80	A	1-164	[»]
4L8G	X-ray	1.52	A	1-169	[»]
4LDJ	X-ray	1.15	A	1-164	[»]
4LPK	X-ray	1.50	A/B	1-169	[»]

Example Questions:

What is known about the protein family, its species distribution, number in humans and residue-wise conservation, etc… ?



PFAM is one of the best protein family databases

Example Questions:

What is known about the protein family, its **species distribution**, number in humans and residue-wise conservation, etc… ?

KRAS - GTPase KRas prec | Pfam: Family: Ras (PF0007

pfam.xfam.org/family/PF00071

EMBL-EBI HOME | SEARC

Family: *Ras* (PF00071)

332 architectures | 21243 sequences | 30 interactions | 1006 species | 663 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to...

enter ID/acc Go

Summary: Ras family

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: Ras subfamily | Wikipedia: Ras superfamily | Pfam | InterPro

This is the Wikipedia entry entitled "Ras subfamily". More...

Ras subfamily

Edit Wikipedia article

This article is about p21/Ras protein. For the p21/waf1 protein, see p21.

Ras is the name given to a family of related proteins which is ubiquitously expressed in all cell lineages and organs. All Ras protein family members belong to a class of protein called small GTPase, and are involved in transmitting signals within cells (cellular signal transduction). Ras is the prototypical member of the Ras superfamily of proteins, which are all related in 3D structure and regulate diverse cell behaviours.

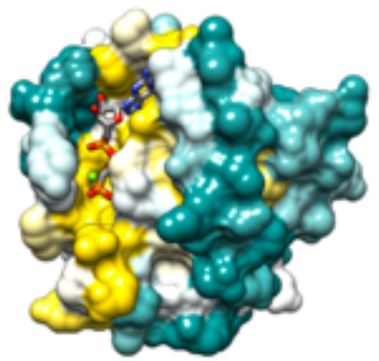
The name 'Ras' is an abbreviation of 'Rat sarcoma', reflecting the way the first members of the protein family were discovered. The name ras is also used to refer to the family of genes encoding those proteins.

When Ras is 'switched on' by incoming signals, it subsequently switches on other proteins, which ultimately turn on genes involved in cell growth, differentiation and survival. As a result, mutations in ras genes can lead to the production of permanently activated Ras proteins. This can cause unintended and overactive signalling inside the cell, even in the absence of incoming signals.

Because these signals result in cell growth and division, overactive Ras signaling can ultimately lead to cancer.[1] The 3 Ras genes in humans (HRAS, KRAS, and NRAS) are the most common oncogenes in human cancer; mutations that permanently activate Ras are found in 20% to 25% of all human tumors and up to 90% in certain types of cancer (e.g., pancreatic cancer).[2] For this reason, Ras inhibitors are being studied as a treatment for cancer, and other diseases with Ras overexpression.

Contents [hide]

1 History
2 Structure
3 Function
  3.1 Activation and deactivation
  3.2 Membrane attachment
4 Members
5 Ras in cancer
  5.1 Inappropriate activation
  5.2 Constitutively active Ras

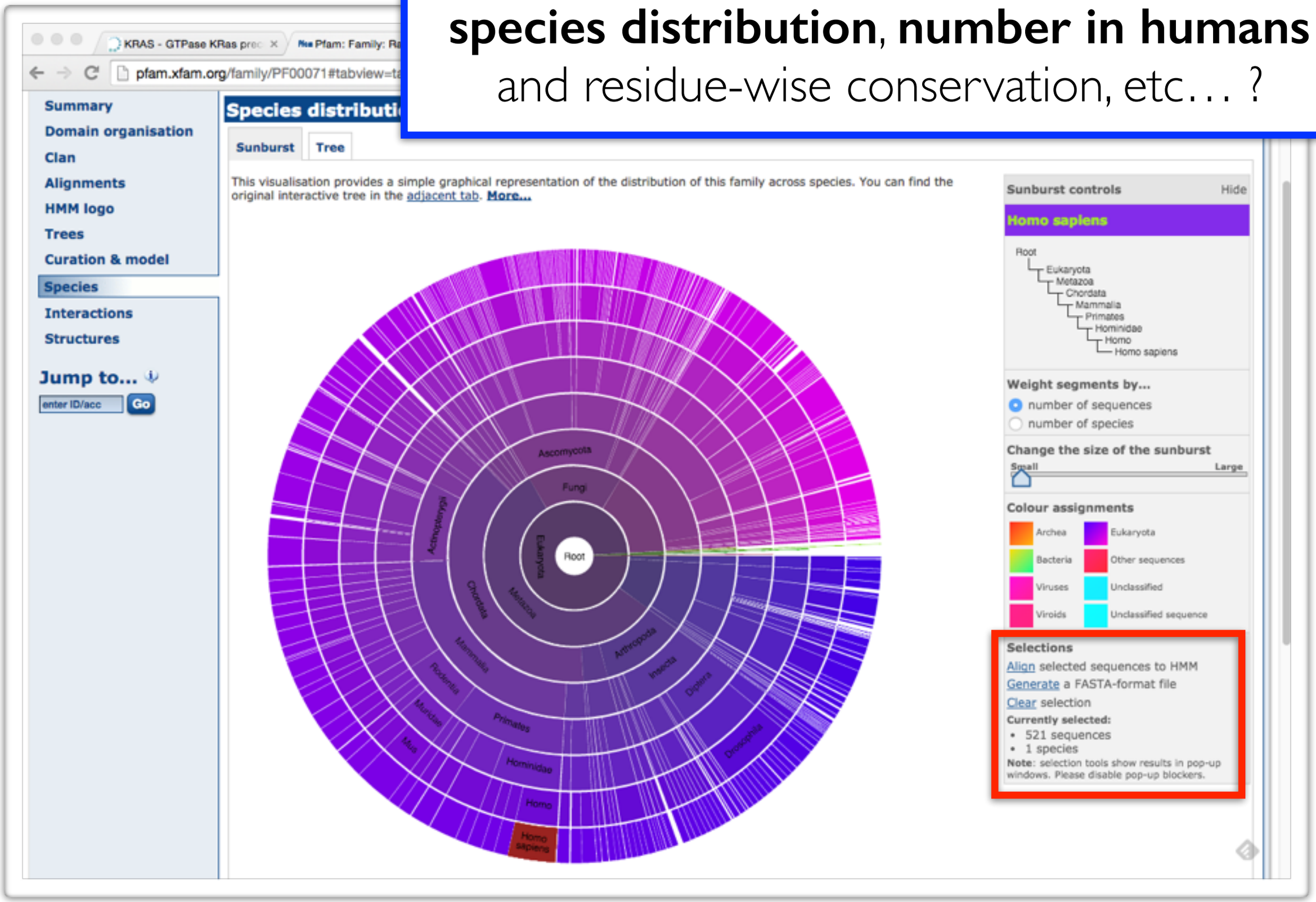


H-Ras structure PDB 121p, surface colored by conservation in Pfam seed alignment: gold, most conserved; dark cyan, least conserved.

Identifiers	
Symbol	Ras
Pfam	PF00071
InterPro	IPR013753
PROSITE	PDOC00017
SCOP	5p21
SUPERFAMILY	5p21

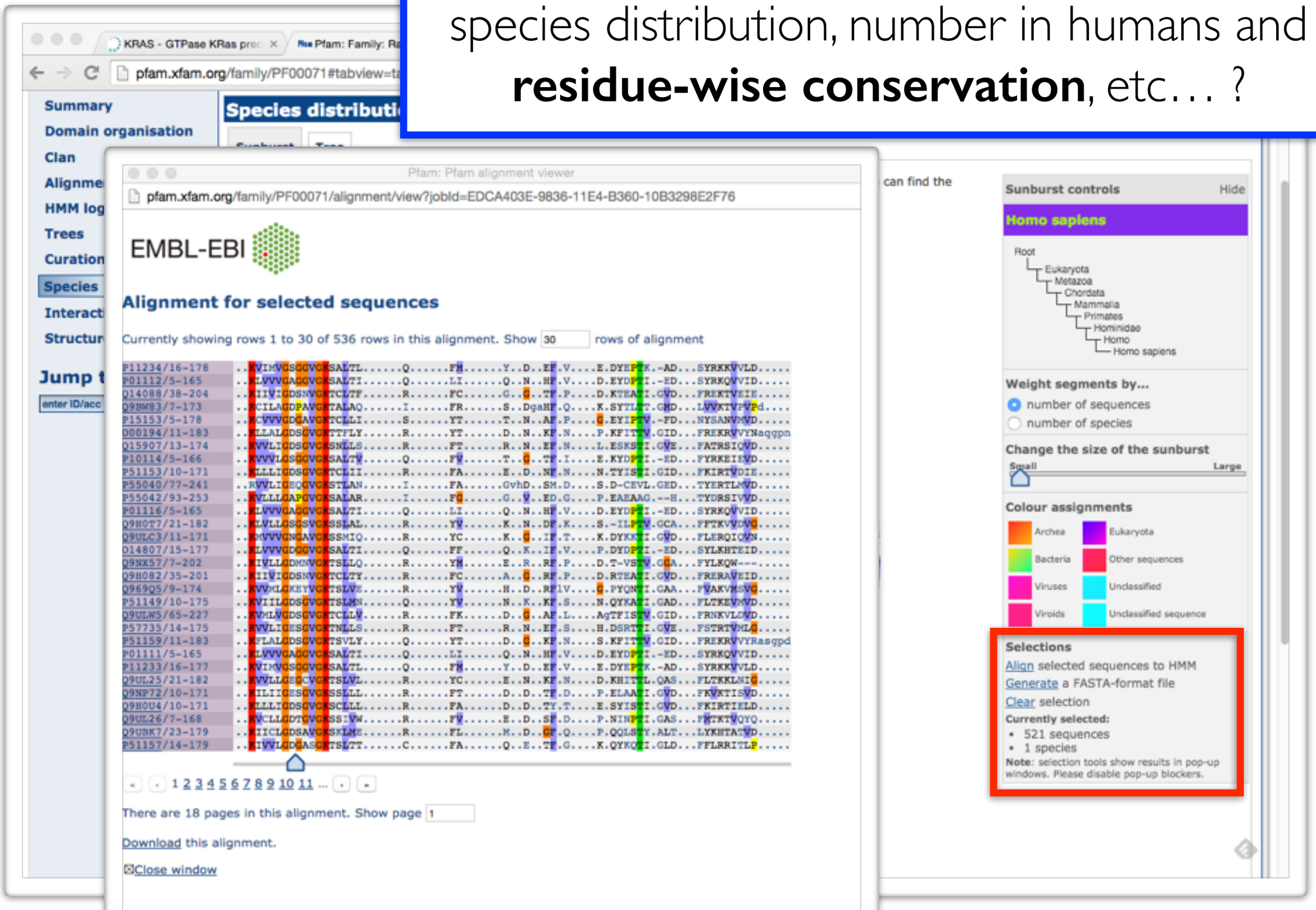
Example Questions:

What is known about the protein family, its **species distribution**, **number in humans** and residue-wise conservation, etc… ?



Example Questions:

What is known about the protein family, its species distribution, number in humans and **residue-wise conservation**, etc… ?



Example Questions:

What is known about the protein family, its species distribution, number in humans and **residue-wise conservation**, etc… ?

pfam.xfam.org/family/PF00071#tabview=tab4

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Family: *Ras* (PF00071)

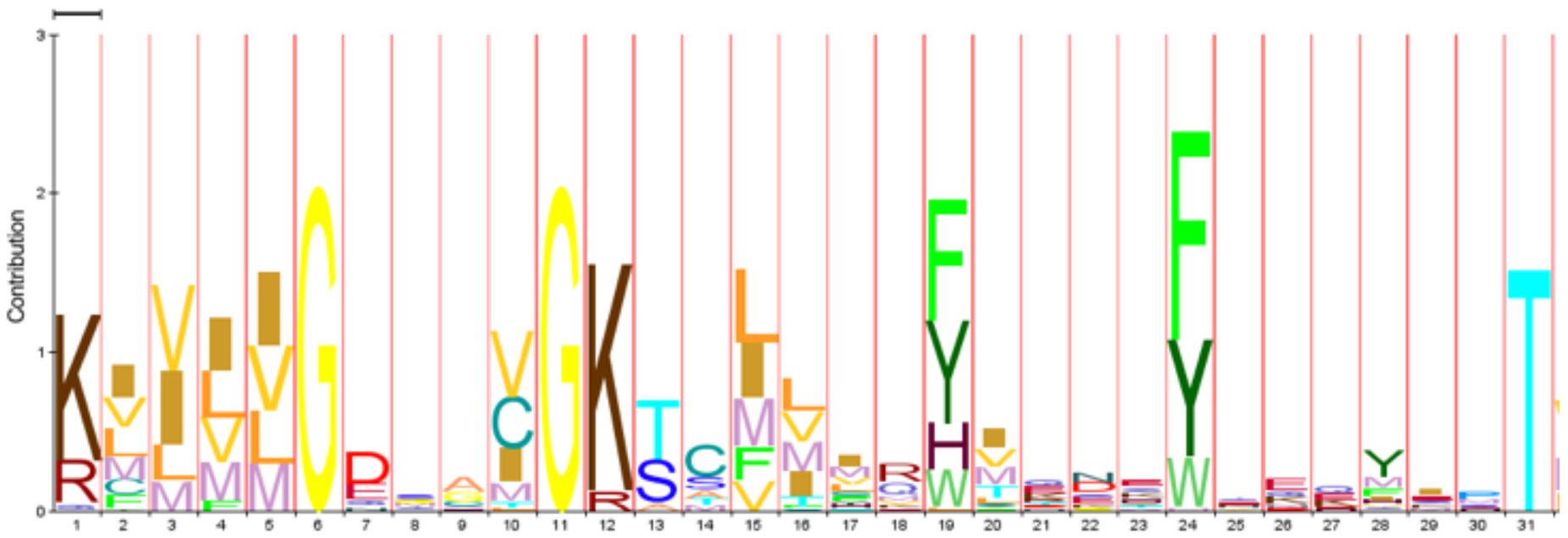
332 architectures 21243 sequences 30 interactions 1006 species 663 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to... enter ID/acc Go

HMM logo

HMM logos is one way of visualising profile HMMs. Logos provide a quick overview of the properties of an HMM in a graphical form. You can see a more detailed description of HMM logos and find out how you can interpret them here. More...



Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.

European Molecular Biology Laboratory

Family: *Kinesin* (PF00225)

Loading page components (1 remaining)...

126 architectures 4150 sequences 6 interactions 248 species 114 structures

Summary
Domain organisation
Clans
Alignments
HMM logo
Trees
Curation & models
Species
Interactions
Structures

Jump to...
enter ID/acc Go

Interactions

There are **6** interactions for this family. **More...**

Tubulin
Tubulin_C
Tubulin_C
Kinesin
Tubulin
Kinesin

Family: *Kinesin* (PF00225)

126 architectures 4150 sequences 6 interactions 248 species 114 structures

Summary
Domain organisation
Clans
Alignments
HMM logo
Trees
Curation & models
Species
Interactions
Structures

Jump to... enter ID/acc Go

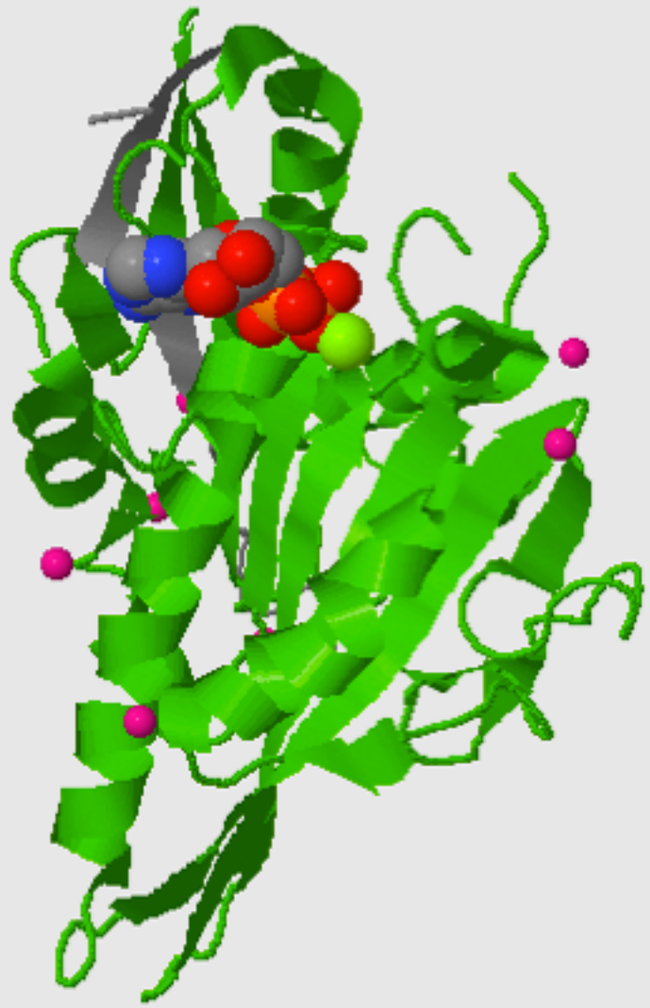
Structures

For those sequences which have a structure in the Protein DataBank, we use the mapping between UniProt, PDB and Pfam coordinate systems from the PDBe group, to allow us to map Pfam domains onto UniProt sequences and three-dimensional protein structures. The table below shows the structures on which the **Kinesin** domain has been found.

UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View
A8BKD1_GIALA	11 - 335	2vvg	A	11 - 335	Jmol AstexViewer SPICE
			B	11 - 335	Jmol AstexViewer SPICE
CENPE_HUMAN	12 - 329	1t5c	A	12 - 329	Jmol AstexViewer SPICE
			B	12 - 329	Jmol AstexViewer SPICE
KAR3_YEAST	392 - 723	1f9t	A	392 - 723	Jmol AstexViewer SPICE
		1f9u	A	392 - 723	Jmol AstexViewer SPICE
		1f9v	A	392 - 723	Jmol AstexViewer SPICE
		1f9w	A	392 - 723	Jmol AstexViewer SPICE
			B	392 - 723	Jmol AstexViewer SPICE
		3kar	A	392 - 723	Jmol AstexViewer SPICE
KI13B_HUMAN	11 - 352	3gbj	A	11 - 352	Jmol AstexViewer SPICE
			B	11 - 352	Jmol AstexViewer SPICE
			C	11 - 352	Jmol AstexViewer SPICE
		1ii6	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		1q0b	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		1x88	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
			A	24 - 359	Jmol AstexViewer SPICE

wellcome trust sanger institute

PDB entry 3bfn



Jmol

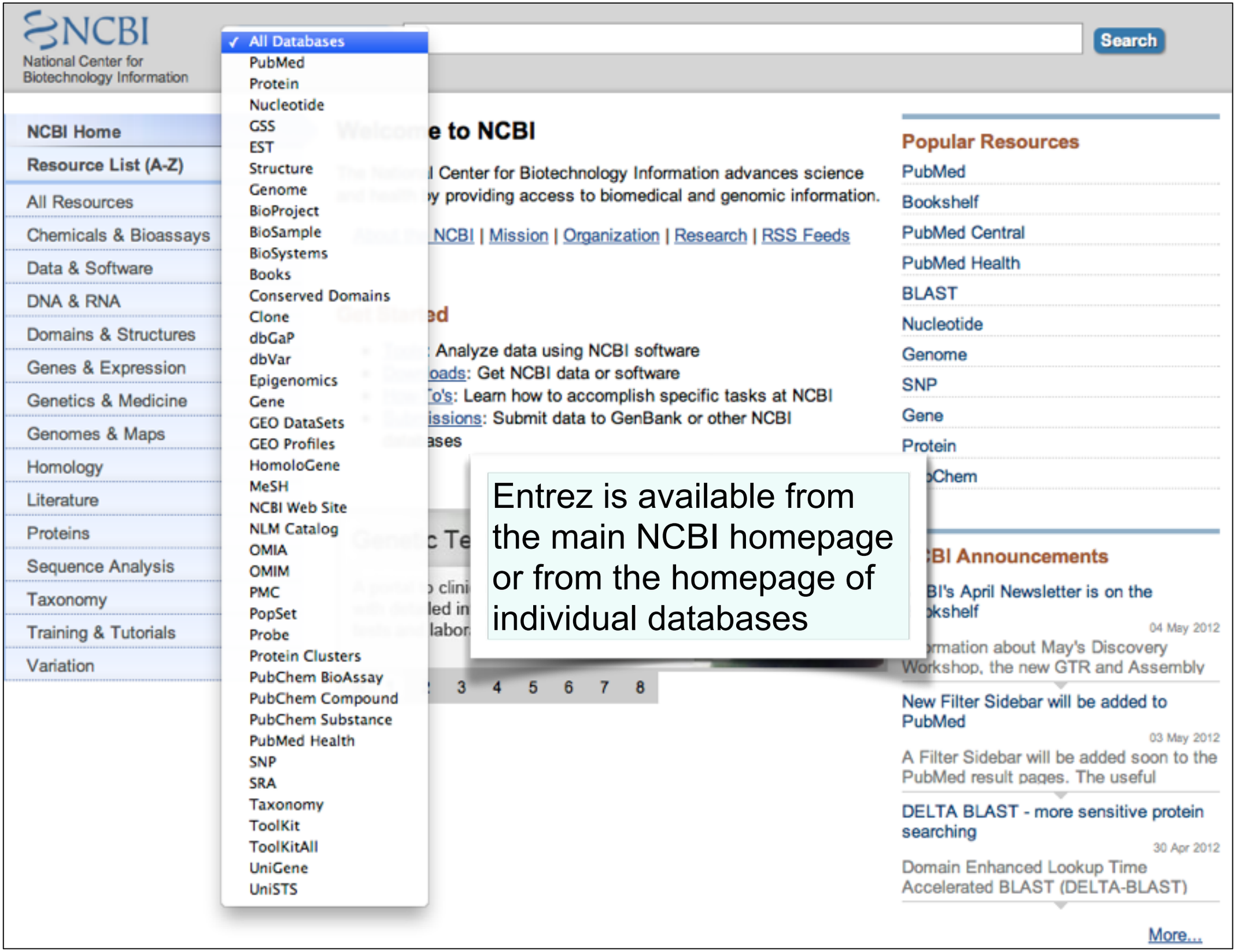
PDB			UniProt			Pfam family	Colour
Chain	Start	End	ID	Start	End		
A	49	368	KIF22_HUMAN	49	368	Kinesin (PF00225)	

☒Close window

ENTREZ & BLAST:

TOOLS FOR SEARCHING AND ACCESSING MOLECULAR DATA AT NCBI

Entrez: Integrated search of NCBI databases



Entrez: navigating across databases

Entrez was setup to allow you to navigate to related data in different databases without having to run additional searches.

Relies on pre-computed and pre-compiled data links:

- Neighbor knowledge based on calculations
- Hard links based on things we know about

Word weight

VAST

Neighbors
Related Structures

Phylogeny

BLAST

Neighbors
Related Sequences

Nucleotide sequences

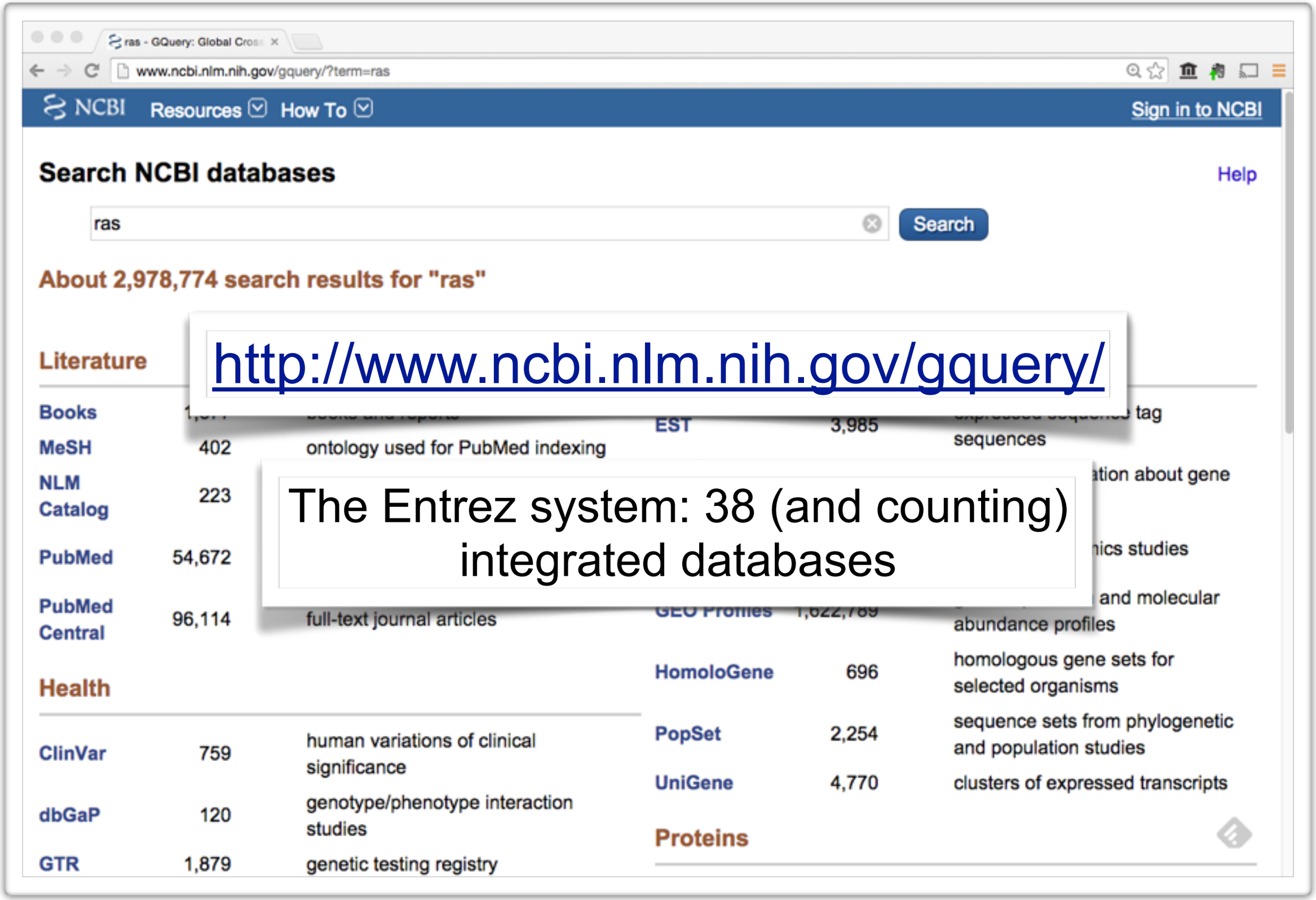
Hard Link

Protein sequences

BLAST

Neighbors
Related Sequences
BLink
Domains

Global Entrez Query: All NCBI Databases



Search Results

Nucleotide | Nucleotide | zebrafish creatine kinase | Search | Save search | Limits | Advanced | Help

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

Found 2324 nucleotide sequences. Nucleotide (35) EST (2289)

Results: 1 to 20 of 35 << First < Prev Page 1 of 2 Next > Last >>

1. **Danio rerio creatine kinase**, muscle b (ckmb), mRNA
   1,463 bp linear mRNA
   Accession: NM_001105683.1 GI: 157787180
   GenBank FASTA Graphics Related Sequences

2. **Danio rerio** zgc:63663 (zgc:63663), mRNA
   2,476 bp linear mRNA
   Accession: NM_200614.1 GI: 41055386
   GenBank FASTA Graphics Related

3. **Danio rerio creatine kinase**, muscle
   1,552 bp linear mRNA
   Accession: NM_130932.1 GI: 18858426
   GenBank FASTA Graphics Related Sequences

4. **Danio rerio creatine kinase**, mitochondrial 2 (sarcomeric), mRNA (cDNA clone MGC:198091 IMAGE:9039080), complete cds
   1,296 bp linear mRNA
   Accession: BC171364.1 GI: 213624628
   GenBank FASTA Graphics Related Sequences

5. **Danio rerio creatine kinase**, mitochondrial 2 (sarcomeric), mRNA (cDNA clone MGC:172259 IMAGE:8798676), complete cds
   1,400 bp linear mRNA
   Accession: BC154617.1 GI: 159155933
   GenBank FASTA Graphics Related Sequences

Discovery Column (sort, filter, link)

Filter your results:

All (35)
Bacteria (0)
INSDC (GenBank) (27)
mRNA (32)
RefSeq (8)

Manage Filters

Top Organisms [Tree]
Danio rerio *(29)*
Ictalurus furcatus *(6)*

Find related data
Database: Select
Find items

Search details
("Danio rerio"[Organism] OR zebrafish[All Fields]) AND creatine kinase[All Fields]
Search
See more...

Recent activity

Limits

Limits

Published in the last

Any Date

Modified in the last

Any Date

Search Field Tags

Field: All Fields

Segmented Sequences

Any

Source database

Any

Molecule

Any

Gene Location

Any

Exclude

- [ ] STSs
- [ ] working draft
- [ ] TPA
- [ ] patents

Reset Search

Search Results

Nucleotide | Nucleotide | zebrafish creatine kinase | Search | Save search | Limits | Advanced | Help

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

Found 2324 nucleotide sequences. Nucleotide (35) EST (2289)

Results: 1 to 20 of 35 << First < Prev Page 1 of 2 Next > Last >>

1. **Danio rerio creatine kinase**, muscle b (ckmb), mRNA
   1,463 bp linear mRNA
   Accession: NM_001105683.1 GI: 157787180
   GenBank FASTA Graphics Related Sequences

2. **Danio rerio** zgc:63663 (zgc:63663), mRNA
   2,476 bp linear mRNA
   Accession: NM_200614.1 GI: 41055386
   GenBank FASTA Graphics Related

3. **Danio rerio creatine kinase**, muscle
   1,552 bp linear mRNA
   Accession: NM_130932.1 GI: 18858426
   GenBank FASTA Graphics Related Sequences

4. **Danio rerio creatine kinase**, mitochondrial 2 (sarcomeric), mRNA (cDNA clone MGC:198091 IMAGE:9039080), complete cds
   1,296 bp linear mRNA
   Accession: BC171364.1 GI: 213624628
   GenBank FASTA Graphics Related Sequences

5. **Danio rerio creatine kinase**, mitochondrial 2 (sarcomeric), mRNA (cDNA clone MGC:172259 IMAGE:8798676), complete cds
   1,400 bp linear mRNA
   Accession: BC154617.1 GI: 159155933
   GenBank FASTA Graphics Related Sequences

Discovery Column (sort, filter, link)

Filter your results:

All (35)
Bacteria (0)
INSDC (GenBank) (27)
mRNA (32)
RefSeq (8)

Manage Filters

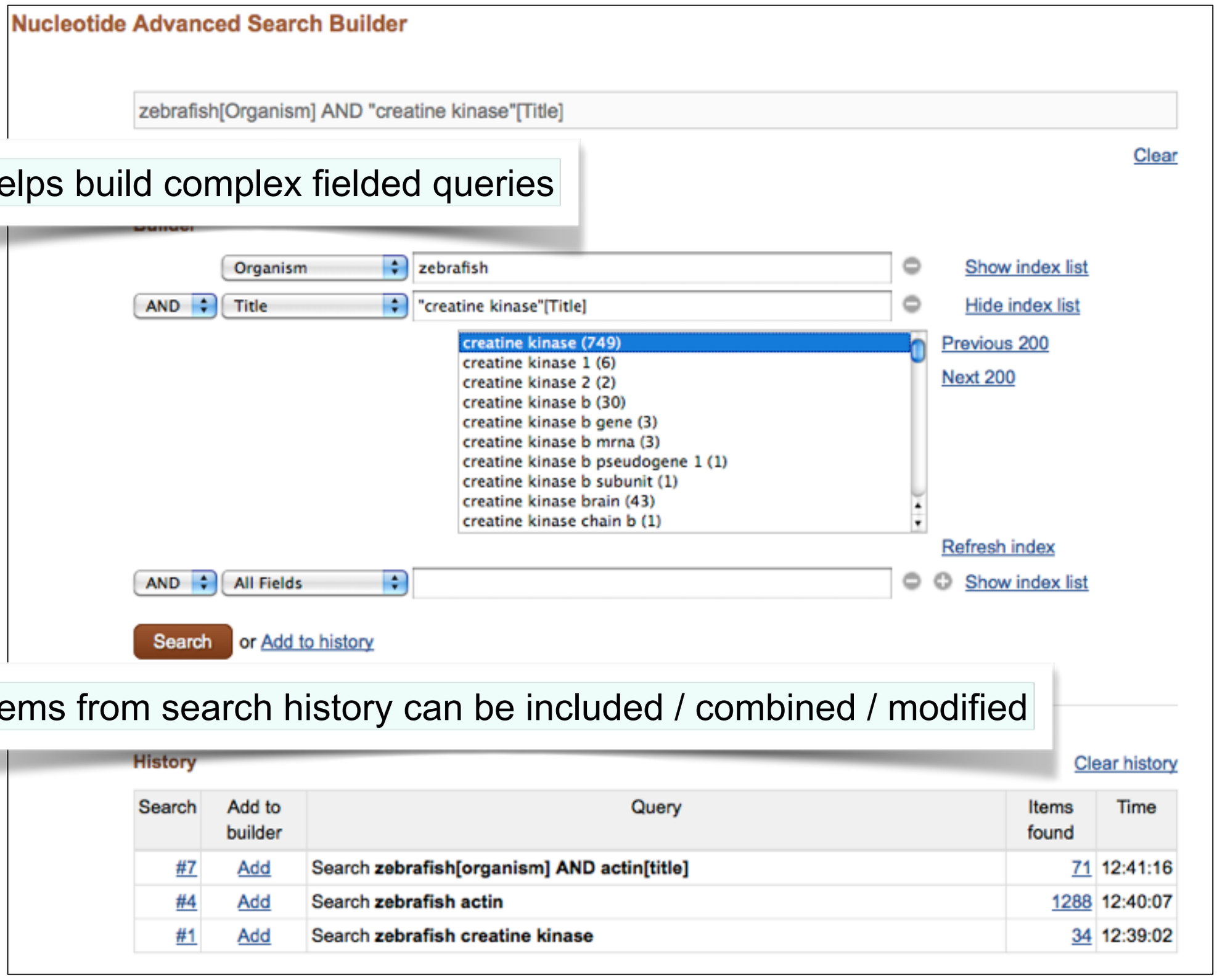
Top Organisms [Tree]
Danio rerio *(29)*
Ictalurus furcatus *(6)*

Find related data
Database: Select
Find items

Search details
("Danio rerio"[Organism] OR zebrafish[All Fields]) AND creatine kinase[All Fields]
Search
See more...

Recent activity

Advanced: Search Builder



Complex Query Results

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

Send to:

Results: 6

1. **Danio rerio creatine kinase**, brain a (ckba), mRNA
   1,481 bp linear mRNA
   Accession: NM_001077163.1 GI: 116004536
   GenBank FASTA Graphics Related Sequences

2. **Danio rerio creatine kinase**, mitochondrial 1 (ckmt1), nuclear gene encoding mitochondrial protein, mRNA

("Danio rerio"[Organism] AND "creatine kinase"[Title]) AND "refseq"[Filter] AND mrna[Filter]

3. 1,552 bp linear mRNA
   Accession: NM_130932.1 GI: 18858426
   GenBank FASTA Graphics Related Sequences

4. **Danio rerio creatine kinase**, mitochondrial 2 (sarcomeric) (ckmt2), nuclear gene encoding mitochondrial protein, mRNA
   1,401 bp linear mRNA
   Accession: NM_200697.1 GI: 41152341
   GenBank FASTA Graphics Related Sequences

5. **Danio rerio creatine kinase**, muscle b (ckmb), mRNA
   1,463 bp linear mRNA
   Accession: NM_001105683.1 GI: 157787180
   GenBank FASTA Graphics Related Sequences

6. **Danio rerio creatine kinase**, brain b (ckbb), mRNA
   1,459 bp linear mRNA
   Accession: NM_173222.1 GI: 27545192
   GenBank FASTA Graphics Related Sequences

Filter your results:

All (6)
Bacteria (0)
INSDC (GenBank) (0)
mRNA (6)
RefSeq (6)

Manage Filters

Analyze these sequences

Run BLAST

Find related data

Database: Select

Find items

Search details

```
("Danio rerio"[Organism]
AND "creatine kinase"
[Title]) AND "refseq"
[Filter]
```

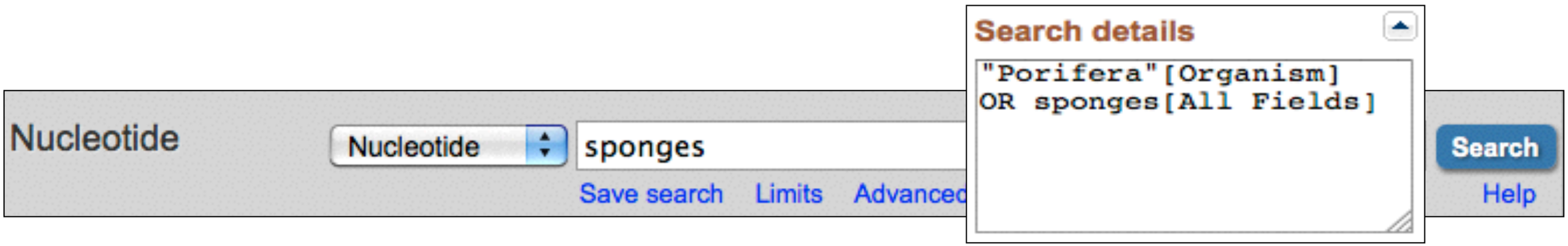
Search

See more...

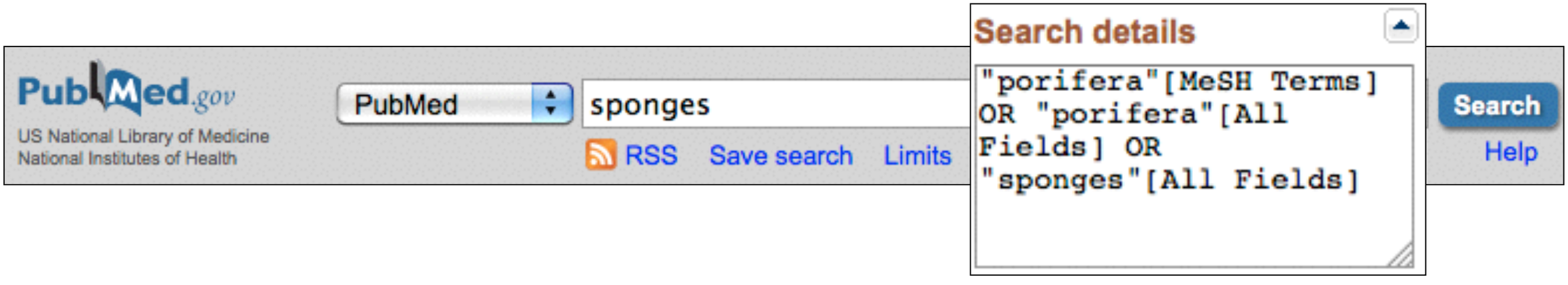
Recent activity

Controlled Vocabularies

- Taxonomy primary controlled vocabulary / classification system for molecular databases at NCBI

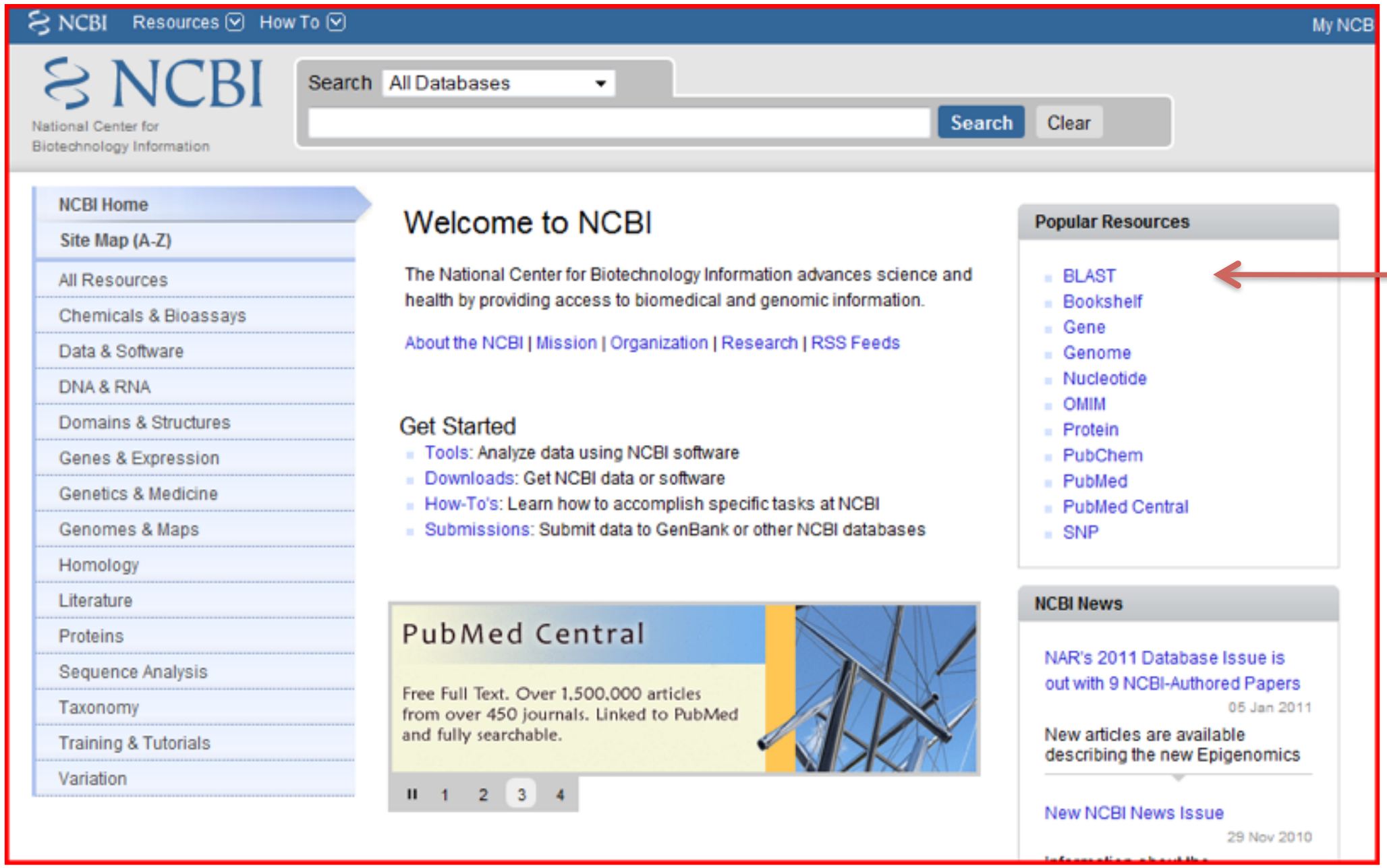


- Medical Subject Headings (MeSH) primary controlled vocabulary / classification system (ontology) for molecular databases at NCBI



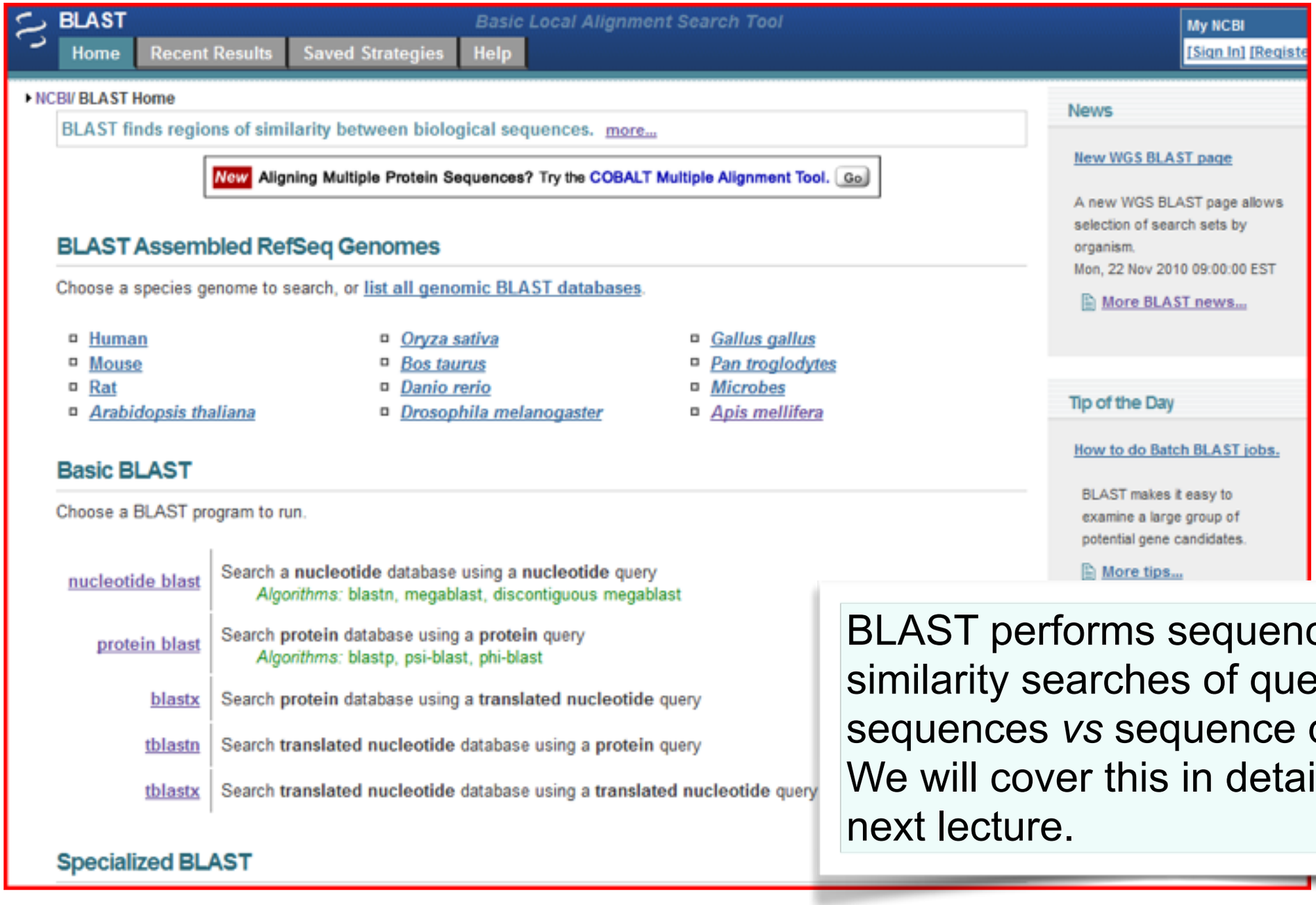
BLAST is a very important tool available from the NCBI Homepage

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



BLAST – Basic Local Alignment Search Tool

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



BLAST performs sequence similarity searches of query sequences *vs* sequence databases. We will cover this in detail in the next lecture.

SUMMARY

- Bioinformatics is computer aided biology.
- Bioinformatics deals with the collection, archiving, organization, and interpretation of a wide range of biological data.
- There are a large number of primary, secondary and tertiary bioinformatics databases.
- The NCBI and EBI are major online bioinformatics service providers.
- Introduced GenBank, RefSeq, UniProt, PDB databases as well as a number of ‘boutique’ databases including PFAM and OMIM.
- Introduced the notion of *controlled vocabularies* and *ontologies.*
- Described the use of ENTREZ and BLAST for searching databases.

HOMEWORK

- ☑ Complete the **initial course questionnaire**:
  http://tinyurl.com/bioinf525-questions

- ☑ Check out the "**Background Reading**" material on Ctools:
  http://tinyurl.com/bioinf525-w16

- ☑ Complete the **lecture 1.1 homework questions**:
  http://tinyurl.com/bioinf525-quiz1

ADDITIONAL DATABASES OF NOTE (SLIDES FOR YOUR REFERENCE)

NCBI Metadatabases

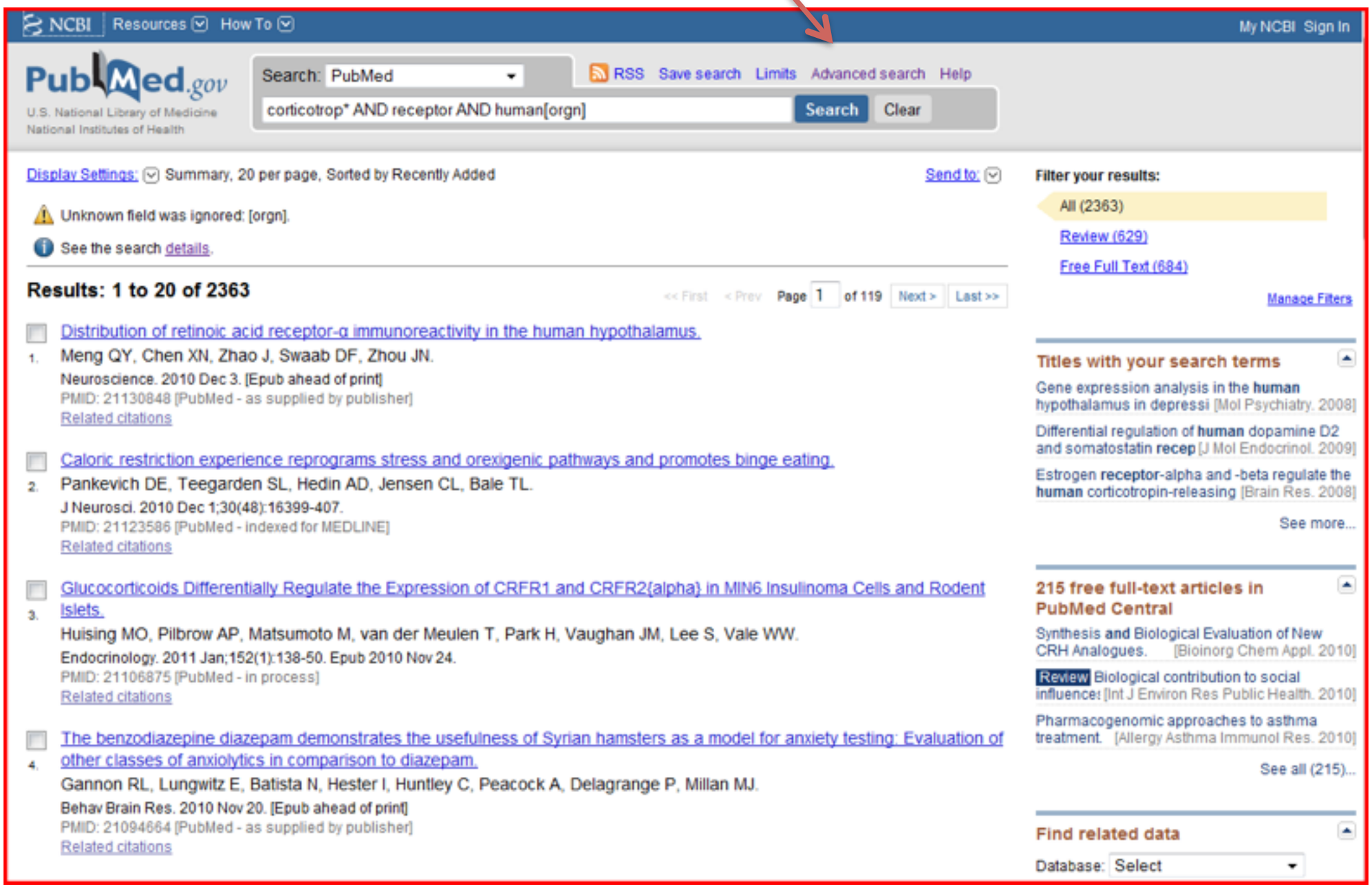
- **Gene**
  - molecular data and literature related to genes
- **HomoloGene**
  - automated collection of homologous genes from selected eukaryotes
- **Taxonomy**
  - access to NCBI data through source organism taxonomic classification
- **PubChem**
  - small organic molecules and their biological activities
- **BioSystems**
  - biochemical pathways and processes linked to NCBI genes, gene products, small molecules, and structures

PubMed

- Curated database of biomedical journal articles
- Data records are annotated with MeSH terms (Medical Subject Headings)
- Contract workers actually read all of the articles and classify them with the MeSH terms
- PubMed entries contain article abstracts
- PubMed Central contains full journal articles, but the majority are not freely re-distributable

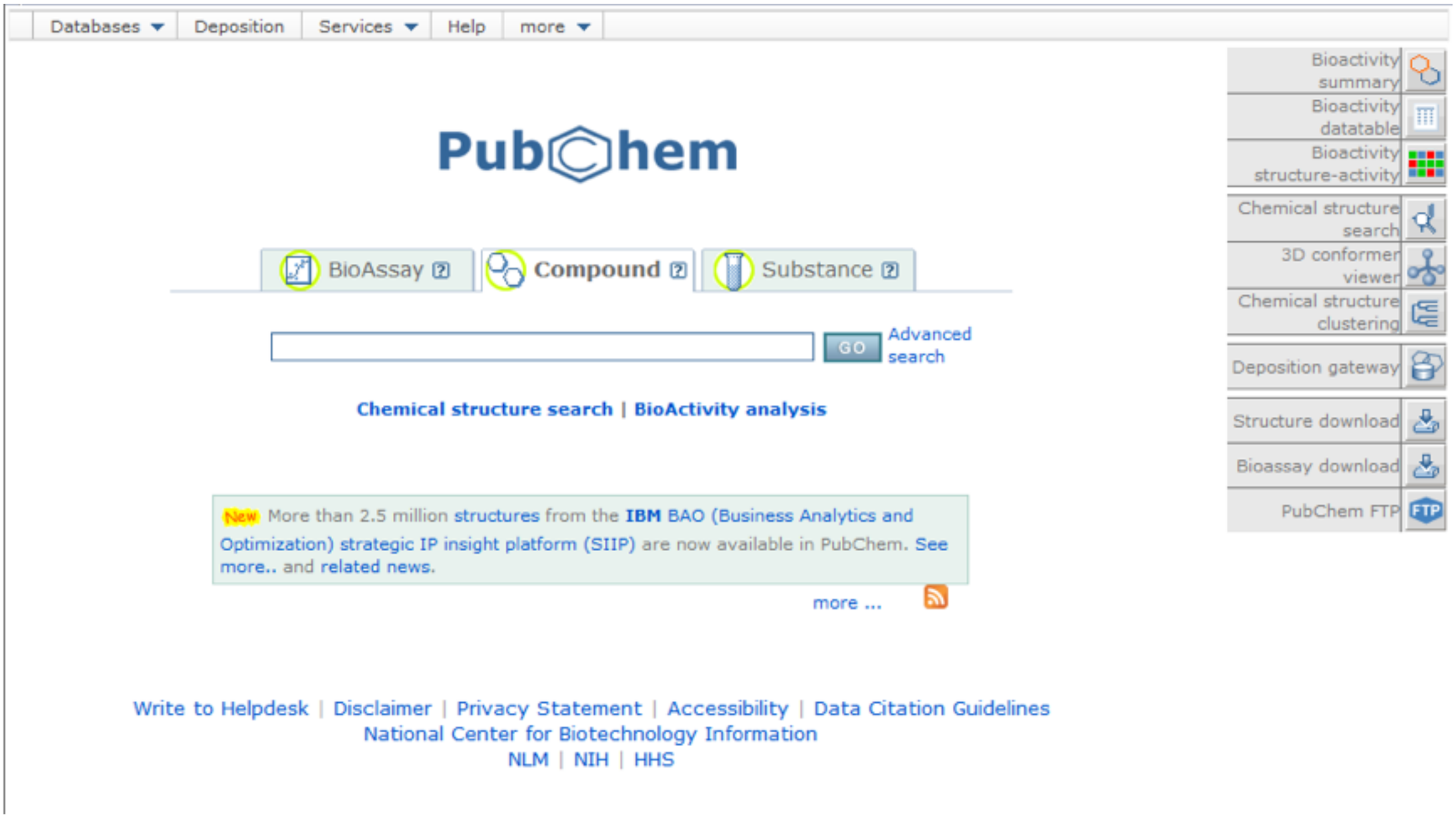
PubMed results

Limits and Advanced search can be used to refine searches



Small molecule databases have been added at NCBI

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



HomoloGene - Homologous genes from different organisms http://www.ncbi.nlm.nih.gov/homologene

NCBI

HomoloGene
Discover Homologs

Help

My NCBI
[Sign In] [Regis

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search HomoloGene for Go Clear

Limits Preview/Index History Clipboard Details

HomoloGene
Homepage
Query Tips
Build Procedure
FTP site

Genome Resources
Homo sapiens
Mus musculus
Rattus norvegicus
Danio rerio

HomoloGene is a system for automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes.

HomoloGene Release 65 Statistics

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

Species	Number of Genes		HomoloGene groups
	Input	Grouped	
Homo sapiens	19,943*	18,981	18,431
Pan troglodytes	25,096	16,850	15,980
Canis familiaris	19,766	16,708	15,951
Bos taurus	22,049	18,180	16,224
Mus musculus	25,388	21,766	19,005
Rattus norvegicus	21,991	19,229	17,473
Gallus gallus	17,959	13,142	11,905
Danio rerio	26,690*	21,084	14,067
Drosophila melanogaster	13,827*	9,282	7,749
Anopheles gambiae	12,460	8,867	7,541
Caenorhabditis elegans	20,132*	8,678	4,810
Schizosaccharomyces pombe	5,043	3,225	2,935
Saccharomyces cerevisiae	5,880	4,851	4,370
Kluyveromyces lactis	5,335	4,459	4,382
Eremothecium gossypii	4,722	3,928	3,884
Magnaporthe grisea	12,832	7,330	6,399
Neurospora crassa	9,821*	6,287	6,144

What's New

HomoloGene release 65 includes updated annotations for the following species: Homo sapiens (NCBI release 37.2), Danio rerio (NCBI release 4.1), Drosophila melanogaster (NCBI release 9.3) Caenorhabditis elegans (NCBI release 9.1), Arabidopsis thaliana (NCBI release 9.1).

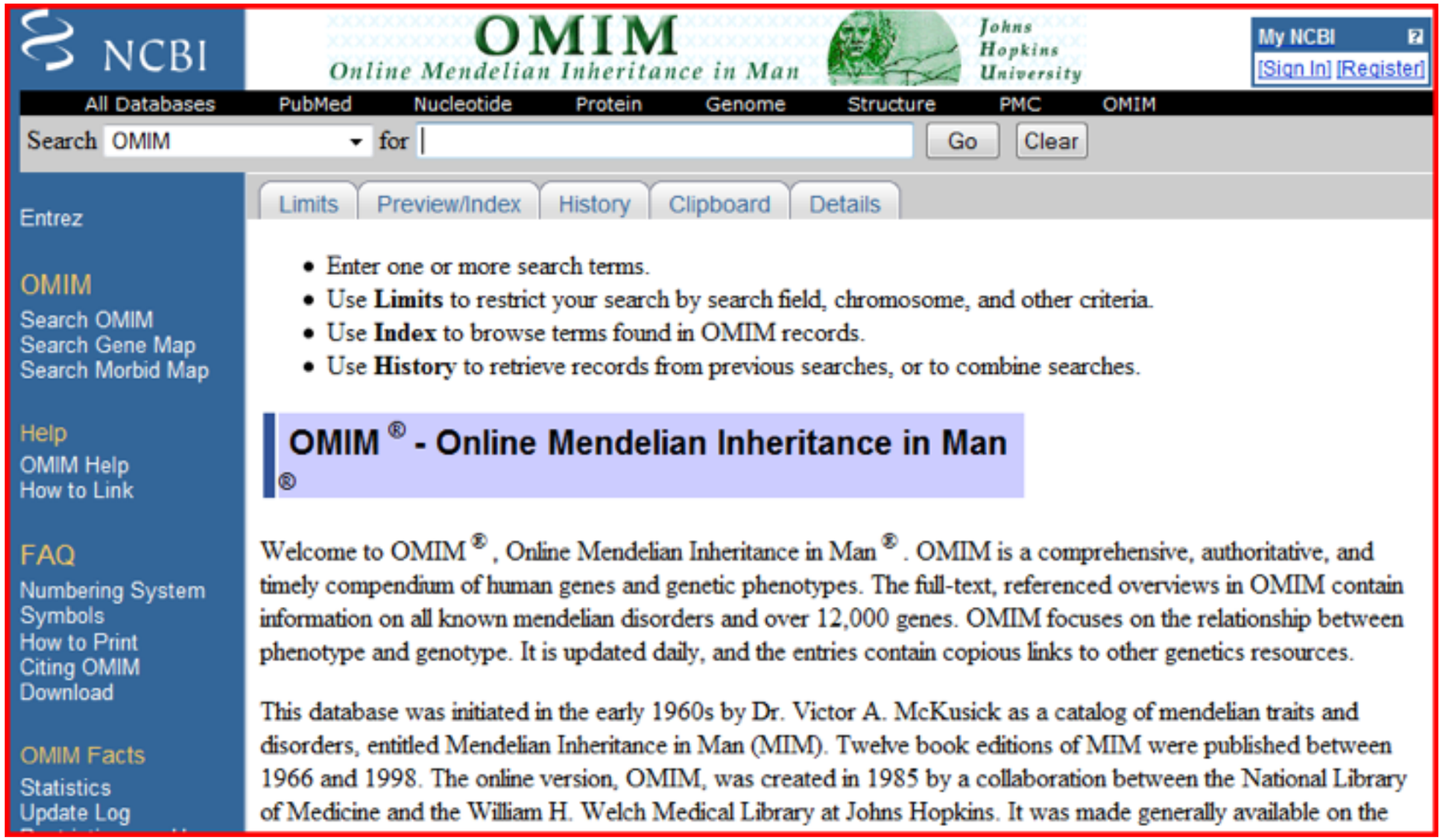
Related Resources

Entrez Genomes
A collection of complete genome sequences that includes more than 1000 viruses and over hundred microbes

- Archaea
- Bacteria
- Eukaryota

Online Mendelian Inheritance in Man – OMIM

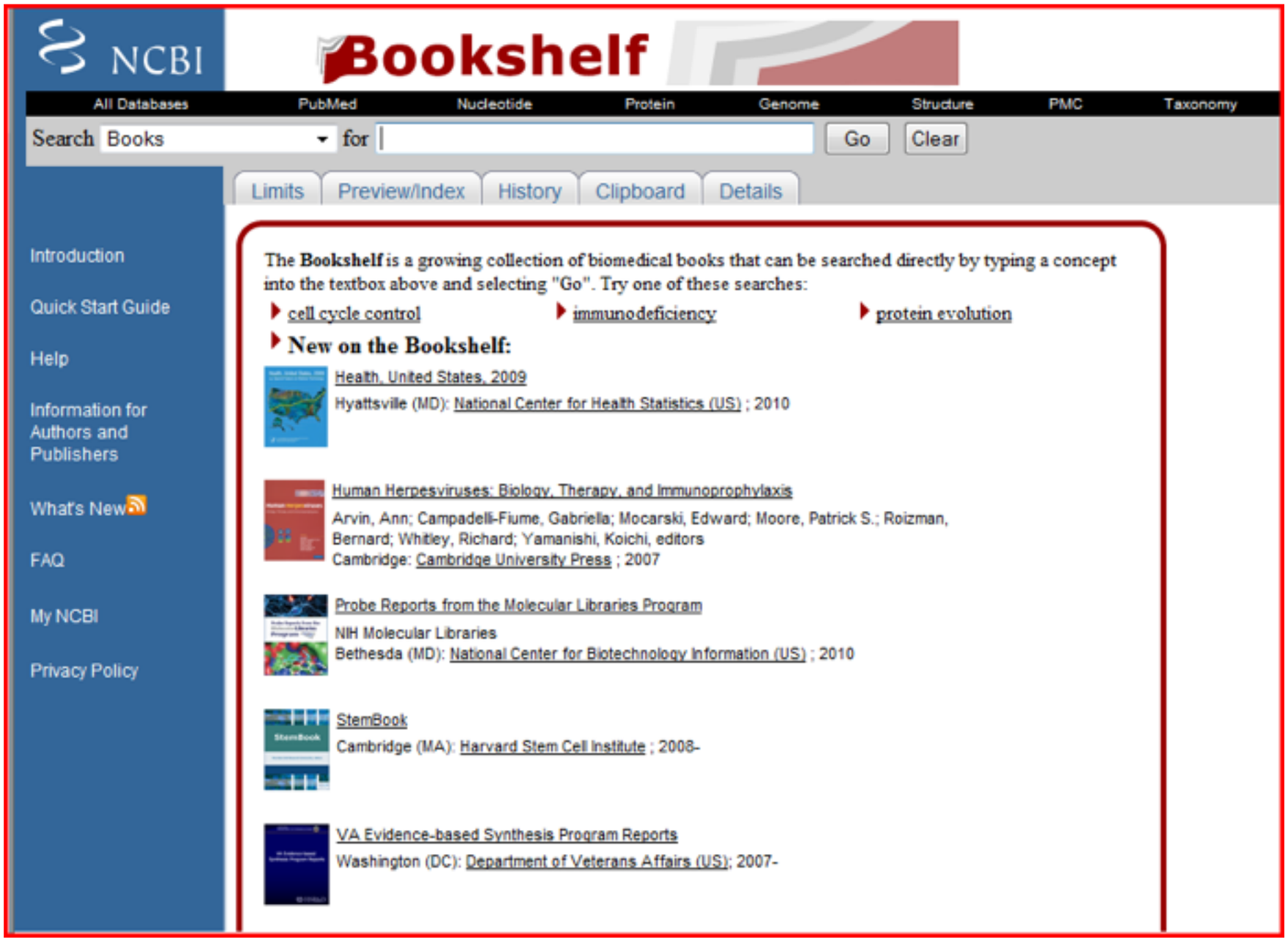
http://www.ncbi.nlm.nih.gov/omim



OMIM is essentially a set of reviews of human genes, gene function and phenotypes. Includes causative mutations where known.

The NCBI Bookshelf includes many well known molecular biology texts.

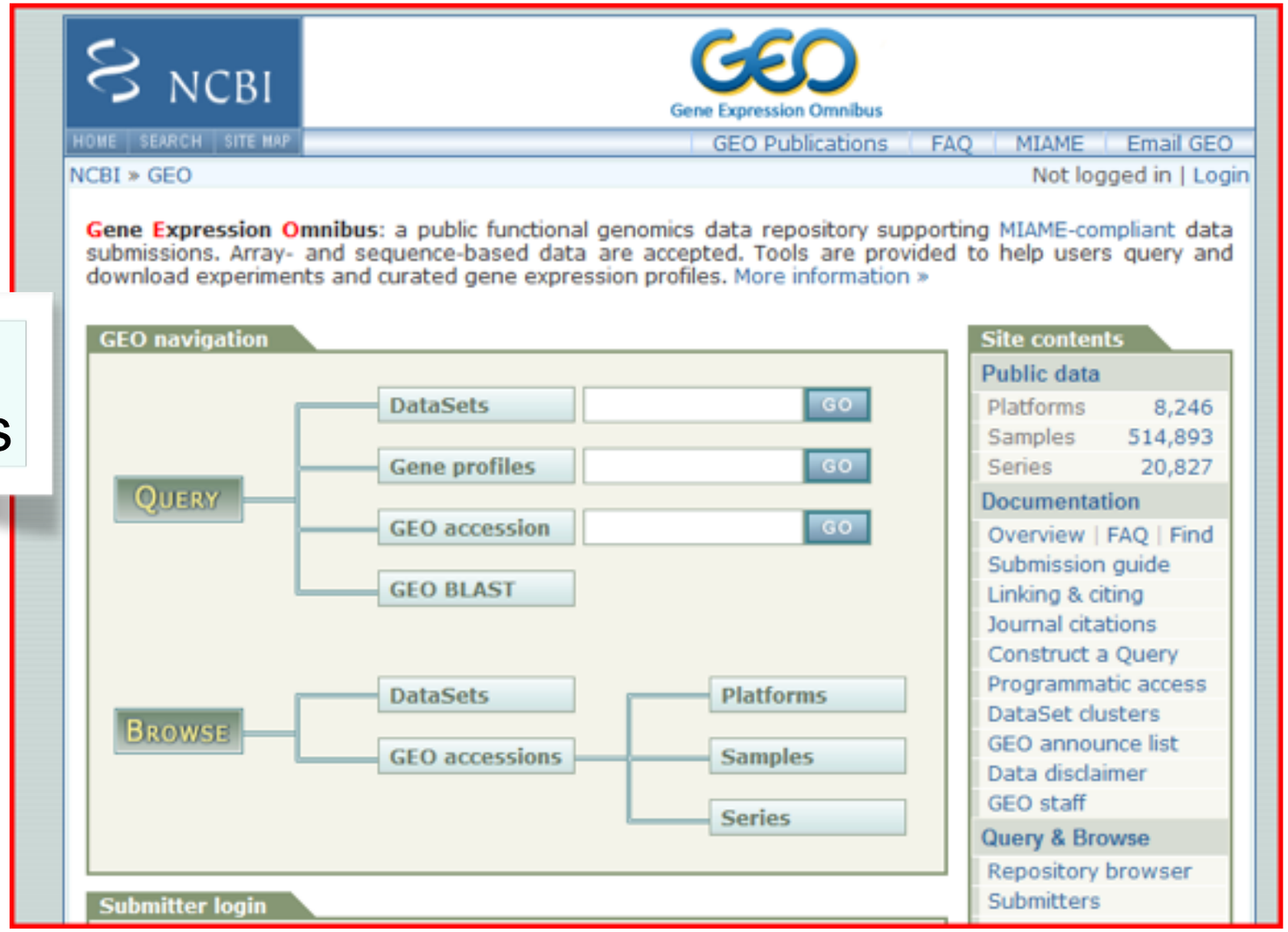
http://www.ncbi.nlm.nih.gov/books/



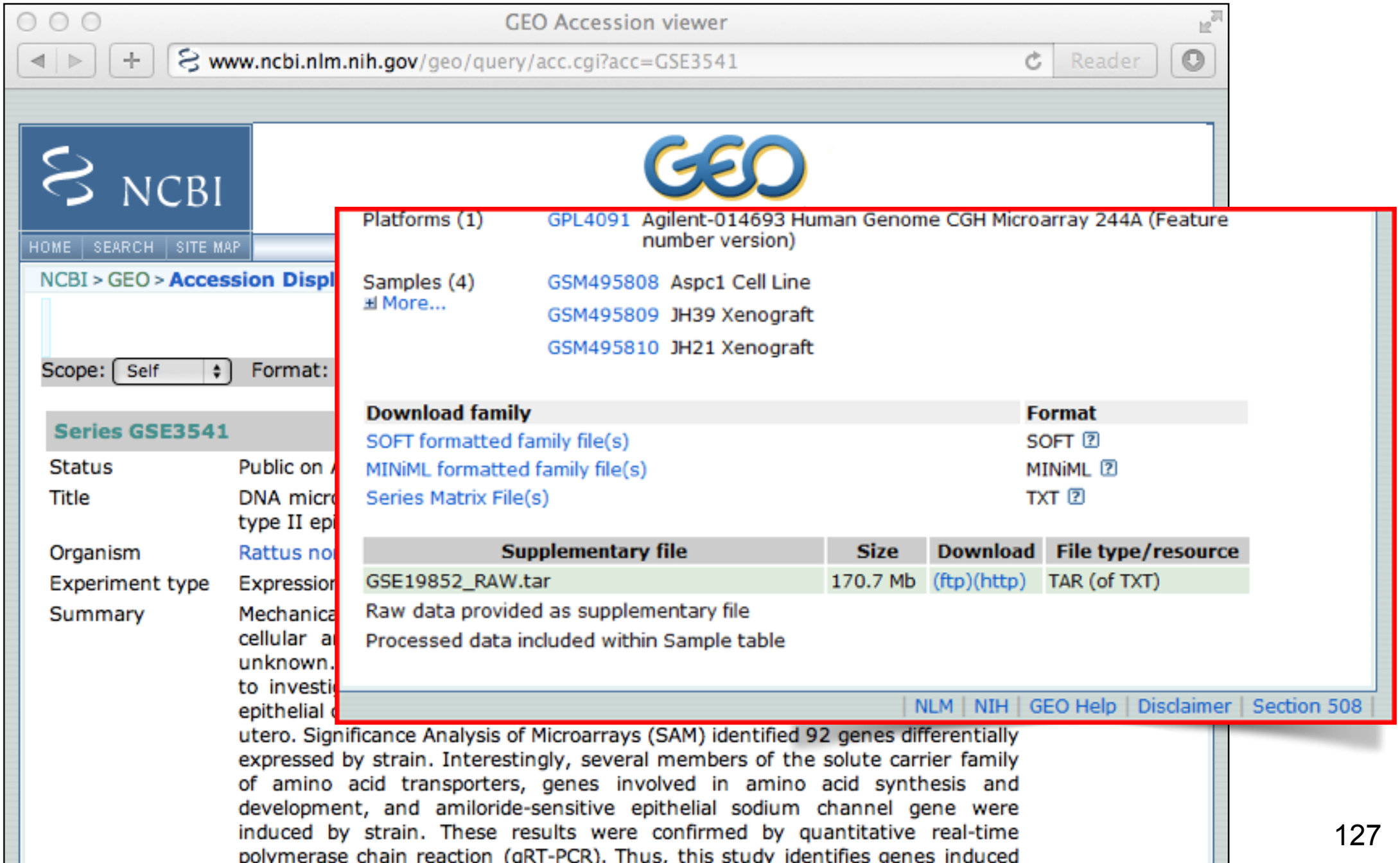
GEO: Gene Expression Omnibus

- Gene expression data (mostly from microarrays but also RNA-seq data, 2 methods for measuring RNA levels)

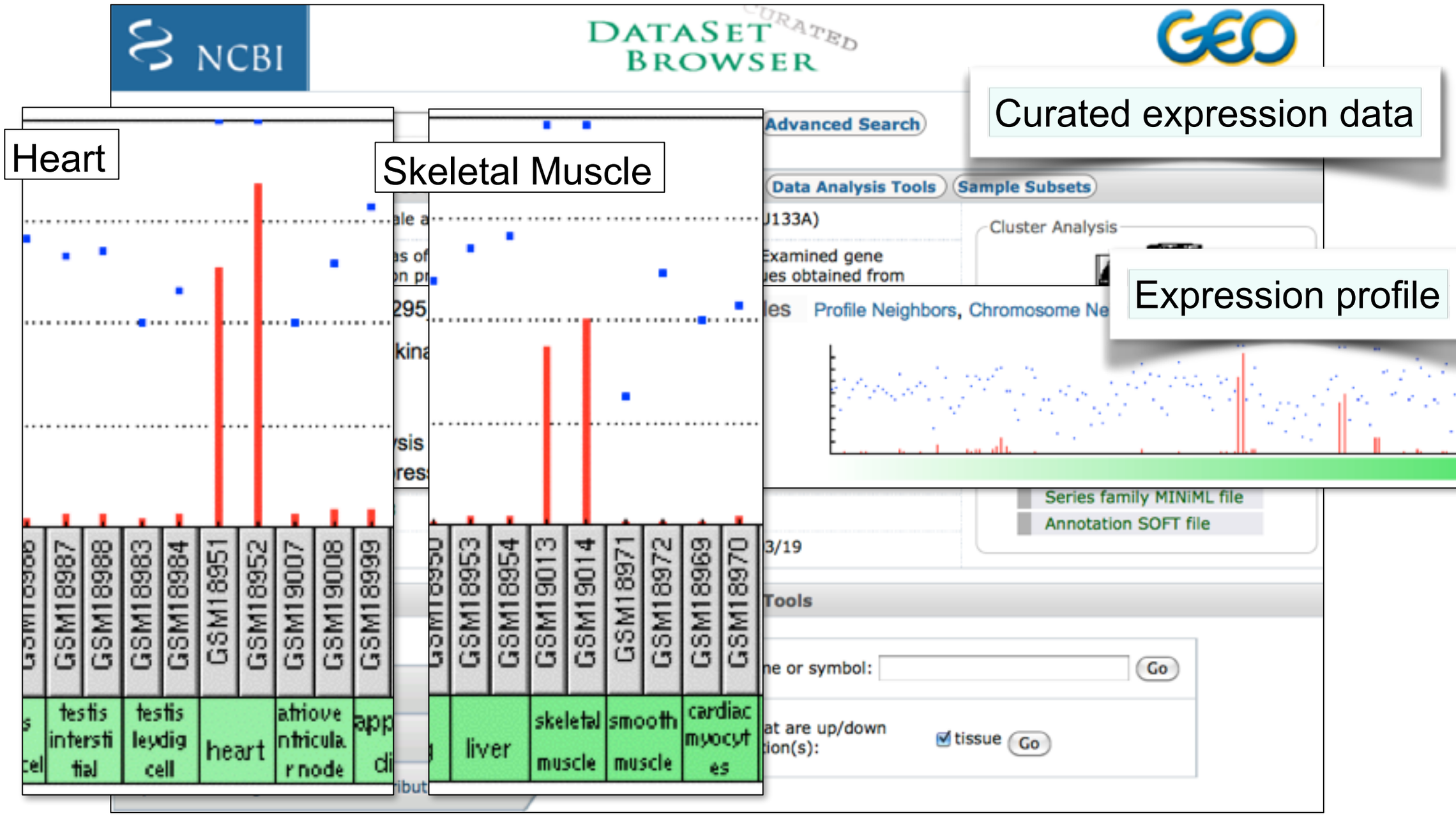
Query browse and download data sets



- **Series** - (GSExxx) is an original submitter-supplied record that summarizes a study. May contain multiple individual **Samples** (GSMxxx).

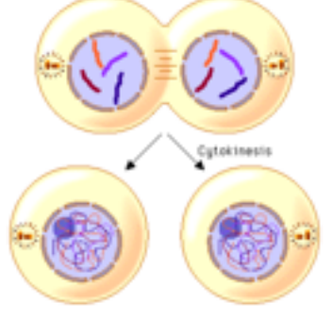


- **DataSets** - (GDSxxx) are curated collections of selected Samples that are biologically and statistically comparable

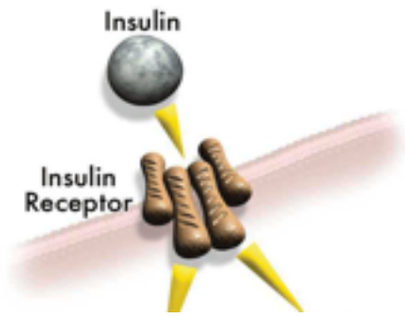


GO Ontologies

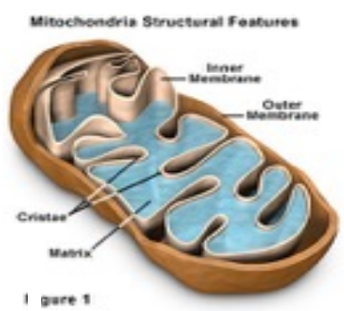
- There are three ontologies in GO:
  - **Biological Process**
    A commonly recognized series of events
    e.g. cell division, mitosis,



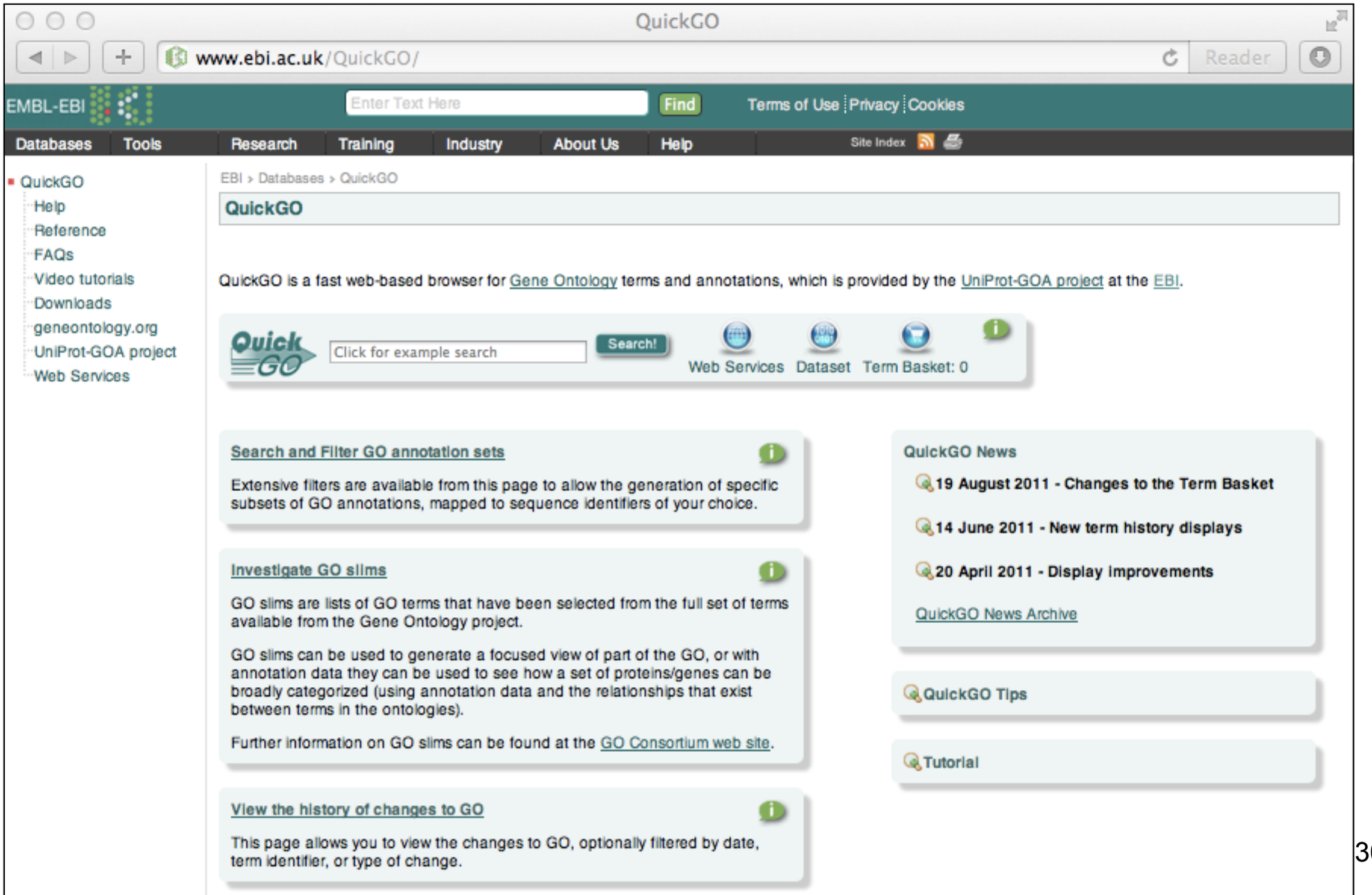
  - **Molecular Function**
    An elemental activity, task or job
    e.g. kinase activity, insulin binding



  - **Cellular Component**
    Where a gene product is located
    e.g. mitochondrion, mitochondrial membrane

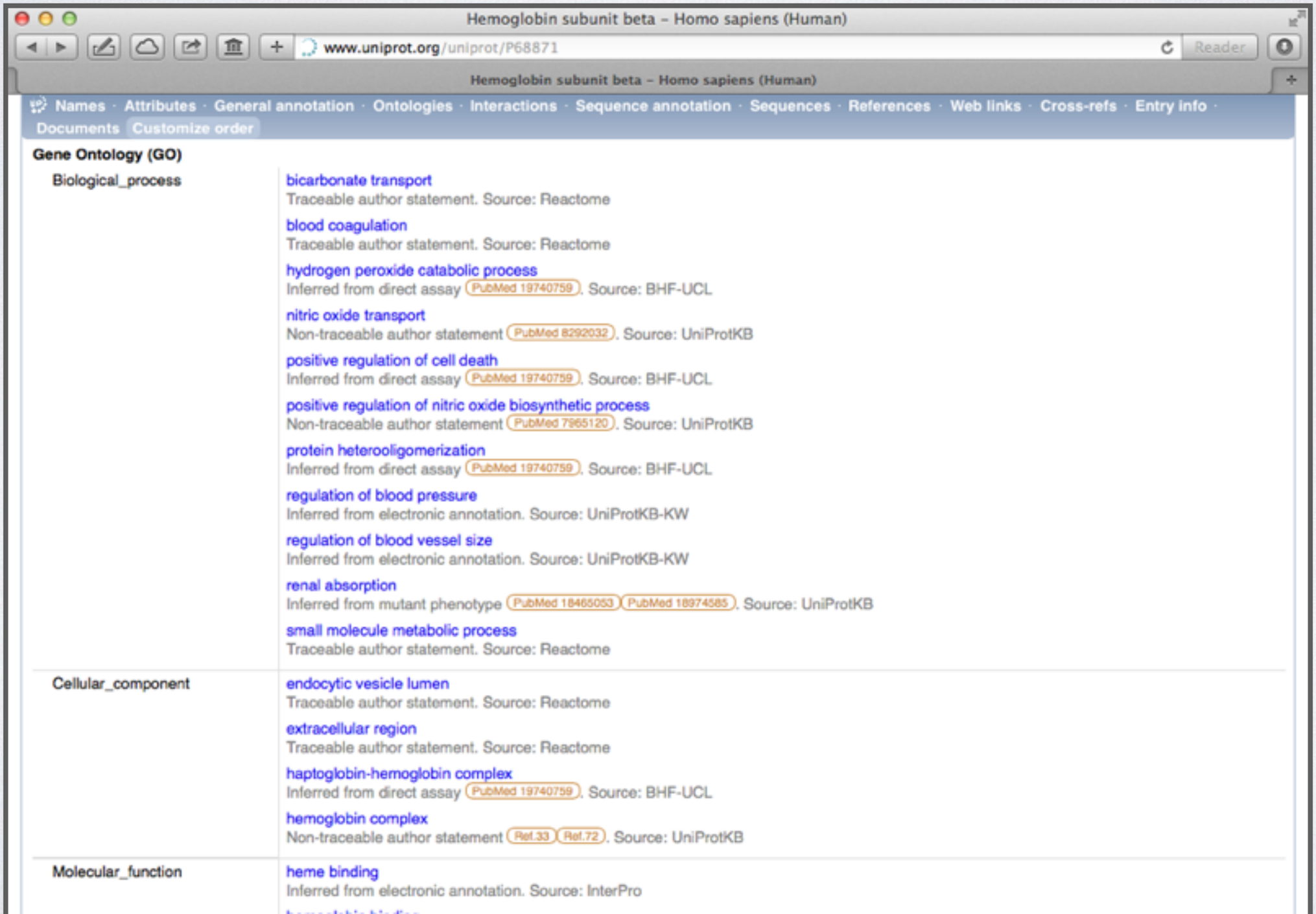


QuickGO is a fast web-based browser of the Gene Ontology and Gene Ontology annotation data



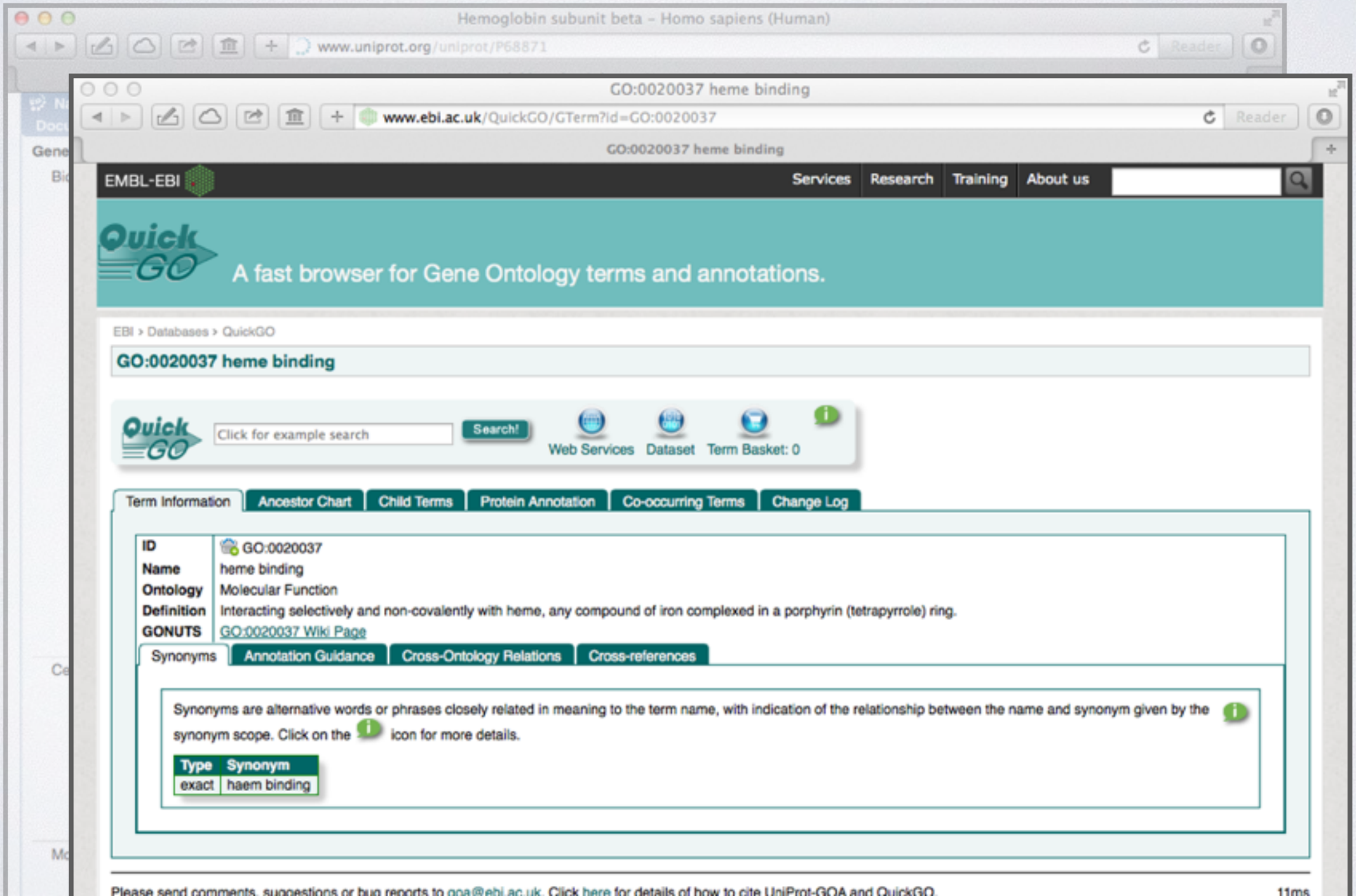
GO annotation in UniProt

An example UniProt entry for hemoglobin beta (HBB_human, P68871) with GO annotation displayed.

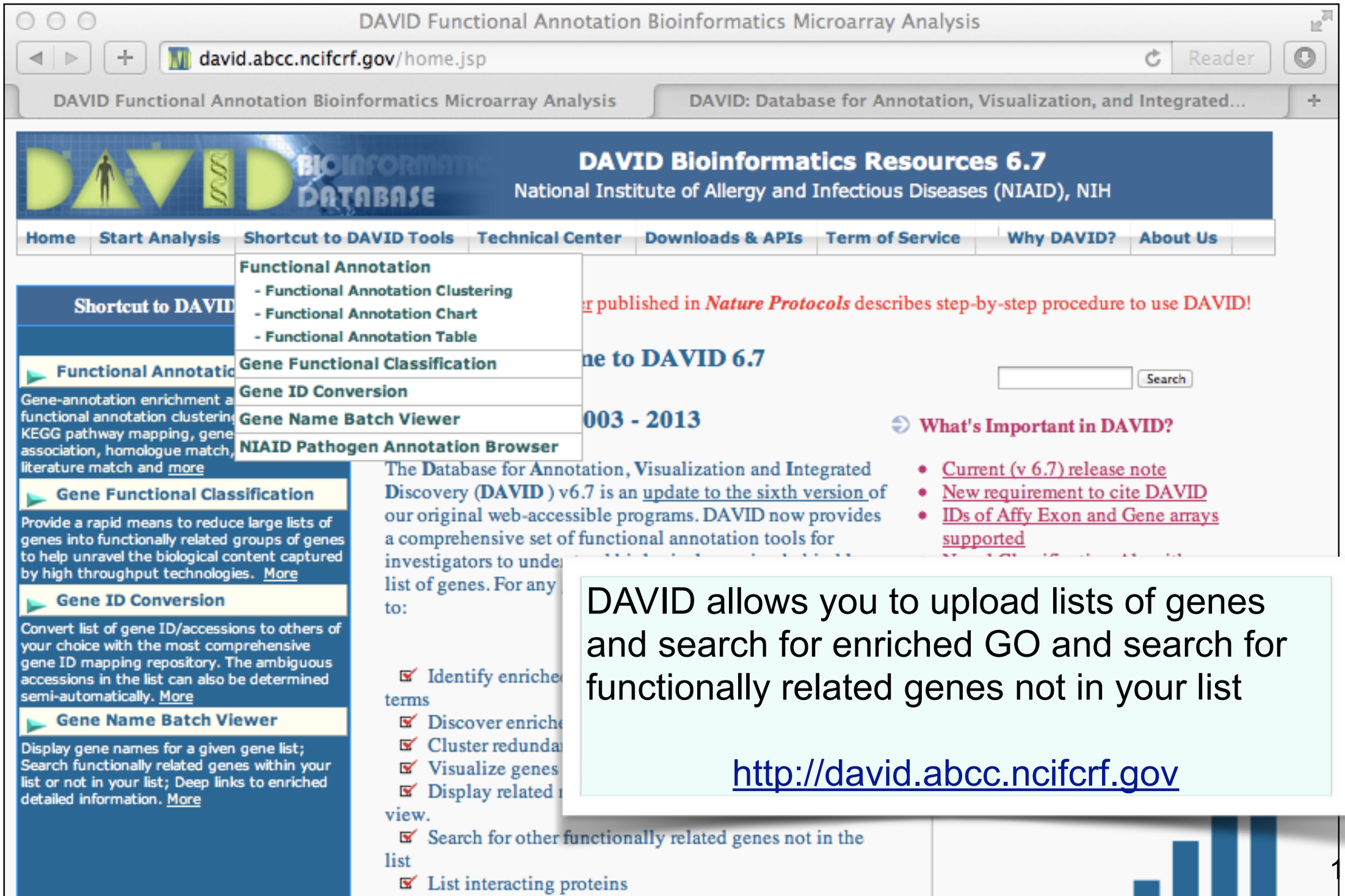


GO annotation in UniProt

An example UniProt entry for hemoglobin beta (HBB_human, P68871) with GO annotation displayed.



DAVID: a online tool for assessing GO term enrichment in gene lists



Example output: enriched functions from GO