

# INTRODUCTION TO BIOINFORMATICS

Please take the initial BIOINF525 questionnaire:  
< <http://tinyurl.com/bioinf525-questions> >

Barry Grant  
**University of Michigan**  
[www.thegrantlab.org](http://www.thegrantlab.org)



**Barry Grant, Ph.D.**  
[bjgrant@umich.edu](mailto:bjgrant@umich.edu)



**Ryan Mills, Ph.D.**  
[remills@umich.edu](mailto:remills@umich.edu)



**Lauren Jepsen (GSI)**  
[ljepsen@umich.edu](mailto:ljepsen@umich.edu)

# COURSE LOGISTICS

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

**Labs:** Thursdays 2:30-4:00 PM  
Rm. 2036 Palmer Commons

**Website:** <http://tinyurl.com/bioinf525-w17>  
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

1.1. ▶ *Introduction to Bioinformatics*

1.2. ▶ *Sequence Alignment and Database Searching*

1.3 ▶ *Structural Bioinformatics*

1.4 ▶ *Genome Informatics: High Throughput Sequencing Applications and Analytical Methods*

# 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 online:  
[PDF1 \(bioinformatics review\)](#),  
[PDF 2 \(bioinformatics challenges\)](#).
- Complete the **lecture 1.1 homework questions**:  
<http://tinyurl.com/bioinf525-quiz1>

## **Q. What is Bioinformatics?**

*“Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.”*

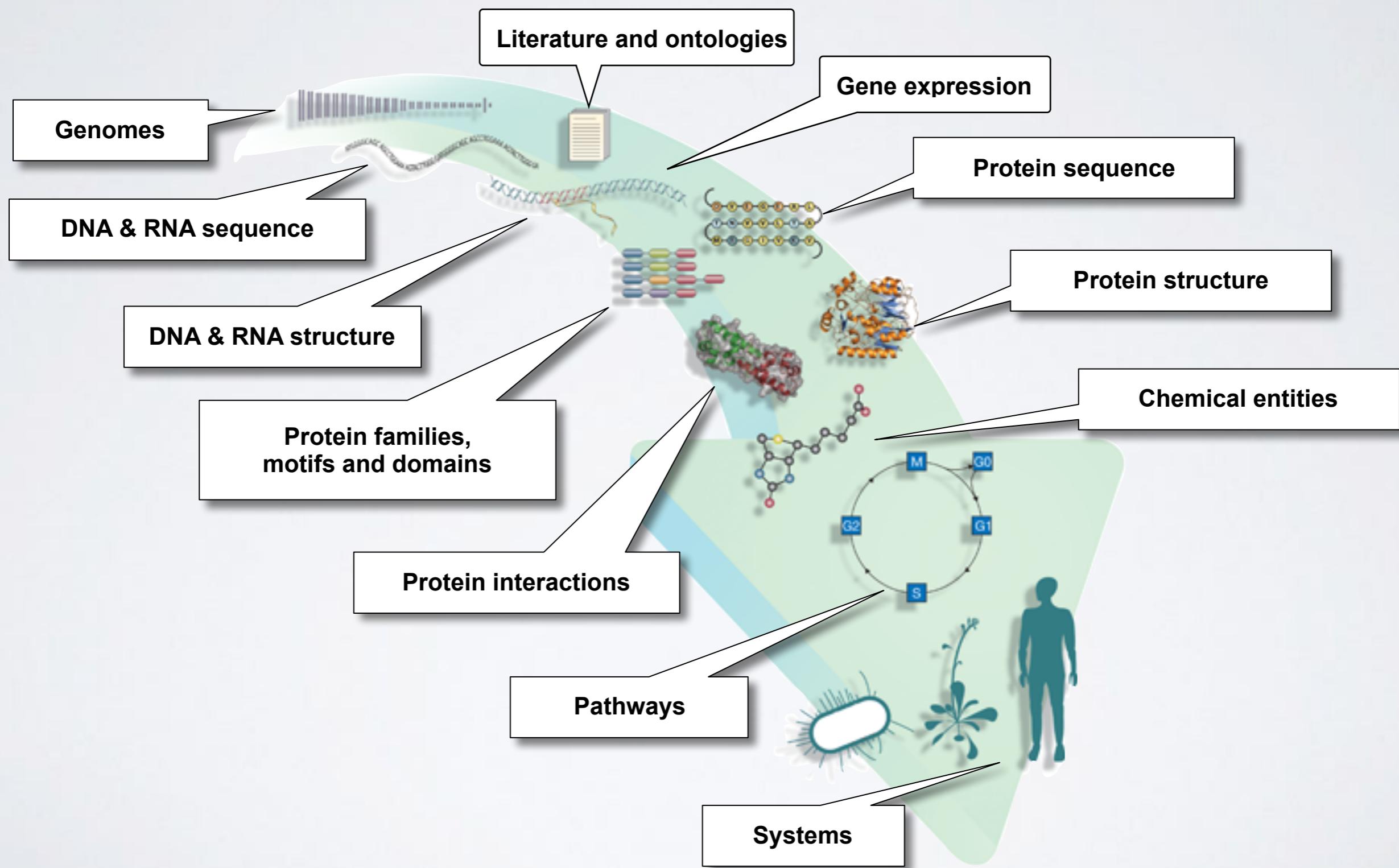
- ... Bioinformatics is a hybrid of biology and computer science
- ... **Bioinformatics is computer aided biology!**

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

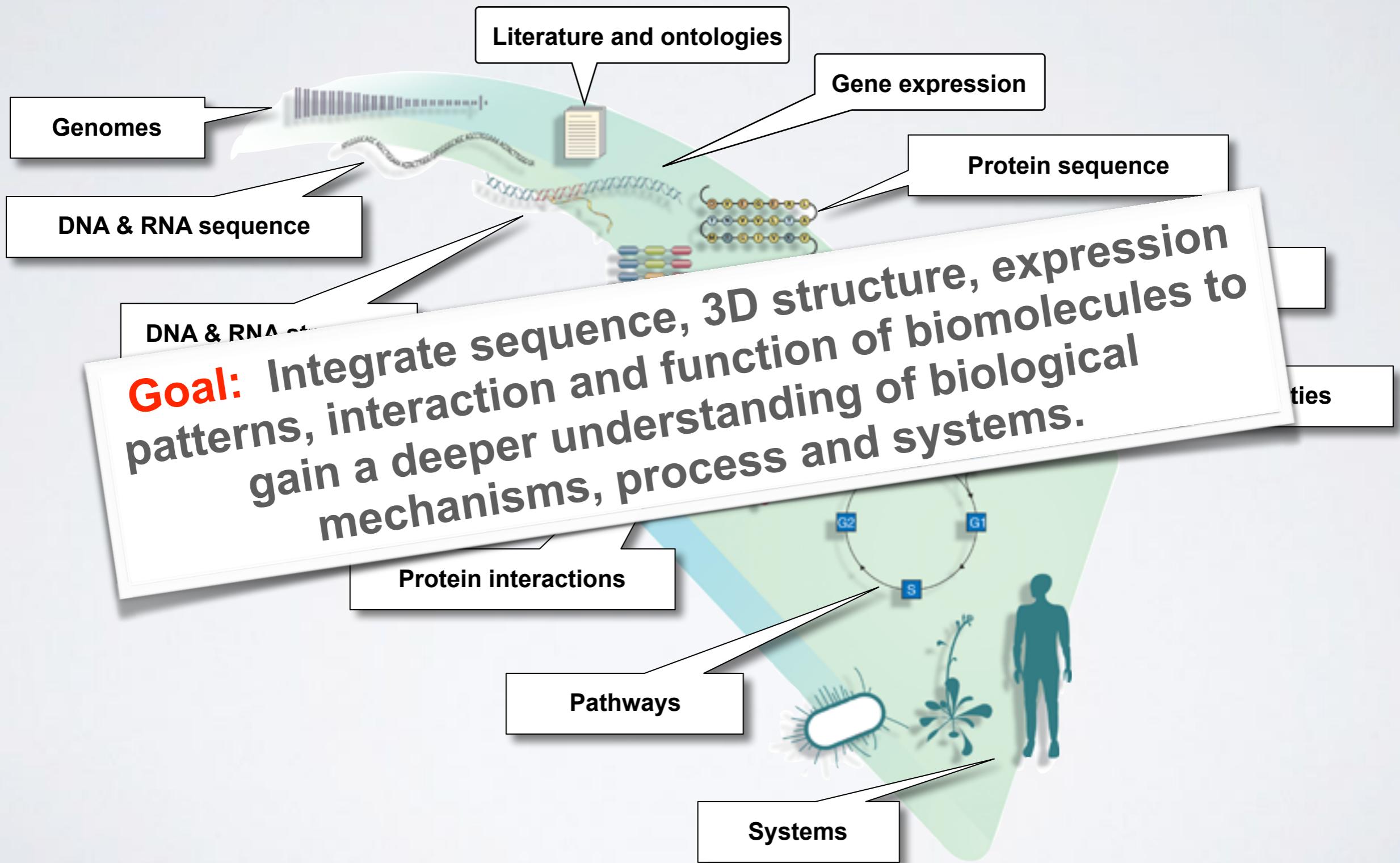
# MORE DEFINITIONS

- ▶ “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> )

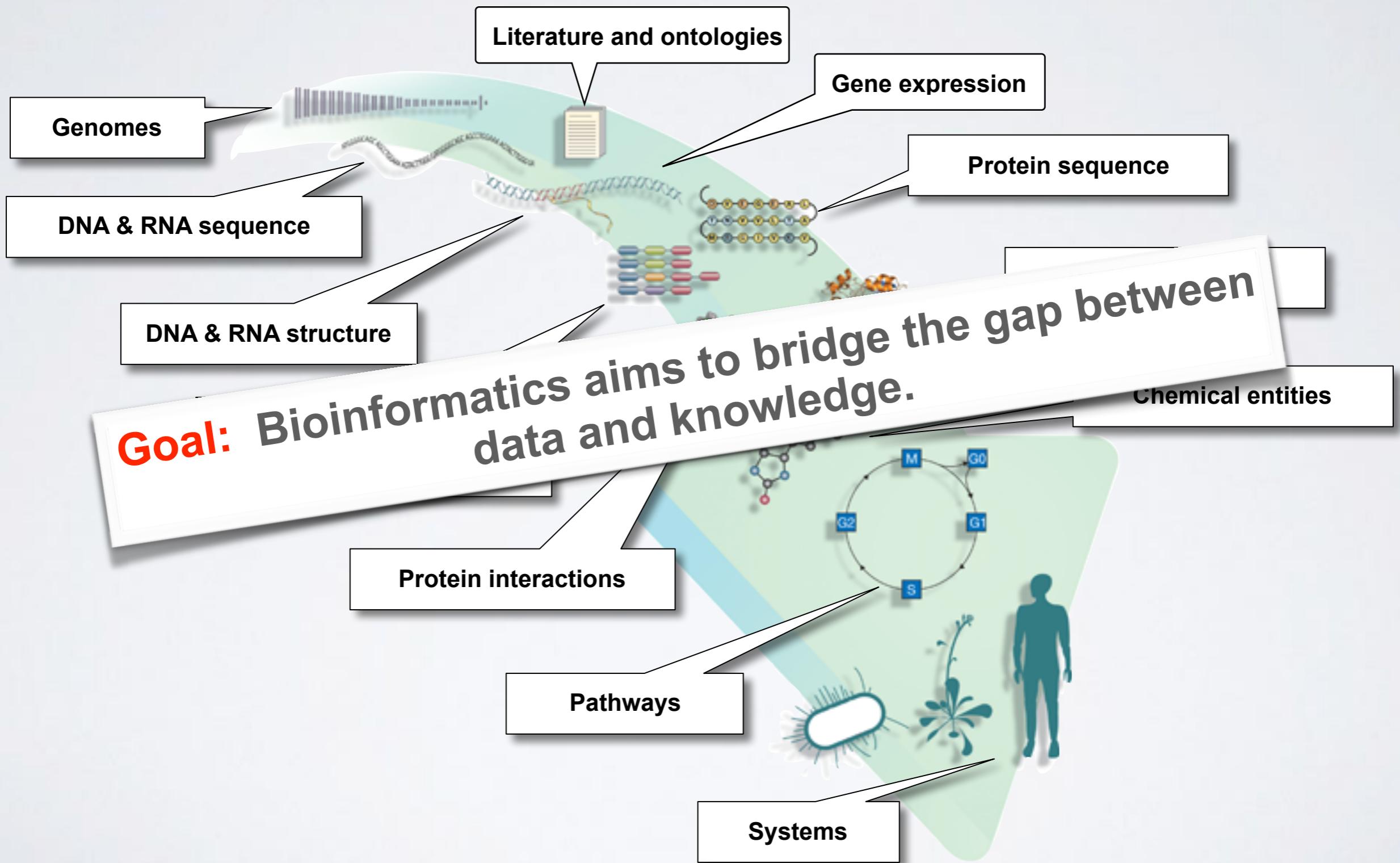
# Major types of Bioinformatics Data



# Major types of Bioinformatics Data



# 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**

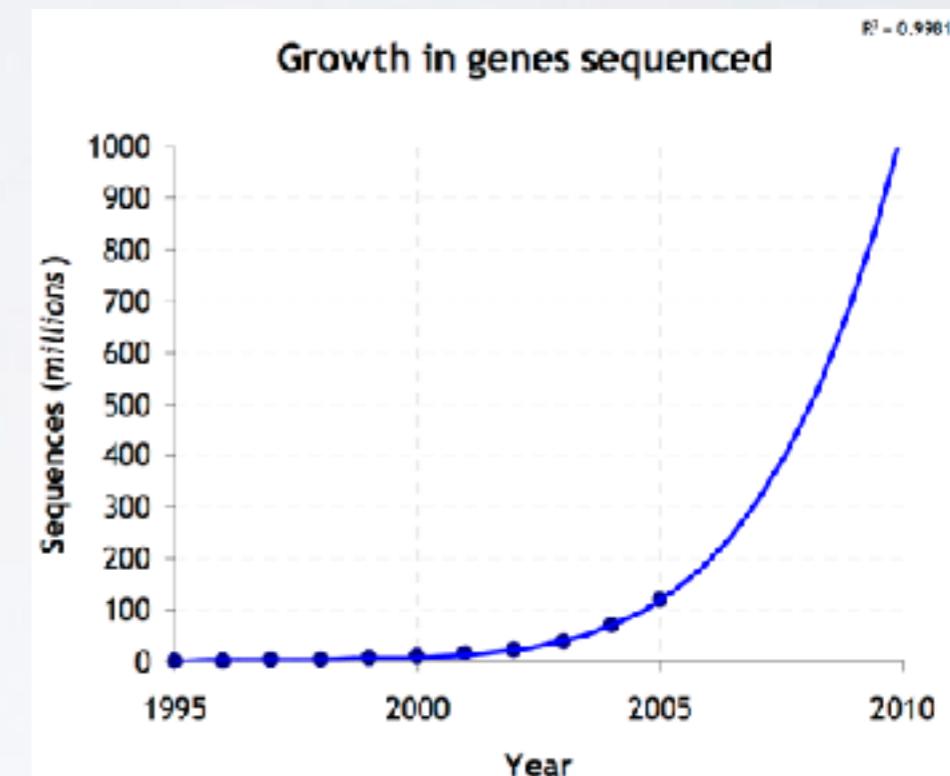
- *DNA sequence determines protein sequence.*
- *Protein sequence determines protein structure.*
- *Protein structure determines protein function.*
- *Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.*

Bioinformatics is now essential for the archiving, organization and analysis of data related to all 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**

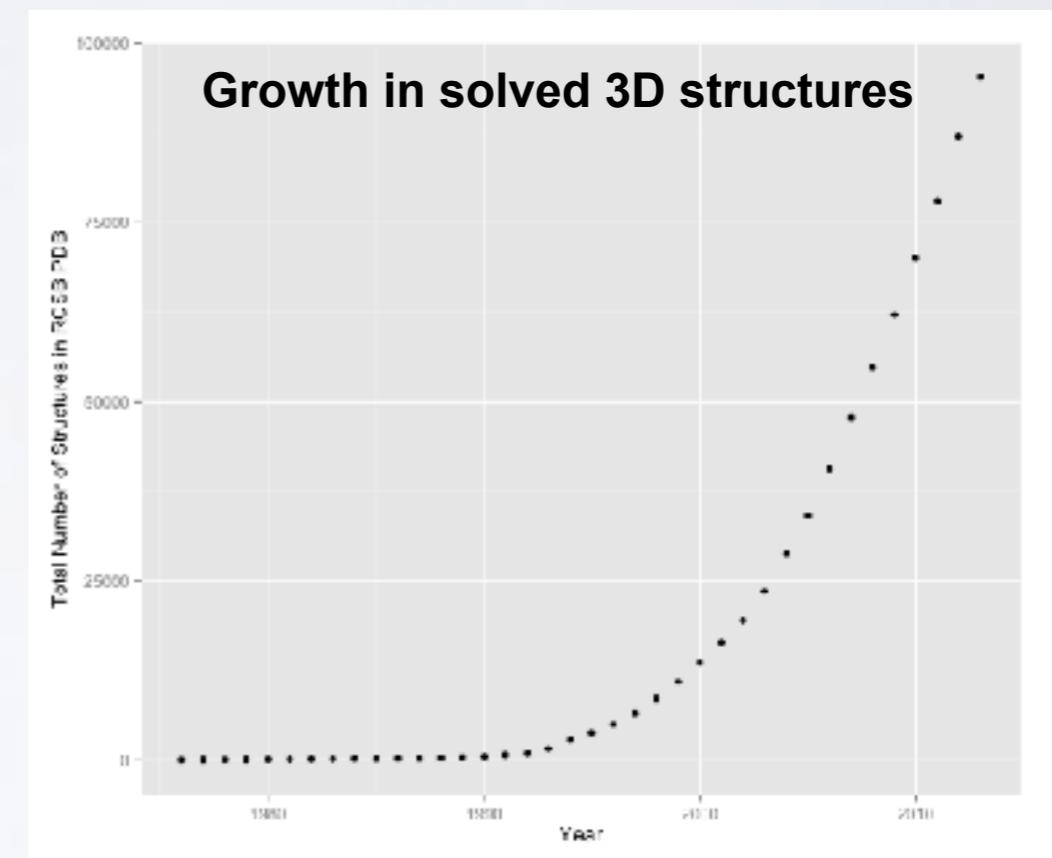


E.G. data from sequencing, structural genomics, microarrays, proteomics, new high throughput assays, etc...

# Why do we need Bioinformatics?

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

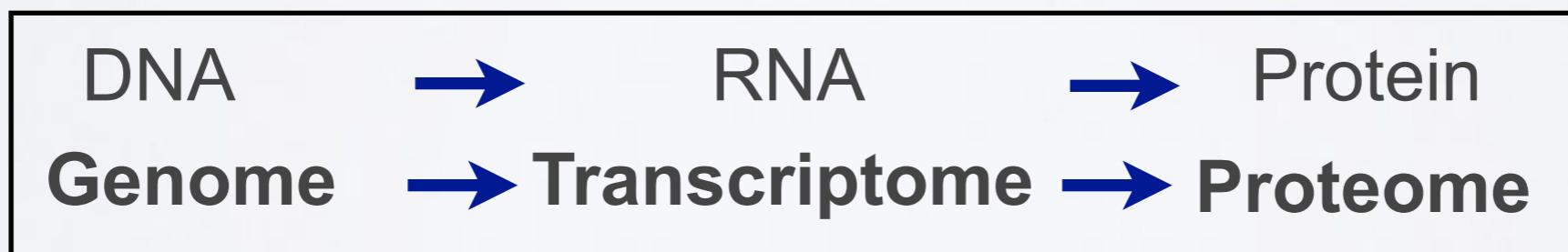
- Bioinformatics provides methods for the efficient:
  - ▶ **storage**
  - ▶ **annotation**
  - ▶ **search and retrieval**
  - ▶ **data integration**
  - ▶ **data mining and analysis**



E.G. data from sequencing, structural genomics, microarrays, proteomics, new high throughput assays, etc...

# 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.



# **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 programming languages frequently required (Python, R, Perl, C Java, Fortran)
- ▶ May require specialized or high performance computing resources...

# 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)

**General Parameters**

**Max target sequences**

Select the maximum number of aligned sequences to display

**Short queries**  Automatically adjust parameters for short input sequences

**Expect threshold**

**Word size**

**Max matches in a query range**

**Scoring Parameters**

**Matrix**

**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**

**Pseudocount**

Even Blast has many settable parameters

**STEP 3 - Set your 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

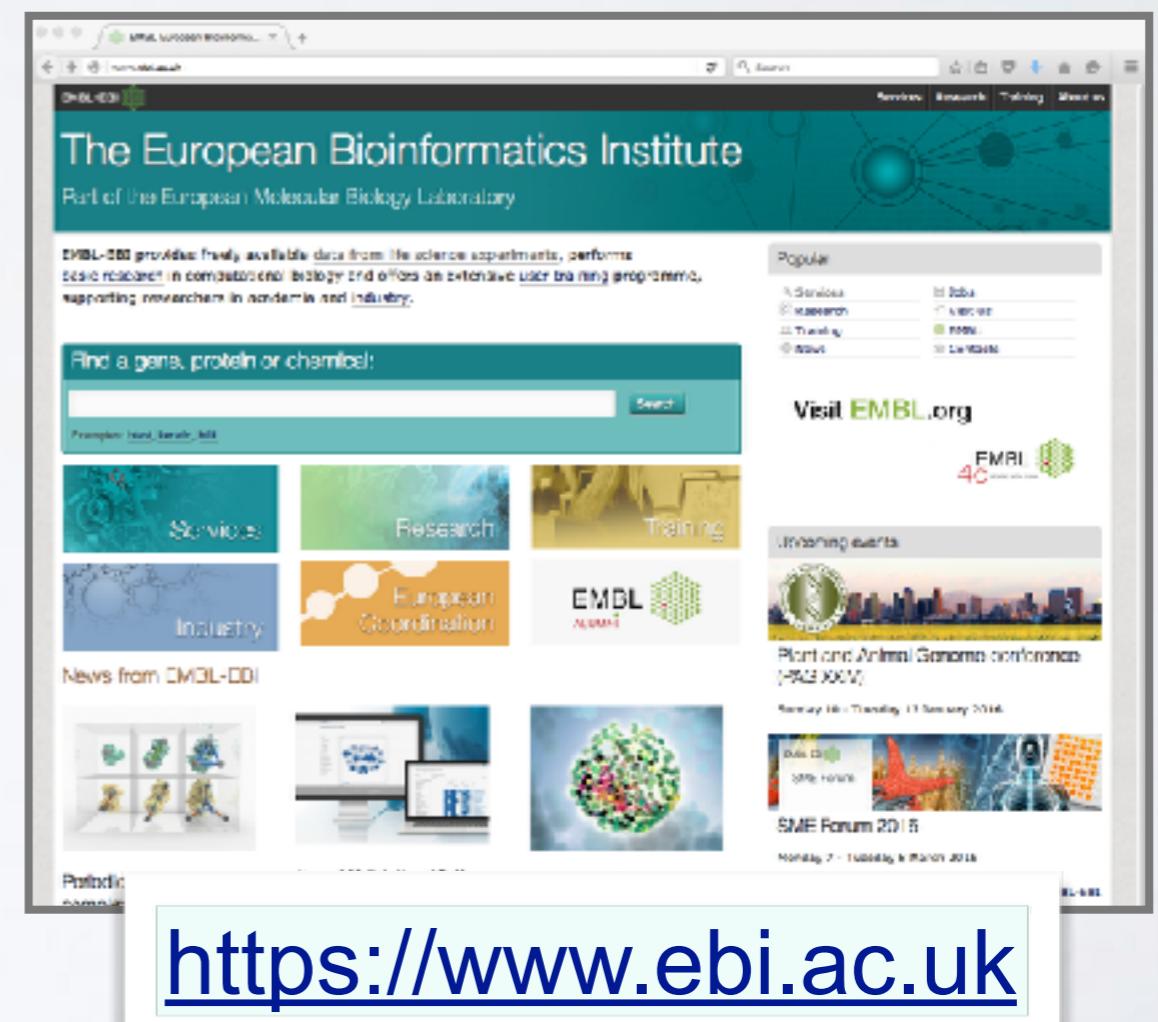
Related tools with different terminology

# Key Online Bioinformatics Resources: NCBI & EBI

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



The screenshot shows the NCBI homepage with a blue header "National Center for Biotechnology Information". Below it is a navigation bar with links to "NCBI Resources", "How To", "Sign in to NCBI", "All Databases", and a search bar. The main content area includes a "Welcome to NCBI" section, a "Get Started" section with links for tools, downloads, how-to guides, and submissions, and a "3D Structures" section featuring a 3D molecular model. On the left, there's a sidebar with links to various NCBI databases like Resource List (A-Z), All Resources, Chemicals & Biosassays, Data & Software, DNA & RNA, Domains & Structures, Genes & Expression, Genetics & Medicine, Genomes & Maps, Homology, Literature, Proteins, Sequence Analysis, Taxonomy, Training & Tutorials, and Variation.



The screenshot shows the EMBL-EBI homepage with a green header "The European Bioinformatics Institute, Part of the European Molecular Biology Laboratory". Below it is a search bar and a "Find a gene, protein or chemical" input field. The main content area features sections for "Services", "Research", "Training", "Industry", "European Coordination", and "News from EMBL-EBI". There are also sections for "Upcoming events" (Plant and Animal Genome conference) and "Past events" (SME Forum 2015). A sidebar on the right lists "Popular" links such as "Advertise", "Research", "Training", "About", "Visit EMBL.org", and "EMBL ALMA".

<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>

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases Search

NCBI Home Resource List (A-Z) All Resources Chemicals & Bioassays Data & Software DNA & RNA Domains & Structures Genes & Expression Genetics & Medicine Genomes & Maps Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

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

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems.

Popular Resources

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

NCBI Announcements

New version of Genome Workbench available 06 Sep

An integrated, downloadable applicati

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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases Search

NCBI Home Resource List (A-Z) All Resources Chemicals & Bioassays Data & Software DNA & RNA Domains & Structures Genes & Expression Genetics & Medicine Genomes & Maps Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

Welcome to NCBI

The National Center for Biotechnology Information provides access to unique information, tools and resources in the fields of medicine, health and biology.

About the NCBI | Mission | Our History

Get Started

- Tools: Analyze data using NCBI's bioinformatics tools
- Downloads: Get NCBI data files and software
- How-To's: Learn how to access and use NCBI resources
- Submissions: Submit data to NCBI's databases

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals and associated biosystems.

Popular Resources

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

Resources

Central Health

Announcements

New version of Genome Workbench available 06 Sep An integrated, downloadable application

A screenshot of the NCBI homepage. The sidebar on the left lists various resources like NCBI Home, Resource List (A-Z), and 3D Structures. The main content area features a 'Welcome to NCBI' section and a 'Get Started' list. On the right, there's a 'Popular Resources' sidebar containing links to PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. A red bracket on the right side of the sidebar groups the 'Gene', 'Protein', and 'PubChem' links, while three separate red arrows point to the 'PubMed', 'BLAST', and 'Gene' links individually.

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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases Search

NCBI Home Resource List (A-Z)

Welcome to NCBI  
The National Center for Biotechnology Information advances science

Popular Resources PubMed

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

and the search tools **ENTREZ** and **BLAST**

Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

databases

3D Structures  
Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals and associated biosystems

Protein PubChem

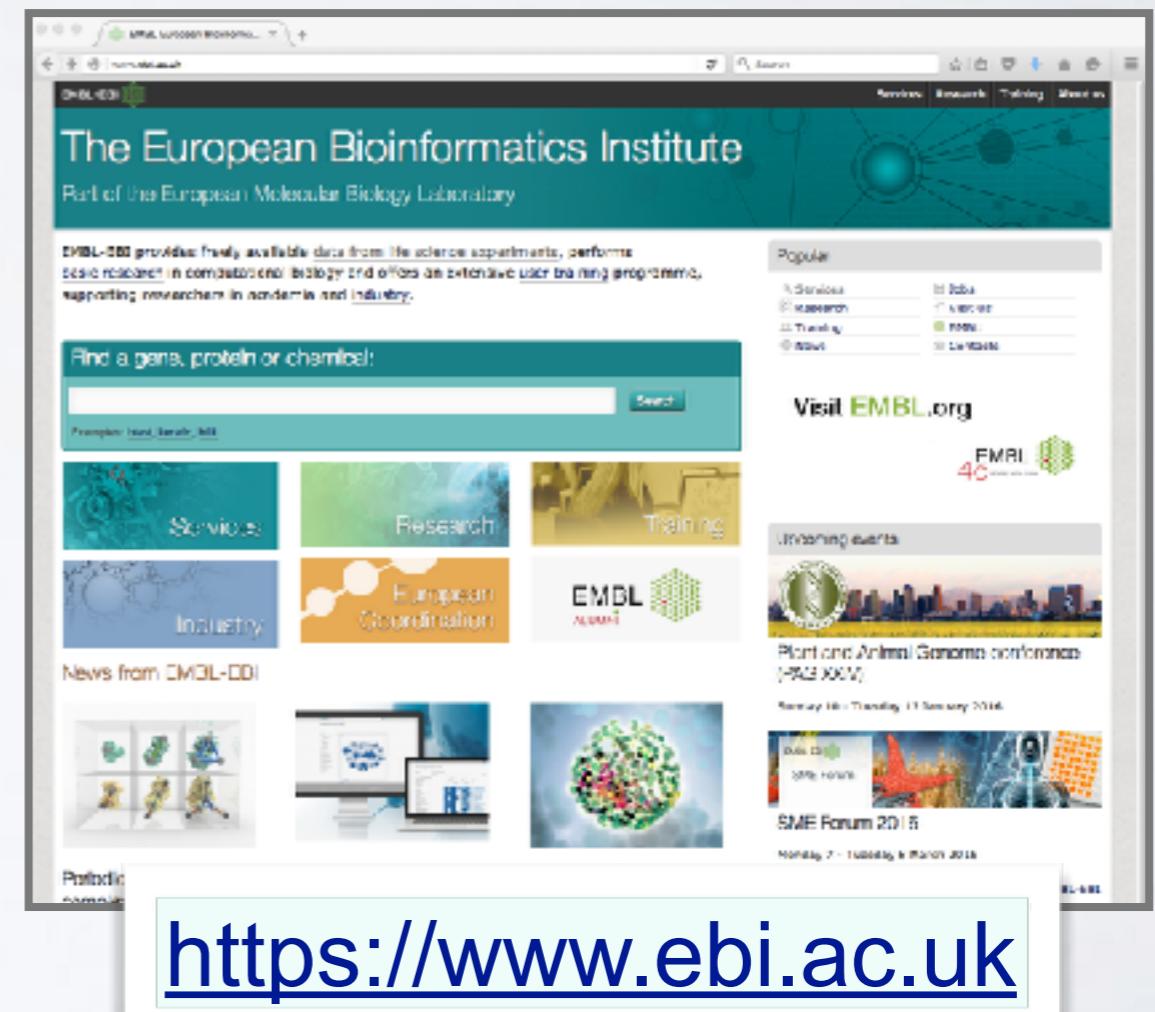
NCBI Announcements  
New version of Genome Workbench available 06 Sep  
An integrated, downloadable applicati

# Key Online Bioinformatics Resources: NCBI & EBI

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



The screenshot shows the NCBI homepage with a blue header bar containing the NCBI logo, a search bar, and a "Sign in to NCBI" button. Below the header, there's a navigation menu with links like "NCBI Home", "Resource List (A-Z)", "All Resources", "Chemicals & Biosassays", "Data & Software", "DNA & RNA", "Domains & Structures", "Genes & Expression", "Genetics & Medicine", "Genomes & Maps", "Homology", "Literature", "Proteins", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation". A main content area features a "Welcome to NCBI" section with a brief description of the center's mission, followed by "Get Started" links for "Tools", "Downloads", "How To's", and "Submissions". There's also a "3D Structures" section with a 3D molecular model and a "NCBI Announcements" box. On the right, there's a sidebar titled "Popular Resources" listing links to PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, NUCOTIDE, GenBank, SNP, Gene, Protein, and PubChem.



The screenshot shows the EBI homepage with a green header bar containing the EBI logo and a search bar. Below the header, there's a main title "The European Bioinformatics Institute" and a subtitle "Part of the European Molecular Biology Laboratory". A central search bar is labeled "Find a gene, protein or chemical:". Below the search bar, there are several colored boxes representing different services: "Services" (blue), "Research" (green), "Training" (yellow), "Industry" (light blue), "European Coordination" (orange), and "News from EMBL-EBI" (grey). To the right, there's a sidebar titled "Popular" with links to "Datasets", "Search", "Training", and "News". Below the sidebar, there are sections for "Visit EMBL.org", "Upcoming events", and "Past and Animal Genome conference (PAG XXV)". At the bottom, there are links for "EMBL-EBI", "EMBL ALMAM", and "SME Forum 2015".

<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 screenshot shows the homepage of the EMBL European Bioinformatics Institute (EBI) at [www.ebi.ac.uk](http://www.ebi.ac.uk). The page features a dark blue header with the EMBL-EBI logo, a search bar, and navigation links for Services, Research, Training, and About us. The main content area has a teal background with the text: "The European Bioinformatics Institute Part of the European Molecular Biology Laboratory". Below this, a paragraph describes the institute's mission: "EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry." A search bar allows users to "Find a gene, protein or chemical" with examples like "blast, keratin, bfl1...". To the right, a "Popular" sidebar lists links for Services, Research, Training, News, Jobs, Visit us, EMBL, and Contacts. A "Visit EMBL.org" section features the EMBL 40th anniversary logo. Another sidebar for "Upcoming events" promotes the "Plant and Animal Genome conference (PAG XXIV)".

EMBL European Bioinforma... [+/-](#)

www.ebi.ac.uk [Search](#) [Services](#) [Research](#) [Training](#) [About us](#)

# The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

Search

Services

Research

Training

EMBL ALUMNI

Industry

European Coordination

News from EMBL-EBI

Popular

Services

Research

Training

News

Jobs

Visit us

EMBL

Contacts

Visit EMBL.org

EMBL 40 YEARS 1974-2014

Upcoming events

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

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

The screenshot shows the EMBL-EBI Services website. At the top, there's a navigation bar with links for Services, Research, Training, and About us. Below the header, a large banner features a blue-toned image of molecular structures and gears. The main content area is titled "Bioinformatics services". It lists several categories: DNA & RNA, Gene expression, Proteins, Structures, Systems, Chemical biology, Ontologies, Literature, and Cross domain. Each category has a small icon and a brief description. To the right, there's a "Popular" sidebar with links to Ensembl, UniProt, PDBc, ArrayExpress, CHEMBL, BLAST, Europe PMC, Reactome, Train online, and Support. Below the sidebar, there's a "Service news" section featuring a photo of a butterfly and a "Training" section.

Services < EMBL-EBI

www.ebi.ac.uk/services

Search

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

**Popular**

- Ensembl
- UniProt
- PDBc
- ArrayExpress
- CHEMBL
- BLAST
- Europe PMC
- Reactome
- Train online
- Support

**Service news**

**Training**

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

The screenshot shows the EBI Services website ([www.ebi.ac.uk/services](http://www.ebi.ac.uk/services)) with a banner for 'Bioinformatics services'. Below the banner, there are nine service categories arranged in a grid:

- DNA & RNA (genes, genomes & variation)
- Gene expression (RNA, protein & metabolite expression)
- Proteins (sequences, families & motifs) - This box is highlighted with a red border.
- 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)

To the right of the grid, a 'Popular' sidebar lists several databases with their logos:

- Ensembl
- UniProt
- PDB
- ArrayExpress
- ChEMBL

At the bottom right, there is a 'Training' section featuring an image of a person working at a computer.

<https://www.ebi.ac.uk>

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

## Proteins

### Popular services

 UniProt	<b>UniProt: The Universal Protein Resource</b> The gold-standard, comprehensive resource for protein sequence and functional annotation data.	<b>Quick links</b> <ul style="list-style-type: none"><li>○ Popular services in this category</li><li>○ All services in this category</li><li>○ Project websites in this category</li></ul>
 InterPro	InterPro A database for the classification of proteins into families, domains and conserved sites.	
 PRIDE	PRIDE: The Proteomics Identifications Database An archive of protein expression data determined by mass spectrometry.	
 Pfam	Pfam A database of hidden Markov models and alignments to describe conserved protein families and domains.	
 Clustal Omega	Clustal Omega Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.	
 HMMER	HMMER - protein homology search Fast sensitive protein homology searches using profile hidden Markov models (HMMs). Variety of different search methods for querying against both sequence and HMM target databases.	
 InterProScan 5	InterProScan 5 searches sequences against InterPro's predictive protein signatures. Please note that InterProScan 4.8 has been retired.	

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

The screenshot shows the homepage of the EMBL European Bioinformatics Institute (EBI) at [www.ebi.ac.uk](http://www.ebi.ac.uk). The page features a dark blue header with the EMBL-EBI logo, a search bar, and navigation links for Services, Research, Training, and About us. Below the header is a teal banner with the text "The European Bioinformatics Institute" and "Part of the European Molecular Biology Laboratory". A large graphic of interconnected nodes is visible on the right side of the banner. The main content area includes a search bar for finding genes, proteins, or chemicals, and several colored boxes for Services, Research, Training, Industry, European Coordination, and EMBL ALUMNI. A red box highlights the "Training" section. To the right, there's a "Popular" sidebar with links to Services, Research, Training, News, Jobs, Visit us, EMBL, and Contacts. A "Visit EMBL.org" section features the EMBL 40th anniversary logo. Below it is an "Upcoming events" section for the Plant and Animal Genome conference (PAG XXIV).

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

Services

Research

Training

Industry

European Coordination

EMBL ALUMNI

Popular

- Services
- Research
- Training
- News
- Jobs
- Visit us
- EMBL
- Contacts

Visit EMBL.org

EMBL 40 YEARS 1974-2014

Upcoming events

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

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

The screenshot shows a web browser displaying the EMBL-EBI Training online course page. The URL in the address bar is [www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-embl-ebi](http://www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-embl-ebi). The page title is "Using sequence similarity searching tools at EMBL-EBI: webinar". The main content area features a video player showing a thumbnail of the webinar presentation. The thumbnail has the same title and includes a photo of Andrew Cowley. Below the video player, a caption reads: "This webinar focuses on how to use tools like BLAST and PSI-Search to find homologous sequences in EMBL-EBI databases, including tips on which tool and database to use, input formats, how to change parameters and how to interpret the results pages." To the left of the video player, there is a sidebar with "Course content" sections for "Using sequence similarity searching tools at EMBL-EBI: webinar" and "Contributors". A "Print Course" link is also present. On the right side, there are "Popular" links for "Train online", "Find us", and "Funding", and a "Find us at..." section with links for "Open days and career days", "Conference exhibitions", "EMBL courses and events", "Genome campus events", and "Science for schools". The top navigation bar includes links for "Services", "Research", "Training" (which is highlighted), and "About us".

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

A screenshot of a web browser displaying the 'Train online | EBI Train online' website. The address bar shows the URL [www.ebi.ac.uk/training/online/](http://www.ebi.ac.uk/training/online/). The page header includes the EMBL-EBI logo and links for 'Find', 'Help', and 'Feedback'. A red 'Beta' badge is visible in the top right corner. The main navigation menu at the top includes 'Databases', 'Tools', 'Research', 'Training', 'Industry', 'About Us', and 'Help'. A secondary navigation bar on the left labeled 'Navigation' contains a link to 'Train online Home'. The central content area features a large heading 'Train online'.

Notable EBI databases include:  
[ENA](#), [UniProt](#), [Ensembl](#)

and the tools [FASTA](#), [BLAST](#), [InterProScan](#),  
[MUSCLE](#), [DALI](#), [HMMER](#)

#### Find a course

##### Browse by subject



[Genes and Genomes](#)



[Gene Expression](#)



[Interactions, Pathways, and Networks](#)

# **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, BiolImage, 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, O-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, RCCP, Beanref, CANSITE, CarbBank, CARBHYD, CATH, CAZy, ChickGBASE, Colibri, COPE, CottonDB, dbSTS, DDBJ, DGP, DictyDb, ECGC, EC02DBASE, FlyBase, GDB, HEPG2, HumanProteome, KEGG, MHCDB, MycoDB, PDBe, PDB, 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, TelDB, 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 online handout

< Handout\_Major\_Databases.pdf >

# Side-note: Databases come in all shapes and sizes



Databases can be of variable quality and often there are multiple databases with overlapping content.

# Finding Bioinformatics Databases

The screenshot shows a web browser displaying the Oxford Journals | Life Sciences | Nucleic Acids Research | Database Summary Paper Categories page. The URL in the address bar is [www.oxfordjournals.org/nar/database/cat/8](http://www.oxfordjournals.org/nar/database/cat/8). The page features a header with the journal's logo and navigation links for CONTACT US, MY BASKET, and MY ACCOUNT. Below the header, there is a large image of a DNA double helix. The main content area is titled "Nucleic Acids Research" and includes links for ABOUT THIS JOURNAL, CONTACT THIS JOURNAL, SUBSCRIPTIONS, CURRENT ISSUE, ARCHIVE, and SEARCH. A breadcrumb trail indicates the current location: Oxford Journals > Life Sciences > Nucleic Acids Research > Database Summary Paper Categories. The main content is titled "2014 NAR Database Summary Paper Category List" and lists various database categories. To the right of the list is a sidebar with links to compilation papers, category lists, alphabetical lists, category/paper lists, and search summary papers. At the bottom of the page is a footer with a link to the full URL.

Oxford Journals | Life Sciences | Nucleic Acids Research | Database Summary Paper Categories  
www.oxfordjournals.org/nar/database/cat/8

Oxford Journals | Life Sciences | Nucleic Acids Research | Database Summary Pa... The 2014 Nucleic Acids Research Database Issue and an updated NAR online M...  
OXFORD JOURNALS CONTACT US MY BASKET MY ACCOUNT

## Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Oxford Journals > Life Sciences > Nucleic Acids Research > Database Summary Paper Categories

### 2014 NAR Database Summary Paper Category List

- [Nucleotide Sequence Databases](#)
- [RNA sequence databases](#)
- [Protein sequence databases](#)
- [Structure Databases](#)
- [Genomics Databases \(ncrn-vertebrate\)](#)
- [Metabolic and Signaling Pathways](#)
- [Human and other Vertebrate Genomes](#)
- [Human Genes and Diseases](#)
  - [CancerResource](#)
  - [Protein Mutant Database](#)
  - [General human genetics databases](#)
  - [General polymorphism databases](#)
  - [Cancer gene databases](#)
  - [Gene-, system- or disease-specific databases](#)
- [Microarray Data and other Gene Expression Databases](#)
- [Proteomics Resources](#)
- [Other Molecular Biology Databases](#)
- [Organelle databases](#)
- [Plant databases](#)
- [Immunological databases](#)
- [Cell biology](#)

- [Compilation Paper](#)
- [Category List](#)
- [Alphabetical List](#)
- [Category/Paper List](#)
- [Search Summary Papers](#)

<http://www.oxfordjournals.org/nar/database/cat/8>

- [Compilation Paper](#)
- [Category List](#)
- [Alphabetical List](#)
- [Category/Paper List](#)
- [Search Summary Papers](#)

Oxford University Press is not responsible for the content of external internet sites.

# 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](#))

There are also many “niche” databases for:

- Classifying 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)

*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 sequence record

Homo sapiens kinesin family member 5A (KIF5A), mRNA – Nucleotide – NCBI  
www.ncbi.nlm.nih.gov/nuccore/NM\_004984.2

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

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide (KIF5A) AND "Homo sapiens" Search Help

Display Settings: GenBank Send: Change region shown

Customize view

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

NCBI Reference Sequence: NM\_004984.2

FASTA Graphics

Go to:

Locus NM\_004984 3897 bp mRNA linear PRI 10-JAN-2014

Definition Homo sapiens kinesin family member 5A (KIF5A), mRNA.

Accession **NM\_004984**

Version NM\_004984.2 GI:45446748

Keywords RefSeq.

Source Homo sapiens (human)

Organism Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea; Homo.

Reference 1 (bases 1 to 3897)

Author Kawaguchi, K.

Title Role of kinesin-1 in the pathogenesis of SDC10, a rare form of hereditary

Journal Neuropathol Appl Neurobiol 22(2):161-71

PubMed 227

Remark GenBank Rev

Reference Authors 2 Prokopenko, Boh

Title alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay

Journal J. Biol. Chem. 288 (30), 21742-21754 (2013)

PubMed 23244021

Analyze this sequence Run BLAST Pick Primers Highlight Sequence Features Find In this Sequence

Articles about the KIF5A gene alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay [J Biol Chem. 2013]

Peptide hormone metabolism

MHC class II antigen presentation

**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)	
NM_006744	RefSeq DNA sequence (from a transcript)	RNA
NP_007635	RefSeq protein	
AAC02945	GenBank protein	
Q28369	UniProtKB/SwissProt protein	Protein
1KT7	Protein Data Bank structure record	
PMID: 12205585	PubMed IDs identify articles at NCBI/NIH	Literature

# GenBank sequence record

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

www.ncbi.nlm.nih.gov/nuccore/NM\_004984.2

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

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide (KIF5A) AND "Homo sapiens" Search Help

Display Settings: GenBank Send: Change region shown

Customize view

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

NCBI Reference Sequence: NM\_004984.2

FASTA Graphics

Go to:

Locus NM\_004984 3897 bp mRNA linear PRI 10-JAN-2014

Definition Homo sapiens kinesin family member 5A (KIF5A), mRNA.

Accession NM\_004984

Version NM\_004984.2 GI:45446748

Keywords RefSeq.

Source Homo sapiens (human)

Organism Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea; Homo.

Reference 1 (bases 1 to 3897)

Authors Kawaguchi, K.

Title Role of kinesin-1 in the pathogenesis of SPC10, a rare form of hereditary spastic paraparesis

Journal Neuroscientist 19 (4), 336-344 (2013)

PubMed 22785106

Remark GeneRIF: A review of the mechanism of pathogenesis involved in spastic paraparesis type 10 when KIF5A is inactivated by mutations. Review article

Reference 2 (bases 1 to 3897)

Authors Prots, I., Veber, V., Brey, S., Campioni, S., Buder, K., Riek, R., Bohm, K.J. and Winner, B.

Title alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay

Journal J. Biol. Chem. 288 (30), 21742-21754 (2013)

PubMed 23244021

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find In this Sequence

Articles about the KIF5A gene

alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay [J Biol Chem. 2013]

Molecular motor KIF5A is essential for GABA(A) receptor transport, a [Neuron. 2012]

Systems-wide analysis of ubiquitylation dynamics reveals a key role for KIF5A in [Nat Cell Biol. 2012]

See all..

Pathways for the KIF5A gene

Peptide hormone metabolism

MHC class II antigen presentation

# GenBank sequence record

Homo sapiens kinesin family member 5A (KIF5A), mRNA – Nucleotide – NCBI  
www.ncbi.nlm.nih.gov/nuccore/NM\_004984.2

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

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide (KIF5A) AND "Homo sapiens" Search Help

Display Settings: GenBank Send: Change region shown

FASTA Graphics Can set different display formats here

NCBI Reference Sequence: NM\_004984.2

Go to:

Locus NM\_004984 3897 bp mRNA linear PRI 10-JAN-2014

Definition Homo sapiens kinesin family member 5A (KIF5A), mRNA.

Accession NM\_004984

Version NM\_004984.2 GI:45446748

Keywords RefSeq.

Source Homo sapiens (human)

Organism Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea; Homo.

Reference 1 (bases 1 to 3897)

Authors Kawaguchi, K.

Title Role of kinesin-1 in the pathogenesis of SPC10, a rare form of hereditary spastic paraparesis

Journal Neuroscientist 19 (4), 336-344 (2013)

PubMed 22785106

Remark GeneRIF: A review of the mechanism of pathogenesis involved in spastic paraparesis type 10 when KIF5A is inactivated by mutations. Review article

Reference 2 (bases 1 to 3897)

Authors Prots, I., Veber, V., Brey, S., Campioni, S., Buder, K., Riek, R., Bohm, K.J. and Winner, B.

Title alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay

Journal J. Biol. Chem. 288 (30), 21742-21754 (2013)

PubMed 23744071

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find In this Sequence

Articles about the KIF5A gene

alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay [J Biol Chem. 2013]

Molecular motor KIF5A is essential for GABA(A) receptor transport, a [Neuron. 2012]

Systems-wide analysis of ubiquitylation dynamics reveals a key role [Nat Cell Biol. 2012]

See all..

Pathways for the KIF5A gene

Peptide hormone metabolism

MHC class II antigen presentation

# FASTA sequence record

Homo sapiens kinesin family member 5A (KIF5A), mRNA – Nucleotide – NCBI  
www.ncbi.nlm.nih.gov/nuccore/45446748?report=fasta

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

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Help

Limits Advanced

Display Settings:  FASTA Send:

Change region shown

Customize view

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

NCBI Reference Sequence: NM\_004984.2

[GenBank](#) [Graphics](#)

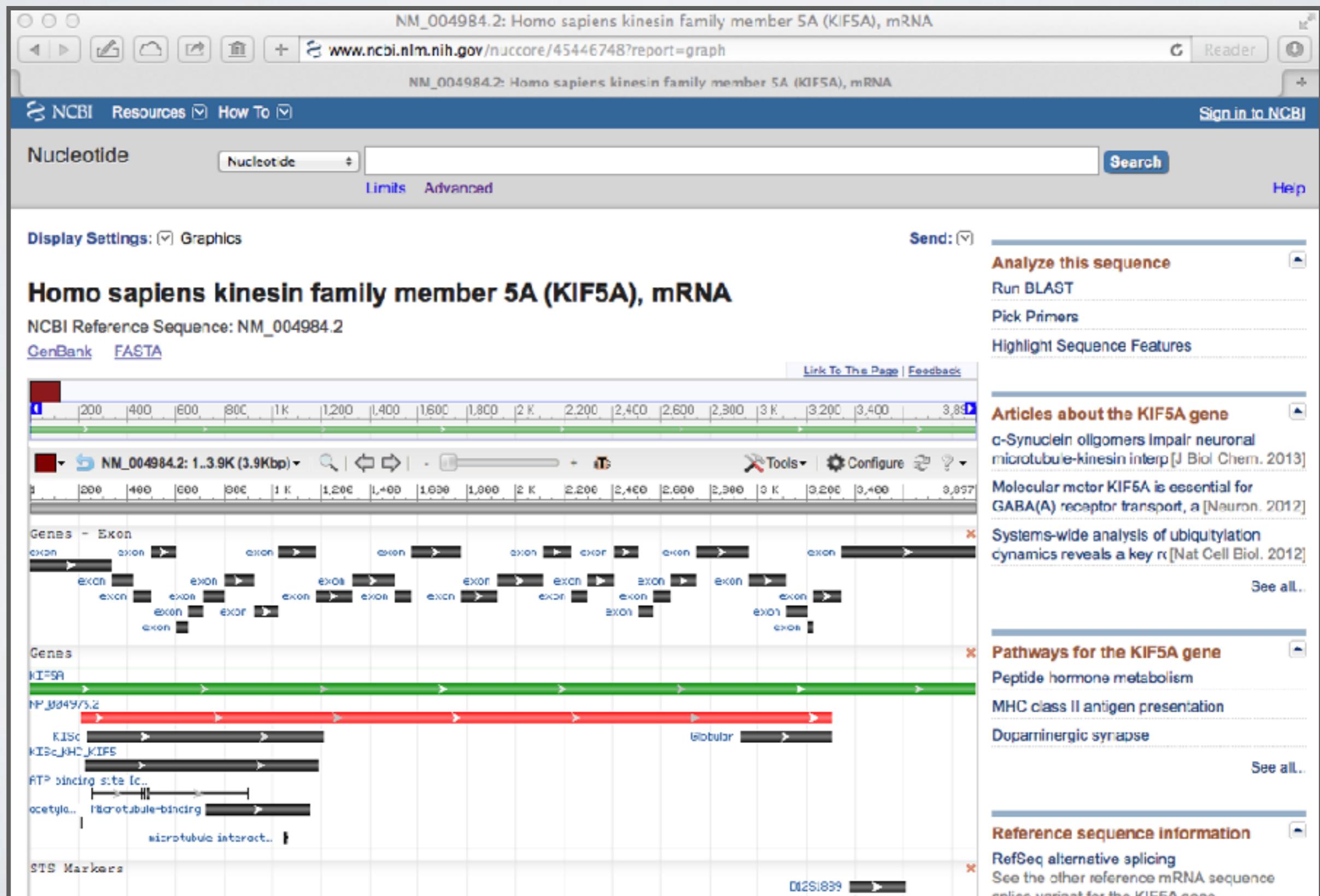
```
>gi|45446748|ref|NM_004984.2| Homo sapiens kinesin family member 5A (KIF5A), mRNA
ACGGCCAGGTGCCCGCATCCCGCTGCCGAGGAGAGACAGCGCCCCCGGCCCTGCTCCCCAGGCTT
CGCCCGGGGGCGCCTCAACTCTGTCCCCAGAGACTGAGCACCTGTCCTCCGCCCTCGCTGAGAGC
CCTCTCCTCTGGAGCACACACCACCCCTGCAGGCCAAGAAAGCTCCAGCCCCACGGGGCTACCACCAT
GGCGGAGACCAACAACGAATGTAGCATCAAGGTGCTCTGCCGATTCCCCCCCCCTGAACCAGGCTGAGATT
CTGCGGGGAGACAAGTTCATCCCCATTTCAGGGGACGACAGCGTCGTTATTGGGGGGAAAGCCATATG
TTTTGACCGTGTATTCCCCCAAACACGACTCAAGAGCAAGTTATCATGCAATGCCATGAGATG
CAAAGATCTCCTTCCCTGGCTACAATGCCACCATTTCTTATGCCACACATCCTCACCCAAAACACAT
ACCATGGACGGAAAGCTGCCACGACCCCTCACCTGTCGAAATCATTCCTCAATTGGCCCCACACATCTCA
ACCACATCTACTCCATGGATGAGAACCTTGAGTTCCACATCAAGGTTCTTACITTGAAATTACCTGG
AAAAATTCTGACCTCTGGATGTGACCAAGACAAATCTGTCGTGCAAGGAGACAAGAACGGGTGCCA
TTTGTCAAGGGTTGTACTGAACGCTTGTGATGAGGAGATTCTGGAGTGATGAGGAGA
AAATCAAATCTCATCTCCCTGCAACCAACATGCAACACACCCCTCCGACCCACACCATCTCCCTCAT
CAACATCAACCGAGAACATGGAACCCAGCAGAGCTCACTGGAAACCTGTATCTGGTGACCTGGCA
GGGAGTGGAGAACGGTACCAAGACTGGAGCAGAGGGAGCCGGTGGACGGAGGAAAGAATAACAAAGT
CACTGTCAGCTCTGGCAATGTGATCTCCGCACTGGCTGAGGGCACTAAAGCTATGTTCCATATCGTGA
CAGCAAAATGACAAGGATTCTCCACGGACTCTCTGGGGAAACTGCCCCACGACTATGTTCATCTGTG
TCACCATCCACTTATAATGTCACACGACCAACTCCACCCCTGATGTTGGCCACGGGGAAAGACCATTA
AGAACACTCCCTCACTAAATTGGACTTCACTGCTGACCAAGAACAAATATGACAACGACANCGA
GAAGACAAAGGGCCCACAGGAGACGATTGGCAACCTGGAGGGCTGAGCTGAGGGGGTGGCGCAATGGAGAG
AATGTGCCCTGAGACAGAGCGCTGGCTGGGGAGGGAGGAGCAGCCCTGGGAGCCGAGCTCTGTGAGGGAGACCC
CTGTGAATGACAACCTCATCCATCGTGGCGCATCGGCCCGAGGAGCCAGAAATACGAGGAGGAGAT
CCCCCGCTCTATAACCCACCTCACCACAAACGATGATGAAATCAACCAACAAACCGCAACTCATACGAAAC
CTCAACCCACCAAAATGCTGGACCCACCAAGACCTGCTGCTCCACCCGACGACAAACGAGAACGCTCCACG
GGGAGCTGAGGCCACCTGCAATCAGAGAACGATGCCGCTAAGGATGAGGTGAAGGAAGTGTGCAAGGCCCT
GGAGGAGCTGGCTGTGAACATATGACCGAGAAGTCCCAGGGAGGAGGAGAGGCCAGCAGAACCGAGCTT
CTGGTGGATGAGCTGTCTCAGAAGCTGGCCACCATGCTGTCTGGAGTCTGAGTTGAGCAGGCTACAGG
ACGTCACTGGACACACCACCCAAAACCAATTGCTGACCTGCTGACCCGCTGATGAAACGATCTGACCCGACTT
```

FASTA sequence files consist of records where each record begins with a “>” and header information on that same line. Each subsequent line of the record is sequence information.

This format is commonly used by sequence analysis programs.

Pathways for the KIF5A gene  
Peptide hormone metabolism  
MHC class II antigen presentation

# GenBank ‘graphics’ sequence record



# GenBank sequence record, cont.

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

www.ncbi.nlm.nih.gov/nuccore/NM\_004984.2

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

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide (KIF5A) AND "Homo sapiens" Search Help

Display Settings: GenBank Send: Change region shown

Customize view

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

NCBI Reference Sequence: NM\_004984.2

FASTA Graphics

Go to:

Locus NM\_004984 3897 bp mRNA linear PRI 10-JAN-2014

Definition Homo sapiens kinesin family member 5A (KIF5A), mRNA.

Accession NM\_004984

Version NM\_004984.2 GI:45446748

Keywords RefSeq.

Source Homo sapiens (human)

Organism Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea; Homo.

Reference 1 (bases 1 to 3897)

Authors Kawaguchi, K.

Title Role of kinesin-1 in the pathogenesis of SPC10, a rare form of hereditary spastic paraparesis

Journal Neuroscientist 19 (4), 336-344 (2013)

PubMed 22785106

Remark GeneRIF: A review of the mechanism of pathogenesis involved in spastic paraparesis type 10 when KIF5A is inactivated by mutations. Review article

Reference 2 (bases 1 to 3897)

Authors Prots, I., Veber, V., Brey, S., Campioni, S., Buder, K., Riek, R., Bohm, K.J. and Winner, B.

Title alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay

Journal J. Biol. Chem. 288 (30), 21742-21754 (2013)

PubMed 23244021

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find In this Sequence

Articles about the KIF5A gene

alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay [J Biol Chem. 2013]

Molecular motor KIF5A is essential for GABA(A) receptor transport, a [Neuron. 2012]

Systems-wide analysis of ubiquitylation dynamics reveals a key role [Nat Cell Biol. 2012]

See all..

Pathways for the KIF5A gene

Peptide hormone metabolism

MHC class II antigen presentation

# 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

NUC

Disp

Ho

NCB

FAST

Go to

LOCU

DEPI

ACCE

VERS

KEYW

SCUP

OR

REFS

AU

TI

JU

RE

REFS

AU

TI

JO

FEATUR

source

gene

exon

misc\_feature

CDS

OMIM

Probe

Protein

PubMed

PubMed (RefSeq)

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"  
134..136  
/gene="KIF5A"  
/gene\_synonym="D12S1889; MY050; NKHC; SPG10"  
/note="upstream in-frame stop codon"

misc\_feature 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"

CDS /translation="MAETNNECSIKVLCRFRPLNQAEILRGDKFIFIFQGDDSVVIGG  
KPYVFDRVFPPTTQEQVYHACAMQIVKDVLASYNTIFAYGQTSSGKTHTMECKLHD  
PQLMCIIIPRIARDIFNHIYSMDENLEPHIKVSYFEIYLDKIRDLLDTKTNLSVHEDK  
NRVPFVKGCTERFVSSPEEILDVIDEGKSNRHVAVTNMNEHSSRSHSIFLINIKQENM  
ETEOKLSGKLYLVLAGSEKSKTGAEGAVLDEAKNINKSLSALGNVISALAEGTKSY  
VPYRDSLKMTRILQDSLGGMCRTTMFICCSPPSYNDAETKSTLMFGQRAKTIKNIASVN

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"[orgn] (1351) Nucleotide

kinesin (37064) Nucleotide

See more..

# GenBank sequence record, cont.

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

[1] "The quaking gene product necessary in embryogenesis and myelination combines features of RNA binding and signal transduction proteins."  
Ebersole T.A., Chen Q., Justice M.J., Artzt K.  
Nat. Genet. 12:260-265 (1996) [PubMed: 8589716] [Abstract]  
Cited for: NUCLEOTIDE SEQUENCE [mRNA] (ISOFORM 3), INVOLVEMENT IN QKV, TISSUE SPECIFICITY, MUTAGENESIS

[2] "Genomic organization and alternative splicing of the mouse *Qkv* gene."  
Kondo T., Furuta T., Misunaga K., Ebersole T.A., Shichiri M., Wu J., Artzt K., Yamamura K., Abe K.  
Mamm. Genome 10:662-669 (1999) [PubMed: 10384037] [Abstract]  
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA / mRNA] (ISOFORMS 2; 3; 4 AND 7), ALTERNATIVE SPlicing (ISOFORM 1).  
Strain: 129/J.

**Literature**

**Annotations**

General annotation (Comments)

Function: RNA-binding protein that plays a central role in myelination. Also required for visceral endoderm function and blood vessel development. Binds to the 5'-ACUAUAY-N12-CAUA-3' RNA core sequence. Acts by regulating pre-mRNA splicing, mRNA export, mRNA stability and protein translation, as well as cellular processes including apoptosis, cell cycle, glial cell fate and development. Required to protect and promote the integrity of the MBP and CbP2 mRNAs to maintain oligodendrocyte differentiation. Participates in mRNA transport by regulating the nuclear export of MBP mRNA. Isoform 1 is involved in regulation of mRNA tailoring of MAG pre-mRNA by acting as a negative regulator of MAG exon 12 alternative splicing. Isoform 3 can induce apoptosis, while heterodimerization with other isoforms result in nuclear translocation of isoform 3 and suppression of apoptosis. Isoform 4 acts as a translational repressor for GLT. May also play a role in smooth muscle development.

Subunit structure: Homodimer. Does not require RNA to homodimerize. Able to heterodimerize with BICC1.

Subcellular location: Cytoplasm, Nucleus. Note: Isoform 1 localizes predominantly in the nucleus and at lower level in cytoplasm. It shuttles between the cytoplasm and the nucleus. Isoform 3 localizes predominantly in the cytoplasm and at much lower level in nucleus. Isoform 4 localizes both in the cytoplasm and nucleus.

Tissue specificity: High expression in brain and peripheral nerve. Expresses in Schwann cells and oligodendrocytes in the central nervous system as well as Schwann cells. Isoform 1 is predominantly expressed in the mesodermal site of developing blood vessels. Isoform 3 is predominantly expressed in the peripheral nervous system, but that expression is down-regulated during myelination. Isoform 4 is predominantly located in the peripheral nervous system, but that expression is down-regulated during myelination. Isoforms 2 and 5 are predominantly expressed in the brain, but isoform 2 is expressed in the peripheral nervous system and migrate away into the emerging nervous system. These have characteristics consistent with the acquisition of a glial rather than neuronal fate (at protein level). First detected in the neuroepithelium of the head folds at E7.5. Expression is strongly present ventrally in the forebrain and neural tube of E8.5 and E10.5 and in the heart. Isoform 1 is expressed in early embryos, while isoform 3 and 4 are expressed later in the embryo. Isoform 5 is expressed in the peripheral nervous system.

Developmental stage: Increased expression of isoform 1 is observed during the first postnatal week (P1-P7). During the vigorous accumulation of MBP mRNA between P7 and P20, phosphorylation in the developing myelin drastically declined. By the end of the fourth postnatal week (P28), phosphorylation is reduced approximately 90%.

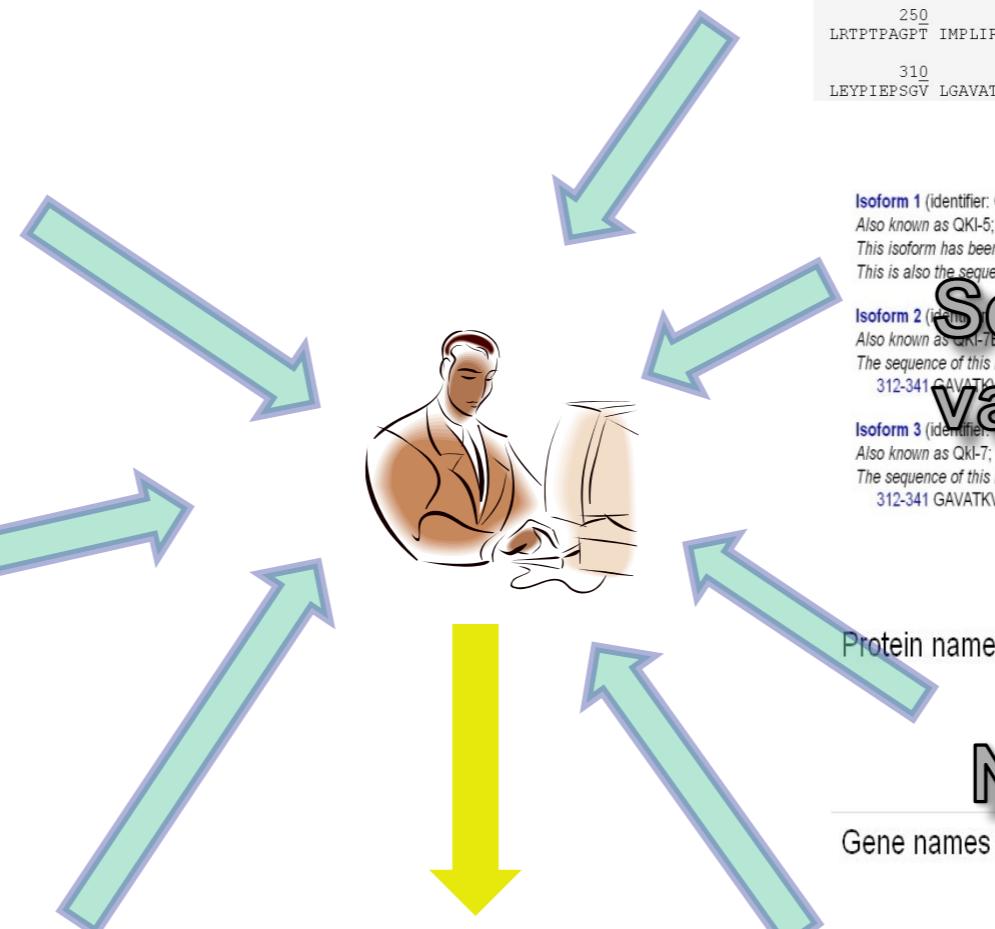
Post-translational modification: The level of tyrosine phosphorylation of the protein is increased during myelination, suggesting a change in binding affinity, affecting transport and stabilization of MBP mRNA. The level of tyrosine phosphorylation in the developing myelin is highest in the first postnatal week (P7). During the vigorous accumulation of MBP mRNA between P7 and P20, phosphorylation in the developing myelin drastically declined. By the end of the fourth postnatal week (P28), phosphorylation is reduced approximately 90%.

Involvement in disease: Defects in *Qkv* are the cause of quaking (qkv). *Qkv* is a spontaneous mutation resulting in hypomyelination of the central and peripheral nervous systems. Mutant mice develop normally until postnatal day 10 when they display rapid tremors or 'quaking' that is especially pronounced in hindlimbs and experience convulsive tonic-clonic seizures as they mature. Mice with *qkv* specifically lack isoform 3.

**Ontologies**

Cell cycle  
DNA damage  
DNA repair  
Fatty acid biosynthesis  
Lipid synthesis  
Nucleus  
Polymorphism  
Disease mutation  
Repeat  
Zinc-finger  
DNA-binding  
Metal-binding  
Zinc  
Anti-oncogene  
Phosphorylation  
3D-structure

Regulation of cell proliferation  
Traceable author statement. Source: UniProtKB  
Regulation of transcription from RNA polymerase II promoter  
Traceable author statement. Source: Protic  
Regulation of transcription from RNA polymerase III promoter  
Traceable author statement. Source: UniProtKB  
Response to estrogen stimulus  
Traceable author statement. Source: UniProtKB  
BRCA1-BARD1 complex  
Inferred from direct assay. Source: UniProtKB  
Gamma-tubulin ring complex  
Non-traceable author statement. Source: UniProtKB  
DNA binding  
Traceable author statement. Source: Protic  
Androgen receptor binding  
Non-traceable author statement. Source: UniProtKB  
Enzyme binding  
Inferred from physical interaction. Source: UniProtKB



**UniProt /Swiss-Prot**

10	20	30	40	50	60
MVGEMETKEK	PKPTPDYLMQ	LMNDKKLMSS	LPNFCGIFNH	LERLLDEEIS	RVRKDMYNDT
70	80	90	100	110	120
LNGSTEKRSA	ELPDAVGIV	QLQEKLKYPV	KEYPDFNFVG	RILGPRGLTA	KQLEAETGCK
130	140	150	160	170	180
IMVRGKGSMR	DYKQKQNG	KENWEHLNED	IHLVLLTVEDA	ONRAEIKLKR	AVEEVKKLLV
190	200	210	220	230	240
PAAEGEDSLK	KMQLMELAIL	NGT	RDANIK	SPALAFSIAA	TAQAAPIIT
250	260	270	280	290	300
LRTPTFAGT	IMPLRQIQT	AVMPNGTPHP	AAIVPPGPE	AGLIYTPYEY	PTYLAPATSI
310	320	330	340		
LEYPIEPSGV	LGAVATKVRR	HDMRVHPYQR	IVTADRAATG	N	

## Sequence

### Isoform 1 (identifier: Q9QYS9-1)

Also known as QKI-5;  
This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it.  
This is also the sequence that appears in the downloadable versions of the entry.

## Sequence variants

### Isoform 2 (identifier: Q9QYS9-2)

Also known as QKI-7B;  
The sequence of this isoform differs from the canonical sequence as follows:  
312-341 GAVATKVRHHDMRVPYQRIVTADRAATGN → VWLSQRKAKNSRTVLTEPSSDLNLNTA

### Isoform 3 (identifier: Q9QYS9-3)

Also known as QKI-7;  
The sequence of this isoform differs from the canonical sequence as follows:  
312-341 GAVATKVRHHDMRVPYQRIVTADRAATGN → EWIEMPVMPDISAH

## Protein names

### Protein quaking

Also known as:

Mqkl

## Gene names

Name: **Qki**

Synonyms: Qk, Qk1, Qka1

## Molecule processing

Chain

1 – 341      341 Protein quaking

Regions

87 – 153      67 KH

Domain

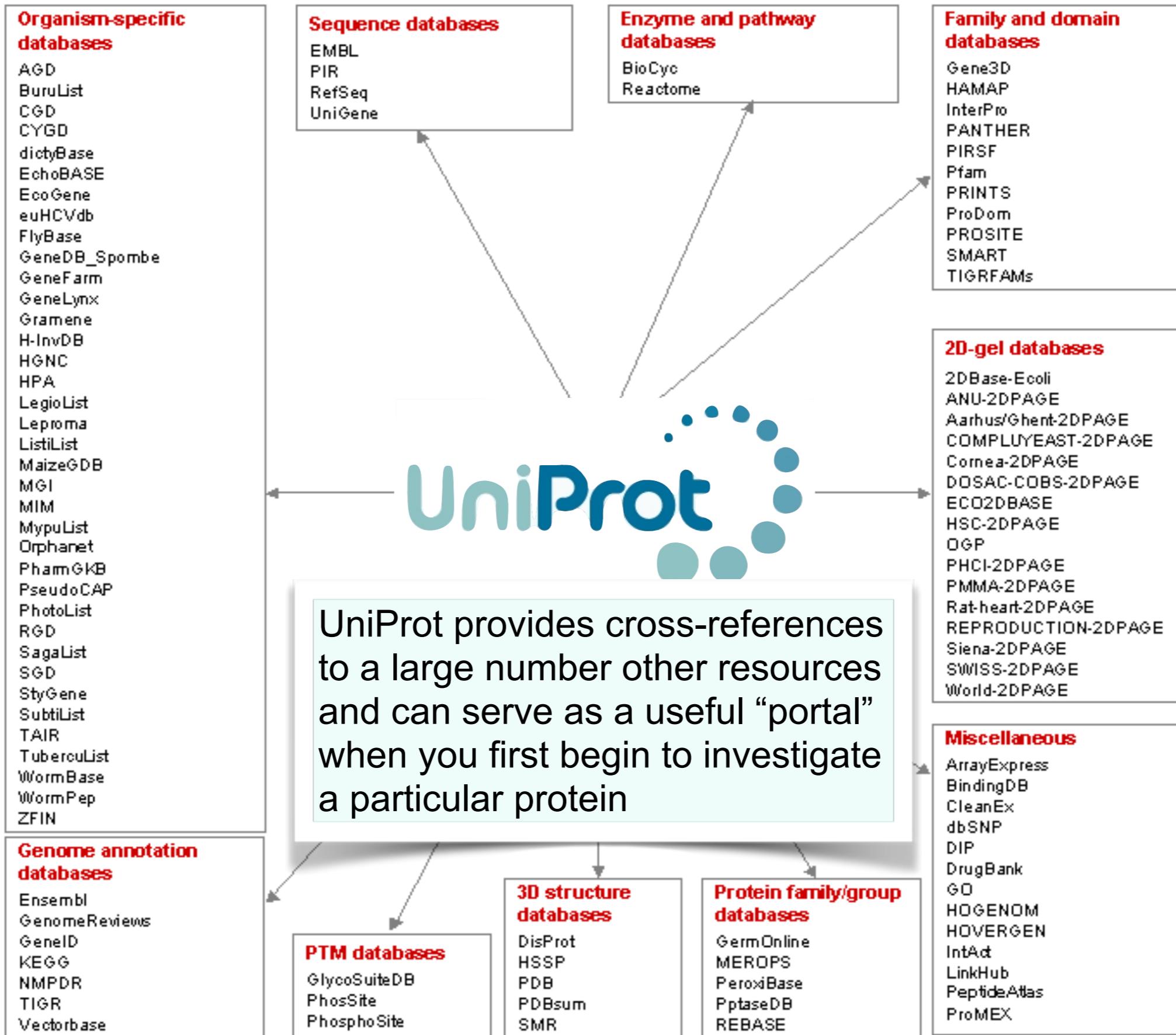
276 – 286      10 NLS

Motif

324 – 330      7 Nuclear localization signal

## Sequence features

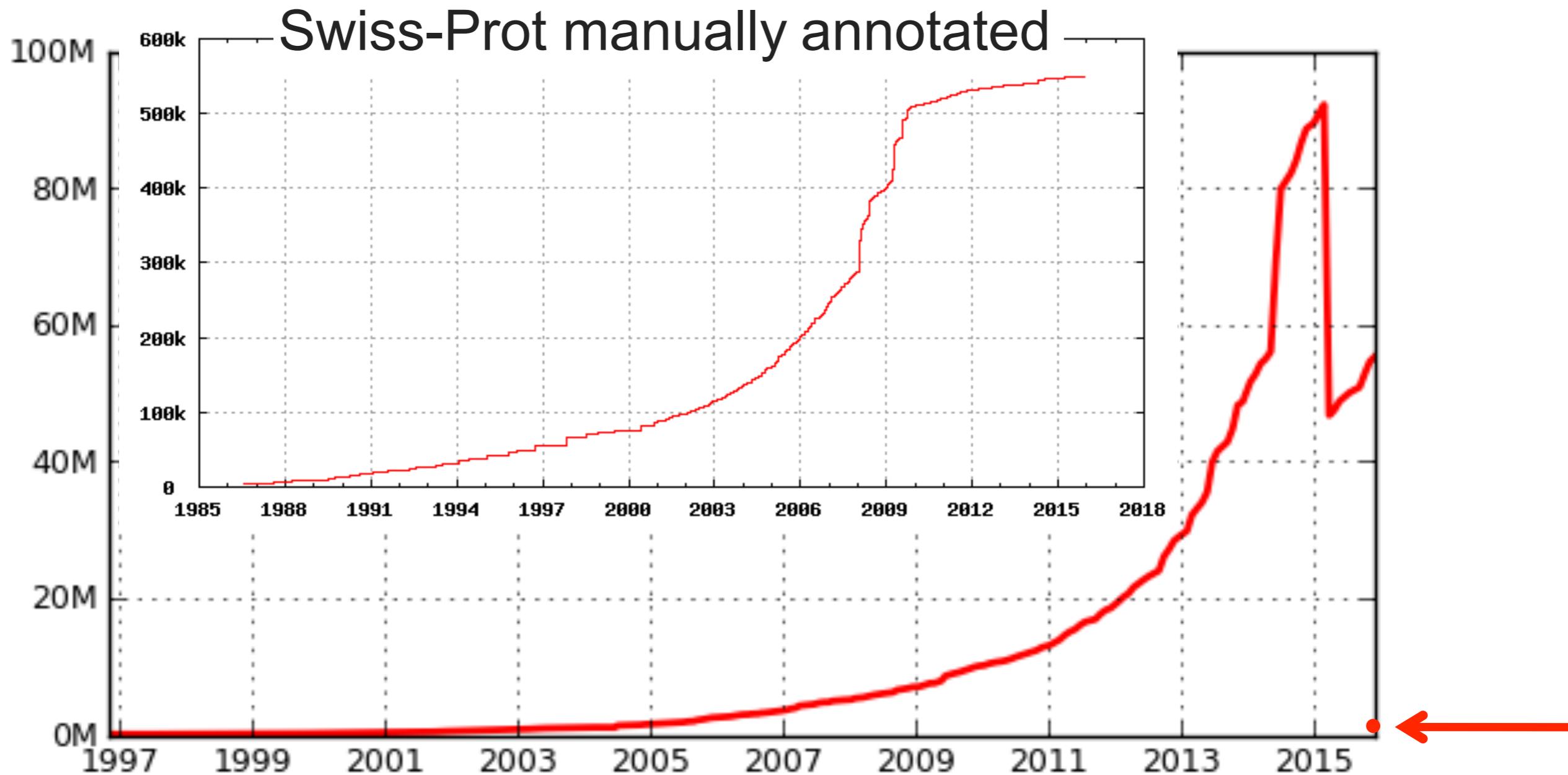




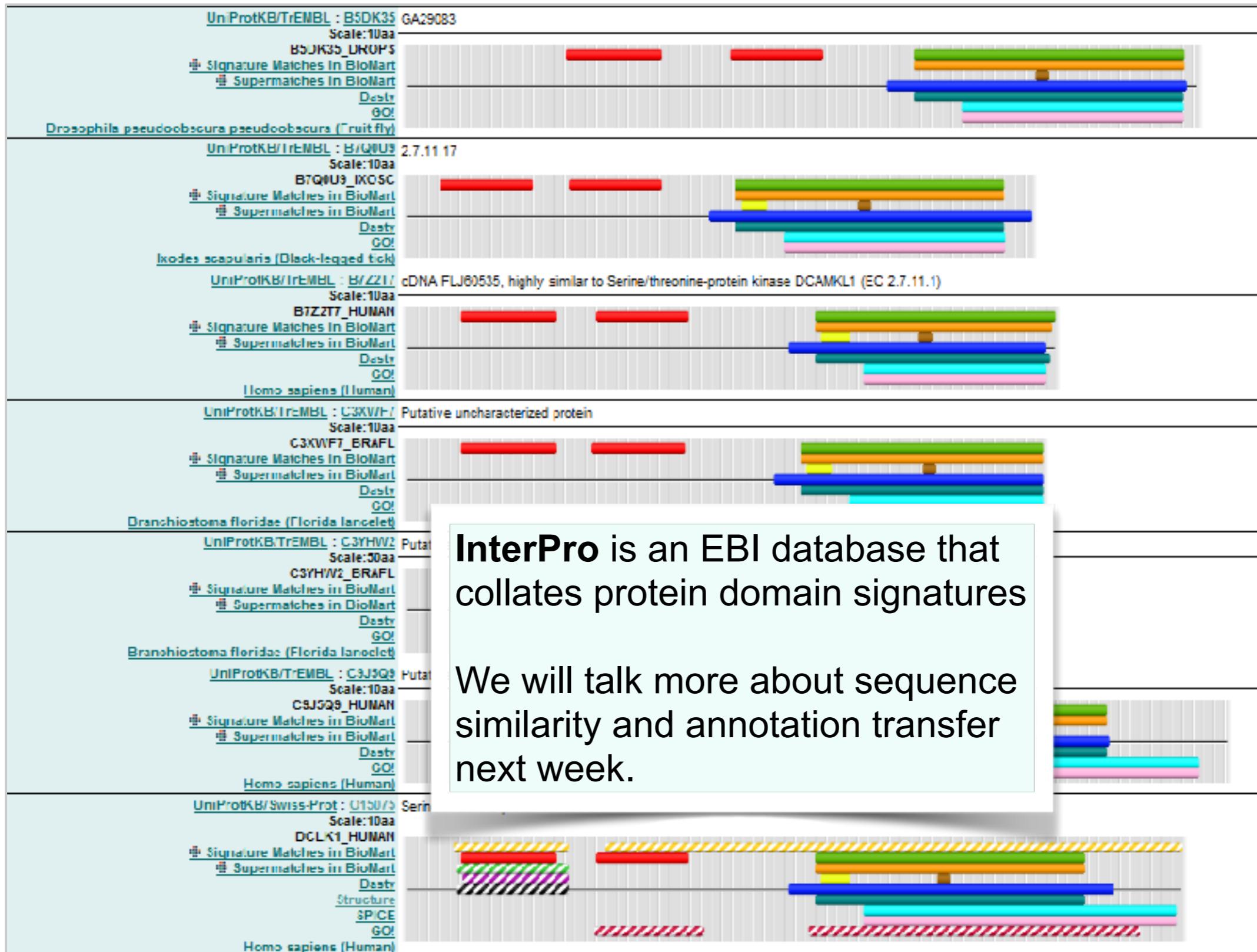
# UniProt/Swiss-Prot vs UniProt/TrEMBL

- 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



# 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/>

NCBI Resources How To Sign in to NCBI

All Databases ras Search

NCBI Home Resource List (A-Z) All Resources Chemicals & Bioassays Data & Software DNA & RNA Domains & Structures Genes & Expression Genetics & Medicine Genomes & Maps Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

Welcome to NCBI

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

About the NCBI | Mission | Organization | NCBI News

Get Started

- Data: Find data using NCBI software
- Tools: Get NCBI data or software
- How Tos: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Genotypes and Phenotypes

Data from Genome Wide Association studies that link genes and diseases. See study variables, protocols, and analysis.

Resources

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

NCBI Announcements

RefSeq release 69 available on

The full RefSeq release 69 is now available on the FTP site with 74 records describing 52,378,420 ...

Hands on demo (or see following slides)

ms - GQuery: Global Cross X

www.ncbi.nlm.nih.gov/gquery/?term=ras

NCBI Resources How To Sign in to NCBI

## Search NCBI databases

Help

ras

About 2,978,774 search results for "ras"

Literature		Genes			
Books	1,677	books and reports	EST	3,985	expressed sequence tag sequences
MeSH	402	ontology used for PubMed indexing	Gene	87,165	collected information about gene loci
NLM Catalog	223	books, journals and more in the NLM Collections	GEO DataSets	3,732	functional genomics studies
PubMed	54,672	scientific & medical abstracts/citations	GEO Profiles	1,622,789	gene expression and molecular abundance profiles
PubMed Central	96,114	full-text journal articles	HomoloGene	696	homologous gene sets for selected organisms
Health		PopSet	2,254	sequence sets from phylogenetic and population studies	
ClinVar	759	human variations of clinical significance	UniGene	4,770	clusters of expressed transcripts
dbGaP	120	genotype/phenotype interaction studies	Proteins		
GTR	1,879	genetic testing registry			

75

ms - 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 Hide sidebar >>

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

Filters: Manage Filters

Did you mean ras as a gene symbol?  
Search Gene for ras as a symbol.

<< First < Prev Page 1 of 4282 Next > Last >>

**Results: 1 to 20 of 85633**

i Filters activated: Current only. [Clear all](#) to show 87165 items.

Name/Gene ID	Description	Location	Aliases
<input type="checkbox"/> <a href="#">ras</a> ID: 19412	resistance to audiogenic seizures [Mus musculus (house mouse)]		asr
<input type="checkbox"/> <a href="#">ras</a> ID: 43873	rasberry [Drosophila melanogaster (fruit fly)]	Chromosome X, NC_004354.4 (10744502..10749097)	Dmel_CG1799, CG11485, CG1799, DmelCG1799, EP(X)1093,

▼ Top Organisms [Tree]

- Homo sapiens (1126)**
- Mus musculus (823)
- Rattus norvegicus (625)
- Oreochromis niloticus (533)
- Neolamprologus brichardi (507)
- All other taxa (82019)

More...

Find related data

Database: Select

Find items

Search details

ras [All Fields] AND alive [property]

(ras) AND "Homo sapiens"[porgn:txid9606]

**Gene**

Gene (ras) AND "Homo sapiens"[porgn:txid9606] Search Help

Show additional filters Hide sidebar >>

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

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

① Filters activated: Current only. Clear all to show 1499 items.

Name/Gene ID	Description	Location	Aliases
<input type="checkbox"/> <a href="#">NRAS</a> 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
<input type="checkbox"/> <a href="#">KRAS</a> 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, K-RAS1, KRAS2, NS, NS2, RASK2

**Filters:** Manage Filters

**Find related data**

**Database:** Select

Find Items

**Search details**

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

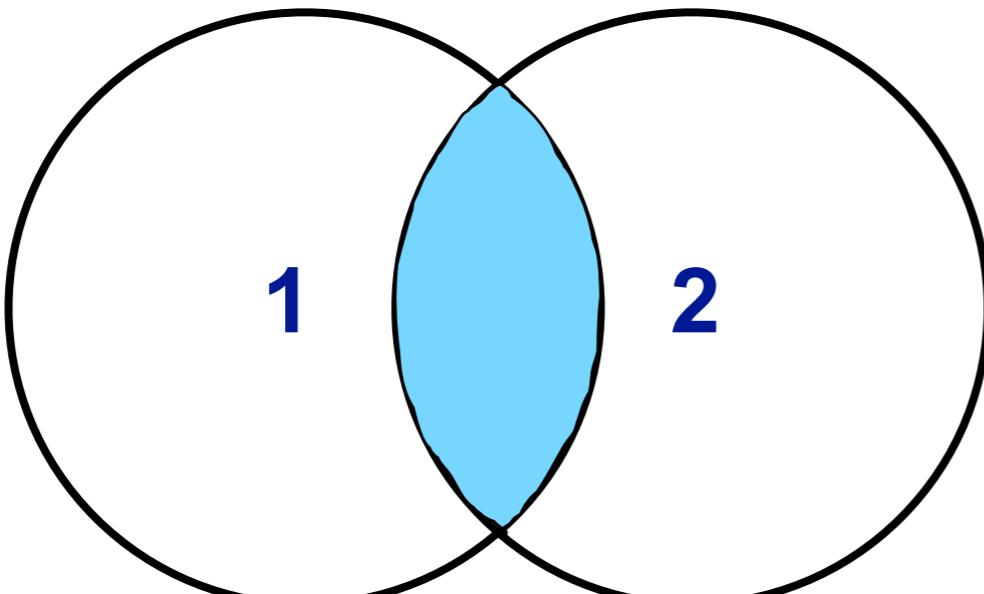
Search See more...

**Recent activity**

Turn Off Clear

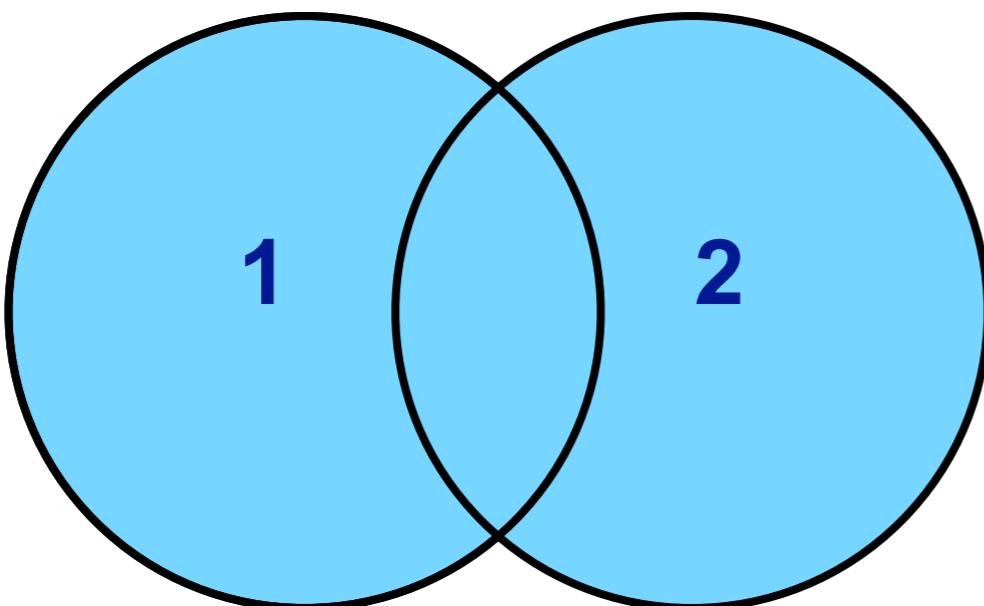
Chromosome locations Select

**1 AND 2**



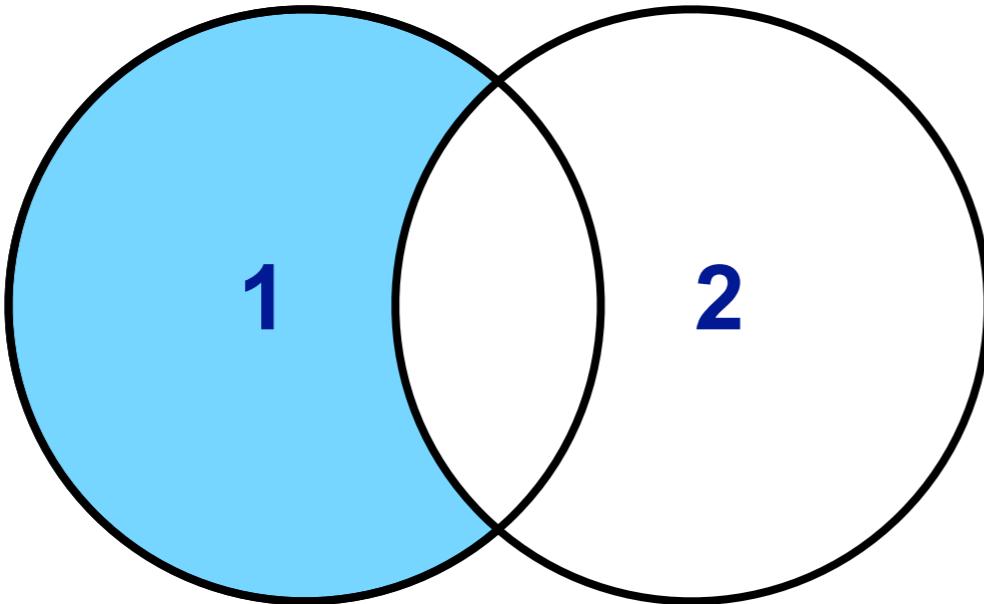
**ras AND disease  
(1185 results)**

**1 OR 2**



**ras OR disease  
(134,872 results)**

**1 NOT 2**



**ras NOT disease  
(84,448 results)**

(ras) AND "Homo sapiens"[porgn:txid9606]

Gene Gene (ras) AND "Homo sapiens"[porgn:txid9606] Search Save search Advanced Help

Show additional filters Hide sidebar >>

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

① Filters activated: Current only. Clear all to show 1499 items.

	Name/Gene ID	Description	Location	Aliases
Categories	<input type="checkbox"/> <a href="#">NRAS</a> 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
Sequence content	<input type="checkbox"/> <a href="#">KRAS</a> 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, K-RAS1, KRAS2, NS, NS2, RASK2

Gene sources Genomic

Clear all

Filters: Manage Filters

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

Chromosome locations Select

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**

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)

30

KRAS Kirsten rat sarcoma viral oncogene homolog

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

NCBI Resources How To Sign in to NCBI

Gene Search Help

Display sidebar >> 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-RAS2B

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)

Related

81

KRAS KRas gene summary

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)

Chromosome 12 - NC\_000012.12

Genomic regions, transcripts, and products

Genomic Sequence: NC\_000012.12 chromosome 12 reference GRCh38 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank

Go to reference sequence details

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 viral oncogene homolog

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

NCBI Resources How To Sign in to NCBI

Gene Search Help

Display Settings Hide sidebar >>

**KRAS** **Ki**  
**(human)** ]

Gene ID: 3845

**Summary**

**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-

**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)  
Related

83

KRAS KRas protein

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 signalling 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

The screenshot shows the UniProt-GOA database homepage. At the top, there's a navigation bar with tabs for 'Services', 'Research', 'Training', and 'About us'. Below the navigation bar is a search bar with examples like 'GO:0006915, tropomyosin, P08727' and a 'Search' button. The main content area features a large title 'UniProt-GOA' and a sub-section titled 'Gene Ontology Annotation (UniProt-GOA) Database'. A detailed description of the program follows, mentioning UniProt GO annotation, UniProt Knowledgebase, UniProt biocuration, manual and electronic annotations, and external collaborating groups. At the bottom, it states 'UniProt is a member of the GO Consortium'. On the right side, there's a 'Menu' sidebar with links to 'Downloads', 'Searching UniProt-GOA', 'Annotation Methods', 'Annotation Tutorial', 'Manual Annotation Efforts' (which is currently selected), 'Reference Genome Annotation Initiative', 'Cardiovascular Gene Ontology Annotation Initiative', 'Renal Gene Ontology Annotation Initiative', and 'Exosome Gene'.

KRAS Kiraten rat sarcoma x UniProt-GOA < EMBL-EBI x

www.ebi.ac.uk/GOA

EMBL-EBI

Services Research Training About us

# UniProt-GOA

Search

Examples: GO:0006915, tropomyosin, P08727

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

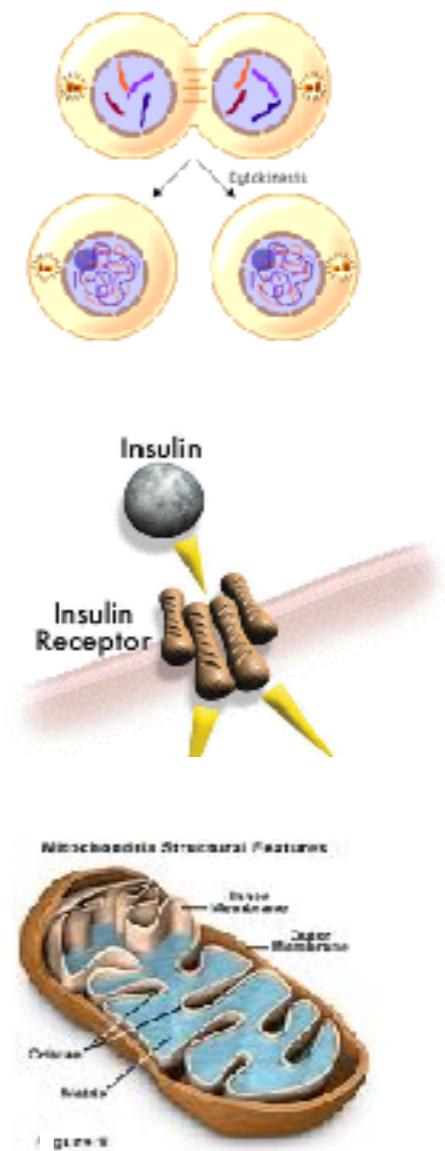
- Downloads
- Searching UniProt-GOA
- Annotation Methods
- Annotation Tutorial
- Manual Annotation Efforts
  - Reference Genome Annotation Initiative
  - Cardiovascular Gene Ontology Annotation Initiative
  - Renal Gene Ontology Annotation Initiative
  - Exosome Gene

# 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 KRasG12D rat isoform 1

Gene Ontology Provided by GOA

Function Evidence Code Pubs

GDP binding TAS

GMP binding IEA

GTP binding TAS

LRR domain binding TAS

protein binding IEA

protein complex binding TAS

Process Code

Fc-epsilon receptor signalling 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

genomic X01669.1 CAA25828.1  
Items 1 - 25 of 43 < Prev Page 1 of 2 Next >

Protein Accession	Links
P01116.1	<a href="#">GenPept Link</a> <a href="#">UniProtKB Link</a> <a href="#">GenPept</a> <a href="#">UniProtKB/Swiss-Prot:P01116</a>

**Additional links**

You are here: NCBI > Genes & Expression > Gene Write to the Help Desk

**GETTING STARTED** **RESOURCES** **POPULAR** **FEATURED** **NCBI INFORMATION**  
[NCBI Education](#) [Chemicals & Bioassays](#) [PubMed](#) [Genetic Testing Registry](#) [About NCBI](#)  
[NCBI Help Manual](#) [Data & Software](#) [Bookshelf](#) [PubMed Health](#) [Research at NCBI](#)  
[NCBI Handbook](#) [DNA & RNA](#) [PubMed Central](#) [GenBank](#) [NCBI News](#)  
[Training & Tutorials](#) [Domains & Structures](#) [PubMed Health](#) [Reference Sequences](#) [NCBI FTP Site](#)  
[Genes & Expression](#) [BLAST](#) [Gene Expression Omnibus](#) [NCBI on Facebook](#)  
[Genetics & Medicine](#) [Nucleotide](#) [Map Viewer](#) [NCBI on Twitter](#)  
[Genomes & Maps](#) [Genome](#) [Human Genome](#) [NCBI on YouTube](#)  
[Homology](#) [SNP](#) [Mouse Genome](#)  
[Literature](#) [Gene](#) [Influenza Virus](#)  
[Proteins](#) [Protein](#) [Primer-BLAST](#)  
[Sequence Analysis](#) [PubChem](#) [Sequence Read Archive](#)

Scroll down to  
UniProt link

UniProt will detail much more information for protein coding genes

P01116 - RASK\_HUMAN

Protein: GTPase KRas  
Gene: KRAS  
Organism: Homo sapiens (Human)  
Status: Reviewed - Experimental evidence at protein level<sup>i</sup>

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 <sup>i</sup>	10 – 18	9	GTP 2 Publications			
Nucleotide binding <sup>i</sup>	29 – 35	7	GTP 2 Publications			
Nucleotide binding <sup>i</sup>	59 – 60	2	GTP 2 Publications			

KRAS - GTPase KRas protein

www.uniprot.org/uniprot/P01116

UniProtKB Advanced

BLAST Align Retrieve/ID Mapping Help Contact Basket

P01116 - RASK\_HUMAN

Protein: GTPase KRas  
Gene: KRAS  
Organism: Homo sapiens (Human)  
Status: Reviewed - ●●●●●

Display: None

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

**Example Questions:**  
What positions in the protein are responsible for GTP binding?

**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 <sup>i</sup>	10 – 18	9	GTP 2 Publications			
Nucleotide binding <sup>i</sup>	29 – 35	7	GTP 2 Publications			
Nucleotide binding <sup>i</sup>	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 protein

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

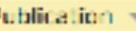
SIMILAR PROTEINS

Top

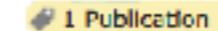
**Pathology & Biotech**

Involvement in disease<sup>1</sup>

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. 

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 <sup>1</sup>	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. 		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?

The screenshot shows the UniProt protein details page for KRAS (P01116). The 'STRUCTURE' tab is selected in the sidebar, highlighted with a red box. The main content includes a secondary structure diagram and a table of 3D structure databases.

**Secondary structure**  
1 [Helix] [Turn] [Beta strand] 189

Legend: Helix Turn Beta strand

Show more details

**3D structure databases**

Select the link destinations:	Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
<input checked="" type="radio"/> PDBi	1D8D	X-ray	2.00	P	178-188	[*]
<input checked="" type="radio"/> RCSB PDBi	1D8E	X-ray	3.00	P	178-188	[*]
<input type="radio"/> PDBj	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	1-154	[*]
	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... ?

The screenshot shows the UniProt protein details page for KRAS (P01116). The 'Display' section has 'None' selected. Under 'Family and domain databases', the Pfam entry is highlighted with a red box. The text 'PFAM is one of the best protein family databases' is overlaid in a red box.

**PFAM** is one of the best protein family databases

**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

OrthoDB  
PhylomeDB<sup>i</sup> P01116

TreeFam<sup>i</sup> TF3

Family and domain databases

Gene3D <sup>i</sup>	3.40.50.300. 1 hit.
InterPro <sup>i</sup>	IPR027417. P-loop_NTPase. IPR005225. Small_GTP-bd_dom. IPR001806. Small_GTPase. IPR020849. Small_GTPase_Ras. [Graphical view]
PANTHER <sup>i</sup>	PTHR24070. PTHR24070. 1 hit.
Pfam <sup>i</sup>	PF00071. Ras. 1 hit. [Graphical view]
PRINTS <sup>i</sup>	PR00449. RASTRNSFRMNG.
SMART <sup>i</sup>	SM00173. RAS. 1 hit. [Graphical view]
SUPERFAMILY <sup>i</sup>	SSF52540. SSF52540. 1 hit.
TIGRFAMs <sup>i</sup>	TIGR00231. small_GTP. 1 hit.
PROSITE <sup>i</sup>	PS51421. RAS. 1 hit. [Graphical view]

Sequences (2)<sup>i</sup>

Sequence status<sup>i</sup>: Complete.

Sequence processing<sup>i</sup>: The displayed sequence is further processed into a mature form.

This entry describes 2 isoforms<sup>i</sup> produced by alternative splicing.

# Example Questions:

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

KRAS - GTPase KRas protein | Pfam: Family: Ras (PF00071)

EMBL-EBI  pfam.xfam.org/family/PF00071

HOME | SEARCH

## Family: Ras (PF00071)

**Summary**

**Domain organisation**

**Clan**

**Alignments**

**HMM logo**

**Trees**

**Curation & model**

**Species** (selected)

**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.

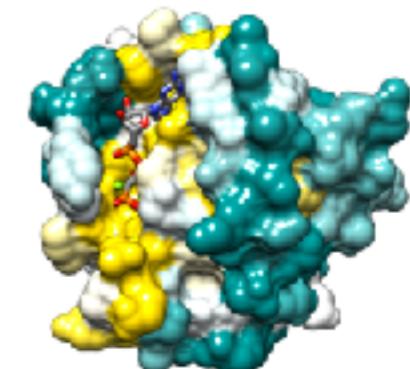
Because these signals result in cell growth and division, overactive Ras signaling can ultimately lead to cancer.<sup>[1]</sup> The 3 Ras genes in humans (HRAS, KRAS, and NRAS) are the most common oncogenes in human cancer; mutations that permanently activate Ras are found in 20% to 25% of all human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

[Contents](#) [edit]

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

**Identifiers**

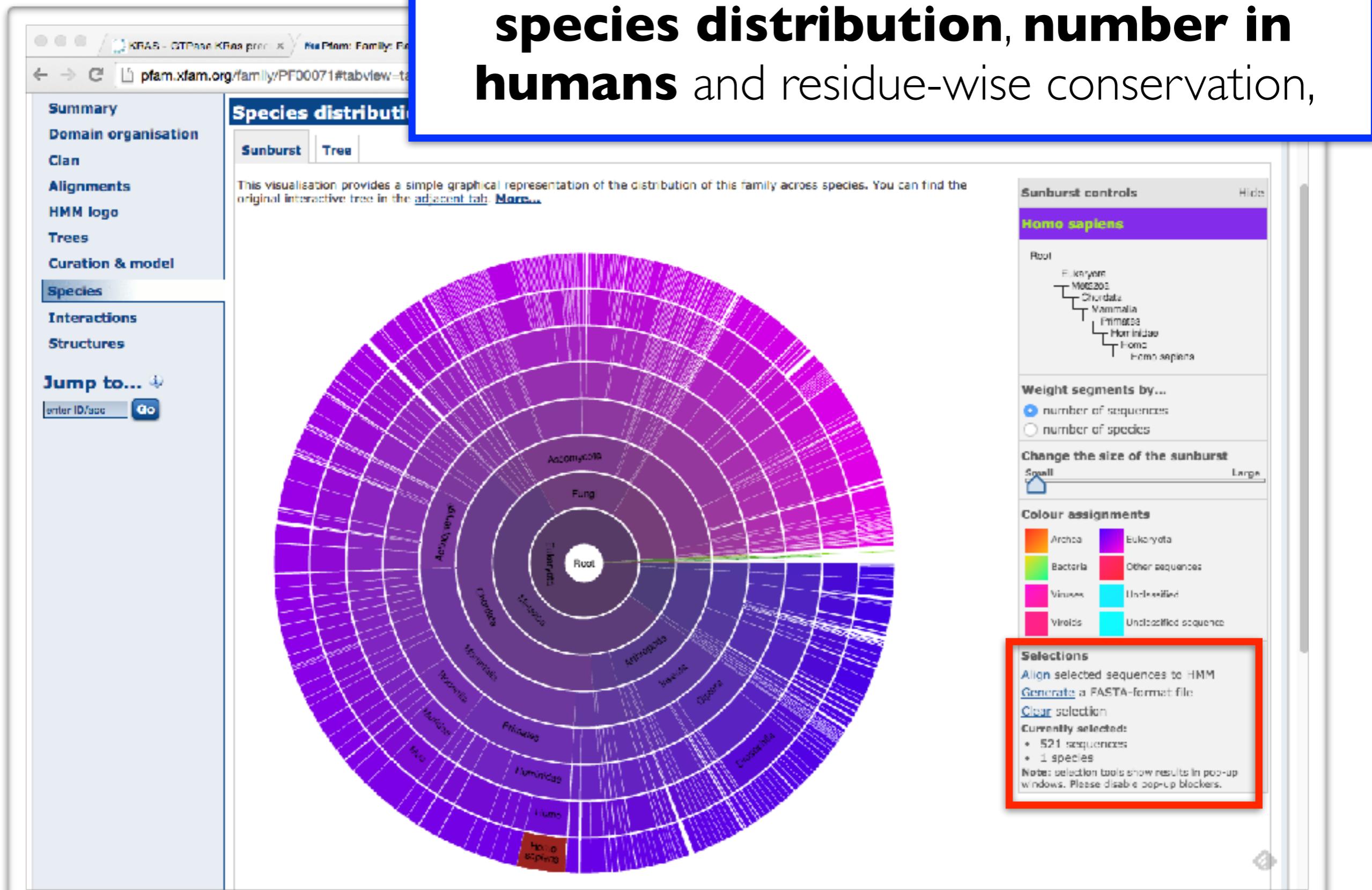
Symbol	Ras
Pfam	PF00071_5
InterPro	IPR013753
PROSITE	PS000017
SCOP	Sp21
SUPERFAMILY	Sp21



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

# Example Questions:

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



# Example Questions:

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

KRAS - GTPase KRas protein  
No Pfam: Family: Pro

← → C pfam.xfam.org/family/PF00071#tabview=tab1

Summary Species distribution

Domain organisation

Clan

Alignment

HMM log

Trees

Curation

Species

Interaction

Structure

Jump to another ID/acc

EMBL-EBI

Pfam: Pfam alignment viewer

pfam.xfam.org/family/PF00071/alignment/view?jobId=EDCA403E-9836-11E4-B360-10D3298E2F76

Alignment for selected sequences

Currently showing rows 1 to 30 of 536 rows in this alignment. Show 30 rows of alignment.

P11234/16-178	...KIVVVGCCVGRGADIL...	Q...	FN...	T...	D...	E...	F...	V...	E...	DYDFPK...	-AD...	SYRENWLD...
P01112/5-163	...KLIVVGGCCVGRGADIL...	Q...	LI...	Q...	N...	H...	V...	D...	DYDFPK...	-ED...	SYREQWVID...	
Q14088/38-204	...KIVVVGCCVGRGADIL...	R...	FC...	G...	E...	D...	F...	V...	DYDFPK...	GID...	TYERKVWIDE...	
Q9RN83/7-173	...KIVVVGCCVGRGADIL...	T...	FR...	S...	D...	N...	K...	AVTIT...	GID...	TAUVTVPED...		
P15153/7-178	...KIVVVGCCVGRGADIL...	S...	YT...	E...	N...	A...	P...	TYIFPV...	-ED...	YISAKWVHD...		
Q00194/11-183	...KLLALGCCVGRGADIL...	R...	YT...	D...	N...	E...	P...	XPFICW...	GID...	TYERKVWVSDPN...		
Q11907/13-174	...KIVVVGCCVGRGADIL...	R...	FT...	R...	N...	E...	N...	L...	DSRITI...	GVE...	FATREIQWT...	
P10114/5-165	...KIVVVGCCVGRGADIL...	Q...	FV...	G...	E...	D...	F...	I...	KYMPIT...	-ED...	TYRKEFTRWD...	
P51153/10-171	...KLLLGDDCCVGRGADIL...	R...	FA...	E...	D...	N...	N...	TYISHI...	GID...	TKIRTWDIE...		
P53040/77-241	...RVLVIGCCVGRGADIL...	I...	FA...	Gvhd...	SM...	D...	S...	D-CEVL...	GID...	TYERTLMVD...		
P5042/93-203	...RVLVIGCCVGRGADIL...	I...	FO...	G...	V...	E...	G...	EEAAB...	--H...	TYDREIWWI...		
P01116/5-163	...KIVVVGCCVGRGADIL...	Q...	LI...	Q...	N...	H...	V...	D...	DYDFPK...	-ED...	SYREKWVID...	
Q9J0W7/21-182	...KIVVVGCCVGRGADIL...	R...	YI...	R...	N...	D...	N...	TYIFPV...	GID...	TYTKUVVDE...		
Q9ULC3/11-171	...KIVVVGCCVGRGADIL...	R...	YC...	R...	E...	I...	T...	K...	DYDFPK...	GID...	TYERGVIN...	
Q11807/15-177	...KIVVVGCCVGRGADIL...	C...	FT...	G...	K...	I...	V...	P...	DYDFPK...	-ED...	SYLKHED...	
Q9NKD7/7-202	...KIVVVGCCVGRGADIL...	R...	YN...	E...	R...	R...	F...	D...	T-V3W...	GID...	TYLEQW...	
Q9A062/35-201	...KIVVVGCCVGRGADIL...	R...	YI...	R...	R...	R...	F...	D...	DRTEATI...	GID...	TYERKVWID...	
Q95905/9-174	...KIVVVGCCVGRGADIL...	R...	YI...	H...	D...	H...	V...	D...	TYQATI...	GAA...	TYAKUMVDE...	
P51149/10-175	...KIVVVGCCVGRGADIL...	C...	YN...	H...	K...	S...	N...	OYKMTI...	GID...	TYTKEWVHD...		
Q9ULN5/65-227	...KIVVVGCCVGRGADIL...	R...	FK...	D...	G...	A...	D...	ATFISW...	GID...	TYNEVLDV...		
P57710/14-175	...KIVVVGCCVGRGADIL...	R...	FT...	R...	N...	E...	S...	H...	DSRITI...	GVE...	TYSTRVWHL...	
P51153/11-183	...KIVVVGCCVGRGADIL...	Q...	YT...	D...	G...	K...	N...	APKTVW...	GID...	TYERKVWYRas...		
P01111/5-165	...KIVVVGCCVGRGADIL...	Q...	LI...	Q...	N...	H...	V...	D...	DYDFPK...	-ED...	SYREKWVID...	
P11233/16-177	...KIVVVGCCVGRGADIL...	C...	FN...	T...	D...	E...	F...	V...	DYDFPK...	-AD...	SYRENWLD...	
Q9UL25/21-182	...KIVVVGCCVGRGADIL...	R...	YC...	E...	N...	K...	N...	D...	XHITL...	GID...	TYTERKLNIC...	
Q9NP72/10-171	...KIVVVGCCVGRGADIL...	R...	FT...	D...	D...	T...	D...	P...	BLANTI...	GID...	TYVKTISV...	
Q9J0L4/10-171	...KIVVVGCCVGRGADIL...	R...	FA...	D...	D...	T...	V...	E...	DYDFPK...	GID...	TYRIMELD...	
Q9J0W6/7-165	...KIVVVGCCVGRGADIL...	R...	FT...	R...	D...	S...	D...	P...	NINMII...	GID...	TYPTWYOQ...	
Q9UBX7/23-179	...KIVVVGCCVGRGADIL...	R...	FL...	H...	D...	Q...	O...	P...	QQLSTY...	ACT...	TYXHTATV...	
P51157/14-179	...KIVVVGCCVGRGADIL...	C...	FA...	G...	E...	T...	G...	K...	TYKGTI...	GID...	TYLRRITL...	

There are 18 pages in this alignment. Show page 1

Download this alignment.

Close window

can find the

Sunburst controls Hide

**Homo sapiens**

Root

Eukaryota

M002203

Chordata

Mammalia

Primates

Hominidae

Homo

Homo sapiens

Weight segments by...

number of sequences

number of species

Change the size of the sunburst

Small Large

Colour assignments

Archaea

Bacteria

Miniviruses

Viruses

Eukaryota

Other sequences

Unclassified

Unclassified sequence

Selections

Align selected sequences to HMM

Generate a FASTA-format file

Clear selection

Currently selected:

- + 521 sequences
- + 1 species

Note: selection tools show results in pop-up windows. Please disable pop-up blockers.

# Example Questions:

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

KRAS - GTPase KRas protein | No Pfam Family: Ras

← → C pfam.xfam.org/family/PF00071#tabview=tab4

EMBL-EBI 

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

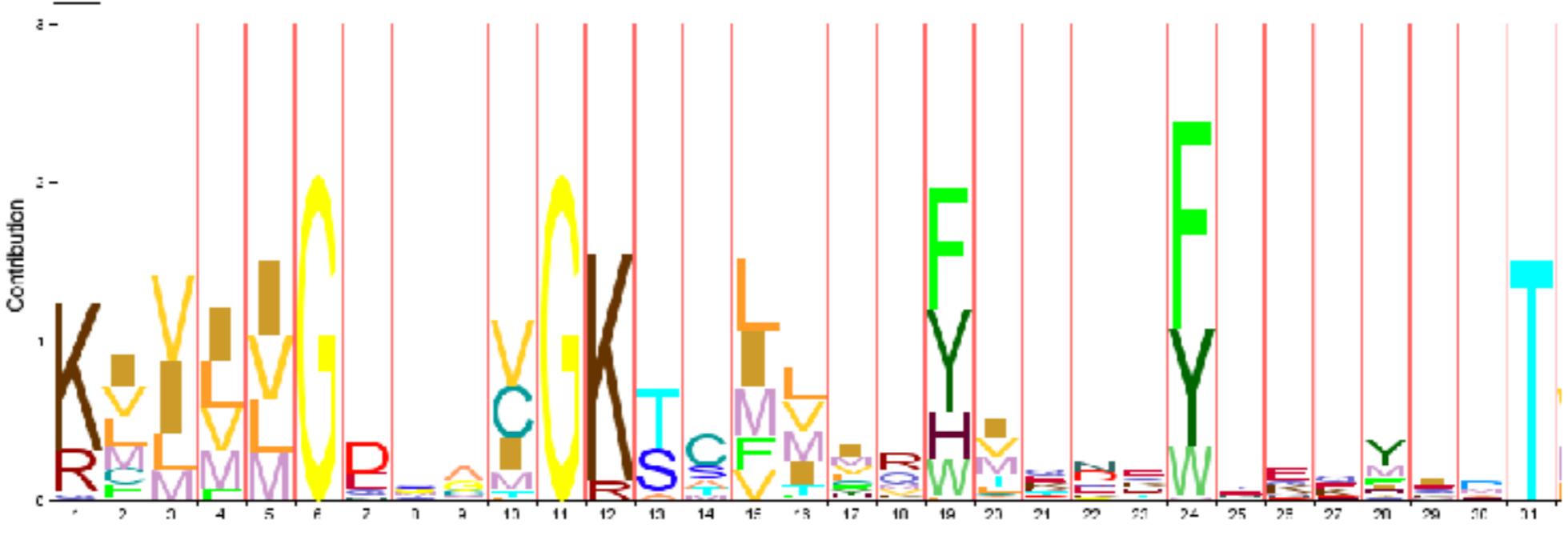
## Family: Ras (PF00071)

Summary Domain organisation Clan Alignments **HMM logo** (highlighted with a red box) 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](mailto: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... !\[\]\(6a328a8fd0f2bf5d258b15376751e449\_img.jpg\)](#) [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

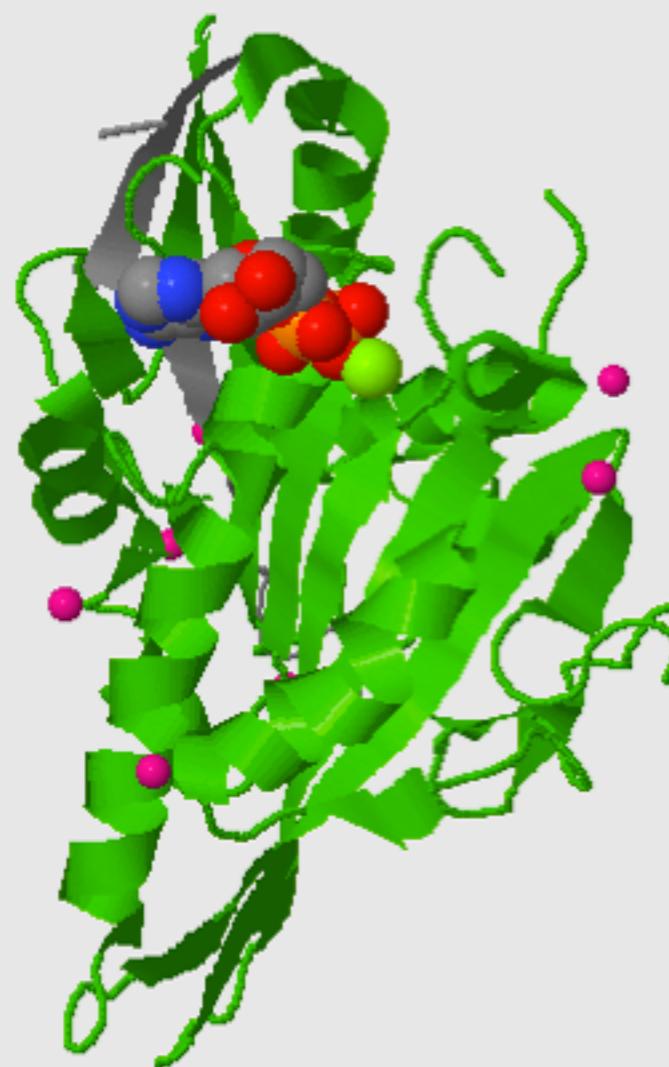
## Structures

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
<a href="#">A8BKD1_GIALA</a>	11 - 335	<a href="#">2vvg</a>	A	11 - 335	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	11 - 335	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
<a href="#">CENPE_HUMAN</a>	12 - 329	<a href="#">1t5c</a>	A	12 - 329	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	12 - 329	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
<a href="#">KAR3_YEAST</a>	392 - 723	<a href="#">1f9t</a>	A	392 - 723	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
		<a href="#">1f9u</a>	A	392 - 723	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
		<a href="#">1f9v</a>	A	392 - 723	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
		<a href="#">1f9w</a>	A	392 - 723	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
		<a href="#">3kar</a>	A	392 - 723	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
<a href="#">KI13B_HUMAN</a>	11 - 352	<a href="#">3qbj</a>	A	11 - 352	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	11 - 352	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			C	11 - 352	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
		<a href="#">1ii6</a>	A	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
		<a href="#">1q0b</a>	A	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
		<a href="#">1x88</a>	A	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
		1	A	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>



PDB entry 3bfm



Jmol

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

 [Close window](#)

# 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 online:  
[PDF1 \(bioinformatics review\)](#),  
[PDF 2 \(bioinformatics challenges\)](#).
- Complete the **lecture 1.1 homework questions**:  
<http://tinyurl.com/bioinf525-quiz1>

THANK YOU

# ADDITIONAL DATABASES OF NOTE (SLIDES FOR YOUR REFERENCE)

# **ENTREZ & BLAST:**

## TOOLS FOR SEARCHING AND ACCESSING MOLECULAR DATA AT NCBI

# Entrez: Integrated search of NCBI databases

The screenshot shows the NCBI homepage with the Entrez sidebar open. The sidebar is titled "All Databases" and lists various NCBI databases. A callout box highlights the text: "Entrez is available from the main NCBI homepage or from the homepage of individual databases".

**NCBI Home**

**Resource List (A-Z)**

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

**All Databases**

- PubMed
- Protein
- Nucleotide
- CGS
- EST
- Structure
- Genome
- BioProject
- BioSample
- BioSystems
- Books
- Conserved Domains
- Clone
- dbGaP
- dbVar
- Eigenomics
- Gene
- GEO DataSets
- GEO Profiles
- HomoloGene
- MeSH
- NCBI Web Site
- NLM Catalog
- CMIA
- CMIM
- PMC
- PopSet
- Probe
- Protein Clusters
- PubChem BioAssay
- PubChem Compound
- PubChem Substance
- PubMed Health
- SNP
- SRA
- Taxonomy
- ToolKit
- ToolKitAll
- UniGene
- UniSTS

Welcome to NCBI

NCBI advances science by providing access to biomedical and genomic information.

[NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Entrez

Analyze data using NCBI software

Downloads: Get NCBI data or software

Tutorials: Learn how to accomplish specific tasks at NCBI

Missions: Submit data to GenBank or other NCBI databases

**Popular Resources**

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

**NCBI Announcements**

NCBI's April Newsletter is on the Bookshelf

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

New Filter Sidebar will be added to PubMed

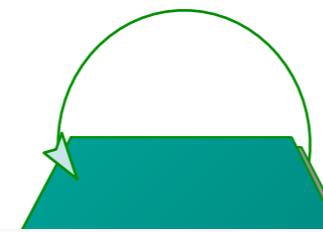
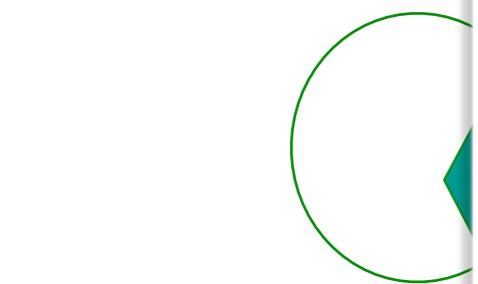
A Filter Sidebar will be added soon to the PubMed result pages. This useful

DELTA BLAST - more sensitive protein searching

Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST)

[More...](#)

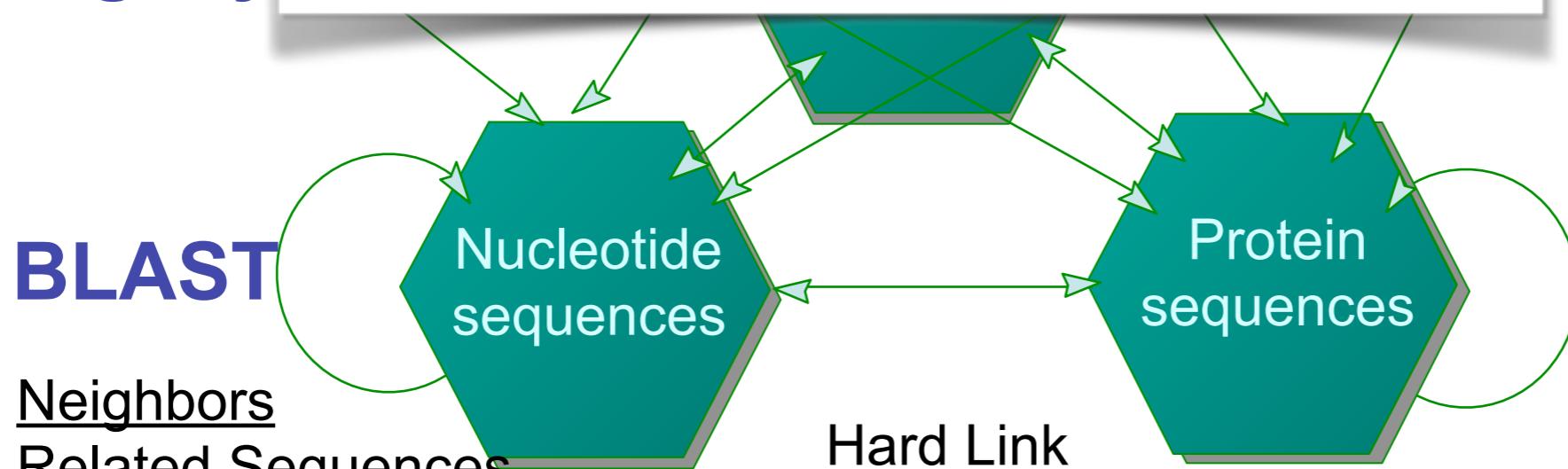
# Entrez: navigating across databases



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



# Global Entrez Query: All NCBI Databases

ma - Global Entrez Query: All NCBI Databases

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 1,000,000 full-text journal articles

MeSH 402 ontology used for PubMed indexing

NLM Catalog 223 books and reports

PubMed 54,672 full-text journal articles

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

The Entrez system: 38 (and counting) integrated databases

EST 3,985 expressed sequence tag sequences

GEO Profiles 1,022,789 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

<http://www.ncbi.nlm.nih.gov/gquery/>

# Search Results

Nucleotide      Nucleotide  zebrafish creatine kinase

Save search   Limits   Advanced

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

Send to:  Filter your results:

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

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

[Danio rerio creatine kinase, muscle b \(ckmb\), mRNA](#)  
1. 1,463 bp linear mRNA  
Accession: NM\_001105683.1 GI: 157787180  
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Danio rerio zgc:63663 \(zgc:63663\), mRNA](#)  
2. 2,478 bp linear mRNA  
Accession: NM\_200614.1 GI: 41055386  
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Danio rerio creatine kinase, muscle](#)  
3. 1,552 bp linear mRNA  
Accession: NM\_130932.1 GI: 18858426  
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

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

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

**Discovery Column (sort, filter, link)**

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:

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

See more...

**Recent activity**

# Advanced: Search Builder

## Nucleotide Advanced Search Builder

zebrafish[Organism] AND "creatin kinase"[Title]

[Clear](#)

Helps build complex fielded queries

Organism

zebrafish

[Show index list](#)

AND

Title

"creatin kinase"[Title]

[Hide index list](#)

creatin kinase (749)

creatin kinase 1 (6)  
creatin kinase 2 (2)  
creatin kinase b (30)  
creatin kinase b gene (3)  
creatin kinase b mrna (3)  
creatin kinase b pseudogene 1 (1)  
creatin kinase b subunit (1)  
creatin kinase brain (43)  
creatin kinase chain b (1)

[Previous 200](#)

[Next 200](#)

[Refresh index](#)

AND

All Fields

[Show Index list](#)

[Search](#)

or [Add to history](#)

Items from search history can be included / combined / modified

## History

[Clear history](#)

Search	Add to builder	Query	Items found	Time
#7	<a href="#">Add</a>	Search zebrafish[organism] AND actin[title]	71	12:41:16
#4	<a href="#">Add</a>	Search zebrafish actin	1288	12:40:07
#1	<a href="#">Add</a>	Search zebrafish creatine kinase	34	12:39:02

# Complex Query Results

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

Send to:  Filter your results:

All (6)

Bacteria (0)

INSDC (GenBank) (0)

mRNA (6)

RefSeq (6)

[Manage Filters](#)

Results: 6

[Danio rerio creatine kinase, brain a \(ckba\), mRNA](#)  
1. 1,481 bp linear mRNA  
Accession: NM\_001077163.1 GI: 116004536  
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Danio rerio creatine kinase, mitochondrial 1 \(ckmt1\), nuclear gene encoding mitochondrial protein, mRNA](#)  
2. mRNA

[Danio rerio creatine kinase, muscle a \(ckma\), mRNA](#)  
3. 1,552 bp linear mRNA  
Accession: NM\_130932.1 GI: 18856426  
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Danio rerio creatine kinase, mitochondrial 2 \(sarcomeric\) \(ckmt2\), nuclear gene encoding mitochondrial protein, mRNA](#)  
4. 1,401 bp linear mRNA  
Accession: NM\_200697.1 GI: 41162341  
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Danio rerio creatine kinase, muscle b \(ckmb\), mRNA](#)  
5. 1,463 bp linear mRNA  
Accession: NM\_001105683.1 GI: 157787180  
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[Danio rerio creatine kinase, brain b \(ckbb\), mRNA](#)  
6. 1,459 bp linear mRNA  
Accession: NM\_173222.1 GI: 27546192  
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

("Danio rerio"[Organism] AND "creatin kinase"[Title])  
AND "refseq"[Filter] AND mRNA[Filter]

Analyze these sequences

Run BLAST

Find related data

Database:

[Find items](#)

Search details

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

See more...

Recent activity

# BLAST is a very important tool available from the NCBI Homepage

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

The screenshot shows the NCBI homepage with a red border around the central content area. At the top, there's a navigation bar with links for NCBI, Resources, How To, and My NCBI. Below that is the NCBI logo and search bar. The main content area features a "Welcome to NCBI" section with a brief description and links to About, Mission, Organization, Research, and RSS Feeds. A "Get Started" section lists links for Tools, Downloads, How To's, and Submissions. To the right is a "Popular Resources" sidebar with a red arrow pointing to the BLAST link. Below that is a "NCBI News" sidebar with news items about NAR's 2011 Database Issue and new articles in Epigenomics.

NCBI Home

Site Map (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

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](#)

PubMed Central

Free Full Text. Over 1,500,000 articles from over 450 journals. Linked to PubMed and fully searchable.

Popular Resources

- [BLAST](#)
- [Bookshelf](#)
- [Cancer](#)
- [Cenome](#)
- [Nucleotide](#)
- [OMIM](#)
- [Protein](#)
- [PubChem](#)
- [PubMed](#)
- [PubMed Central](#)
- [SNP](#)

NCBI News

NAR's 2011 Database Issue is out with 9 NCBI-Authored Papers 05 Jan 2011

New articles are available describing the new Epigenomics

New NCBI News Issue 29 Nov 2010

# BLAST – Basic Local Alignment Search Tool

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

The screenshot shows the NCBI BLAST homepage. At the top, there's a navigation bar with links for Home, Recent Results, Saved Strategies, and Help. On the right, there's a "My NCBI" section with links for Sign In and Register. Below the navigation, there's a banner for "NCBI BLAST Home" with a link to "BLAST finds regions of similarity between biological sequences." A red box highlights a "New" announcement about aligning multiple protein sequences using the COBALT Multiple Alignment Tool. The main content area is divided into sections: "BLAST Assembled RefSeq Genomes" (listing species like Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera), "Basic BLAST" (listing nucleotide blast, protein blast, tblastx, and blastx), and "Specialized BLAST". To the right, there's a sidebar with "News" (mentioning a new WGS BLAST page) and "Tip of the Day" (mentioning how to do batch BLAST jobs). A large callout box in the bottom right corner states: "BLAST performs sequence similarity searches of query sequences vs sequence databases. We will cover this in detail in the next lecture."

# 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

The screenshot shows the PubMed search results for the query "corticotropin\* AND receptor AND human[orgn]". The results page is framed by a red border. A red arrow points from the text above to the "Limits" link in the top navigation bar.

**Display Settings:** Summary, 20 per page, Sorted by Recently Added

**Send to:** [Email](#) [Print](#) [Clipboard](#)

**Filter your results:**

- All (2383)
- Review (529)
- Free Full Text (681)

**Manage Filters**

**Results: 1 to 20 of 2363**

**Titles with your search terms**

- Gene expression analysis in the human hypothalamus in depression [Mol Psychiatry 2008]
- Differential regulation of human dopamine D<sub>2</sub> and somatostatin receptor [J Mol Endocrinol. 2009]
- Estrogen receptor-alpha and -beta regulate the human corticotropin-releasing [Brain Res. 2008]

[See more...](#)

**215 free full-text articles in PubMed Central**

- Synthesis and biological evaluation of New CRH Analogs. [Bioinorg Chem Appl. 2010]
- Review: Biological contribution to social influence. [Int J Environ Res Public Health. 2010]
- Pharmacogenomic approaches to asthma treatment. [Allergy Asthma Immunol Res. 2010]

[See all \(215\)](#)

**Find related data**

Database: [Select](#)

**Results:**

- Distribution of retinoic acid receptor  $\alpha$  immunoreactivity in the human hypothalamus.**  
1. Meng QY, Chen XN, Zhao J, Swaab DF, Zhou JN.  
Neuroscience. 2010 Dec 3. [Epub ahead of print]  
PMID: 21130848 | PubMed - as supplied by publisher  
[Related citations](#)
- Caloric restriction experience reprograms stress and orexinergic pathways and promotes binge eating.**  
Pankevich DL, Teegardin GL, Heden AJ, Jensen CL, Dale TL.  
J Neurosci. 2010 Dec 1;30(48):16399-407.  
PMID: 21123506 | PubMed - Indexed for MEDLINE  
[Related citations](#)
- Glucocorticoids Differentially Regulate the Expression of CRFR1 and CRFR2 $\alpha$  in MN9 Insulinoma Cells and Rodent Islets.**  
Huising MO, Pilbrow AP, Matsumoto M, van der Meulen I, Park H, Vaughan JM, Lee S, Vale WW.  
Endocrinology. 2011 Jan;152(1):138-50. Epub 2010 Nov 24.  
PMID: 21108975 | PubMed - in process  
[Related citations](#)
- The benzodiazepine diazepam demonstrates the usefulness of Syrian hamsters as a model for anxiety testing: Evaluation of other classes of anxiolytics in comparison to diazepam.**  
Gammie RI, Lüngwitz F, Balista N, Hesler J, Huntley C, Peacock A, Delagrange P, Millan MJ.  
Behav Brain Res. 2010 Nov 20. [Epub ahead of print]  
PMID: 21081001 | PubMed - as supplied by publisher  
[Related citations](#)

# Small molecule databases have been added at NCBI

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

The screenshot shows the main interface of the PubChem database. At the top, there is a navigation bar with links for "Databases", "Deposition", "Services", "Help", and "more". Below the navigation bar is the "PubChem" logo. Underneath the logo are three main search tabs: "BioAssay", "Compound", and "Substance". A search bar with a "GO" button and a link to "Advanced search" is located below these tabs. To the right of the search area is a vertical sidebar containing ten links with corresponding icons:

Bioactivity summary	
Biactivity datatable	
Biactivity structure-activity	
Chemical structure search	
3D conformer viewer	
Chemical structure clustering	
Deposition gateway	
Structure download	
Bioassay download	
PubChem FTP	

Below the sidebar, there is a "New" announcement box stating: "More than 2.5 million structures from the IBM BAO (Business Analytics and Optimization) strategic IP insight platform (SIIP) are now available in PubChem. See more... and related news." At the bottom of the page, there are links to "Write to Helpdesk", "Disclaimer", "Privacy Statement", "Accessibility", "Data Citation Guidelines", "National Center for Biotechnology Information", "NLM", "NIH", and "HHS".

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

NCBI

HomoloGene Discover Homologs

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 <sup>*</sup>	18,981	18,431
Pan troglodytes	25,096	16,050	15,900
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 <sup>*</sup>	21,084	14,067
Drosophila melanogaster	13,027 <sup>*</sup>	9,282	7,749
Anopheles gambiae	12,460	8,867	7,541
Caenorhabditis elegans	20,132 <sup>*</sup>	8,678	1,810
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 <sup>*</sup>	6,287	6,144
Arabidopsis thaliana	27,000 <sup>*</sup>	19,961	11,242

**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**

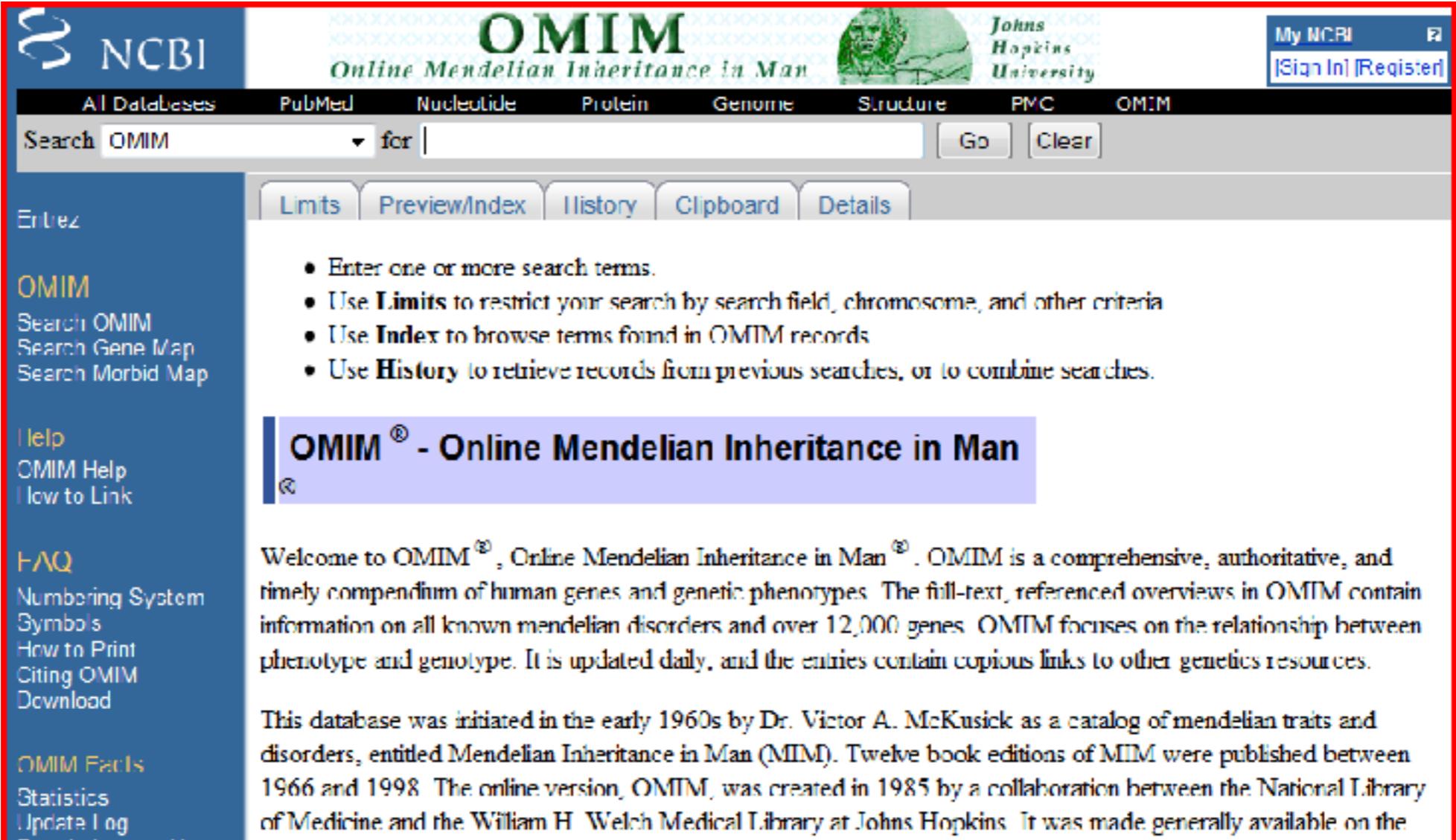
**Entrez Genomes**

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>



The screenshot shows the homepage of OMIM (Online Mendelian Inheritance in Man). The top navigation bar includes links for All Databases, PubMed, Nucleotide, Protein, Genome, Structure, PMC, and OMIM. A search bar is present with the text "Search OMIM" and a dropdown menu set to "for". To the right of the search bar are buttons for "Go" and "Clear". The Johns Hopkins University logo is visible. On the left, a sidebar provides links for Entrez, OMIM (Search OMIM, Search Gene Map, Search Morbid Map), Help (OMIM Help, How to Link), FAQ (Numbering System, Symbols, How to Print, Citing OMIM, Download), and OMIM Facts (Statistics, Update Log). The main content area features a purple header "OMIM® - Online Mendelian Inheritance in Man". Below it, a welcome message states: "Welcome to OMIM®, Online Mendelian Inheritance in Man®. OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources." Another paragraph describes the history: "This database was initiated in the early 1960s by Dr. Victor A. McKusick as a catalog of mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM). Twelve book editions of MIM were published between 1966 and 1998. The online version, OMIM, was created in 1985 by a collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins. It was made generally available on the

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/>

NCBI Bookshelf

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy

Search Books for  Go Clear

Limits Preview/Index History Clipboard Details

Introduction  
Quick Start Guide  
Help  
Information for Authors and Publishers  
What's New  
FAQ  
My NCBI  
Privacy Policy

The Bookshelf is a growing collection of biomedical books that can be searched directly by typing a concept into the textbox above and selecting "Go". Try one of these searches:

► [cell cycle control](#) ► [immuno deficiency](#) ► [protein evolution](#)

► **New on the Bookshelf:**

[Health, United States, 2009](#)  
Hyattsville (MD): National Center for Health Statistics (US) ; 2010

[Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis](#)  
Arvin, Ann; Campadell-Flume, Gabriela; Mocarski, Edward; Moore, Patrick S.; Rotzman, Bennett; Whitley, Richard; Yamashita, Kuniichi, editors  
Cambridge: Cambridge University Press ; 2007

[Probe Reports from the Molecular Libraries Program](#)  
NIH Molecular Libraries  
Bethesda (MD): National Center for Biotechnology Information (US) ; 2010

[StemBook](#)  
Cambridge (MA): Harvard Stem Cell Institute ; 2008-

[VA Evidence-based Synthesis Program Reports](#)  
Washington (DC): Department of Veterans Affairs (US); 2007-

# 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

**GEO navigation**

**QUERY**

- DataSets
- Gene profiles
- GEO accession
- GEO BLAST

**BROWSE**

- DataSets
  - Platforms
  - Samples
  - Series
- GEO accessions

**Site contents**

**Public data**

Platforms	8,246
Samples	514,893
Series	20,827

**Documentation**

- Overview | FAQ | Find
- Submission guide
- Linking & citing
- Journal citations
- Construct a Query
- Programmatic access
- DataSet clusters
- GEO announce list
- Data disclaimer
- GEO staff

**Query & Browse**

- Repository browser
- Submitters

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

GEO Accession viewer  
[www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3541](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3541) Reader

**NCBI** **GEO**

Platforms (1) GPL4091 Agilent-014693 Human Genome CGH Microarray 244A (Feature number version)

Samples (4) [More...](#)

GSM495808	Aspc1 Cell Line
GSM495809	JH39 Xenograft
GSM495810	JH21 Xenograft

**Download family**

SOFT formatted family file(s)	Format
MINIML formatted family file(s)	MINIML
Series Matrix File(s)	TXT

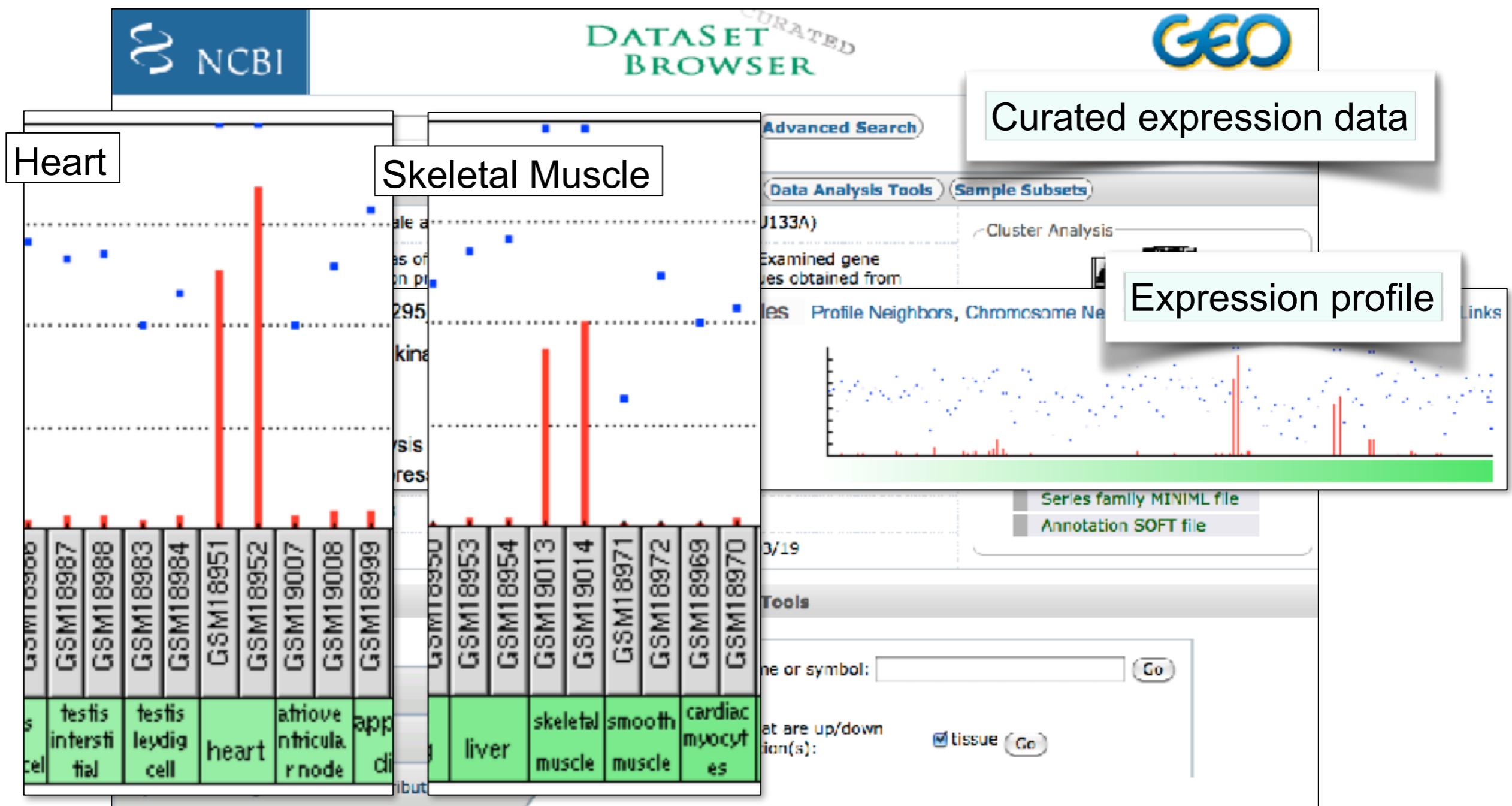
**Supplementary file**

GSE19052_RAW.tar	Size	Download	File type/resource
	170.7 Mb	(ftp)(http)	TAR (of TXT)

NLM | NIH | GEO Help | Disclaimer | Section 508 |

uter. Significance Analysis of Microarrays (SAM) identified 92 genes differentially expressed by strain. Interestingly, several members of the solute carrier family of amino acid transporters, genes involved in amino acid synthesis and development, and amiloride-sensitive epithelial sodium channel gene were induced by strain. These results were confirmed by quantitative real-time polymerase chain reaction (qRT-PCR). Thus, this study identifies genes induced by strain that may be important for amino acid signaling pathways, protein

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



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

The screenshot shows the QuickGO homepage on a web browser. The header includes the EMBL-EBI logo, a search bar with 'Enter Text Here' and a 'Find' button, and links for 'Terms of Use', 'Privacy', and 'Cookies'. A navigation menu at the top has tabs for 'Databases', 'Tools', 'Research', 'Training', 'Industry', 'About Us', and 'Help', along with links for 'Site Index', 'RSS', and 'Print'.

The main content area features a sidebar with links to 'QuickGO', 'Help', 'Reference', 'FAQs', 'Video tutorials', 'Downloads', 'geneontology.org', 'UniProt-GOA project', and 'Web Services'. The main content area displays a brief introduction to QuickGO, a search bar with placeholder 'Click for example search' and a 'Search!' button, and icons for 'Web Services', 'Dataset', and 'Term Basket: 0'. Below this are three callout boxes: one about 'Search and Filter GO annotation sets', one about 'Investigate GO slims', and one about 'View the history of changes to GO'. To the right, there is a 'QuickGO News' section with links to news items from August 2011, April 2011, and a news archive, as well as sections for 'QuickGO Tips' and 'Tutorial'.

**QuickGO**

www.ebi.ac.uk/QuickGO/

EMBL-EBI

Databases Tools Research Training Industry About Us Help Site Index RSS Print

QuickGO

EBI > Databases > QuickGO

QuickGO

QuickGO is a fast web-based browser for [Gene Ontology](#) terms and annotations, which is provided by the [UniProt-GOA project](#) at the [EBI](#).

**QuickGO**

Click for example search **Search!**

Web Services Dataset Term Basket: 0

**Search and Filter GO annotation sets**

Extensive filters are available from this page to allow the generation of specific subsets of GO annotations, mapped to sequence identifiers of your choice.

**Investigate GO slims**

GO slims are lists of GO terms that have been selected from the full set of terms available from the Gene Ontology project.

GO slims can be used to generate a focused view of part of the GO, or with annotation data they can be used to see how a set of proteins/genes can be broadly categorized (using annotation data and the relationships that exist between terms in the ontologies).

Further information on GO slims can be found at the [GO Consortium web site](#).

**View the history of changes to GO**

This page allows you to view the changes to GO, optionally filtered by date, term identifier, or type of change.

**QuickGO News**

19 August 2011 - Changes to the Term Basket

14 June 2011 - New term history displays

20 April 2011 - Display Improvements

[QuickGO News Archive](#)

**QuickGO Tips**

**Tutorial**

# GO annotation in UniProt

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

Hemoglobin subunit beta – Homo sapiens (Human)

www.uniprot.org/uniprot/P68871

Hemoglobin subunit beta – Homo sapiens (Human)

Names · Attributes · General annotation · **Ontologies** · Interactions · Sequence annotation · Sequences · References · Web links · Cross-refs · Entry Info · Documents · Customize order

**Gene Ontology (GO)**

**Biological\_process**

- bicarbonate transport  
Traceable author statement. Source: Reactome
- blood coagulation  
Traceable author statement. Source: Reactome
- hydrogen peroxide catabolic process  
Inferred from direct assay ([PubMed 19740759](#)). Source: BHF-UCL
- nitric oxide transport  
Non-traceable author statement ([PubMed 8292032](#)). Source: UniProtKB
- positive regulation of cell death  
Inferred from direct assay ([PubMed 19740759](#)). Source: BHF-UCL
- positive regulation of nitric oxide biosynthetic process  
Non-traceable author statement ([PubMed 7965120](#)). Source: UniProtKB
- protein heterooligomerization  
Inferred from direct assay ([PubMed 19740759](#)). Source: BHF-UCL
- regulation of blood pressure  
Inferred from electronic annotation. Source: UniProtKB-KW
- regulation of blood vessel size  
Inferred from electronic annotation. Source: UniProtKB-KW
- renal absorption  
Inferred from mutant phenotype ([PubMed 18465050](#)) ([PubMed 10974565](#)). Source: UniProtKB
- small molecule metabolic process  
Traceable author statement. Source: Reactome

**Cellular\_component**

- endocytic vesicle lumen  
Traceable author statement. Source: Reactome
- extracellular region  
Traceable author statement. Source: Reactome
- haptoglobin-hemoglobin complex  
Inferred from direct assay ([PubMed 19740759](#)). Source: BHF-UCL
- hemoglobin complex  
Non-traceable author statement ([Ref.33](#)) ([Ref.72](#)). Source: UniProtKB

**Molecular\_function**

- heme binding  
Inferred from electronic annotation. Source: InterPro
- hemoglobin binding

# GO annotation in UniProt

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

The screenshot shows a web browser window with two tabs open. The main tab displays the UniProt entry for hemoglobin subunit beta - Homo sapiens (Human) (P68871). The second tab shows the QuickGO annotation for GO:0020037 haeme binding. The QuickGO page includes the EBI logo, navigation links (Services, Research, Training, About us), and search functionality. The term information section for GO:0020037 haeme binding provides details such as ID, Name, Ontology (Molecular Function), Definition (Interacting selectively and non-covalently with heme, any compound of iron complexed in a porphyrin [tetrapyrrole] ring), and GONUTS link. It also lists Synonyms (haem binding) and provides annotation guidance and cross-references.

Hemoglobin subunit beta – Homo sapiens (Human)

www.uniprot.org/uniprot/P68871

Reader

GO:0020037 haeme binding

www.ebi.ac.uk/QuickGO/GTerm?id=GO:0020037

Reader

EMBL-EBI

Quick GO

A fast browser for Gene Ontology terms and annotations.

EBI > Databases > QuickGO

GO:0020037 haeme binding

Quick GO Click for example search Search! Web Services Dataset Term Basket: 0

Term Information | Ancestor Chart | Child Terms | Protein Annotation | Co-occurring Terms | Change Log

ID	GO:0020037
Name	haeme binding
Ontology	Molecular Function
Definition	Interacting selectively and non-covalently with heme, any compound of iron complexed in a porphyrin [tetrapyrrole] ring.
GONUTS	GO:0020037 Wiki Page

Synonyms | Annotation Guidance | Cross-Ontology Relations | Cross-references

Synonyms are alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope. Click on the icon for more details.

Type	Synonym
exact	haem binding

Please send comments, suggestions or bug reports to [goa@ebi.ac.uk](mailto:goa@ebi.ac.uk). Click here for details of how to cite UniProt-GOA and QuickGO.

11ms

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

The screenshot shows the DAVID Bioinformatics Resources 6.7 homepage. The top navigation bar includes links for Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, Why DAVID?, and About Us. The main content area features the DAVID logo and the text "DAVID Bioinformatics Resources 6.7" and "National Institute of Allergy and Infectious Diseases (NIAID), NIH". A sidebar on the left provides links to Functional Annotation, Gene Functional Classification, Gene ID Conversion, Gene Name Batch Viewer, and NIAID Pathogen Annotation Browser. The central content area discusses the latest release (v6.7) and highlights its features, including functional annotation clustering, charting, and tables. A sidebar on the right lists "What's Important in DAVID?" such as current release notes and citation requirements. A large callout box in the bottom right corner states: "DAVID allows you to upload lists of genes and search for enriched GO and search for functionally related genes not in your list". Below this is a URL link: <http://david.abcc.ncifcrf.gov>.

DAVID Functional Annotation Bioinformatics Microarray Analysis

david.abcc.ncifcrf.gov/home.jsp

DAVID Functional Annotation Bioinformatics Microarray Analysis DAVID: Database for Annotation, Visualization, and Integrated...

**DAVID Bioinformatics Resources 6.7**  
National Institute of Allergy and Infectious Diseases (NIAID), NIH

**Functional Annotation**

- Functional Annotation Clustering
- Functional Annotation Chart
- Functional Annotation Table

**Gene Functional Classification**

**Gene ID Conversion**

**Gene Name Batch Viewer**

**NIAID Pathogen Annotation Browser**

DAVID 6.7  
003 - 2013

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 is an update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological themes or disease pathways across a list of genes. For any list of genes, DAVID can:

- Identify enriched biological processes, molecular functions, and cellular components.
- Discover enriched pathways.
- Cluster redundant genes.
- Visualize genes.
- Display related genes.
- Search for other functionally related genes not in the list.
- List interacting proteins.
- Explore gene names in batch.

DAVID allows you to upload lists of genes and search for enriched GO and search for functionally related genes not in your list

<http://david.abcc.ncifcrf.gov>

# Example output: enriched functions from GO

DAVID: Database for Annotation, Visualization, and Integrat...ID); Science Applications International Corporation (SAIC)  david.abcc.ncifcrf.gov/chartReport.jsp?annot=25  Reader 

DAVID: Functional Annotation Result Summary Database for Annotation, Visualization, and Integrated Discov...

**DAVID Bioinformatics Resources 6.7**  
National Institute of Allergy and Infectious Diseases (NIAID), NIH

## Functional Annotation Chart

Help and Manual

Current Gene List: List\_1  
Current Background: Homo sapiens  
14 DAVID IDs  
 Options

Rerun Using Options Create Sublist

10 chart records 

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">regulation of osteoclast differentiation</a>			2	14.3	2.1E-2	9.9E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">response to organic substance</a>			4	28.6	2.9E-2	9.6E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">regulation of myeloid leukocyte differentiation</a>			2	14.3	3.9E-2	9.5E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">positive regulation of transcription from RNA polymerase II promoter</a>			3	21.4	4.8E-2	9.4E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">regulation of myeloid cell differentiation</a>			2	14.3	6.5E-2	9.5E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">cartilage development</a>			2	14.3	6.9E-2	9.3E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">positive regulation of transcription, DNA-dependent</a>			3	21.4	7.5E-2	9.2E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">positive regulation of RNA metabolic process</a>			3	21.4	7.6E-2	8.9E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">response to protein stimulus</a>			2	14.3	9.8E-2	9.3E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">positive regulation of transcription</a>			3	21.4	1.0E-1	9.1E-1

8 gene(s) from your list are not in the output.