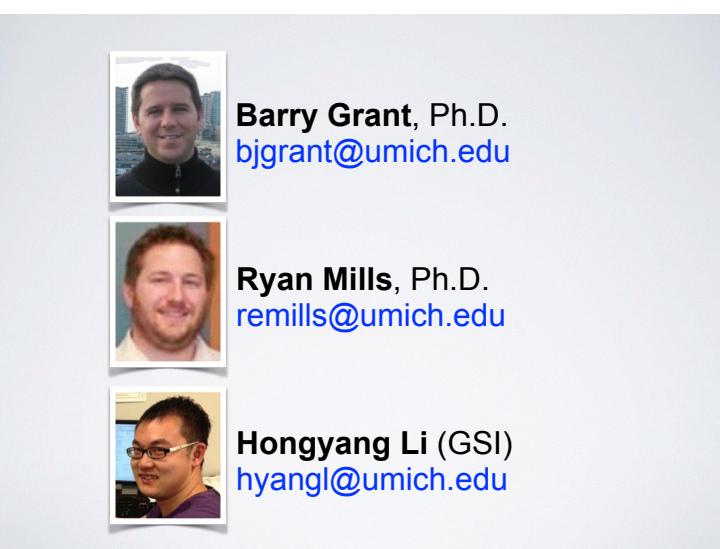


**INTRODUCTION TO BIOINFORMATICS**

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

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

BIOINF 525 [http://bioboot.github.io/bioinf525\\_w16/](http://bioboot.github.io/bioinf525_w16/) 12-Jan-2016



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

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

**Hongyang Li (GSI)**  
[hyangl@umich.edu](mailto:hyangl@umich.edu)

## COURSE LOGISTICS

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

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

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

## MODULE OVERVIEW

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

- 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 on Ctools:  
<http://tinyurl.com/bioinf525-w16>
- Complete the **lecture 1.1 homework questions**:  
<http://tinyurl.com/bioinf525-quiz1>

## Q. What is Bioinformatics?

## Q. What is Bioinformatics?

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

[After Orengo, 2003]

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

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

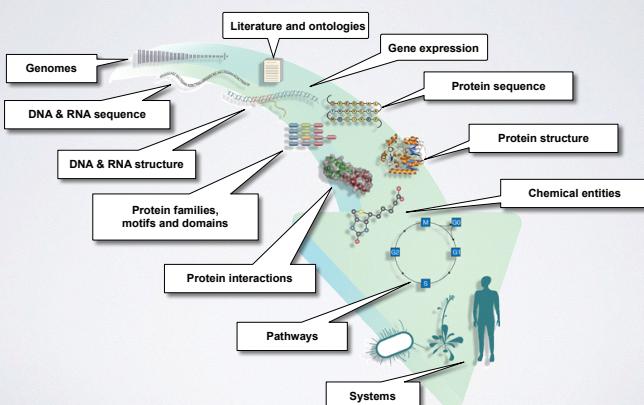
## MORE DEFINITIONS

- “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” techniques (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**. Luscombe NM, et al. Methods Inf Med. 2001;40:346.
- “Bioinformatics is research, development, or application of **computational approaches** for expanding the use of **biological, medical, behavioral or health data**, including those to acquire, store, organize and analyze such data.” National Institutes of Health (NIH) (<http://tinyurl.com/l3gxr6b>)

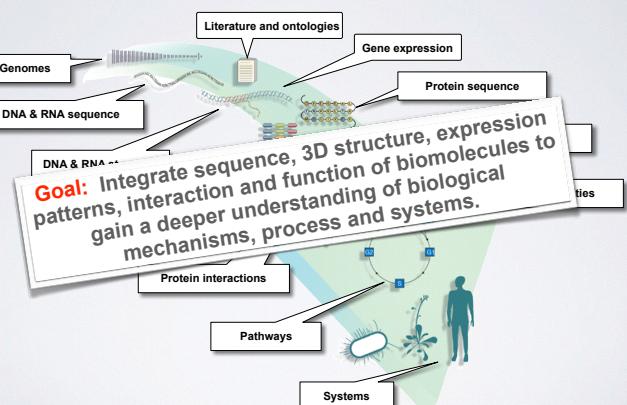
## MORE DEFINITIONS

- “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” techniques (derived from disciplines such as applied 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>)

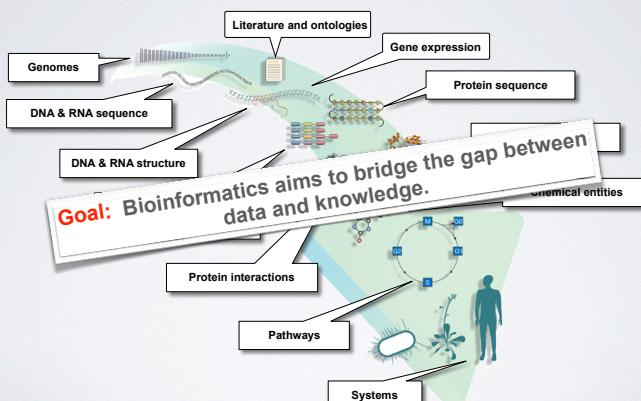
## Major types of Bioinformatics Data



## 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

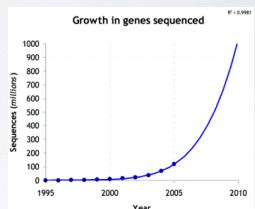
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.

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

## Why do we need Bioinformatics?

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

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

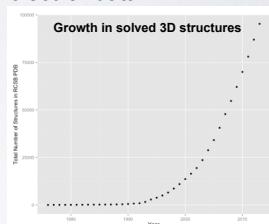


Bioinformatics is essential for the archiving, organization and analysis of data from sequencing, structural genomics, microarrays, proteomics and new high throughput assays.

## Why do we need Bioinformatics?

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

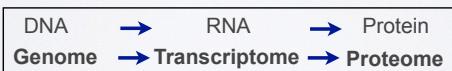
- Bioinformatics provides methods for the efficient:
  - storage
  - annotation
  - search and retrieval
  - data integration
  - data mining and analysis



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

## How do we do Bioinformatics?

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



## How do we actually do Bioinformatics?

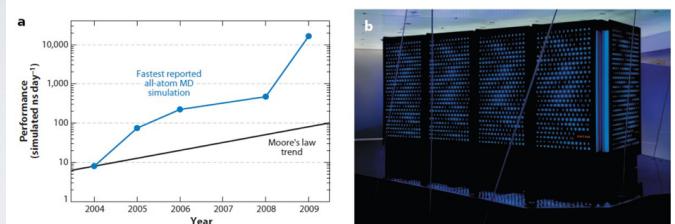
### Pre-packaged tools and databases

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

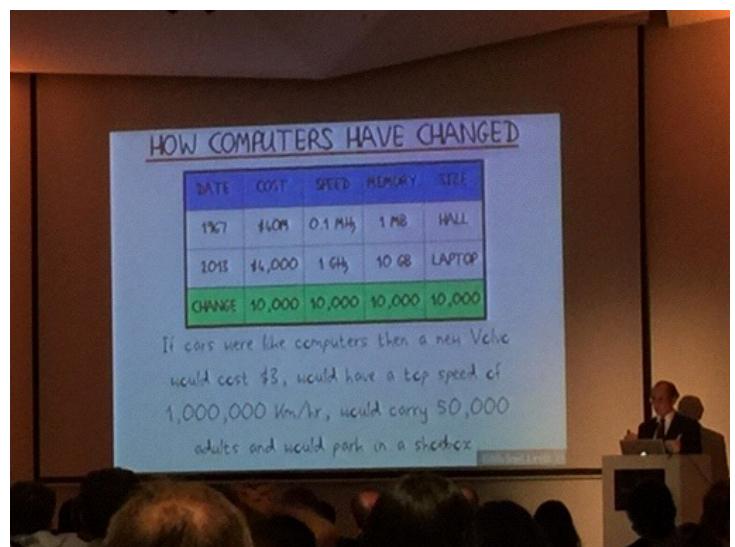
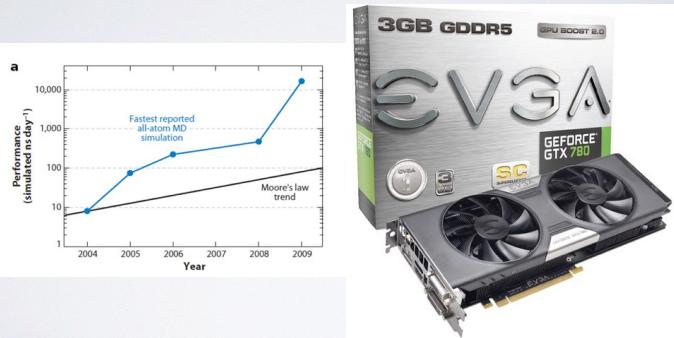
### Tool development

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

### SIDE-NOTE: SUPERCOMPUTERS AND GPUs



### SIDE-NOTE: SUPERCOMPUTERS AND GPUs



## 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)

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST\_PROGRAMS=blastp&PAGE\_TYPE=BlastSearch&SHOW\_DEFAULTS=on&LINK\_LOC=blasthome

**General Parameters**

- Max target sequences: 500 (Select the maximum number of aligned sequences to display)
- Short queries: Automatically adjust parameters for short input sequences
- Expect threshold: 10
- Word size: 3
- Max matches in a query range: 0

**Scoring Parameters**

- Matrix: BLOSUM62

**Gap Costs**

- Existence: 11 Extension: 1

**Compositional adjustments**

- Conditional compositional scoring

**Filters and Masking**

- Filter: Low complexity regions
- Mask: Mask for lookup table only, Mask lower case letters

**PSI/PHI/DELTA BLAST**

- Upload PSSM: Choose File (no file selected)
- Optional PSI-BLAST Threshold: 0.005
- Pseudocount: 0

**Even Blast has many settable parameters**

**Related tools with different terminology**

STEP 3 - Set your PROGRAM  
Fasta

MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION	UPPER VALUE	LOWER VALUE
BLOSUM60	-10	-2	2	10	0 (default)	
DNA STRAND	none	none	none	STATISTICAL ESTIMATES		
N/A	no	none	none	Regress		
SCORES	50	50	START-END	START-END	no	
ALIGNMENTS						
SEQUENCE RANGE						
DATABASE RANGE						
MULTIHSPS						

SCORE FORMAT: Default

## Key Online Bioinformatics Resources: NCBI & EBI

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

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

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

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

## Key Online Bioinformatics Resources: NCBI & EBI

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

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

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

## National Center for Biotechnology Information (NCBI)

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

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

National Center for Biotechnology Information

**Welcome to NCBI**

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

**Get Started**

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-Tos: Learn how to accomplish specific tasks at NCBI databases

**3D Structures**

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine receptor-ligand relationships, protein-protein interactions, molecular interactions, biological activities of small chemicals, and associated annotations.

**NCBI Announcements**

New version of Genome Workbench available

**Popular Resources**

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

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

National Center for Biotechnology Information

**Welcome to NCBI**

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

**Get Started**

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-Tos: Learn how to accomplish specific tasks at NCBI databases

**3D Structures**

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine receptor-ligand relationships, protein-protein interactions, molecular interactions, biological activities of small chemicals, and associated annotations.

**Popular Resources**

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

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

The screenshot shows the NCBI homepage with a search bar at the top. Below it, a banner highlights "Notable NCBI databases include: GenBank, RefSeq, PubMed, dbSNP" and search tools ENTREZ and BLAST. On the left, there's a sidebar with links to Homology, Literature, Proteins, Sequence Analysis, Taxonomy, Training & Tutorials, and Variation. The main content area features sections for 3D Structures, Protein, PubChem, and NCBI Announcements.

## Key Online Bioinformatics Resources: NCBI & EBI

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

This image compares the NCBI and EBI websites. It shows two side-by-side screenshots. The left one is the NCBI homepage, and the right one is the EBI homepage. Both pages feature search bars and links to various databases and services. The EBI page includes a "Services" section highlighted with a red box.

## 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 EBI homepage with a search bar and a "Services" section highlighted with a red box. Other sections include Research, Training, Industry, European Coordination, and EMBL ALUMNI. A sidebar on the right provides links to Services, Research, Training, News, and Contacts, along with information about the Plant and Animal Genome conference (PAG XXIV).

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

The screenshot shows the EBI Services page with a "Bioinformatics services" section. It lists various services: DNA & RNA, Gene expression, Proteins, Structures, Systems, Chemical biology, Ontologies, Literature, and Cross domain. A "Popular" section on the right highlights Ensembl, UniProt, PDB, ArrayExpress, and ChEMBL.

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

The screenshot shows the EBI Services page with a "Bioinformatics services" section. It lists various services: DNA & RNA, Gene expression, Proteins, Structures, Systems, Chemical biology, Ontologies, Literature, and Cross domain. A "Popular" section on the right highlights Ensembl, UniProt, PDB, ArrayExpress, and ChEMBL. The "Proteins" service is also highlighted with a red box.

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

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

This screenshot shows the 'Proteins' section of the EBI website. It lists several popular services under the heading 'Popular services': UniProt (The Universal Protein Resource), InterPro (A database for the classification of proteins into families, domains and conserved sites), PRIDE (The Proteomics Identifications Database), Pfam (A database of hidden Markov models and alignments to describe conserved protein families and domains), Clustal Omega (Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools), 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), and 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

This screenshot shows the main homepage of the European Bioinformatics Institute (EMBL-EBI). At the top, it features the EBI logo and navigation links for Services, Research, Training, and About us. Below the header, there is a search bar with the placeholder 'Find a gene, protein or chemical:' and a 'Search' button. To the right of the search bar, there is a 'Popular' section with links for Services, Research, Training, News, and Contacts. A prominent red box highlights the 'Training' link in the Popular section. Further down the page, there is a 'Services' section with icons for Services, Research, European Coordination, and EMBL ALUMNI. On the right side, there is a 'Visit EMBL.org' section with a link to the EMBL 40th anniversary logo and information about the Plant and Animal Genome conference (PAG XXIV) taking place from Sunday 10 - Tuesday 12 January 2016.

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

This screenshot shows a specific online training course titled 'Using sequence similarity searching tools at EMBL-EBI: webinar'. The course is part of the 'Train online' series. The page includes a video player showing a recording of the webinar, which features Andrew Cowley. The video player displays the title 'Using sequence similarity searching tools at EMBL-EBI: webinar' and the duration '0:00 / 37:42'. To the right of the video, there is a sidebar with links for Popular, Find us at..., and Navigation.

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

This screenshot shows the 'Train online' page of the EBI website. The page features a large callout box in the center with the text 'Notable EBI databases include: ENA, UniProt, Ensembl and the tools FASTA, BLAST, InterProScan, ClustalW, T-Coffee, MUSCLE, DALI, HMMER'. Below this box, there is a 'Find a course' section with a 'Browse by subject' dropdown menu. The subjects listed are Genes and Genomes, Gene Expression, and Interactions, Distances 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

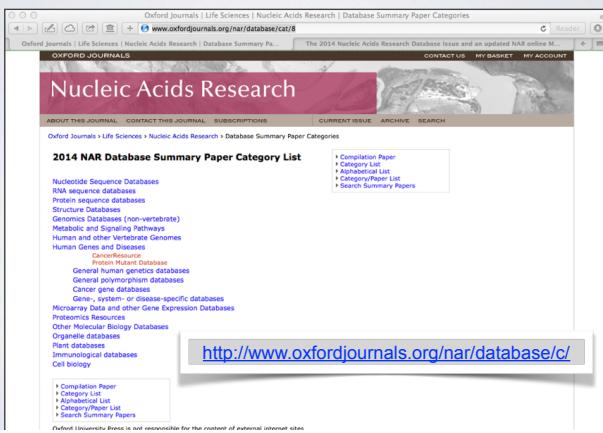
## Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, Biolmage, 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, DPLinteract, ECDc, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genllesene, GenLink, GENOTk, GenProtEC, GIFTS, GPCRDB, GRAP GRBase, gRNAsdb, GRR GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXADb, HGMD, HIDB, HIDC, HIVdb, HotMoleBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGenbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSub, 0-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtList, 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 ..... !!!

## Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, Biolmage, 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, DPLinteract, ECDc, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genllesene, GenLink, GENOTk, GenProtEC, GIFTS, GPCRDB, GRAP GRBase, gRNAsdb, GRR GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXADb, HGMD, HIDB, HIDC, HIVdb, HotMoleBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGenbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSub, 0-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtList, 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 ..... !!!

## Finding Bioinformatics Databases



## Major Molecular Databases

The most popular bioinformatics databases focus on:

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

The are also many popular "boutique" databases for:

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

## Major Molecular Databases

The most popular bioinformatics databases focus on:

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

The are also many "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](#))

## Primary, secondary & composite databases

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

- **Primary databases** (or *archival databases*) consist of data derived experimentally.
  - **GenBank**: NCBI's primary nucleotide sequence database.
  - **PDB**: Protein X-ray crystal and NMR structures.
- **Secondary databases** (or *derived databases*) contain information derived from a primary database.
  - **RefSeq**: non redundant set of curated reference sequences primarily from GenBank
  - **PFAM**: protein sequence families primarily from UniProt and PDB
- **Composite databases** (or *metadatabases*) join a variety of different primary and secondary database sources.
  - **OMIM**: catalog of human genes, genetic disorders and related literature
  - **GENE**: molecular data and literature related to genes with extensive links to other databases.

## GENBANK & REFSEQ: NCBI'S NUCLEOTIDE SEQUENCE DATABASES

### What is GenBank?

- GenBank is NCBI's primary nucleotide only sequence database
  - ▶ Archival in nature - reflects the state of knowledge at time of submission
  - ▶ Subjective - reflects the submitter point of view
  - ▶ Redundant - can have many copies of the same nucleotide sequence
- GenBank is actually three collaborating international databases from the US, Japan and Europe
  - ▶ GenBank (US)
  - ▶ DNA Database of Japan (DDBJ)
  - ▶ European Nucleotide Archive (ENA)

### GenBank, ENA and DDBJ Share and synchronize data



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

### GenBank sequence record

The screenshot shows a detailed view of a GenBank sequence record for NM\_004984.2. It includes fields such as Locus, Definition, Version, Reference, Source, Organization, Reference, Authors, Title, Journal, PMID, Remark, and References. A callout box highlights the "GenBank flat file format has defined fields including unique identifiers such as the ACCESSION number." Another callout box notes that "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 NT_030059	GenBank genomic DNA sequence Genomic contig	DNA
N91759.1 NM_006744	An expressed sequence tag (1 of 170) RefSeq DNA sequence (from a transcript)	RNA
NP_007635 AAC02945 Q28369 1KT7	RefSeq protein GenBank protein UniProtKB/SwissProt protein Protein Data Bank structure record	Protein
PMID: 12205585	PubMed IDs identify articles at NCBI/NIH	Literature

### GenBank sequence record

This screenshot shows another view of the same GenBank sequence record for NM\_004984.2. It includes fields such as Locus, Definition, Version, Reference, Source, Organization, Reference, Authors, Title, Journal, PMID, Remark, and References. Callout boxes highlight various features of the interface, including links to PubMed, Entrez, and other NCBI databases.

## GenBank sequence record

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

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

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

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide : [KIF5A] & "Homo sapiens" Search Help

Display Settings GenBank

Send: Change region shown

Fasta Graphics **Can set different display formats here**

PMID: 504984.2

NCBI Reference Sequence: NM\_004984.2

**Fasta** Graphics

Send: Change region shown

Optimize view

Display this sequence Run BLAST Pick Primers Highlight Sequence Features Find in This Sequence

LOCUS NM\_004984 3897 bp mRNA linear PR1 10-JAN-2014

DEFINITION Homo sapiens kinesin family member 5A (KIF5A); mRNA.

ACCESSION NM\_004984

VERSION NM\_004984.2 GI:54467448

KEYWORDS BrefSeq

SOURCE Homo sapiens (human)

ORGANISM Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Borelia; Bovariochiroptery; Primates; Haplorrhini; Chiroptera; Bovidae; Bos taurus

REFERENCE 1 (bases 1 to 3897)

TITLE Role of kinesin-1 in the pathogenesis of SPG15, a rare form of hereditary spastic paraparesis

JOURNAL J Neurogenet

YEAR 2013

PAGENO 237-238

REMARK GeneID: A review of the mechanism of pathogenesis involved in the hereditary spastic type 10 when KIF5A is inactivated by mutations. Review Article

REFERENCE 2 (bases 1 to 3897)

AUTHORS Bohn,X.,Zhu,Y.,Stryj,S.,Camilioni,S.,Buder,K.,Rieck,R.,Bohm,X.J.,and Wimmer,B.

TITLE  $\alpha$ -Synuclein oligomers impair neuronal microtubule-kinesin dynamics

JOURNAL J Biol. Chem.

YEAR 2013

PMID: 237422154

ARTICLES about the KIF5A gene  $\alpha$ -Synuclein oligomers impair neuronal microtubule-kinesin dynamics | J Biol Chem. (2013) Molecular cloning of KIF5A for GABA(A) receptor transport, a Neuron. (2012) Systems-wide analysis of ubiquitylation dynamics reveals a key n (Neuron Cell Biol. 2012)

SEE ALL...

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

# FASTA sequence record

## GenBank ‘graphics’ sequence record

The screenshot shows the NCBI mRNA report for NM\_004984.2, Homo sapiens kinesin family member 5A (KIF5A), mRNA. The top navigation bar includes links for Home, Help, and Sign in to NCBI. The main content area displays the sequence details, including the reference sequence (NM\_004984.2), its length (13,890 bp), and its genomic location (chromosome 11, position 120,000,000-133,890,000). Below this, the sequence is shown as a track with exons and introns. A detailed description of the gene's structure and function is provided, mentioning its role in microtubule-kinesin interaction, its essentiality for GABA(A) receptor transport, and its dynamics in ubiquitination pathways. The bottom section contains a 'DTS Markers' table and a 'Reference sequence information' section.

## GenBank sequence record, cont.

Home sapiens kinesin family member 5A (KIF5A); mRNA = Nucleotide – NCBI

[www.ncbi.nlm.nih.gov/nucleotide/NC\\_004984.2](http://www.ncbi.nlm.nih.gov/nucleotide/NC_004984.2)

Home sapiens Kinesin Family member 5A (KIF5A); mRNA = Nucleotide – NCBI

NCBI Resources How To Search Sign in to NCBI Help

Nucleotide (KIF5A) AND "Homo sapiens" Search

Display Settings GenBank

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

NCBI Reference Sequence: NM\_004984.2

FASTA Graphics

Go to: ▾

LOCUS NM\_004984

DEFINITION Home sapiens kinesin family member 5A (KIF5A). mRNA.

ACCESSION NM\_004984.2

VERSION 4.2

KEYWORDS RefSeq; Human; Homo sapiens (human)

SOURCE Human Genome Project

ORGANISM Homo sapiens

Muscaris; Metase; Chodat; Craniata; Vertebrata; Euteleostomi; Mammalia; Primates; Eutheria; Eothenoleopidea; Primates; Haplorhini; Catarrhini; Hominoidea; Homo;

REFERENCE (bases 1 to 3897)

AUTHORS Bader,K., Riek,R.

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

JOURNAL J. Biol. Chem. 286 (13): 336-344 (2011)

PUBMED 22785106

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

REFERENCE (bases 1 to 3897)

AUTHORS Prota,T., Weber,V., Brey,S., Campioni,B., Buder,K., Riek,R., Bohm,J.X., and Winter,B.

TITLE Small-molecular oligomers impair neuronal microtubule-kinesin interplay

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

PubCite

Send: ▾

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in This Sequence

Articles about the KIF5A gene

α-hydroxyl digoxigenin impairs 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 upylation dynamics reveals a key [Nat Cell Biol. 2012]

See all...

Pathways for the KIF5A gene

Peptide hormone metabolism

MHC class II antigen presentation

## GenBank sequence record, cont.

Home sapiens kinesin family member SA (KIFSA), mRNA - Nucleotide - NCBI

Home sapiens kinesin family member SA (KIFSA), mRNA - Nucleotide - NCBI

OMIM  
Probe  
Protein  
PubMed  
PubMed (RefSeq)

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

Recent activity

Turn Off Gear

Home sapiens kinesin family member SA (KIFSA), mRNA

(KIFSA) AND "Home sapiens" [orgn] (1551) Nucleotide

kinsein (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)

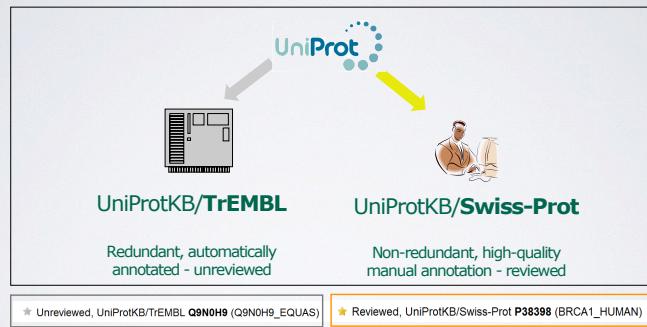
#### 2. UniRef (Reference Clusters)

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

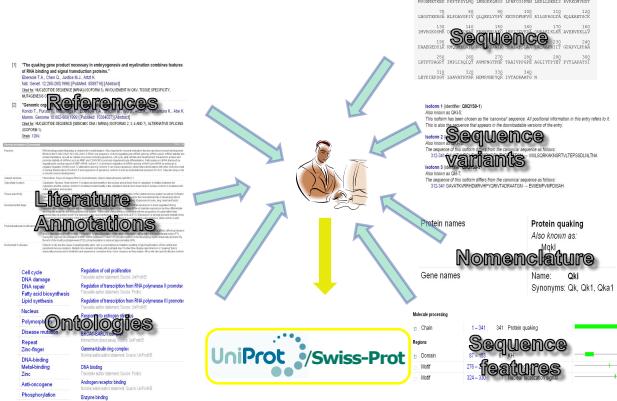
#### 3. UniParc (Archive) with database cross-references to source.

#### 4. UniMES (Metagenomic and Environmental Sequences)

### The two sides of UniProtKB

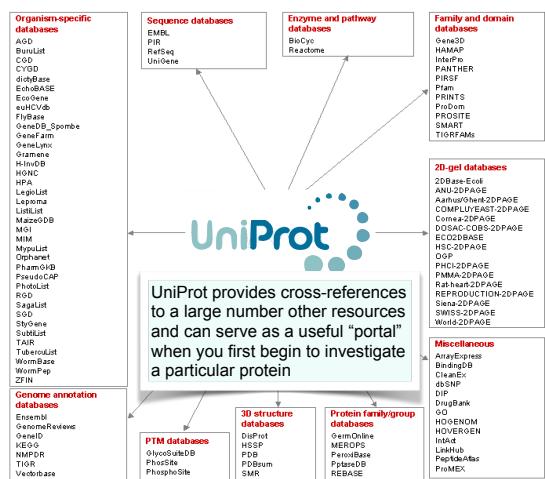


### The main information added to a UniProt/Swiss-Prot entry



70

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

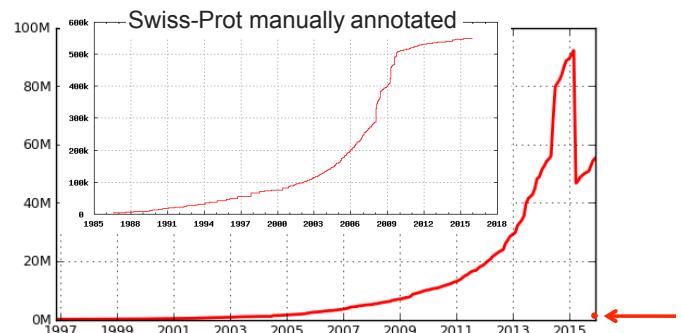


71

## UniProt/Swiss-Prot vs UniProt/TrEMBL

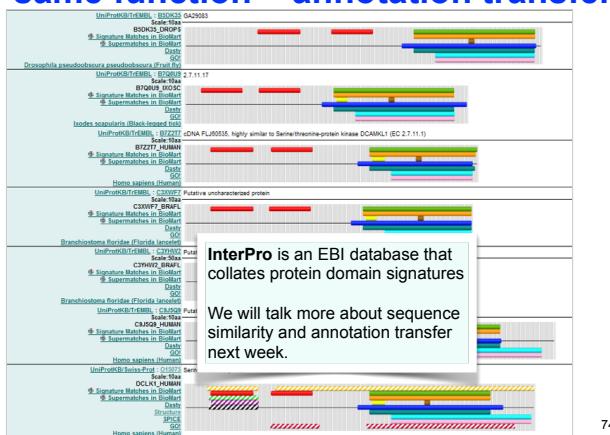
- **UniProtKB/Swiss-Prot** is a **non-redundant** database with one entry per protein
- **UniProtKB/TrEMBL** is a **redundant** database with one entry per translated ENA entry (ENA is the EBI's equivalent of GenBank)
  - Therefore TrEMBL can contain multiple entries for the same protein
  - Multiple 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)



73

## Same domain composition = same function = annotation transfer



74

## 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/>

*Hands on demo (or see following slides)*

	Literature	Genes
Books	1,677	books and reports
MeSH	402	ontology used for PubMed indexing
NLM Catalog	223	books, journals and more in the NLM Collections
PubMed	54,672	scientific & medical abstracts/citations
PubMed Central	96,114	full-text journal articles
ClinVar	759	human variations of clinical significance
dbGaP	120	genotype/phenotype interaction studies
GTR	1,879	genetic testing registry

77

[www.ncbi.nlm.nih.gov/gene/ras](http://www.ncbi.nlm.nih.gov/gene/ras)

NCBI Resources How To Sign in to NCBI

Gene Gene ras Search Help

Show additional filters Save search Advanced

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

Filters: Manage Filters

Clear all Gene sources Genomic Mitochondria Organelles Plasmids Plastids Categories Alternatively spliced Annotated genes Non-coding Protein-coding Pseudogene Sequence content CCDS Ensembl RefSeq Status Current only Chromosome locations

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

Results: 1 to 20 of 85633 << First < Prev Page 1 of 4282 Next > Last >>

Filters activated: Current only. Clear all to show 87165 items.

Name/Gene ID	Description	Location	Aliases
r8s ID: 19412	resistance to audiogenic seizures (Mus musculus (house mouse))	asr	
r8s ID: 43873	rasberry [Drosophila melanogaster (fruit fly)]	Chromosome X, NC_004354.4, CG11485, CG1799, DmelCG1799, EP1X1093,	Dmel_CG1799, EP1X1093,

78

[www.ncbi.nlm.nih.gov/gene](http://www.ncbi.nlm.nih.gov/gene)

NCBI Resources How To Sign in to NCBI

Gene Gene (ras) AND "Homo sapiens"[porgn:\_txid9606] Search Help

Show additional filters Display Settings: Tabular, 20 per page, Sorted by Relevance Send to: Hide sidebar >>

Filters: Manage Filters

Clear all Gene sources Genomic Top Organisms [Tree]

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

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

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4893	neuroblastoma, RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCM51, NS6, NRAS
KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, N-ras, K-RAS2

Find related data Database: Select Find items Search details ras[All Fields] AND "Homo sapiens"[porgn] AND alive[property] See more... Recent activity Turn Off Clear 79

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)

80

[www.ncbi.nlm.nih.gov/gene](http://www.ncbi.nlm.nih.gov/gene)

NCBI Resources How To Sign in to NCBI

Gene Gene (ras) AND "Homo sapiens"[porgn:\_txid9606] Search Help

Show additional filters Display Settings: Tabular, 20 per page, Sorted by Relevance Send to: Hide sidebar >>

Filters: Manage Filters

Clear all Gene sources Genomic Top Organisms [Tree]

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

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

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4893	neuroblastoma, RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCM51, NS6, NRAS
KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, N-ras, K-RAS2

Find related data Database: Select Find items Search details ras[All Fields] AND "Homo sapiens"[porgn] AND alive[property] See more... Recent activity Turn Off Clear 81

[www.ncbi.nlm.nih.gov/gene/3845](http://www.ncbi.nlm.nih.gov/gene/3845)

NCBI Resources How To Sign in to NCBI

Gene Gene KRAS Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)] Search Help

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

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

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-RAS2; K-

32

[www.ncbi.nlm.nih.gov/gene/3845](http://www.ncbi.nlm.nih.gov/gene/3845)

NCBI Resources How To Sign in to NCBI

Gene Gene KRAS Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)] Search Help

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)

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-RAS2; K-

Example Questions:  
What chromosome location and what genes are in the vicinity?

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)

83

**Genomic context**

Location: 12p12.1  
Exon count: 6

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

Go to nucleotide: Graphics FASTA GenBank

**Example Questions:**  
What 'molecular functions', 'biological processes', and 'cellular component' information is available?

**KRAS KRAS (human) [Summary]**  
Gene ID: 3845

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; Eucarchontoglires; Primates; Haplorhini; Catarrhini; Hominoidea; Homo  
Also known as: NS; NS3; CFC2; KRAS1; KRAS2; RASK2; K-RAS; C-K-RAS; K-RAS2A; K-RAS2B

**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

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

## GO: Gene Ontology

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

**UniProt-GOA**

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 biocurator . 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 .

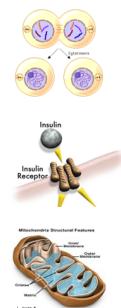
## Why do we need Ontologies?

- 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

88

## 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



89

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

UniProt will detail much more information for protein coding genes

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

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

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

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

PFAM is one of the best protein family databases

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

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

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

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

**Family: Kinesin (PF00225)**

There are 6 interactions for this family. [More...](#)

Tubulin	Tubulin_C	Kinesin	Tubulin	Kinesin
Tubulin	Tubulin_C	Kinesin	Tubulin	Kinesin

Interactions

Questions or comments: pfam@janelia.hhmi.org  
Howard Hughes Medical Institute

Pfam: Family: Kinesin (PF00225)

HHMI  
janelia farm research campus

**Family: Kinesin (PF00225)**

**Structures**

For those sequences which have a structure in the Protein DataBank<sup>9</sup>, we use the mapping between UniProt<sup>5</sup>, PDB and Pfam coordinate systems from the PDB<sup>6</sup> group, to allow us to map Pfam domains onto UniProt sequences and three-dimensional protein structures. The table below shows the structures on which the Kinesin domain has been found.

UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View
A8BKD1_G1AL	11 - 335	2vg	A	11 - 335	Jmol AstexViewer SPICE <sup>8</sup>
			B	11 - 335	Jmol AstexViewer SPICE <sup>8</sup>
CENPE_HUMAN	12 - 329	15c	A	12 - 329	Jmol AstexViewer SPICE <sup>8</sup>
			B	12 - 329	Jmol AstexViewer SPICE <sup>8</sup>
KAR3_YEAST	392 - 723	1pq	A	392 - 723	Jmol AstexViewer SPICE <sup>8</sup>
		1pu	A	392 - 723	Jmol AstexViewer SPICE <sup>8</sup>
		1pv	A	392 - 723	Jmol AstexViewer SPICE <sup>8</sup>
		1pw	A	392 - 723	Jmol AstexViewer SPICE <sup>8</sup>
KI13B_HUMAN	11 - 352	3gb	A	11 - 352	Jmol AstexViewer SPICE <sup>8</sup>
			B	11 - 352	Jmol AstexViewer SPICE <sup>8</sup>
			C	11 - 352	Jmol AstexViewer SPICE <sup>8</sup>
		1q6	A	24 - 359	Jmol AstexViewer SPICE <sup>8</sup>
			B	24 - 359	Jmol AstexViewer SPICE <sup>8</sup>
		1q0b	A	24 - 359	Jmol AstexViewer SPICE <sup>8</sup>
			B	24 - 359	Jmol AstexViewer SPICE <sup>8</sup>
		1x88	A	24 - 359	Jmol AstexViewer SPICE <sup>8</sup>
			B	24 - 359	Jmol AstexViewer SPICE <sup>8</sup>
			A	24 - 359	Jmol AstexViewer SPICE <sup>8</sup>

Pfam: Jm1

Pfam: Family: Kinesin (PF00225)

welcome to sanger institute

PDB entry 3bf

Jmol

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

Close window

## ENTREZ & BLAST:

### TOOLS FOR SEARCHING AND ACCESSING MOLECULAR DATA AT NCBI

**Entrez: Integrated search of NCBI databases**

NCBI National Center for Biotechnology Information

**All Databases**

- PubMed
- Nucleotide
- EST
- Genome
- BioProject
- Sample
- Biosystems
- Books
- Conserved Domains
- dbGaP
- dbVar
- Geographic
- GEO Datasets
- GEO Profiles
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

**to NCBI**

I Center for Biotechnology Information advances science by providing access to biomedical and genomic information.

**Popular Resources**

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

**Entrez is available from the main NCBI homepage or from the homepage of individual databases**

4 May 2012

Information about May's Discovery Workshop, the new GTR and Assembly New Filter Sidebar will be added to PubMed

04 May 2012

A Filter Sidebar will be added next to the PubMed result pages. The useful DELTA BLAST - more sensitive protein searching

20 Apr 2012

Domain Enhanced Lookup Time Accelerated BLAST (DELTABLAST)

More...

### Entrez: navigating across databases

**Word weight**

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

**Phylogeny**

**BLAST**

**Nucleotide sequences**

**Protein sequences**

**VAST**

**Neighbors**

**Related Structures**

**Hard Link**

**Neighbors**

**Related Sequences**

**BLLink**

**Domains**

### Global Entrez Query: All NCBI Databases

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

NCBI Resources How To

Search NCBI databases

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

**Literature**

- Books
- MeSH
- NLM Catalog
- PubMed
- PubMed Central

**Health**

- ClinVar
- dbGaP
- GTR

**Other**

- EST
- full-text journal articles
- GEO Profiles
- HomoloGene
- PopSet
- UniGene
- Proteins

**The Entrez system: 38 (and counting) integrated databases**

<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

Results: 1 to 20 of 35

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:53663 (zgc:53663), mRNA  
2. 2,476 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  
1,269 bp linear mRNA  
Accession: BC11364.1 GI: 213824628  
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: BC114617.1 GI: 15915933  
GenBank FASTA Graphics Related Sequences

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

Send to: Filter your results:  
All (35) Bacteria (0) NSDC (GenBank) (27)  
mRNA (32) RefSeq (6) Manage Filters

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

Find related data Database: Select Find items

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

Recent activity

## Limits

Limits

Published in the last Any Date

Search Field Tags Field: All Fields

Source database Any

Gene Location Any

Modified in the last Any Date

Segmented Sequences Any

Molecule Any

Exclude STS, working draft, TPA, Patents

Reset Search

## Search Results

Nucleotide Nucleotide zebrafish creatine kinase Save search Limits Advanced

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

Results: 1 to 20 of 35

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:53663 (zgc:53663), mRNA  
2. 2,476 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  
1,269 bp linear mRNA  
Accession: BC11364.1 GI: 213824628  
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: BC114617.1 GI: 15915933  
GenBank FASTA Graphics Related Sequences

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

Send to: Filter your results:  
All (35) Bacteria (0) NSDC (GenBank) (27)  
mRNA (32) RefSeq (6) Manage Filters

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

Find related data Database: Select Find items

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

Recent activity

## Advanced: Search Builder

Nucleotide Advanced Search Builder

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

Helps build complex fielded queries

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

Creative kinase (749)  
creative kinase 1 (50)  
creative kinase 2 (2)  
creative kinase b (30)  
creative kinase chain (3)  
creative kinase b mRNA (3)  
creative kinase b pseudogene 1 (1)  
creative kinase brain (43)  
creative kinase brain b (1)

Items from search history can be included / combined / modified

History	Search	Add to builder	Query	Items found	Time
#1	Add		Search zebrafish[organism] AND actin[title]	71	12:41:16
#4	Add		Search zebrafish actin	1288	12:40:07
#1	Add		Search zebrafish creatine kinase	34	12:39:02

## Complex Query Results

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

Results: 6

Danio rerio creatine kinase, brain a (ckba), mRNA  
1. 1,481 bp linear mRNA  
Accession: NM\_001071718.1 GI: 16004536  
GenBank FASTA Graphics Related Sequences

Danio rerio creatine kinase, mitochondrial 1 (ckmtt), nuclear gene encoding mitochondrial protein, mRNA  
2. mRNA  
("Danio rerio"[Organism] AND "creatin kinase"[Title]) AND "refseq"[Filter] AND mRNA[Filter]

Danio rerio creatine kinase, brain b (ckbb), mRNA  
3. 1,552 bp linear mRNA  
Accession: NM\_130932.1 GI: 18858426  
GenBank FASTA Graphics Related Sequences

Danio rerio creatine kinase, mitochondrial 2 (sarcomeric) (ckmt2), nuclear gene encoding mitochondrial protein, mRNA  
4. mitochondrial protein, mRNA  
1,401 bp linear mRNA  
Accession: NM\_200614.1 GI: 41152441  
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\_17322.1 GI: 27545192  
GenBank FASTA Graphics Related Sequences

Send to: Filter your results:  
All (6) Bacteria (0) NSDC (GenBank) (0)  
mRNA (6) RefSeq (6) Manage Filters

Analyze these sequences Run BLAST

Find related data Database: Select Find items

Search details ("Danio rerio"[Organism] AND "creatin kinase"[Title] AND "refseq"[Filter])  
Search See more...

Recent activity

## Controlled Vocabularies

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

Nucleotide Nucleotide sponges Save search Limits Advanced

Search details "Porifera"[Organism] OR sponges[All Fields]

PubMed PubMed sponges RSS Save search Limits

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

Search details "porifera"[MeSH Terms] OR "porifera"[All Fields] OR "sponges"[All Fields]

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

Homepage

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

The screenshot shows the NCBI homepage. In the top right corner, there is a search bar with dropdown menus for 'Search' and 'Clear'. Below the search bar is a 'My NCBI' button. On the left, there's a sidebar with links like 'NCBI Home', 'Site Map (A-Z)', and 'Get Started'. The main content area features a 'Welcome to NCBI' banner, a 'Popular Resources' sidebar with 'BLAST' highlighted (indicated by a red arrow), and a 'PubMed Central' section.

## BLAST – Basic Local Alignment Search Tool

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

The screenshot shows the BLAST search tool homepage. It has a header with tabs for 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. Below the header, there's a news section about a new WGS BLAST page. The main area is titled 'BLAST Assembled RefSeq Genomes' and shows a list of organisms for search. To the right, there's a 'Tip of the Day' and a callout box containing the text: 'BLAST performs sequence similarity searches of query sequences vs sequence databases. We will cover this in detail in the next lecture.'

## SUMMARY

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

## HOMEWORK

- Complete the **initial course questionnaire**:  
<http://tinyurl.com/bioinf525-questions>
- Check out the **"Background Reading"** material on Ctools:  
<http://tinyurl.com/bioinf525-w16>
- Complete the **lecture 1.1 homework questions**:  
<http://tinyurl.com/bioinf525-quiz1>



## ADDITIONAL DATABASES OF NOTE (SLIDES FOR YOUR REFERENCE)

## NCBI Metadatabases

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

## PubMed

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

120

## PubMed results

Limits and Advanced search can be used to refine searches

NCBI Resources | How To | NCBI Sign In

Search: PubMed | RSS | Save search | Limits | Advanced search | Help | Search | Clear

Search Settings | Summary: 20 per page, Sorted by Recently Added

Unsorted field was ignored (orgn)

See the search details

Results: 1 to 20 of 2363

Distribution of retinol, acyl retinol, and retinol ester in the human hypothalamus. [abstract] [published online ahead of print] Meier DR, Chaitin J, Seelen DR, Zhou JT. Neurosci Lett. 2010 Dec 3; 480(3):407-10. PMID: 21130649 [PubMed - as supplied by publisher] [CrossRef]

Cortisol, restricted experience, hippocampus stress, and cognitive pathways and promotes binge-eating. [abstract] [published online ahead of print] Patekovich DF, Teeguarden SL, Hedin AD, Jensen CL, Bell TL. J Neurosci. 2010 Dec 1; 30(48):16399-407. PMID: 21106173 [PubMed - as supplied by publisher] [CrossRef]

Glycocorticoids Differentially Regulate the Expression of CRFR1 and CRFR2(α/β) in MH6 Insulinoma Cells and Rodent Islets. [abstract] [published online ahead of print] Huang MO, Pitrow AP, Matsumoto M, van der Meulen T, Park H, Vaughan JM, Lee S, Vale WW. Endocrinology. 2011 Jan; 152(1):15-20. Epub 2010 Nov 24. PMID: 21106173 [PubMed - as supplied by publisher] [CrossRef]

Benzodiazepine receptor agonism demonstrates the usefulness of Syrian hamsters as a model for anxiety testing. Evaluation of other classes of anxiolytics in comparison to clonazepam. [abstract] [published online ahead of print] Gannon RL, Lungquist E, Balista N, Hester I, Huntley C, Peacock A, Delargy P, Milan MJ. Behav Brain Res. 2010 Nov 20; [Epub ahead of print]. PMID: 21106173 [PubMed - as supplied by publisher] [CrossRef]

Related citations

215 free full-text articles in PubMed Central

Searches and Biological Evaluation of New Compounds for Anticancer Activity Chem Biol. 2010; 17(10):1111-1120.

Biological contributions to social behavior and cognition. J Neurosci. 2010; 30(48):16399-16407.

Pharmacogenomic approaches to asthma treatment. J Allergy Immunol. 2010; 184(4):1251-1258.

Find related data Database: Select

## Small molecule databases have been added at NCBI <http://pubchem.ncbi.nlm.nih.gov/>

Databases | Deposition | Services | Help | more

**PubChem**

BioAssay | Compound | Substance

Advanced search

Chemical structure search | BioActivity analysis

More ...

More than 2.5 million structures from the IBM BAO (Business Analytics and Optimization) strategic IP insight platform (SIIIP) are now available in PubChem. See more... and related news.

Write to Helpdesk | Disclaimer | Privacy Statement | Accessibility | Data Citation Guidelines  
National Center for Biotechnology Information  
NCI | NIH | HHS

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

NCBI | HomoloGene | Discover Homologs | Help | NCBI Sign In | Register

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Search: HomoloGene | for | Go | Clear

Limits | Preview/Index | History | Clipboard | Details

HomoloGene Homepage | Help | Contact | Build Procedure | FTP site

Genome Resources

Homologous genes

Mus musculus

Rattus norvegicus

Danio rerio

Homologous genes

Homo sapiens

Pan troglodytes

Canis familiaris

Bos taurus

Mus musculus

Rattus norvegicus

Gallus gallus

Danio rerio

Drosophila melanogaster

Anopheles gambiae

Candida albicans

Schizosaccharomyces pombe

Saccharomyces cerevisiae

Kluyveromyces lactis

Eremothecium gossypii

Magnaporthe grisea

Neurospora crassa

Azotobacter vinelandii

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

Species	Number of Genes	HomologGene Input	Grouped	HomologGene groups
Homo sapiens	19,943	18,361		18,431
Pan troglodytes	2,045	18,450		18,450
Canis familiaris	19,766	16,708		15,951
Bos taurus	22,049	19,180		16,224
Mus musculus	25,389	21,766		19,005
Rattus norvegicus	21,991	19,229		17,473
Gallus gallus	17,958	13,142		11,905
Danio rerio	26,690 <sup>*</sup>	21,084		14,067
Drosophila melanogaster	9,827 <sup>*</sup>	7,782		7,782
Anopheles gambiae	12,469 <sup>*</sup>	9,867		9,867
Candida albicans	20,000 <sup>*</sup>	9,778		4,910
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
Azotobacter vinelandii	***	***		***

What's New

HomoloGene release 65 includes updated annotations for the following species: Human sapiens (NCBI release 4.1), Drosophila melanogaster (NCB release 9.3) and Anopheles gambiae (NCBI release 9.1). Arabidopsis thaliana (NCBI release 9.1).

Related Resources

Entrez Genome

Collects 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>

NCBI | OMIM | Online Mendelian Inheritance in Man | Johns Hopkins University | My NCBI | Sign In | Register

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | OMIM

Search: OMIM | for | Go | Clear | Details

Entrez

OMIM

Search OMIM

Search Gene Map

Search Mapped Map

Help

FAQ

Numbering System

Symbols

How to Print

Print

Download

OMIM Facts

Statistics

Update Log

OMIM® - Online Mendelian Inheritance in Man

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.

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 Institute of Medicine and the William H. Welch Medical Library at Johns Hopkins. It was made generally available on the Internet in 1992.

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 | Books | Help | NCBI Sign In | Register

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy

Search: Books | for | Go | Clear

Introduction

Quick Start Guide

Help

Information for Authors and Publishers

What's New

FAQ

My NCBI

Privacy Policy

Limits | Preview/Index | History | Clipboard | Details

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

• cell cycle control • protein synthesis • protein evolution

New on the Bookshelf:

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

Human Herpesviruses: Biology, Therapy, and Immunopathogenesis (Arvin, Ann Campbell-Furneaux, Georges, Michael, Edward, Moore, Patrick S., Reznikoff, William, Rozenblatt, Michael, Strober, Wolfgang, Wadsworth, John, Weller, Robert, et al. Cambridge: Cambridge University Press; 2007)

Probe Reports from the Molecular Libraries Program (Bethesda (MD): National Center for Toxicology Research (NCTR); 2010)

GenBank (Cambridge (MA): Harvard Stem Cell Institute; 2008)

VIS-Evidence-Based Therapeutics Project Results (Washington (DC): Department of Veterans Affairs (VA); 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

126

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

127

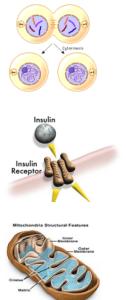
- DataSets** - (GDSxxx) are curated collections of selected Samples that are biologically and statistically comparable

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

30

## 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



129

## GO annotation in UniProt

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

## GO annotation in UniProt

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

The screenshot shows the UniProt QuickGO browser interface. At the top, it displays the UniProt entry for hemoglobin beta (HBB). Below this, the QuickGO search results for 'GO:0003037 heme binding' are shown. The results include a table with columns for ID, Name, Ontology, Definition, GONUTS, and Synonyms. The 'Definition' column describes the term as 'Interacting selectively and non-covalently with heme, any compound of iron complexed in a porphyrin (tetrapyrrole) ring.' The 'Synonyms' column lists 'hemoprotein' and 'heme binding'.

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

The screenshot shows the DAVID Bioinformatics Resources 6.7 homepage. It features a sidebar with links like 'Home', 'Start Analysis', 'Shortcut to DAVID Tools', 'Functional Annotation', 'Gene ID Conversion', 'Gene Functional Classification', and 'Gene ID Conversion'. The main content area includes a 'What's Important in DAVID?' section with a list of features, a '003 - 2013' news item, and a 'DAVID allows you to upload lists of genes and search for enriched GO and search for functionally related genes not in your list' section. At the bottom right, there is a small bar chart and the number '33'.

## Example output: enriched functions from GO

The screenshot shows the DAVID Functional Annotation Result Summary page. It displays a 'Functional Annotation Chart' for a current gene list. The chart lists various biological processes (e.g., 'regulation of transcription, promoter', 'positive regulation of transcription, DNA-dependent') along the y-axis, and their enrichment scores along the x-axis. A legend indicates that darker shades of blue represent higher enrichment. The chart includes a 'Download File' button at the top right.

134