



## Introduce Yourself!

Your preferred name,  
Place you identify with,  
Major area of study/research,  
Favorite joke (optional)!



## Today's Menu

<b>Course Logistics</b>	Website, screencasts, survey, ethics, assessment and grading.
<b>Learning Objectives</b>	What you need to learn to succeed in this course.
<b>Course Structure</b>	Major lecture topics and specific learning goals.
<b>Introduction to Bioinformatics</b>	<b>Introducing the <i>what, why and how</i> of bioinformatics?</b>
<b>Computer Setup</b>	Ensuring your laptop is all set for future sections of this course.

<http://thegrantlab.org/bggn213/>

The screenshot shows the homepage of the course website. At the top left is the UC San Diego logo and the course title "BGGN 213". Below the title is a brief description of the course: "A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD". A navigation bar on the left includes links for "Overview", "Lectures", "Computer Setup", "Learning Goals", "Assignments & Grading", "Ethics Code", and "Screen Cast Videos". Social media icons for Twitter, GitHub, and LinkedIn are at the bottom. The main content area features a section titled "Foundations of Bioinformatics (BGGN 213, Fall 2017)" with a DNA helix icon. It lists the "Course Director" (Prof. Barry J. Grant) and "Instructional Assistant" (Ileena Mitra). Below this are links for "Course Syllabus" (Fall 2017 PDF) and "Overview". The "Overview" section contains a brief description of the course's purpose and major topics.

<http://thegrantlab.org/bggn213/>

This screenshot is identical to the one above, but the "Learning Goals" link in the left sidebar is highlighted with a red box. The rest of the page content is the same, showing the course title, staff, syllabus, and overview sections.

What essential concepts and skills should  
YOU attain from this course?

The screenshot shows the "Learning Goals" page. The left sidebar has the "Learning Goals" link highlighted with a red box. The main content area starts with a heading "Learning Goals" and a sub-section "At the end of this course students will:". This is followed by a bulleted list of learning objectives. Below this is a summary statement: "In short, students will develop a solid foundational knowledge of bioinformatics and be able to evaluate new biomolecular and genomic information using existing bioinformatic tools and resources."

**At the end of this course students will:**

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources including major biomolecular and genomic databases, search and analysis tools, genome browsers, structure viewers, and select quality control and analysis tools to solve problems in the biological sciences.
- Be able to use the UNIX command line and the R environment to analyze bioinformatics data at scale.
- Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genomics, Transcriptomics and Structural bioinformatics.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genome informatics, Transcriptomics and Structural informatics.

In short, you will develop a solid foundational knowledge of **bioinformatics** and be able to evaluate new biomolecular and genomic information using **existing bioinformatic tools and resources**.

## Specific Learning Goals....

What I want you to know by course end!

**UC San Diego**

**BGGN 213**

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

**Overview**

**Lectures**

**Computer Setup**

**Learning Goals** (highlighted)

**Assignments & Grading**

**Ethics Code**

**Screen Cast Videos**

**Specific Learning Goals**

Teaching toward the specific learning goals below is expected to occupy 60%-70% of class time. The remaining course content is at the discretion of the instructor with student body input. This includes student selected topics for peer presentation as well one student selected guest lecture from an industry based genomic scientist.

All students who receive a passing grade should be able to:

	Lecture(s):
1 Appreciate and describe in general terms the role of computation in hypothesis-driven discovery processes within the life sciences.	1, 2, 20
2 Be able to query, search, compare and contrast the data contained in major bioinformatics databases and describe how these databases intersect (GenBank, GENE, UniProt, PFAM, OMIM, PDB, UCSC, ENSEMBLE).	2, 12, 13
3 Describe how nucleotide and protein sequence and structure data are represented (FASTA, FASTQ, GenBank, UniProt, PDB).	3, 10
4 Be able to describe how dynamic programming works for pairwise sequence alignment and appreciate the differences	..

## Course Structure

Derived from specific learning goals

**UC San Diego**

**BGGN 213**

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

**Overview**

**Lectures** (highlighted)

**Computer Setup**

**Learning Goals**

**Assignments & Grading**

**Ethics Code**

**Screen Cast Videos**

**Lectures**

All Lectures are Tu/Th 9:00-12:00 pm in Warren Lecture Hall 2015 (WLH 2015) ([Map](#)). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, pre-class video screen-casts and required reading material.

#	Date	Topics for Fall 2017
1	Th, 09/28	<b>Welcome to Foundations of Bioinformatics</b> Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student computer setup
2	Tu, 10/03	<b>Bioinformatics databases and key online resources</b> NCBI & EBI resources for the molecular domain of bioinformatics, Focus on GenBank, UniProt, Entrez and Gene Ontology. Hands on with BLAST, GenBank, OMIM, GENE, UniProt, Muscle, PFAM and PDB bioinformatics tools and databases

## Course Structure

Derived from specific learning goals

**UC San Diego**

**BGGN 213**

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

**Overview**

**Lectures** (highlighted)

**Computer Setup**

**Learning Goals**

**Assignments & Grading**

**Ethics Code**

**Screen Cast Videos**

**Lectures**

All Lectures are Tu/Th 9:00-12:00 pm in Warren Lecture Hall 2015 (WLH 2015) ([Map](#)). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, pre-class video screen-casts and required reading material.

#	Date	Topics for Fall 2017
1	Th, 09/28	<b>Welcome to Foundations of Bioinformatics</b> <span style="border: 2px solid red; padding: 2px;">(highlighted)</span> Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student computer setup
2	Tu, 10/03	<b>Bioinformatics databases and key online resources</b> NCBI & EBI resources for the molecular domain of bioinformatics, Focus on GenBank, UniProt, Entrez and Gene Ontology. Hands on with BLAST, GenBank, OMIM, GENE, UniProt, Muscle, PFAM and PDB bioinformatics tools and databases

# Class Details

Goals, Class material, Screencasts & **Homework**

The screenshot shows the course details page for BGGN 213. The main content area is titled "1: Welcome to Foundations of Bioinformatics". It includes sections for "Topics", "Goals", and "Material". The "Goals" section lists several bullet points related to course scope, expectations, logistics, and practical application. The "Material" section lists pre-class screen casts, lecture slides, handouts, and computer setup instructions. On the left sidebar, there are links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos.

# Homework

Goals, Class material, Screencasts & **Homework**

The screenshot shows the homework page for BGGN 213. The main content area is titled "Homework". It includes sections for "Questions", "Readings", and "Screen Casts". The "Questions" section is highlighted with a red box. The "Readings" section lists PDFs and other resources. The "Screen Casts" section features a video thumbnail of Barry Grant introducing the course. On the left sidebar, there are links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos.

# Homework

Goals, Class material, Screencasts & **Homework**

The screenshot shows the homework page for BGGN 213. The main content area is titled "Homework". It includes sections for "Questions" (highlighted with a red box), "Readings", and "Screen Casts". The "Questions" section lists several bullet points. The "Screen Casts" section features a video thumbnail of Barry Grant. On the left sidebar, there are links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos.

# Homework

Goals, Class material, Screencasts & **Homework**

The screenshot shows a Google Forms document titled "BGGN213 Lecture 1 Homework (F17)". The form asks for the user's UCSD username/email address and a response. It also includes a question about the most frequently used operating system for bioinformatics tool development, with options for Windows, iOS, Unix, and Perl. The background of the form is purple.

# Homework

Goals, Class material, Screencasts & **Homework**

BGGN213 Lecture 1 Homework

Please answer the following questions

\* Required

Your name/email address \*

Enter your UCSD email address before the '@ucsd.edu' part

Your answer

Which of the following operating systems is most frequently used for bioinformatics tool development

Windows

iOS

Unix

Perl

# Today's Menu

## Course Logistics

Website, screencasts, survey, ethics, assessment and grading.

## Learning Objectives

What you need to learn to succeed in this course.

## Course Structure

Major lecture topics and specific learning goals.

## Introduction to Bioinformatics

Introducing the *what, why and how* of bioinformatics?

## Computer Setup

Ensuring your laptop is all set for future sections of this course.

# OUTLINE

## Overview of bioinformatics

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

## Online 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

- How-to productively navigate major databases.

## 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

## 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!**

## 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>)

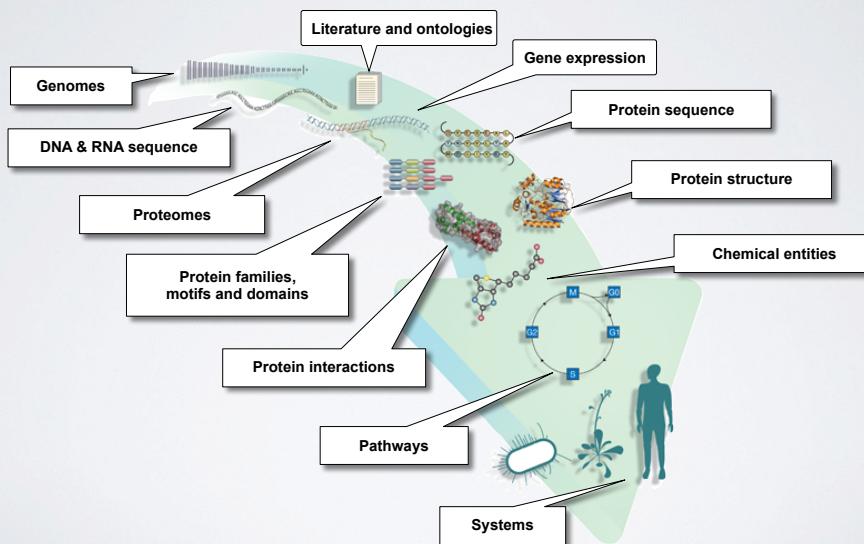
## MORE DEFINITIONS

- “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.

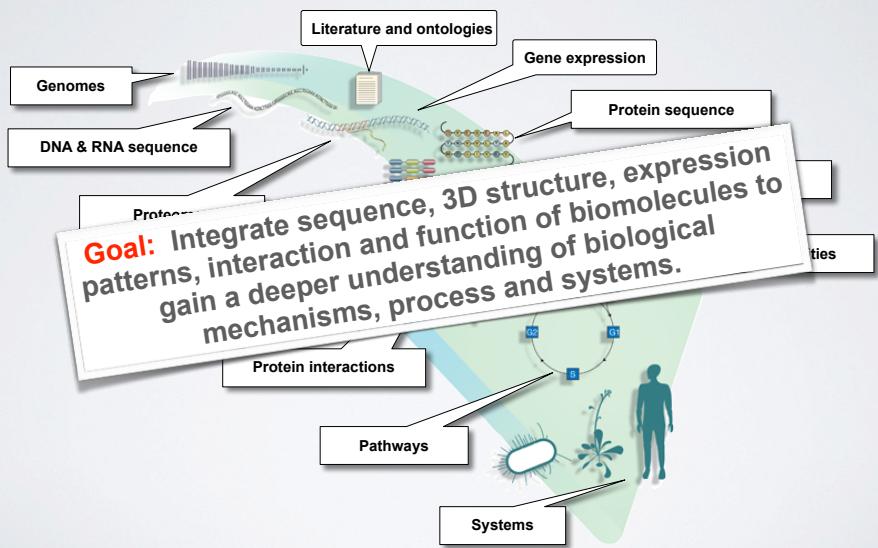
- “Bioinformatics is research, development, or application of **computational approaches** for expanding the use of **biological, medical, behavioral or health data**, including those to **acquire, store, organize and analyze** such data.”  
National Institutes of Health (NIH) (<http://tinyurl.com/l3gxr6b>)

*Key Point: Bioinformatics is Computer Aided Biology*

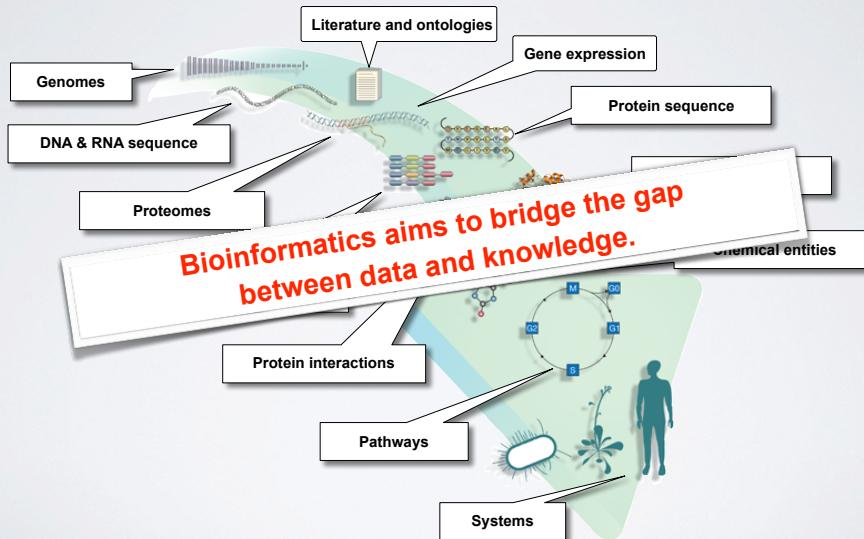
## Major types of Bioinformatics Data



## 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

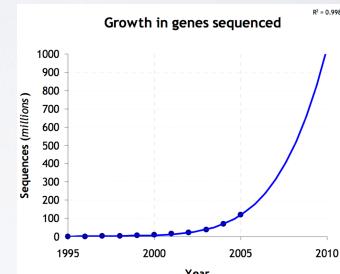
- DNA sequence determines *protein* sequence.
- *Protein* sequence determines *protein* structure.
- *Protein* structure determines *protein* function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular *function in space and time*.

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**

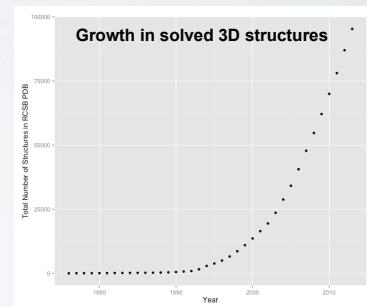


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

## Why do we need Bioinformatics?

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

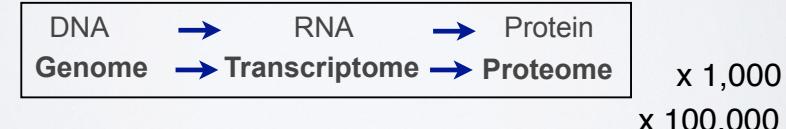
- Bioinformatics provides methods for the efficient:
  - **storage**
  - **annotation**
  - **search and retrieval**
  - **data integration**
  - **data mining and analysis**



E.G. data from sequencing, structural genomics, 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?

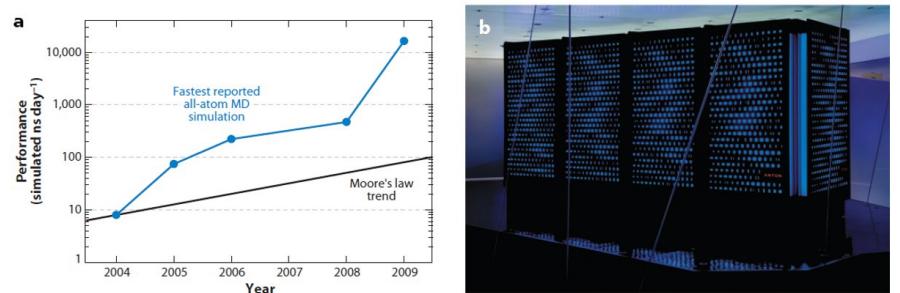
## Pre-packaged tools and databases

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

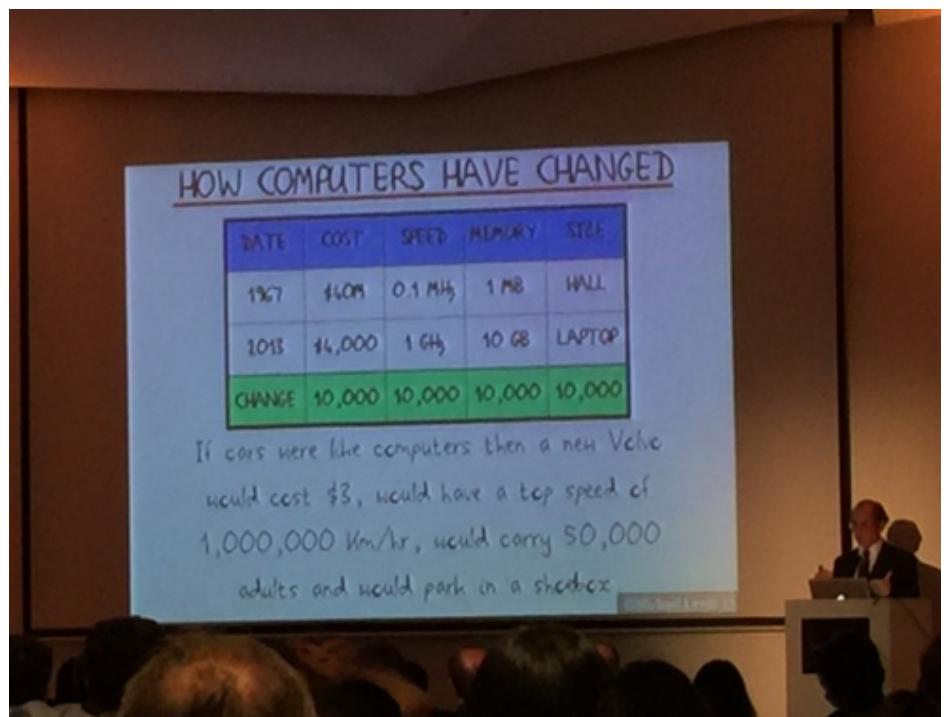
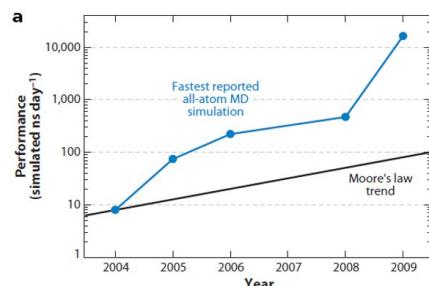
## Tool development

- ▶ 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...

## SIDE-NOTE: SUPERCOMPUTERS AND GPUs



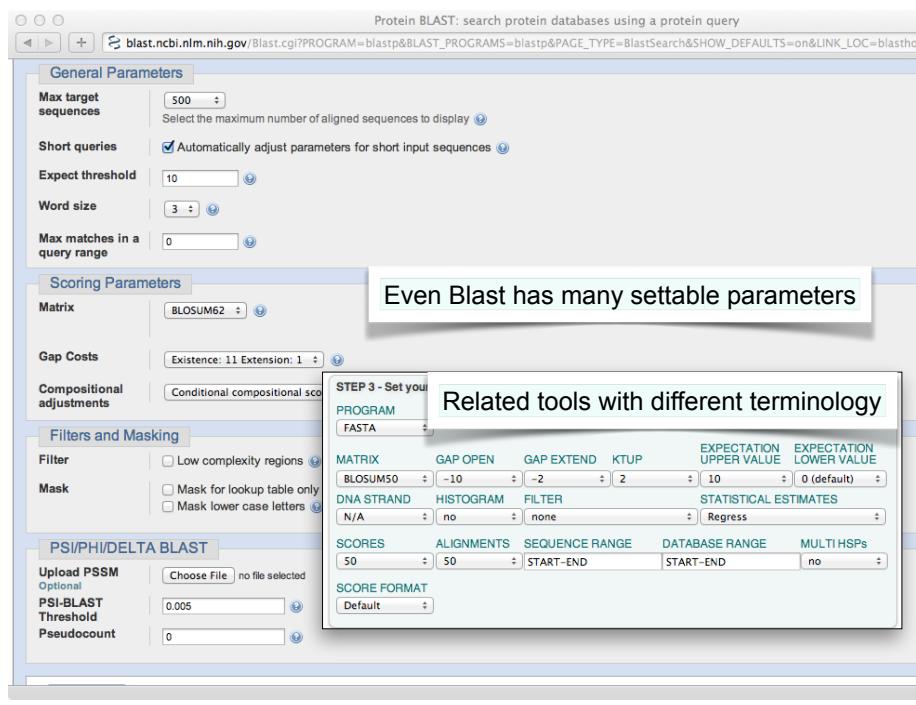
## SIDE-NOTE: SUPERCOMPUTERS AND GPUs



## 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'



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	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					

## 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!

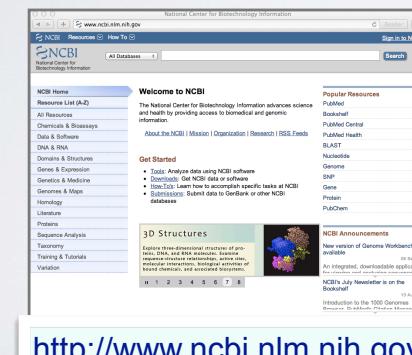
Most are developed independently

Notable exceptions are found at the:

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

## Key Online Bioinformatics Resources: NCBI & EBI

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



Welcome to NCBI

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

Get Started

- Tools: Analyze data using NCBI software
- Databases: Get NCBI data or software
- HowTo's: Learn how to accomplish specific tasks at NCBI
- Resources: Submit data to BioProject or other NCBI databases

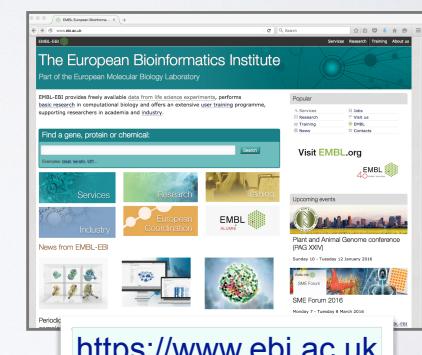
3D Structures

New version of Genome Workbench available

NCBI Announcements

Introduction to the 1000 Genomes Project

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



The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

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

Find a gene, protein or chemical:

Services

Research

European Coordination

EMBL-EBI

News from EMBL-EBI

Upcoming events

Visit EMBL.org

Periodicals

EMBL Forum

EMBL Annual Conference (EMBO XXVII)

Sunday 01 - Tuesday 22 January 2013

Monday 7 - Tuesday 8 March 2013

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>

The screenshot shows the NCBI homepage with a map of Bethesda, MD. The main menu includes links to All Databases, NCBI Home, Resource List (A-Z), Popular Resources (PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem), Welcome to NCBI, Get Started, and various research sections like 3D Structures, Genes & Expression, and Homology.

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

This screenshot is similar to the one above but includes red arrows pointing to specific links: 'PubMed' in the Popular Resources section, 'BLAST' in the Get Started section, and 'SNP' in the 'Gene' section under 'Popular Resources'.

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

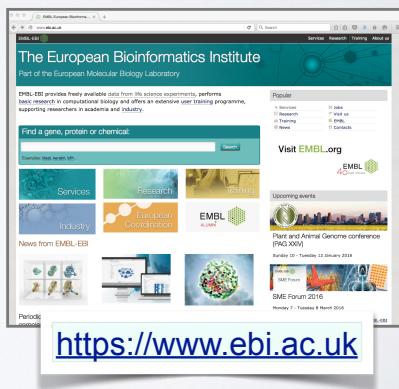
A large central banner on the page reads: "Notable NCBI databases include: GenBank, RefSeq, PubMed, dbSNP and the search tools ENTREZ and BLAST". Below the banner, the page layout is identical to the other screenshots, showing the NCBI logo, menu, and various database links.

# Key Online Bioinformatics Resources: NCBI & EBI

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

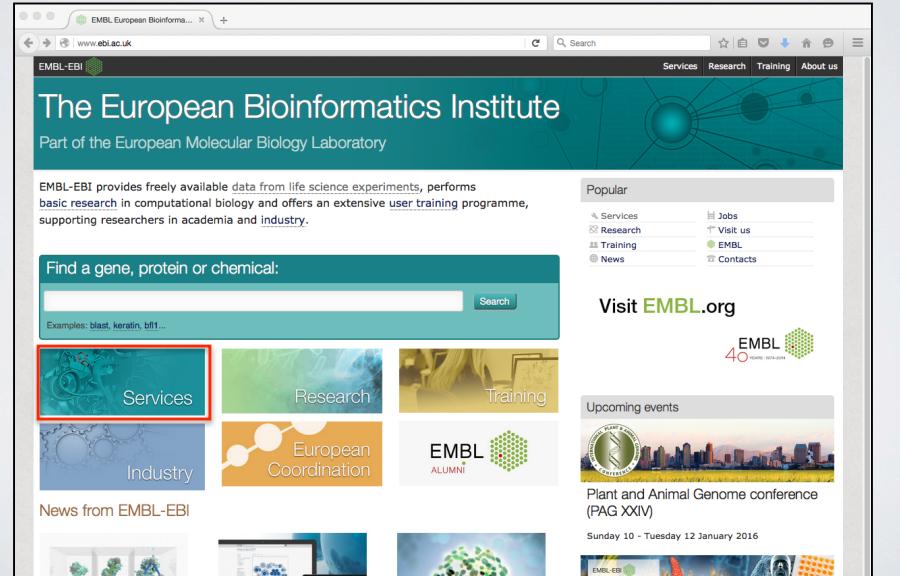


<http://www.ncbi.nlm.nih.gov>



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

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

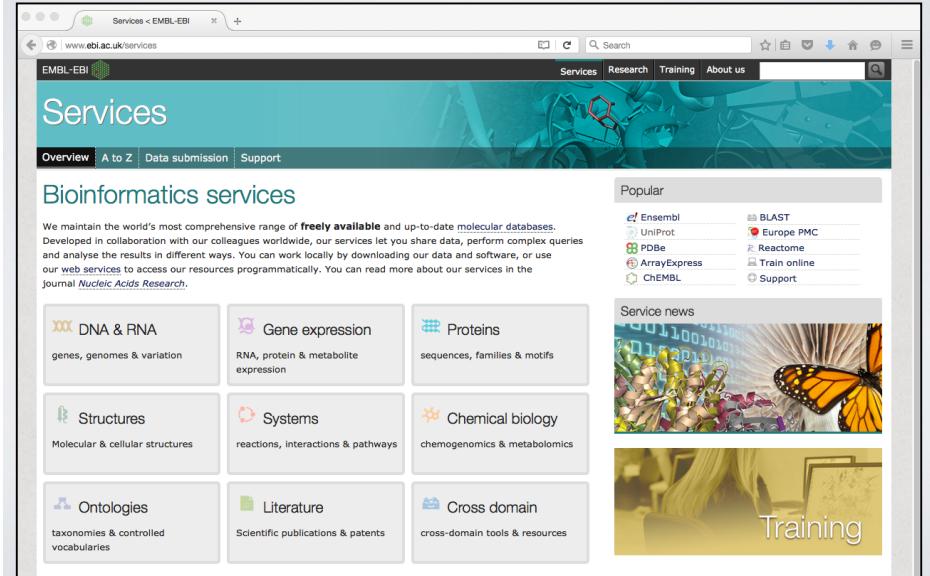


# European Bioinformatics Institute (EBI)

- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
  - ▶ providing freely available **data and bioinformatics services**
  - ▶ and providing advanced **bioinformatics training**
- We will briefly cover several EBI databases and tools that have advantages over those offered at NCBI



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



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

The screenshot shows the EMBL-EBI Services website. In the 'Popular' section, the 'Proteins' link is highlighted with a red box. The 'Proteins' link leads to a page titled 'Proteins: sequences, families & motifs'.

<https://www.ebi.ac.uk>  
The EBI makes available a wider variety of **online tools** than NCBI

The screenshot shows the 'Proteins' page on the EBI website. Under 'Popular services', there is a list of tools including UniProt, InterPro, PRIDE, Pfam, Clustal Omega, HMMER, and InterProScan 5. A 'Quick links' sidebar on the right provides links to popular services, all services, and project websites in this category.

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

The screenshot shows the EMBL European Bioinformatics Institute homepage. The 'Training' section is highlighted with a red box. The 'Training' link leads to a page titled 'Train online'.

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

The screenshot shows a 'Using sequence similarity searching tools at EMBL-EBI: webinar' page. It features a video player showing a presentation by Andrew Cowley. The video title is 'Using sequence similarity search tools at EMBL-EBI - Finding homologous sequences with BLAST, FASTA, PSI-Search etc.' The page includes a 'Course content' section and a 'Popular' sidebar with links to other training resources.

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

Notable EBI databases include:  
ENA, UniProt, Ensembl  
and the tools FASTA, BLAST, InterProScan,  
MUSCLE, DALI, HMMER

## 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, 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, MHCPEPS, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSSub, 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, TelIDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..... !!!

## Next Class...

# MAJOR BIOINFORMATICS DATABASES AND ASSOCIATED ONLINE TOOLS

## 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, 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, MHCPEPS, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSSub, 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, TelIDB, 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.

## 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.

## 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/>

The screenshot shows the NCBI homepage with a search bar containing the term "ras". To the right of the search bar, a red arrow points diagonally upwards and to the left towards the text "Hands on demo (or see following slides)". The page includes a sidebar with links to various NCBI resources like PubMed, Bookshelf, and BLAST, and a main content area with sections for "Welcome to NCBI", "Get Started", "Genotypes and Phenotypes", and "NCBI Announcements".

**ras - GQuery: Global Cross** | www.ncbi.nlm.nih.gov/gquery/?term=ras

NCBI Resources How To Sign In to NCBI

### Search NCBI databases

ras Search

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
<b>Health</b>		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		

62

**ras - Gene - NCBI** | www.ncbi.nlm.nih.gov/gene/?term=ras

NCBI Resources How To Sign In to NCBI

### Gene

Gene ras Search Save search Advanced

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

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

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

Results: 1 to 20 of 85633

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

Name/Gene ID	Description	Location	Aliases
ras	ID: 19412 resistance to audiogenic seizures [ <i>Mus musculus</i> (house mouse)]		asr
ras	ID: 43873 raspberry [ <i>Drosophila melanogaster</i> (fruit fly)]	Chromosome X, NC_004354.4 (10744502..10749097)	Dmel_CG1799, CG11485, CG1799, Dmel_CG1799, EP(X)1093

Find related data Database: Select Find items

Search details ras[All Fields] AND alive[property]

63

**(ras) AND "Homo sapiens"** | www.ncbi.nlm.nih.gov/gene

NCBI Resources How To Sign In to NCBI

### Gene

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

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

Filters: Manage Filters

Find related data Database: Select Find items

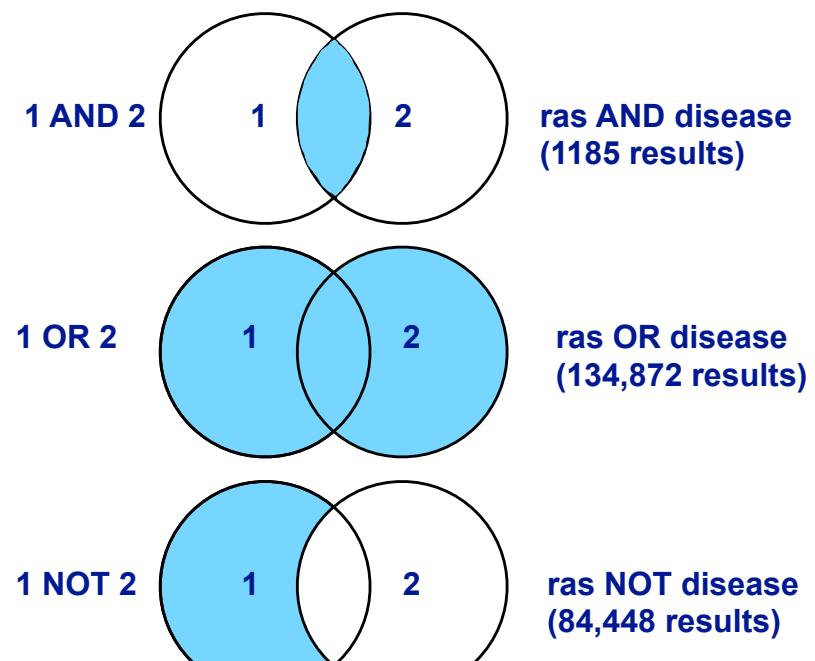
Results: 1 to 20 of 1126 << First < Prev Page 1 of 57 Next > Last >

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 (11470446..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS_N-ras, NCMS1, NS6, NRAS
KRAS	ID: 3845 Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, NS2, RAS2

Search See more... Recent activity Turn Off Clear

64



Search results for "ras AND "Homo sapiens"" (txid9606) showing 1 to 20 of 1126 results.

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (11470446..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
<b>KRAS ID: 3845</b>	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, RAS2

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

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

**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
- General protein information
- NCBI Reference Sequences (RefSeq)

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

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

**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

**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 regions, transcripts, and products**

**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?

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:OTTHUMG000017193  
 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-  
 70

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	

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



## GO: Gene Ontology

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

UniProt-GOA Examples: GO:0006915, tropomyosin, P06727 Search

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

72

## Why do we need Ontologies?

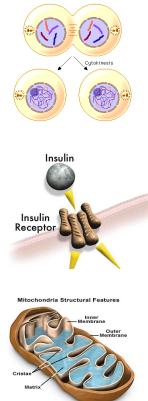
- Annotation is essential for capturing the understanding and knowledge associated with a sequence or other molecular entity
- Annotation is traditionally recorded as “free text”, which is easy to read by humans, but has a number of disadvantages, including:
  - Difficult for computers to parse
  - Quality 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

74

The screenshot shows the 'Gene Ontology' section of the NCBI Gene page for KRAS. A red box highlights the 'Gene Ontology' header and the 'Provided by GOA' link. A blue box contains the text: 'The 'Gene Ontology' or GO is actually maintained by the EBI so lets switch or link over to UniProt also from the EBI.' A red arrow points down to the UniProt link at the bottom of the page.

The screenshot shows the NCBI Gene page for KRAS. A blue box contains the text: 'UniProt will detail much more information for protein coding genes such as this one'. A red box highlights the 'UniProtKB/Swiss-Prot:P01116' link under the 'Additional links' section. A red arrow points down to the UniProt link at the bottom of the page.

The screenshot shows the UniProt protein details page for P01116 - RASK\_HUMAN. A blue box contains the text: 'UniProt will detail much more information for protein coding genes'. A red box highlights the 'Reviewed - Experimental evidence at protein level' status indicator. A red arrow points down to the UniProt link at the bottom of the page.

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

Feature key	Position(s)	Length	Description	Graphical view	Identifier	Actions
Nucleotide binding <sup>i</sup>	10 - 18	9	GTP 2 Publications	graphical icon	VAR_034601	
Nucleotide binding <sup>i</sup>	29 - 35	7	GTP 2 Publications	graphical icon		
Nucleotide binding <sup>i</sup>	59 - 60	2	GTP 2 Publications	graphical icon		

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

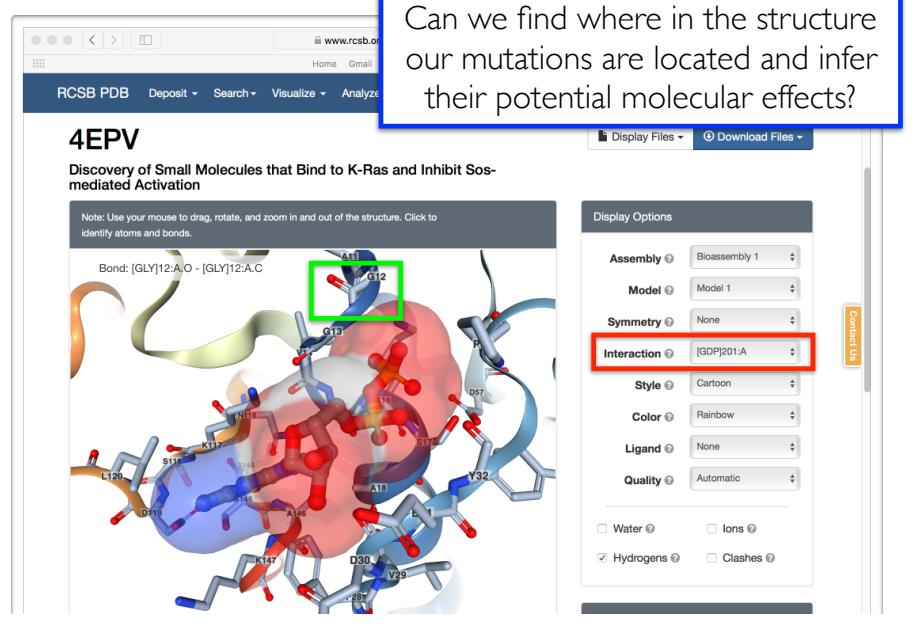
Feature key	Position(s)	Length	Description	Graphical view	Identifier	Actions
Natural variant <sup>i</sup>	10 - 10	1	1 G → GG in one individual with AML; expression in 3T3 cell causes cellular transformation; expression in COS cells activates the Ras-MAPK signaling pathway; lower GTPase activity; faster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

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

Select the link destinations:	Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
<input checked="" type="radio"/> RCSB PDB <sup>ii</sup>	1D8D	X-ray	2.00	P	178-188	[>]
<input type="radio"/> PDB <sup>ii</sup>	1DBE	X-ray	3.00	P	178-188	[>]
<input checked="" type="radio"/> RCSB PDB <sup>ii</sup>	1KZO	X-ray	2.20	C	169-173	[>]
<input type="radio"/> PDB <sup>ii</sup>	1KZP	X-ray	2.10	C	169-173	[>]
<input checked="" type="radio"/> RCSB PDB <sup>ii</sup>	3GFT	X-ray	2.27	A/B/C/D/E/F	1-164	[>]
<input type="radio"/> PDB <sup>ii</sup>	4DSN	X-ray	2.03	A	2-164	[>]
<input checked="" type="radio"/> RCSB PDB <sup>ii</sup>	4DSO	X-ray	1.85	A	2-164	[>]
<input type="radio"/> PDB <sup>ii</sup>	4EPR	X-ray	2.00	A	1-164	[>]
<input checked="" type="radio"/> RCSB PDB <sup>ii</sup>	4EPV	X-ray	2.00	A	1-164	[>]
<input type="radio"/> PDB <sup>ii</sup>	4EPW	X-ray	1.35	A	1-164	[>]
<input checked="" type="radio"/> RCSB PDB <sup>ii</sup>	4EPX	X-ray	1.70	A	1-164	[>]
<input type="radio"/> PDB <sup>ii</sup>	4EPY	X-ray	1.76	A	1-164	[>]
<input checked="" type="radio"/> RCSB PDB <sup>ii</sup>	4L8G	X-ray	1.80	A	1-164	[>]
<input type="radio"/> PDB <sup>ii</sup>	4LDJ	X-ray	1.52	A	1-164	[>]
<input checked="" type="radio"/> RCSB PDB <sup>ii</sup>	4LPK	X-ray	1.15	A/B	1-169	[>]

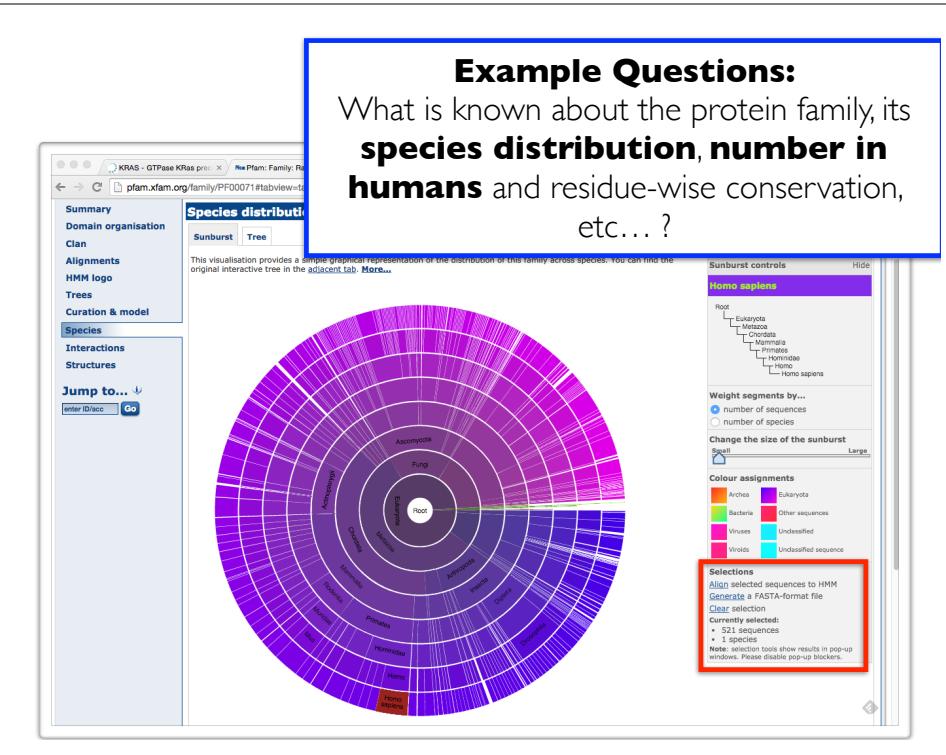
**Open link in a new tab!**

**Lets view the 3D structure:**  
Can we find where in the structure our mutations are located and infer their potential molecular effects?



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

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



## Example Questions:

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

Pfam: Pfam alignment viewer  
pfam.xfam.org/family/PF00071#tabview=tab4

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

**Alignment for selected sequences**

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

**Jump to...** enter ID/acc Go

There are 18 pages in this alignment. Show page 1

Download this alignment.

Close window

Sunburst controls Hide

Home analysis

Root

- Eukaryota
- Metazoa
- Dinoflagellates
- Mammalia
- Primates
- Homo

Weight segments by...

number of sequences (radio button)

number of species (radio button)

Change the size of the sunburst

Small Large

Colour assignments

- Archea
- Eukaryota
- Bacteria
- Other sequences
- Viruses
- Undescribed
- Virions
- Undescribed sequence

**HMM logo**

Selections

Align selected sequences to HMM

Generate a FASTA-formatted file

Clear selection

Currently selected:

- 1 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... ?

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Family: Ras (PF00071)

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

Jump to... enter ID/acc Go

HMM logo

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

Comments or questions on the site? Send a mail to [pfa-help@ebi.ac.uk](mailto:pfa-help@ebi.ac.uk). European Molecular Biology Laboratory

HHMI Janelia Farm research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

**Family: Kinesin (PF00225)**

>Loading page components (1 remaining)...

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

**Interactions**

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

Tubulin Tubulin\_C Kinesin Tubulin Kinesin

126 architectures 4150 sequences 6 interactions 248 species 114 structures

**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 PDB 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
ABBK1_GIALA	11 - 335	2vvg	A	11 - 335	<a href="#">Jmol AstexViewer SPICE</a>
			B	11 - 335	<a href="#">Jmol AstexViewer SPICE</a>
CENPE_HUMAN	12 - 329	1t5c	A	12 - 329	<a href="#">Jmol AstexViewer SPICE</a>
			B	12 - 329	<a href="#">Jmol AstexViewer SPICE</a>
KAR3_YEAST	392 - 723	1f9t	A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			1f9u	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			1f9v	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			1f9w	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			3kar	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
KI13B_HUMAN	11 - 352	3gbi	A	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
			B	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
			C	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
			A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>

Questions or comments: [pfa@janelia.hhmi.org](mailto:pfa@janelia.hhmi.org)  
Howard Hughes Medical Institute

HHMI Janelia Farm research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

**Family: Kinesin (PF00225)**

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

**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 PDB 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
ABBK1_GIALA	11 - 335	2vvg	A	11 - 335	<a href="#">Jmol AstexViewer SPICE</a>
			B	11 - 335	<a href="#">Jmol AstexViewer SPICE</a>
CENPE_HUMAN	12 - 329	1t5c	A	12 - 329	<a href="#">Jmol AstexViewer SPICE</a>
			B	12 - 329	<a href="#">Jmol AstexViewer SPICE</a>
KAR3_YEAST	392 - 723	1f9t	A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			1f9u	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			1f9v	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			1f9w	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			3kar	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
			A	392 - 723	<a href="#">Jmol AstexViewer SPICE</a>
KI13B_HUMAN	11 - 352	3gbi	A	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
			B	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
			C	11 - 352	<a href="#">Jmol AstexViewer SPICE</a>
			A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			B	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>
			A	24 - 359	<a href="#">Jmol AstexViewer SPICE</a>

Your turn:  
What can you find out about "eg5"

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

## 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 Gene, UniProt, PDB databases as well as a number of ‘boutique’ databases including PFAM and OMIM.
- Introduced the notion of *controlled vocabularies* and *ontologies*.

## HOMEWORK

[https://bioboot.github.io/bggn213\\_f17/lectures/#1](https://bioboot.github.io/bggn213_f17/lectures/#1)

- Complete the **initial course questionnaire**:
- Check out the “**Background Reading**” material online:
- Complete the **lecture 1 homework questions**:

THANK YOU