



BGGN 213
Foundations of Bioinformatics
Barry Grant
UC San Diego
<http://thegrantlab.org/bggcn213>



HELLO
my name is
BARRY
bjgrant@ucsd.edu

HELLO
HER name is
ILEENA
ileenamitra@eng.ucsd.edu

Introduce Yourself!

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

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 <i>what, why and how</i> of bioinformatics?
Computer Setup	Ensuring your laptop is all set for future sections of this course.

http://thegrantlab.org/bggn213/

UC San Diego

Foundations of Bioinformatics (BGGN 213, Fall 2017)

BGGN 213

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

Course Director
Prof. Barry J. Grant (Email: bjgrant@ucsd.edu)

Instructional Assistant
Ileena Mitra (Email: ileenamitra@eng.ucsd.edu)

Course Syllabus
[Fall 2017 \(PDF\)](#)

Overview

Bioinformatics is driving the collection and analysis of big data in the biosciences. This course is designed for bioscience graduate students and provides a hands-on introduction to the computer-based analysis of genomic and biomolecular data.

Major topics include:

- Genomic and biomolecular bioinformatic resources,

Navigation: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, Screen Cast Videos

http://thegrantlab.org/bggn213/

UC San Diego

Foundations of Bioinformatics (BGGN 213, Fall 2017)

BGGN 213

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

Course Director
Prof. Barry J. Grant (Email: bjgrant@ucsd.edu)

Instructional Assistant
Ileena Mitra (Email: ileenamitra@eng.ucsd.edu)

Course Syllabus
[Fall 2017 \(PDF\)](#)

Overview

Bioinformatics is driving the collection and analysis of big data in the biosciences. This course is designed for bioscience graduate students and provides a hands-on introduction to the computer-based analysis of genomic and biomolecular data.

Major topics include:

- Genomic and biomolecular bioinformatic resources,

Navigation: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, Screen Cast Videos

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

UC San Diego

Learning Goals

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.

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.

Navigation: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, Screen Cast Videos

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 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.
- 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!

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
	Be able to describe how dynamic programming works for pairwise sequence alignment and appreciate the differences	

Course Structure

Derived from specific learning goals

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	Welcome to Foundations of Bioinformatics 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	Bioinformatics databases and key online resources 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

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	Welcome to Foundations of Bioinformatics 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	Bioinformatics databases and key online resources 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 GitHub page for BGGN 213. The main heading is "1: Welcome to Foundations of Bioinformatics". Under "Topics", it lists course introduction, learning goals, biology as an information science, history of bioinformatics, types of data, application areas, and course segments. Under "Goals", it lists understanding course scope, the increasing necessity for computation in modern life sciences research, and getting introduced to bioinformatics. Under "Material", it lists pre-class screen cast, lecture slides, a handout, and computer setup instructions. A sidebar on the left contains navigation links: 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 section of the BGGN 213 page. It lists "Questions" and "Readings" which include PDF1, PDF2, and a New York Times article. Below this is a "Screen Casts" section featuring a video thumbnail titled "Welcome to 'Foundations of Bioinformatics' (BGGN-21...)" with the BGGN 213 logo and Barry Grant's name.

Homework

Goals, Class material, Screencasts & Homework

This screenshot is identical to the previous one, but with a red box highlighting the "Questions" link in the homework list.

Homework

Goals, Class material, Screencasts & Homework

The screenshot shows a Google Forms page titled "BGGN213 Lecture 1 Homework (F17)". It asks for the user's UCSD username/email address and then poses a multiple-choice question: "Which of the following operating systems is most frequently used for bioinformatics tool development?" with options for Windows, iOS, Unix, and Perl.

Homework

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

BGGN213 Lecture 1 Homework

Please answer the following questions

* Required

Name/email address *

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

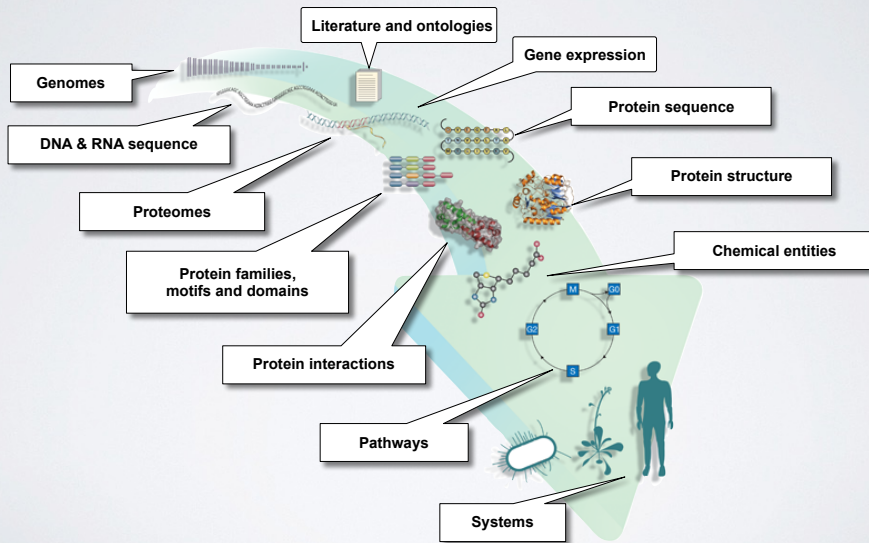
- ▶ “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “informatics” techniques (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**.
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

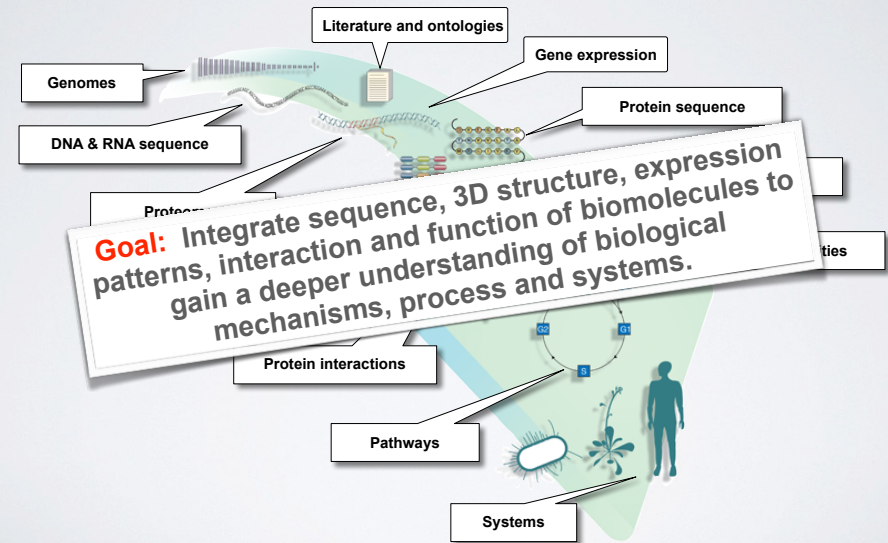
- ▶ “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “informatics” techniques (derived from disciplines such as applied mathematics, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**.
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

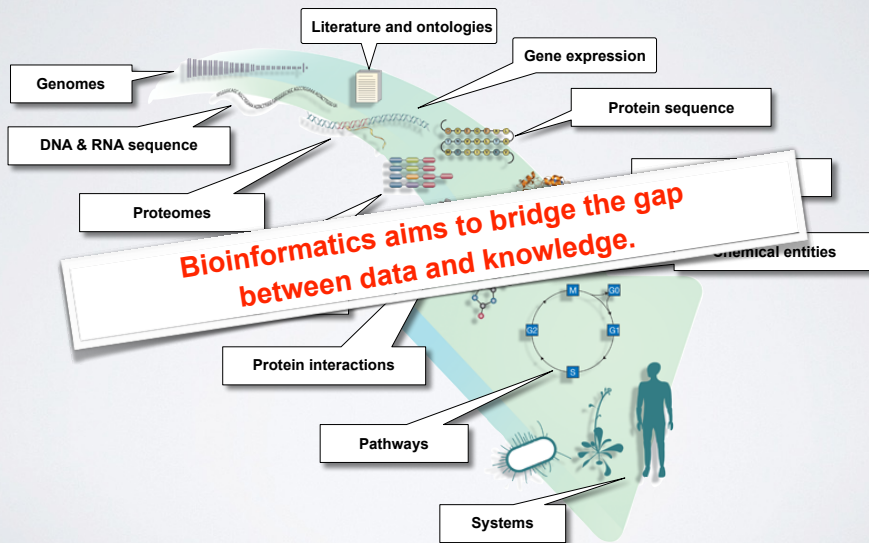
Major types of Bioinformatics Data



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

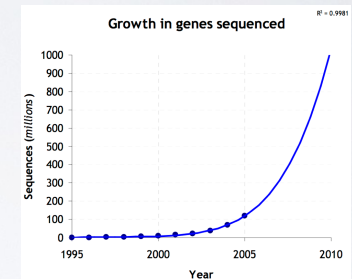
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.

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

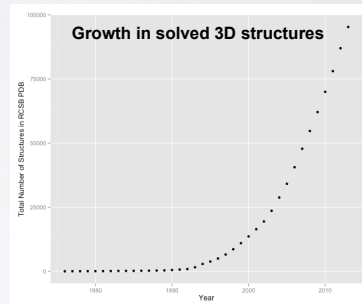


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

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?

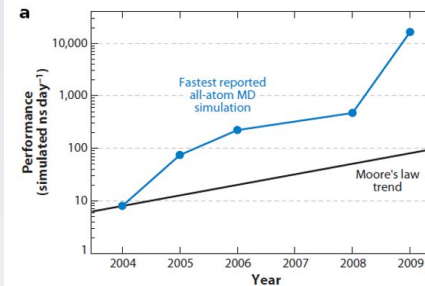
Pre-packaged tools and databases

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

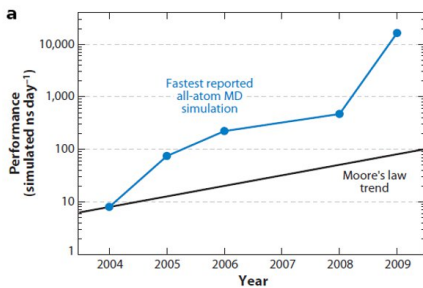
Tool development

- ▶ 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



HOW COMPUTERS HAVE CHANGED

DATE	COST	SPEED	MEMORY	SIZE
1967	\$409	0.1 MHz	1 MB	WALL
2013	\$4,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

If cars were like computers then a new Velve would cost \$3, would have a top speed of 1,000,000 km/hr, would carry 50,000 adults and would park in a shedbox

Skepticism & Bioinformatics

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

- Do they make sense?
- Is it what we expected?
- Do we have adequate controls, and how did they come out?
- Modeling is modeling, but biology is different...
What does this model actually contribute?
- Avoid the miss-use of 'black boxes'

Common problems with Bioinformatics

Confusing multitude of tools available

- Each with many options and settable parameters

Most tools and databases are written by and for nerds

- Same is true of documentation - if any exists!

Most are developed independently

Notable exceptions are found at the:

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

Protein BLAST: search protein databases using a protein query

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

General Parameters

- Max target sequences: 500
- 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)
- PSI-BLAST Threshold: 0.005
- Pseudocount: 0

STEP 3 - Set your PROGRAM

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

Related tools with different terminology

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

SCORE FORMAT

Default

Even Blast has many settable parameters

Key Online Bioinformatics Resources: NCBI & EBI

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

National Center for Biotechnology Information

NCBI Home

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
- **Database:** Get NCBI data for software
- **Site (S):** Learn how to accomplish specific tasks at NCBI
- **Publications:** Search data in GenBank or other NCBI databases

3D Structures

NCBI Announcements

Popular Resources

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

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

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

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

Find a gene, protein or chemical

Visit EMBL.org

Services

- Genomics
- Proteomics
- Metagenomics
- Plant and Animal Genome conference (PAG 2016)
- SME Forum 2016

<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>

A screenshot of the NCBI homepage in a web browser. The page features a navigation menu on the left with categories like 'NCBI Home', 'Resource List (A-Z)', and 'All Resources'. The main content area includes a 'Welcome to NCBI' message, a 'Get Started' section with links to 'Tools', 'Downloads', 'How-To's', and 'Submissions', and a '3D Structures' section. On the right, there are 'Popular Resources' and 'NCBI Announcements'.

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

A screenshot of the NCBI homepage with a white overlay box titled 'Popular Resources'. The overlay lists several key resources: PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Red arrows point to PubMed, BLAST, and SNP. A red bracket groups Nucleotide, Genome, SNP, Gene, and Protein.

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

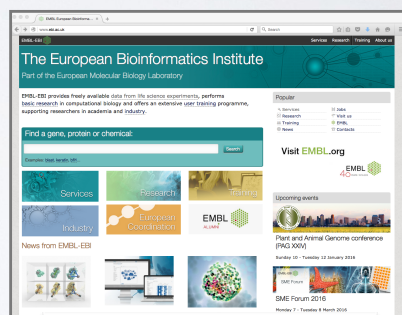
A screenshot of the NCBI homepage with a white text box overlaid on the right side. The text box contains the following text: 'Notable NCBI databases include: **GenBank**, **RefSeq**, **PubMed**, dbSNP and the search tools **ENTREZ** and **BLAST**'. The background shows the same NCBI homepage as in the previous images.

Key Online Bioinformatics Resources: NCBI & EBI

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



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



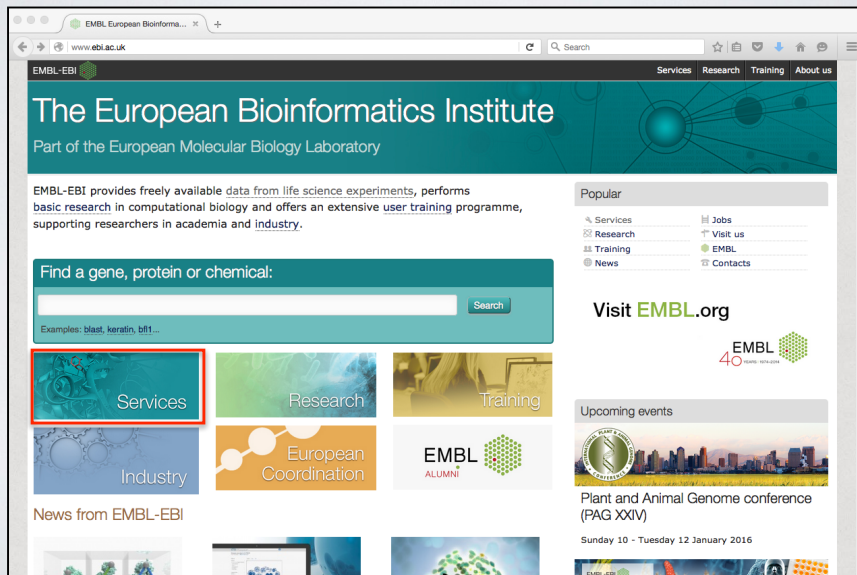
<https://www.ebi.ac.uk>

European Bioinformatics Institute (EBI)

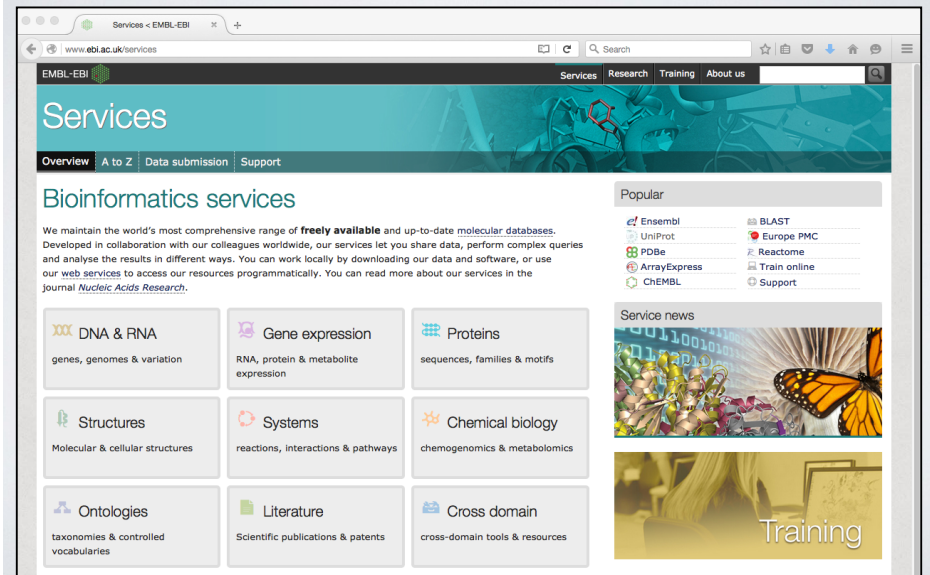
- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
 - ▶ providing freely available **data and bioinformatics services**
 - ▶ and providing advanced **bioinformatics training**
- We will briefly cover several EBI databases and tools that have advantages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



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



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



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

The screenshot shows the EBI Services page. The main heading is 'Services' with sub-links for Overview, A to Z, Data submission, and Support. Below this is a section for 'Bioinformatics services' with a grid of categories: DNA & RNA, Gene expression, Proteins (highlighted with a red box), Structures, Systems, Chemical biology, Ontologies, Literature, and Cross domain. A 'Popular' sidebar lists databases: Ensembl, UniProt (highlighted with a red box), PDB, ArrayExpress, and ChEMBL. A 'Training' banner is visible at the bottom right.

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

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

The screenshot shows the EBI Proteins page. The main heading is 'Proteins' with a sub-heading 'Popular services'. A list of services is provided: UniProt: The Universal Protein Resource, InterPro, PRIDE: The Proteomics Identifications Database, Pfam, Clustal Omega, HMMER - protein homology search, and InterProScan 5. A 'Quick links' sidebar on the right lists: Popular services in this category, All services in this category, and Project websites in this category.

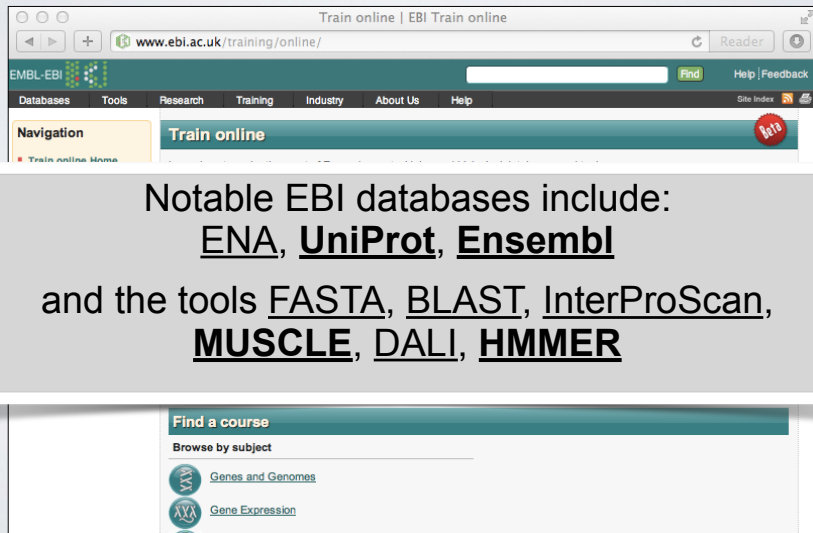
The EBI also provides a growing selection of **online tutorials** on EBI databases and tools

The screenshot shows the EBI website home page. The main heading is 'The European Bioinformatics Institute' with the sub-heading 'Part of the European Molecular Biology Laboratory'. A search bar is present with the text 'Find a gene, protein or chemical:'. Below the search bar are several navigation tiles: Services, Research, Training (highlighted with a red box), Industry, European Coordination, and EMBL ALUMNI. A 'Popular' sidebar lists: Services, Research, Training, EMBL, News, Jobs, Visit us, and Contacts. An 'Upcoming events' section is also visible.

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

The screenshot shows the EBI online training page. The main heading is 'Train online'. Below this is a section for 'Using sequence similarity searching tools at EMBL-EBI: webinar'. A video player is embedded, showing a webinar by Andrew Cowley. The video content includes: 'Using sequence similarity searching tools at EMBL-EBI: webinar', 'Finding homologous sequences with BLAST, FASTA, PSI-Search etc.', and 'Andrew Cowley'. A 'Popular' sidebar lists: Train online, Find us, Funding, Open days and career days, Conference exhibitions, EMBL courses and events, Genome campus events, and Science for schools.

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**

Next Class...

MAJOR BIOINFORMATICS DATABASES AND ASSOCIATED ONLINE TOOLS

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BiImage, 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, DPIinteract, ECDC, ECGC, ECO2DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSUB, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, Subtilist, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, 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, BiImage, 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, DPIinteract, ECDC, ECGC, ECO2DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSUB, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, Subtilist, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc !!!!

There are lots of Bioinformatics Databases
 For an 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 website interface. At the top, there is a search bar with the text 'All Databases' and a dropdown menu showing 'ras'. A red box highlights the search bar. To the right of the search bar is a 'Search' button. Below the search bar, there is a 'Welcome to NCBI' message and a list of resources. A diagonal banner with the text 'Hands on demo (or see following slides)' is overlaid on the page. The page also features a 'Genotypes and Phenotypes' section and 'NCBI Announcements'.

Search NCBI databases

Search: ras

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

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

Gene: ras

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

Results: 1 to 20 of 85633

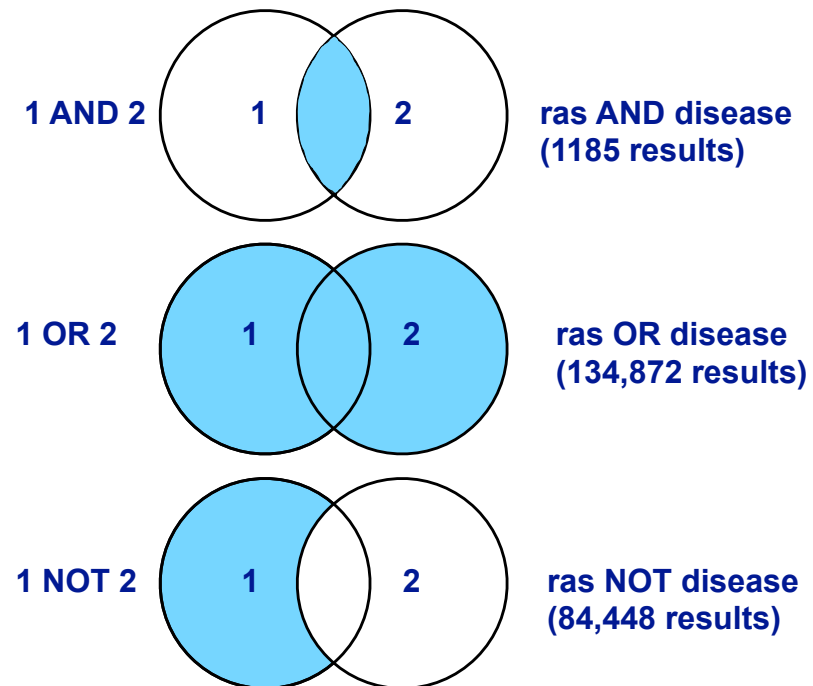
Name/Gen ID	Description	Location	Aliases
ras ID: 19412	resistance to audiogenic seizures [<i>Mus musculus</i> (house mouse)]		asr
ras ID: 43873	raspberry [<i>Drosophila melanogaster</i> (fruit fly)]	Chromosome X, NC_004354.4 (10744502..10749097)	Dmel_CG1799, CG11485, CG1799, DmelCG1799, EP(X)1093,

Top Organisms: Homo sapiens (1126), Mus musculus (823), Rattus norvegicus (625), Oreochromis niloticus (533), Neolamprologus brichardi (507), All other taxa (82019)

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

Results: 1 to 20 of 1126

Name/Gen ID	Description	Location	Aliases
NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS1, KRAS2, NS, NS3, RAS2



NCBI Resources How To Sign in to NCBI

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

Save search Advanced Help

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

Clear all Results: 1 to 20 of 1126 Filters activated: Current only. Clear all to show 1499 items.

Find related data Database: Select Find items

Name/Gen ID	Description	Location	Aliases
<input type="checkbox"/> NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000011.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
<input checked="" type="checkbox"/> KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS1, KRAS2, NS, NS3, RASK2

Search details: ras[All Fields] AND "Homo sapiens"[porgn] AND alive[property] Search See more...

Recent activity Turn Off Clear

66

NCBI Resources How To Sign in to NCBI

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

Advanced Help

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

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

Summary

Official Symbol: KRAS provided by HGNC
 Official Full Name: Kirsten rat sarcoma viral oncogene homolog provided by HGNC
 Primary source: HGNC:HGNC:6407
 See related: Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
 Gene type: protein coding
 RefSeq status: REVIEWED
 Organism: Homo sapiens
 Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominidae; Homo
 Also known as: NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Table of contents

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

67

NCBI Resources How To Sign in to NCBI

Gene KRAS (human) Search

Advanced Help

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

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

Summary

Official Symbol: KRAS provided by HGNC
 Official Full Name: Kirsten rat sarcoma viral oncogene homolog provided by HGNC
 Primary source: HGNC:HGNC:6407
 See related: Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
 Gene type: protein coding
 RefSeq status: REVIEWED
 Organism: Homo sapiens
 Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominidae; Homo
 Also known as: NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Table of contents

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

68

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

NCBI Resources How To Sign in to NCBI

Gene KRAS (human) Search

Advanced Help

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

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

Genomic context

Location: 12p12.1 See KRAS in Epigenomics, MapViewer
 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 reference sequence details

Go to nucleotide: Graphics FASTA GenBank

69

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

Official Symbol KRAS provided by HGNC
Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source HGNC:HGNC:6407
See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193
Gene type protein coding
RefSeq status REVIEWED
Organism [Homo sapiens](#)
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

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

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.

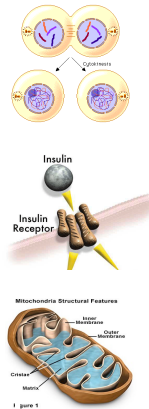
UniProt is a member of the [GO Consortium](#).

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

Gene Ontology Provided by GOA

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

Scroll down to UniProt link

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

Process	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	

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

UniProtKB Link
UniProtKB/Swiss-Prot:P01116

Scroll down to UniProt link

You are here: NCBI > Genes & Expression > Gene

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

UniProt will detail much more information for protein coding genes

UniProtKB - P01116

P01116 - RASK_HUMAN

Protein: **GPase KRas**
Gene: **KRAS**
Organism: *Homo sapiens (Human)*
Status: Reviewed - Experimental evidence at protein level!

Display: None

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

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

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

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding ¹	10 - 18	9	GTP # 2 Publications			
Nucleotide binding ¹	29 - 35	7	GTP # 2 Publications			
Nucleotide binding ¹	59 - 60	2	GTP # 2 Publications			

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

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding ¹	10 – 18	9	GTP # 2 Publications			
Nucleotide binding ¹	29 – 35	7	GTP # 2 Publications			
Nucleotide binding ¹	59 – 60	2	GTP # 2 Publications			

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

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

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

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

Structure Summary | **3D View** | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

Biological Assembly 1

4EPV
Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation

DOI: 10.2210/pdb4epv/pdb
Classification: **HYDROLASE**
Deposited: 2012-04-17 Released: 2012-05-23
Deposition author(s): Sun, Q., Burke, J.R., Phan, J., Burns, M.C., Olejniczak, E.T., Waterson, A.G., Lee, T., Rossanese, O.W., Fesik, S.W.
Organism: Homo sapiens
Expression System: Escherichia coli
Mutation(s): 1

View in 3D: NGL or JSmol (in Browser) | Experimental Data Snapshot | wwPDB Validation | 3D Report | Full Report

Lets view the 3D structure:

Can we find where in the structure our mutations are located and infer their potential molecular effects?

4EPV
Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Click to identify atoms and bonds.

Bond: [GLY]12:A:O - [GLY]12:A:C

Display Options

- Assembly: Bioassembly 1
- Model: Model 1
- Symmetry: None
- Interaction: [GDP]201:A
- Style: Cartoon
- Color: Rainbow
- Ligand: None
- Quality: Automatic

Water Ions

Hydrogens Clashes

Back to UniProt:

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

Display: None

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

Family and domain databases

Gene3D ¹	3.40.50.300. 1 hit.
InterPro ¹	IPR027417. P-loop_NTPase. IPR005225. Small_GTP-bd_dom. IPR01806. Small_GTPase. IPR020849. Small_GTPase_Ras. [Graphical view]
PANTHER ¹	PTHR24070. PTHR24070. 1 hit.
Pfam¹	PF00071. Ras. 1 hit. [Graphical view]
PRINTS ¹	PR00449. RASTRANFRMNG.
SMART ¹	SM00173. RAS. 1 hit. [Graphical view]
SUPFAM ¹	SSF52540. SSF52540. 1 hit.
TIGRFAMs ¹	TIGR00231. small_GTP. 1 hit.
PROSITE ¹	PSS1421. RAS. 1 hit. [Graphical view]

Sequences (2)¹

Sequence status¹: Complete.

Sequence processing¹: The displayed sequence is further processed into a mature form.

This entry describes 2 isoforms¹ produced by **alternative splicing**. [Align](#)

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... ?

EMBL-EBI HOME

Family: **Ras (PF00071)**

Summary: Ras family

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

Species

Interactions

Structures

Jump to...

Summary: Ras family

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

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

This is the Wikipedia entry entitled "Ras subfamily": [More...](#)

Ras subfamily [Edit Wikipedia article](#)

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

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

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

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

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

Contents [hide]

- History
- Structure
- Function
 - 3.1 Activation and deactivation
 - 3.2 Membrane attachment
- Members
- Ras in cancer
 - 5.1 Inappropriate activation
 - 5.2 Constitutively active Ras

Identifiers

Symbol	Ras
Pfam	PF00071 ↗
SisterPro	IPR013753 ↗
PROSITE	PD0C00117 ↗
SCOP	Sp21 ↗
SUPERFAMILY	Sp21 ↗

Example Questions:

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

Summary

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

Species distribution

Interactions

Structures

Jump to...

This visualisation provides a simple graphical representation of the distribution of this family across species. You can find the original interactive tree in the sidebar tab [More...](#)

Sunburst controls

Hide

Human expansion

Root

- Elkayota
- Metazoa
- Chordata
- Mammalia
- Primates
- Hominidae
- Homo
- Homo sapiens

Weight segments by...

number of sequences

number of species

Change the size of the sunburst

Colour assignments

- Archaea
- Bacteria
- Eukaryota
- Viruses
- Unclassified
- Woody
- Unclassified sequence

Selections

Align selected sequences to HMM

Generate a FASTA-format file

Clas: selection

Currently selected:

- 521 sequences
- 3 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... ?

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


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

UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View
ABBK01_GIALA	11 - 335	2vvg	A	11 - 335	Jmol AstexViewer SPICE
			B	11 - 335	Jmol AstexViewer SPICE
CENPE_HUMAN	12 - 329	1t5c	A	12 - 329	Jmol AstexViewer SPICE
			B	12 - 329	Jmol AstexViewer SPICE
KAR3_YEAST	392 - 723	1f9t	A	392 - 723	Jmol AstexViewer SPICE
			A	392 - 723	Jmol AstexViewer SPICE
			A	392 - 723	Jmol AstexViewer SPICE
			B	392 - 723	Jmol AstexViewer SPICE
KI13B_HUMAN	11 - 352	3obj	A	11 - 352	Jmol AstexViewer SPICE
			C	11 - 352	Jmol AstexViewer SPICE
		1iig	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		1a0b	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		1x88	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE

PFam: jmol
 http://pfam.janelia.org/structure/viewer/jmol?id=3bfn
 PFam: Family: Kinesin (PF00225) PFam: Jmol

welcome trust
sanger
 institute

DB entry 3bfn



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

Jmol

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

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/bgg213_f17/lectures/#1

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

THANK YOU