BGGN 213

Foundations of Bioinformatics

Barry Grant
UC San Diego

http://thegrantlab.org/bggn213

HELLO
my name is

BARRY

bjgrant@ucsd.edu

HELLO
HER ~~my~~ name is

ILEENA

ileenamitra@eng.ucsd.edu

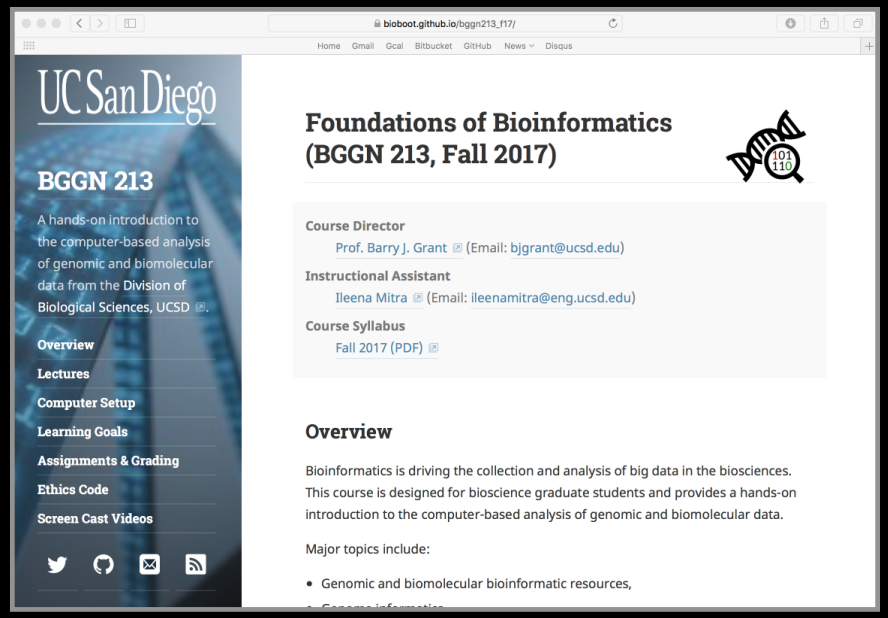
Introduce Yourself!

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

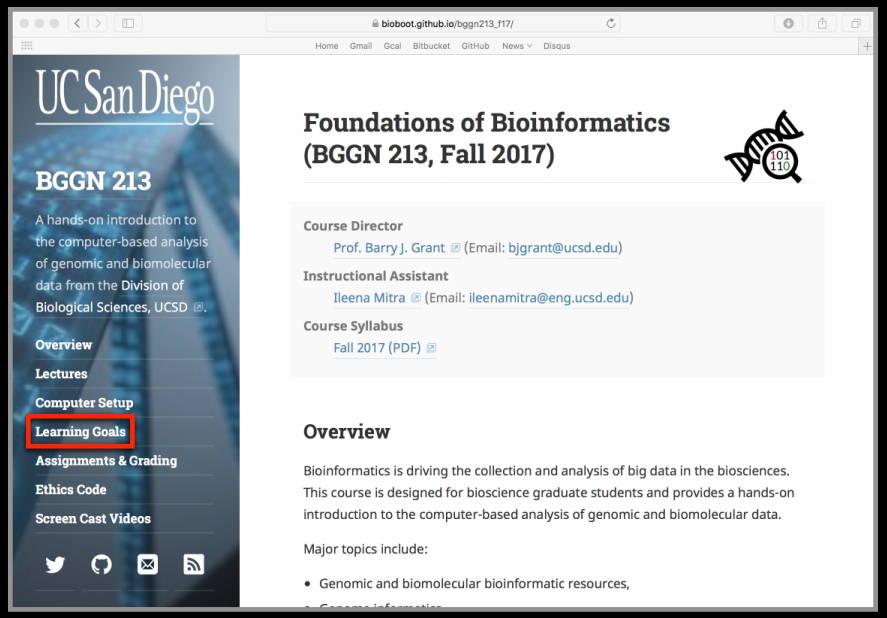
Today's Menu

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 leaning goals.
Introduction to Bioinformatis	**Introducing the *what*, *why* and *how* of bioinformatics?**
Computer Setup	Ensuring your laptop is all set for future sections of this course.

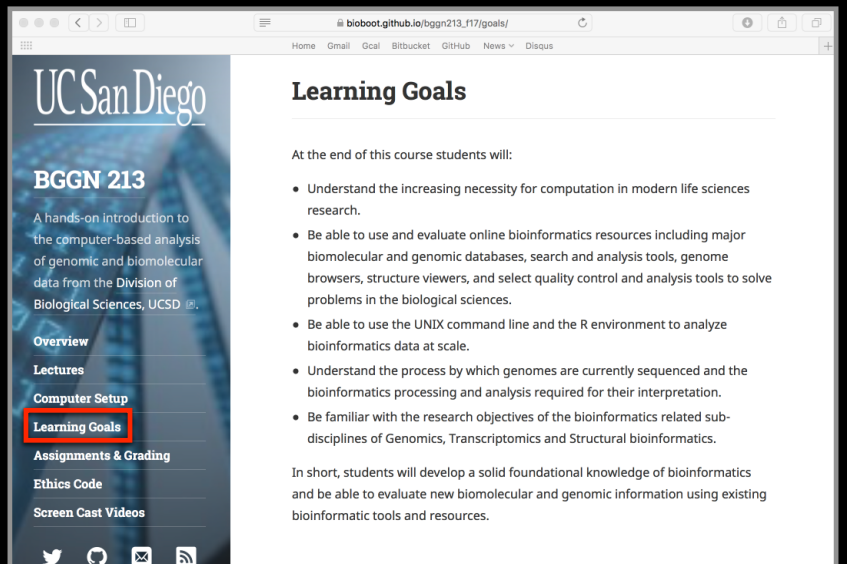
http://thegrantlab.org/bggn213/



http://thegrantlab.org/bggn213/



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



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!

bioboot.github.io/bggn213_f17/goals/

Home Gmail Gcal Bitbucket GitHub News Disqus

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

Course Structure

Derived from specific learning goals

bioboot.github.io/bggn213_f17/lectures/

Home Gmail Gcal Bitbucket GitHub News Disqus

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
- 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	**Welcome to Foundations of Bioinformatics** Course introduction, Leaning 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

bioboot.github.io/bggn213_f17/lectures/

Home Gmail Gcal Bitbucket GitHub News Disqus

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
- Assignments & Grading
- Ethics Code
- Screen Cast Videos

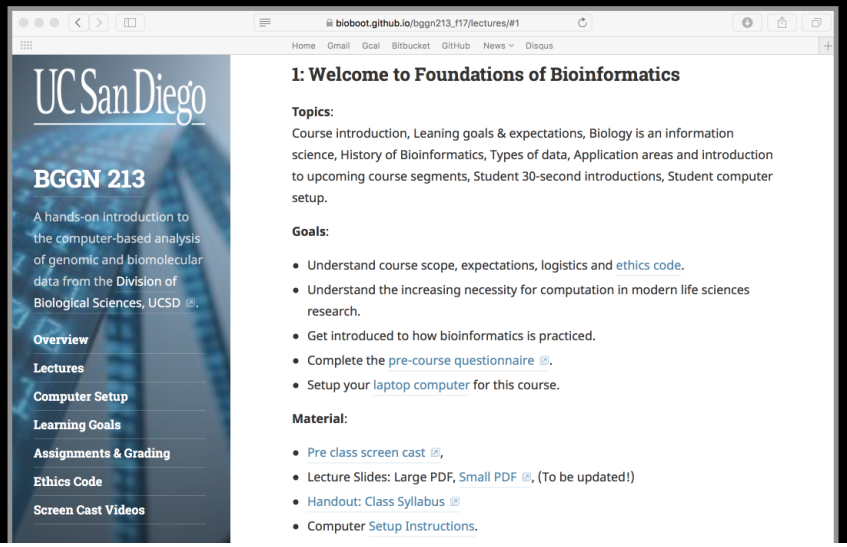
Lectures

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	**Welcome to Foundations of Bioinformatics** Course introduction, Leaning 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

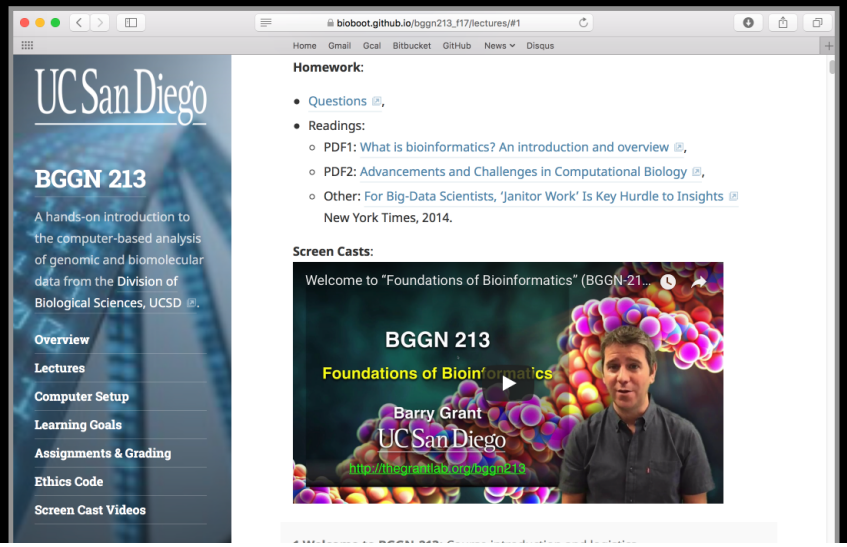
Class Details

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



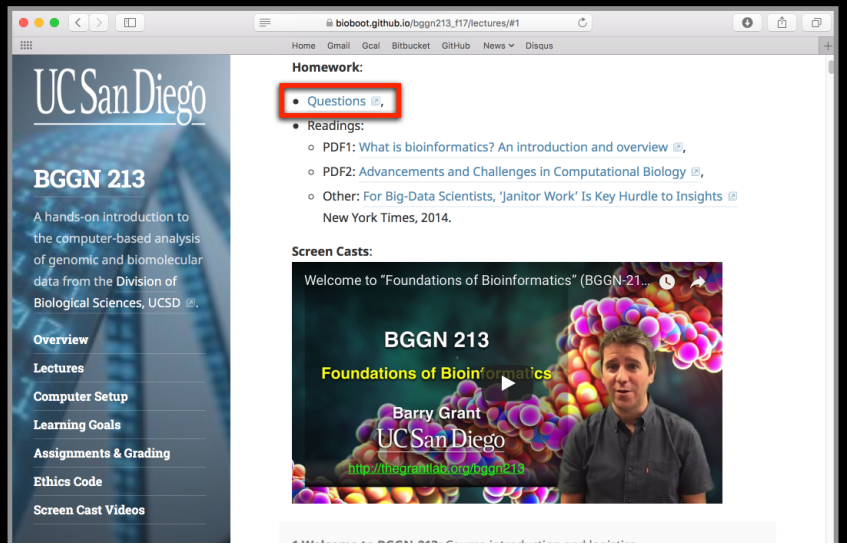
Homework

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



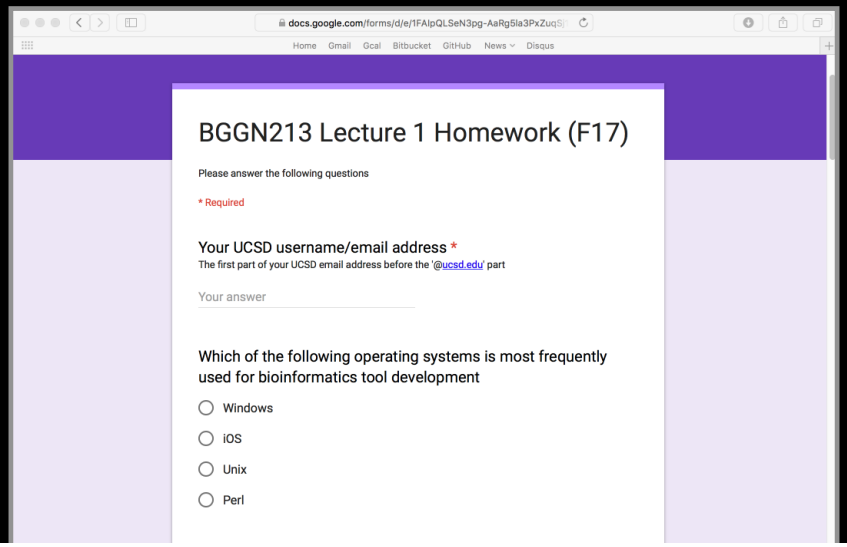
Homework

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



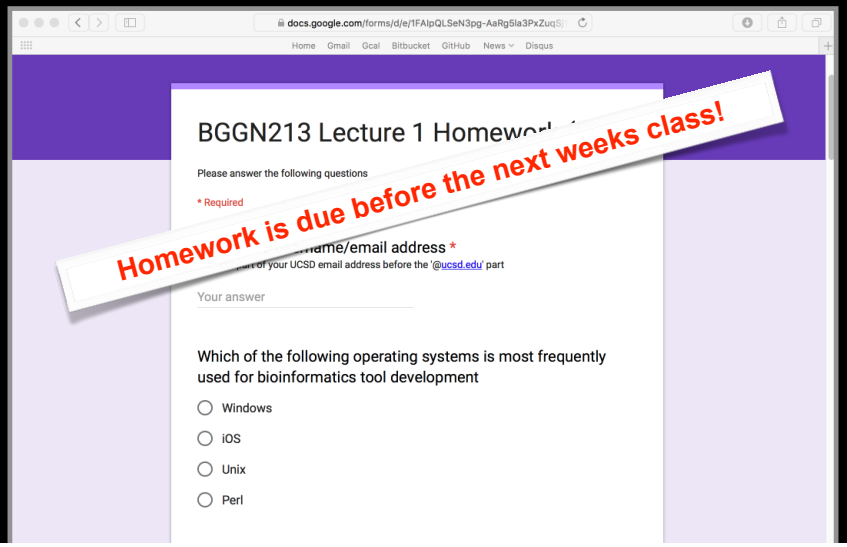
Homework

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



Homework

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



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 leaning goals.
Introduction to Bioinformatis	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)

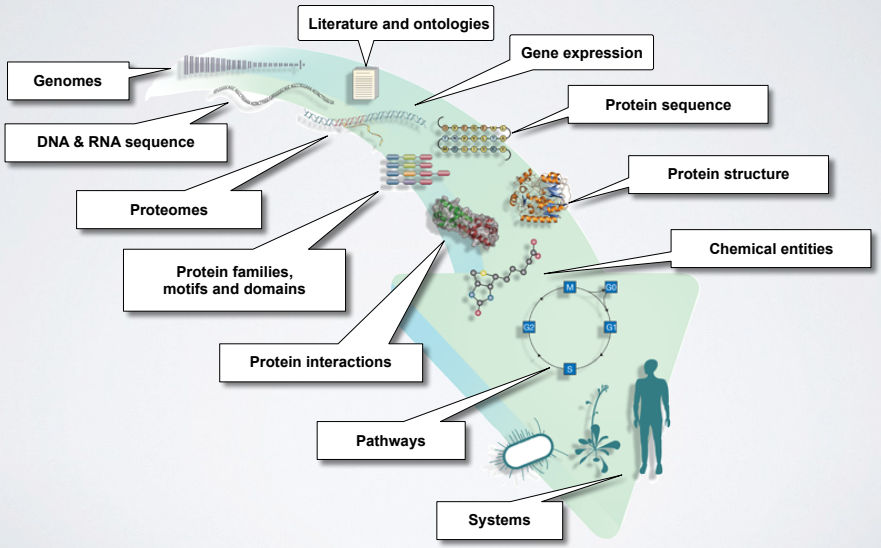
MORE DEFINITIONS

- "Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying **"informatics"** [illegible] **niques** (derived from disciplines such as applied ma[illegible] [illegible]ter science, and statistics) to **understand** a[illegible] **ze** the information associated with these [illegible], on a **large-scale**. Luscombe NM, *et al.* Methods [illegible] 01;40:346.

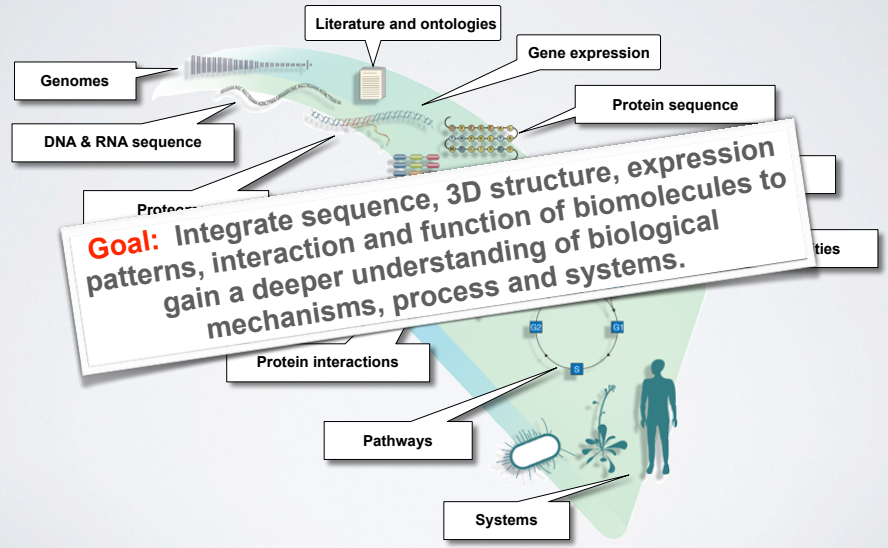
- "Bioinformati[illegible] [illegible]arch, development, or application of **comp[illegible]** **approaches** for expanding the use of **bi[illegible]** **medical**, **behavioral** or **health data**, including those to **[illegible]quire**, **store**, **organize** and **analyze** such data." National Institutes of Health (NIH) (http://tinyurl.com/l3gxr6b)

Key Point: Bioinformatics is Computer Aided Biology

Major types of Bioinformatics Data

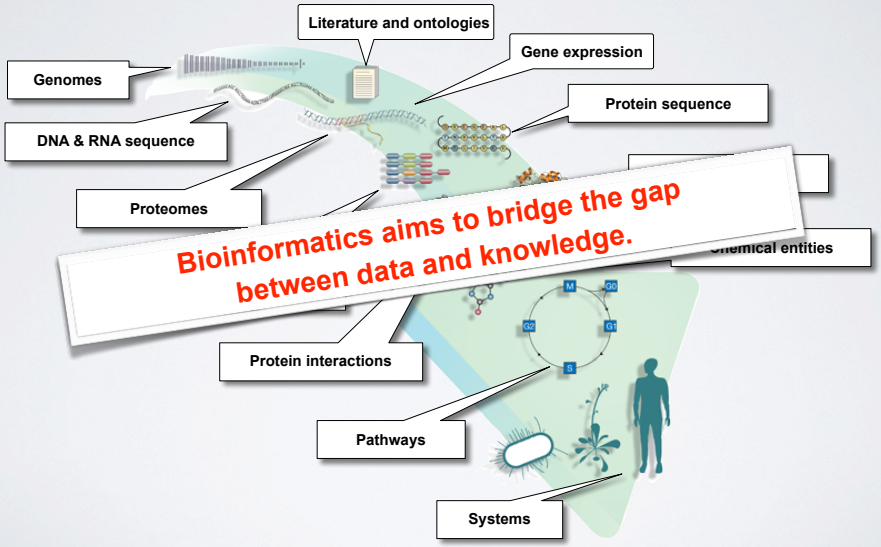


Major types of Bioinformatics Data



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

Major types of Bioinformatics Data



Bioinformatics aims to bridge the gap between data and knowledge.

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

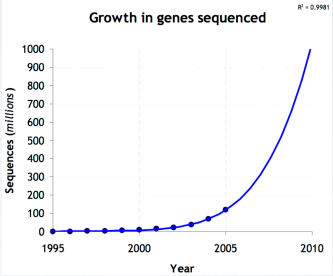
- *DNA sequence* determines *protein sequence.*
- *Protein sequence* determines *protein structure.*
- *Protein structure* determines *protein function.*
- *Regulatory mechanisms* (e.g. gene expression) determine the amount of a particular *function in space and time.*

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**

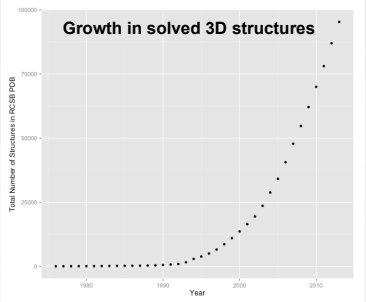


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, *etc…*

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.

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

x 1,000

x 100,000

How do we actually do Bioinformatics?

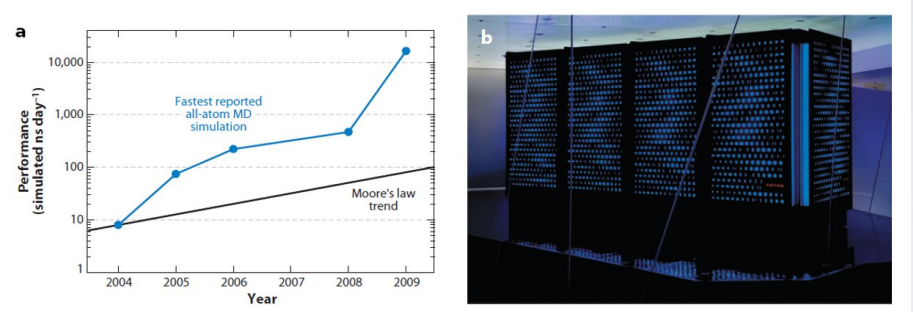
Pre-packaged tools and databases

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

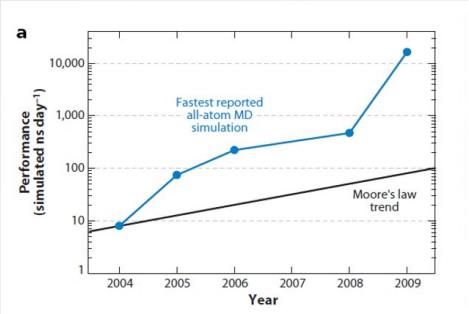
Tool development

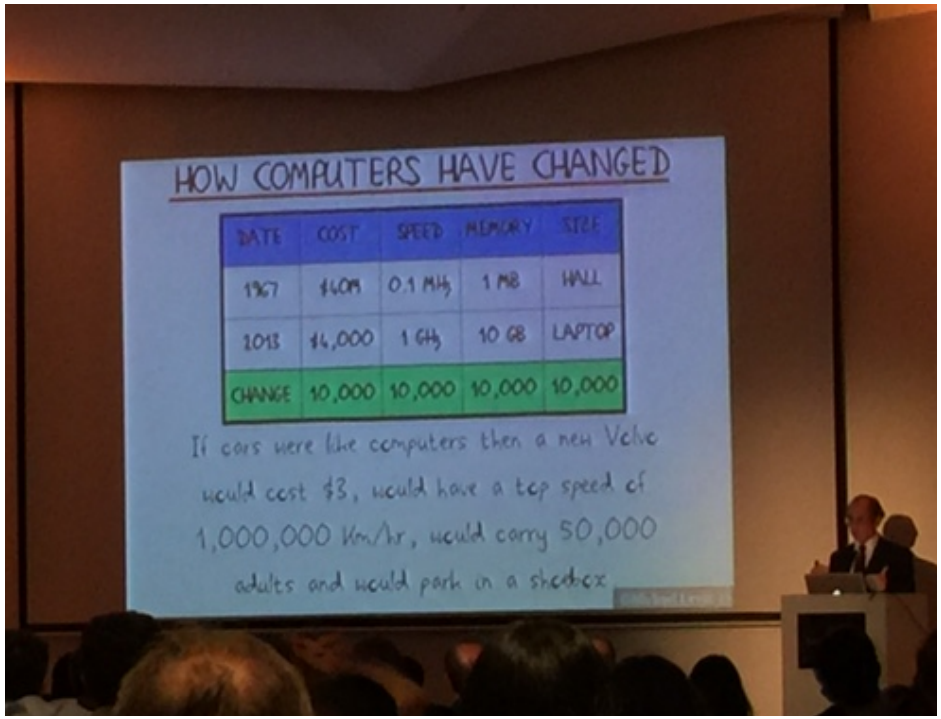
- Mostly on a UNIX environment
- Knowledge of programing languages frequently required (Python, **R**, Perl, C Java, Fortran)
- May require specialized or high performance computing resources…

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'

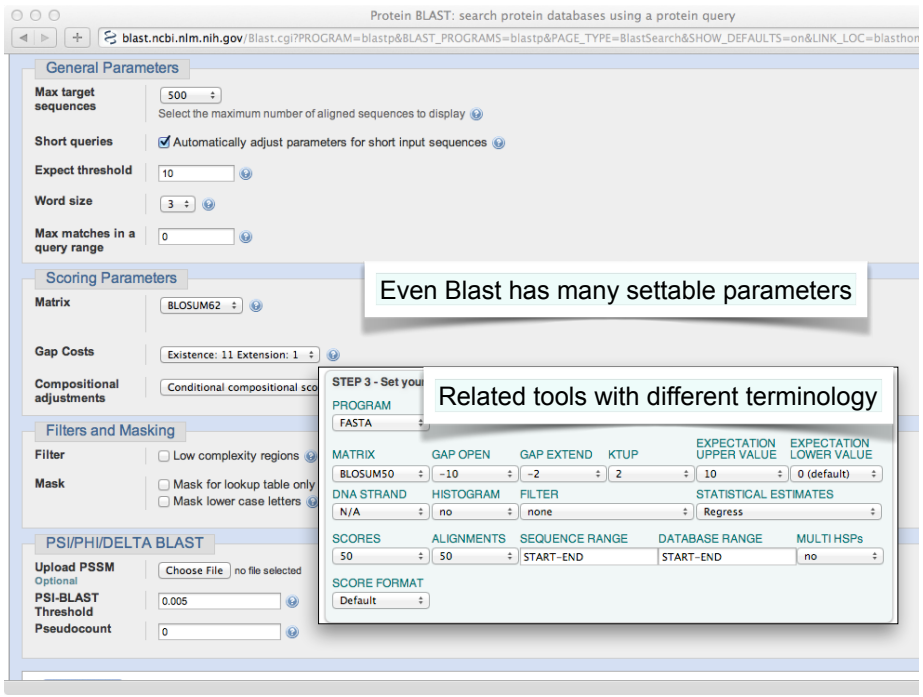
Common problems with Bioinformatics

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!

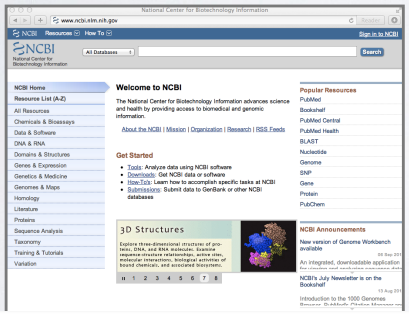
Most are developed independently

Notable exceptions are found at the:
- **EBI** (European Bioinformatics Institute) and
- **NCBI** (National Center for Biotechnology Information)

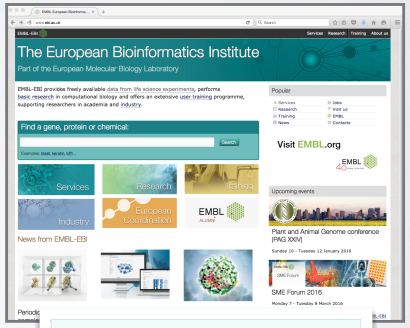


Key Online Bioinformatics Resources: NCBI & EBI

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



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



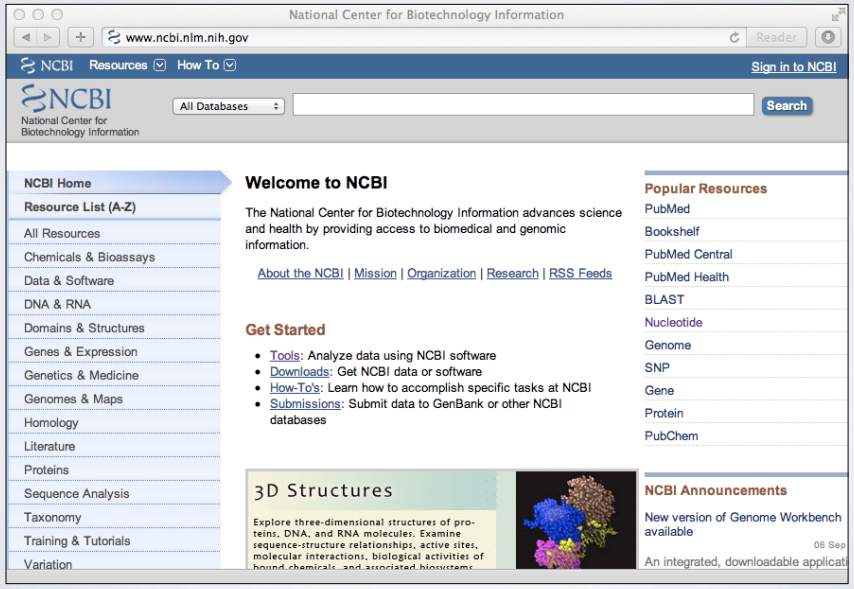
https://www.ebi.ac.uk

National Center for Biotechnology Information (NCBI)

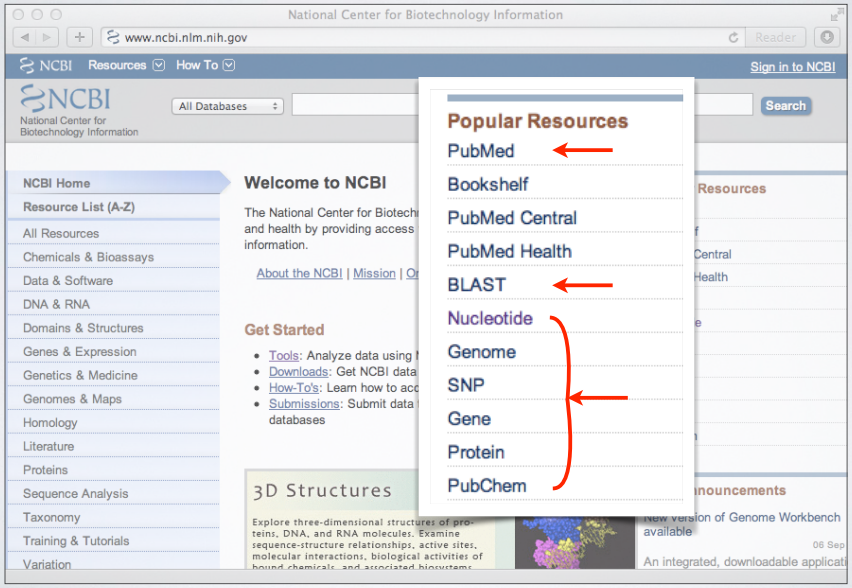
- Created in 1988 as a part of the National Library of Medicine (NLM) at the National Institutes of Health
- NCBI's mission includes:
  - Establish **public databases**
  - Develop **software tools**
  - **Education** on and dissemination of biomedical information

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

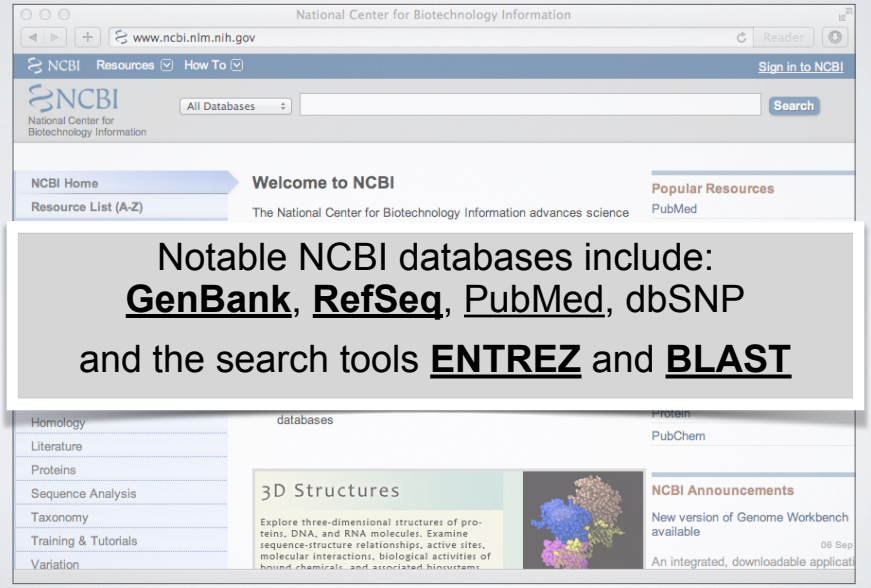
http://www.ncbi.nlm.nih.gov



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



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

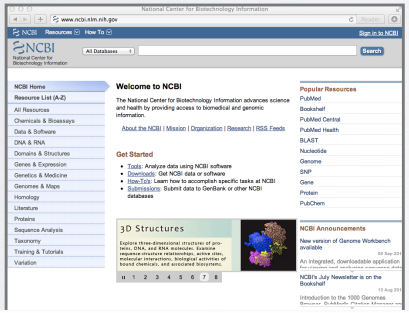


Notable NCBI databases include:
GenBank, **RefSeq**, PubMed, dbSNP

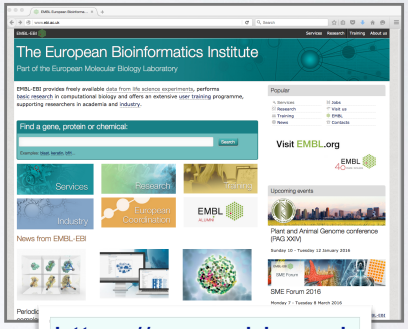
and the search tools **ENTREZ** and **BLAST**

Key Online Bioinformatics Resources: NCBI & EBI

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



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



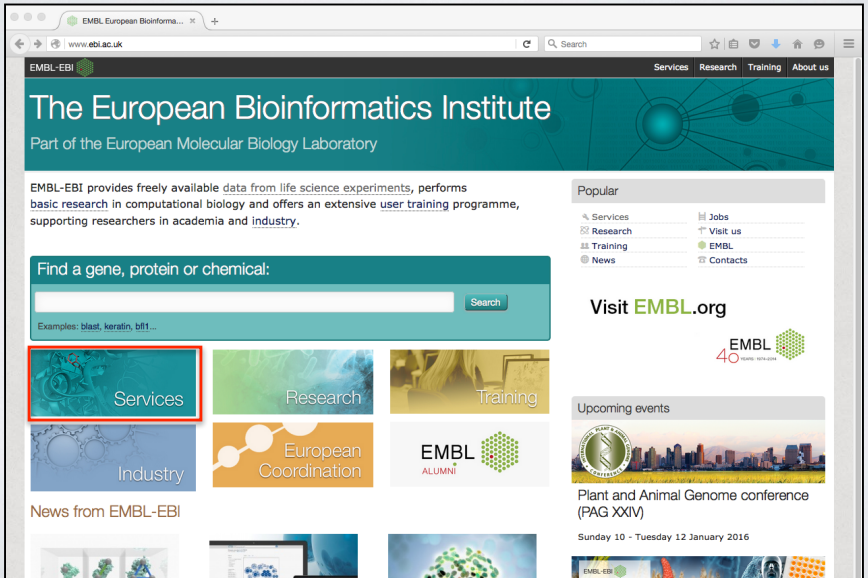
https://www.ebi.ac.uk

European Bioinformatics Institute (EBI)

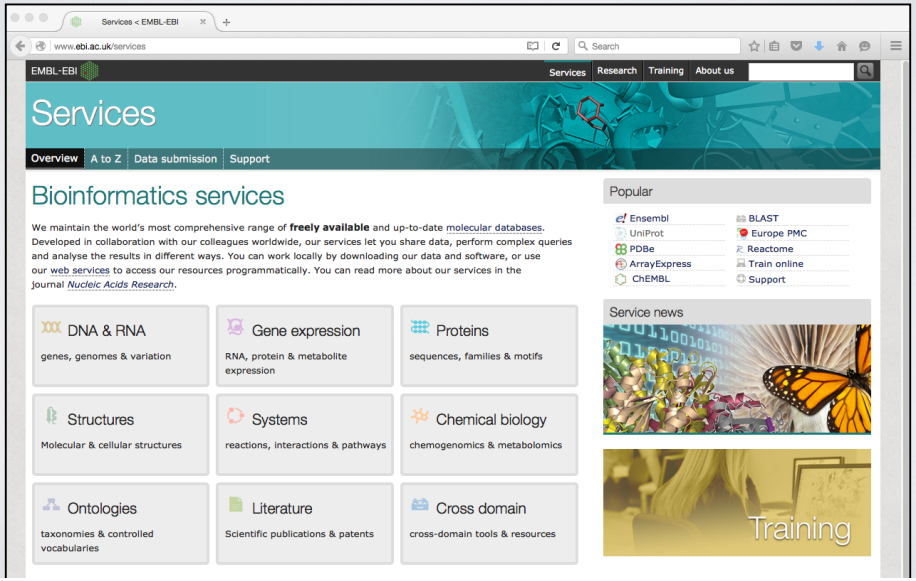
- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
  - providing freely available **data** and **bioinformatics services**
  - and providing advanced **bioinformatics training**

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

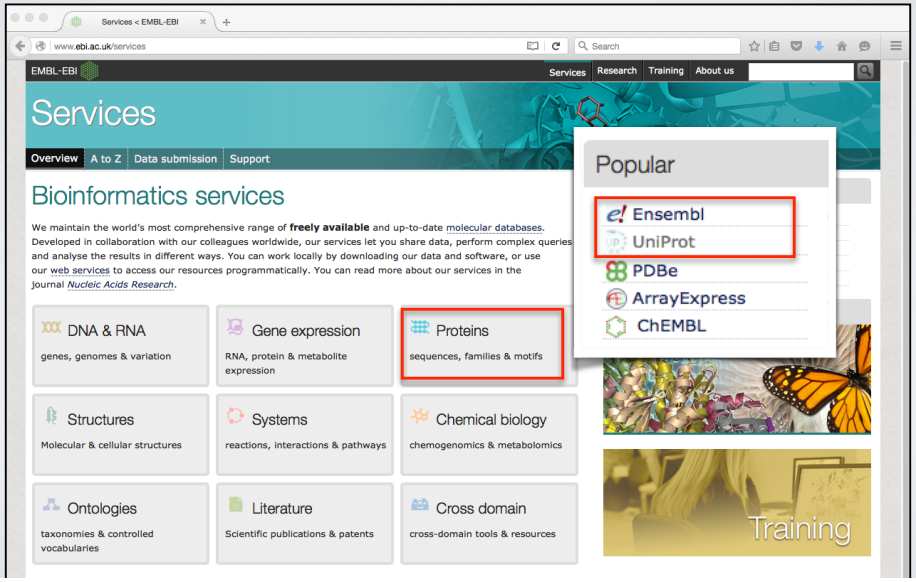
The EBI maintains a number of high quality curated **secondary databases** and associated tools



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

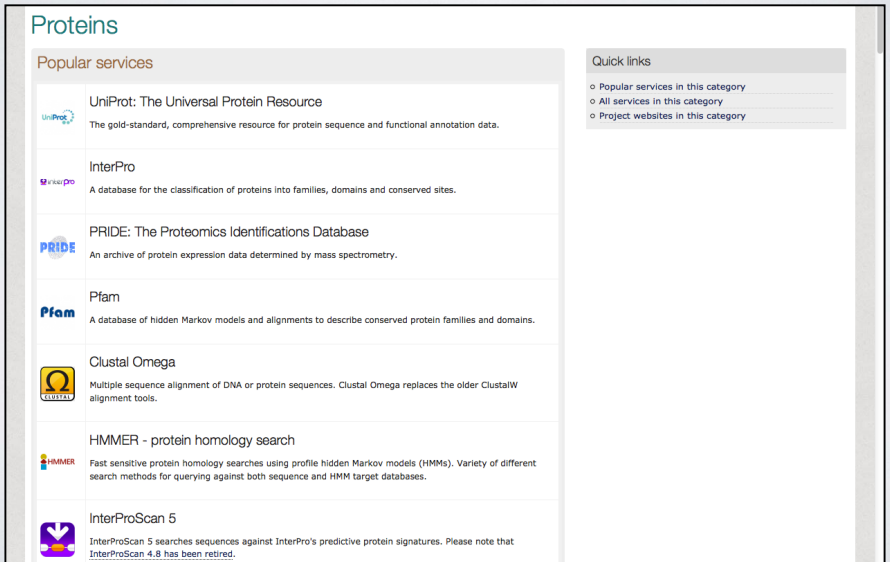


The EBI maintains a number of high quality curated **secondary databases** and associated tools

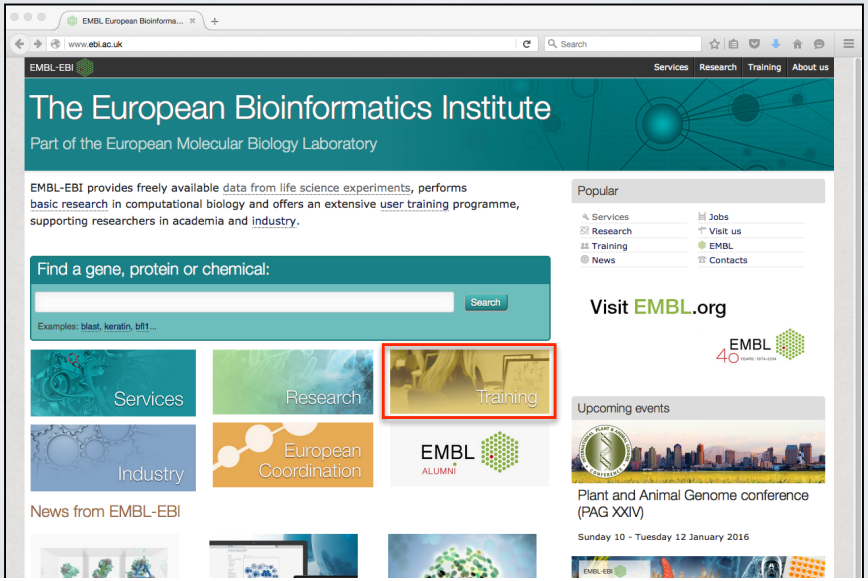


https://www.ebi.ac.uk

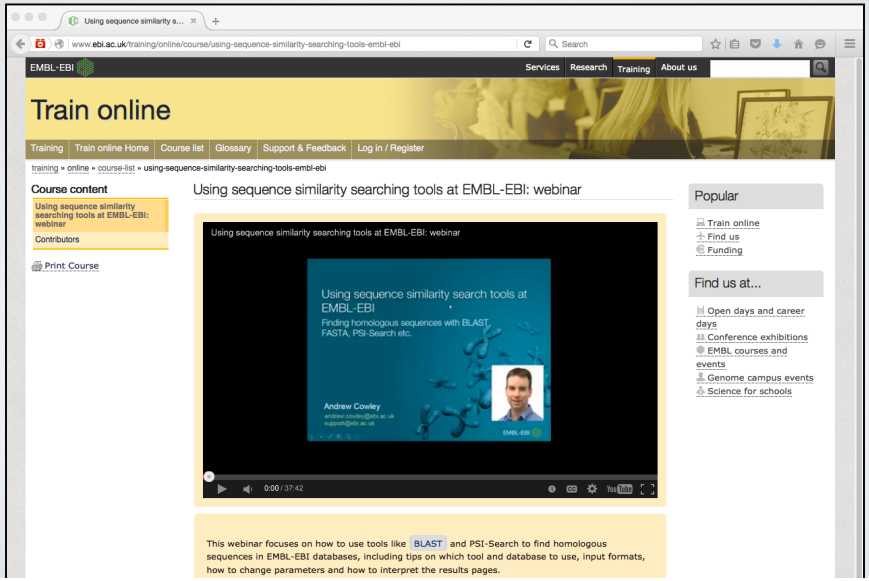
The EBI makes available a wider variety of **online tools** than NCBI



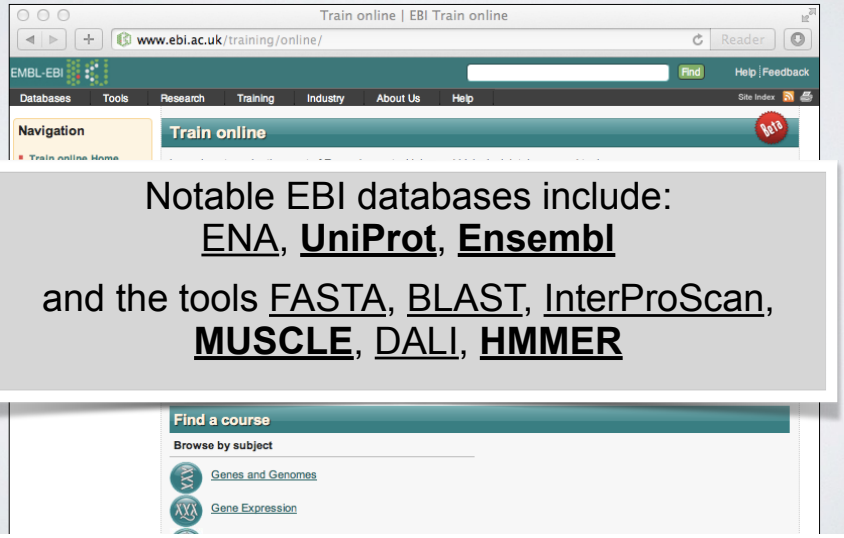
The EBI also provides a growing selection of **online tutorials** on EBI databases and tools



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



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



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

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

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

There are lots of Bioinformatics Databases

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

< Handout_Major_Databases.pdf >

Side-note: Databases come in all shapes and sizes

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

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:

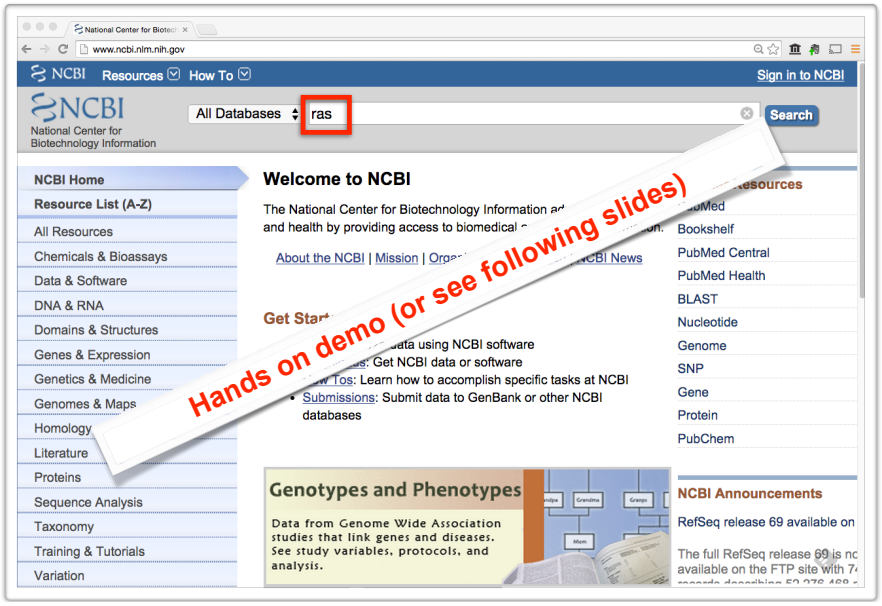
"What do you know about that 'Kras' gene the speaker kept taking about?"

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/



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

NCBI Resources How To Sign in to NCBI

Search NCBI databases Help

ras Search

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

Literature

Books	1,677	books and reports
MeSH	402	ontology used for PubMed indexing
NLM Catalog	223	books, journals and more in the NLM Collections
PubMed	54,672	scientific & medical abstracts/citations
PubMed Central	96,114	full-text journal articles

Health

ClinVar	759	human variations of clinical significance
dbGaP	120	genotype/phenotype interaction studies
GTR	1,879	genetic testing registry

Genes

EST	3,985	expressed sequence tag sequences
Gene	87,165	collected information about gene loci
GEO DataSets	3,732	functional genomics studies
GEO Profiles	1,622,789	gene expression and molecular abundance profiles
HomoloGene	696	homologous gene sets for selected organisms
PopSet	2,254	sequence sets from phylogenetic and population studies
UniGene	4,770	clusters of expressed transcripts

Proteins

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

NCBI Resources How To Sign in to NCBI

Gene Gene ras Search

Save search Advanced Help

Show additional filters

Clear all

Gene sources: Genomic, Mitochondria, Organelles, Plasmids, Plastids

Categories: Alternatively spliced, Annotated genes, Non-coding, Protein-coding, Pseudogene

Sequence content: CCDS, Ensembl, RefSeq

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

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

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

Results: 1 to 20 of 85633

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

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

Filters: Manage Filters

Top Organisms [Tree]
- Homo sapiens (1126)
- Mus musculus (823)
- Rattus norvegicus (625)
- Oreochromis niloticus (533)
- Neolamprologus brichardi (507)
- All other taxa (82019)

More...

Find related data
Database: Select
Find items

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

www.ncbi.nlm.nih.gov/gene

NCBI Resources How To Sign in to NCBI

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

Help

Show additional filters

Clear all

Gene sources: Genomic

Categories: Alternatively spliced, Annotated genes, Non-coding, Protein-coding, Pseudogene

Sequence content: CCDS, Ensembl, RefSeq

Status clear: ✓ Current only

Chromosome locations

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

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

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

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

Filters: Manage Filters

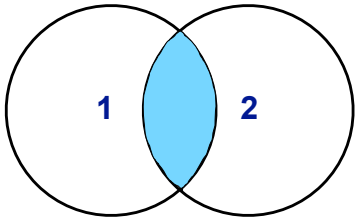
Find related data
Database: Select
Find items

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

Search See more...

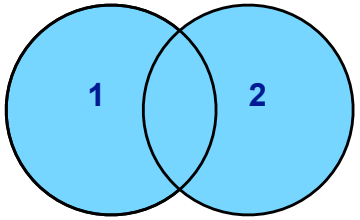
Recent activity Turn Off Clear

1 AND 2



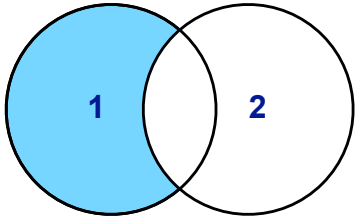
ras AND disease (1185 results)

1 OR 2



ras OR disease (134,872 results)

1 NOT 2



ras NOT disease (84,448 results)

www.ncbi.nlm.nih.gov/gene

NCBI Resources How To Sign in to NCBI

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

Save search Advanced Help

Show additional filters

Clear all

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene

Sequence content
CCDS
Ensembl
RefSeq

Status clear
✓ Current only

Chromosome locations

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

Filters: Manage Filters

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

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

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [*Homo sapiens* (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [*Homo sapiens* (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS1, KRAS2, NS, NS3, RASK2

Find related data
Database: Select
Find items

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

Recent activity
Turn Off Clear

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

NCBI Resources How To Sign in to NCBI

Gene | Gene | Search

Advanced Help

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

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

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

Summary

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

Table of contents
- 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)

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

NCBI Resources How To Sign in to NCBI

Gene | Search

Help

Hide sidebar >>

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

KRAS
(human)

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

Summary

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

Table of contents
- 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)

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

Genomic context

Location: 12p12.1
Exon count: 6

See KRAS in Epigenomics, MapViewer

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

Chromosome 12 - NC_000012.12

[25052101 ▶ [25436297 ▶

LRMP LYRM5 LOC100421617
CASC1 RPL39P27
KRAS ◀

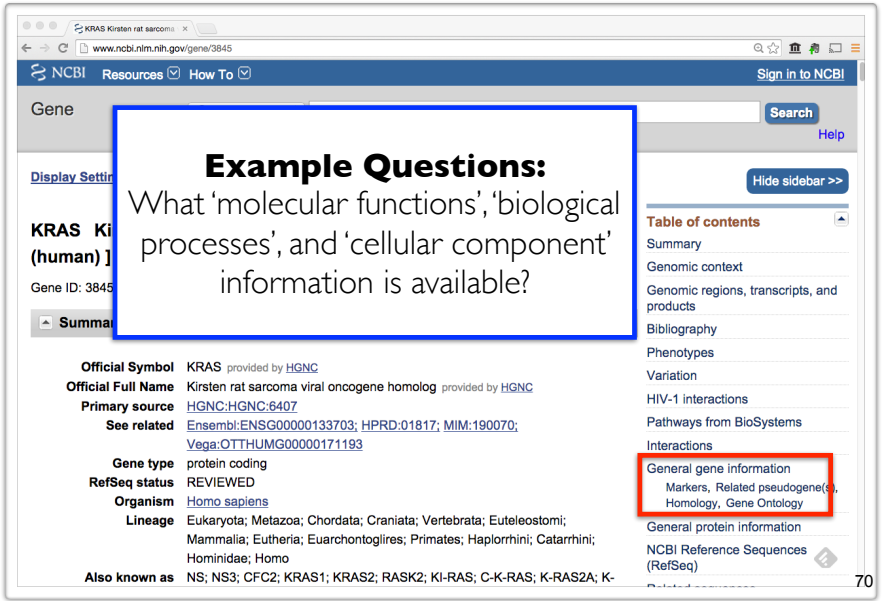
Genomic regions, transcripts, and products

Go to reference sequence details

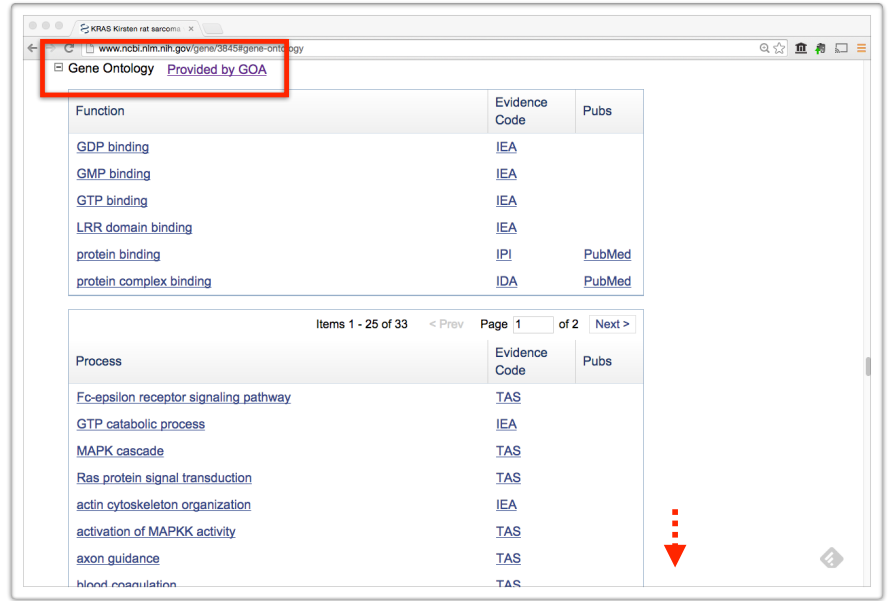
Genomic Sequence: NC_000012.12 chromosome 12 reference GRCh38 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank

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

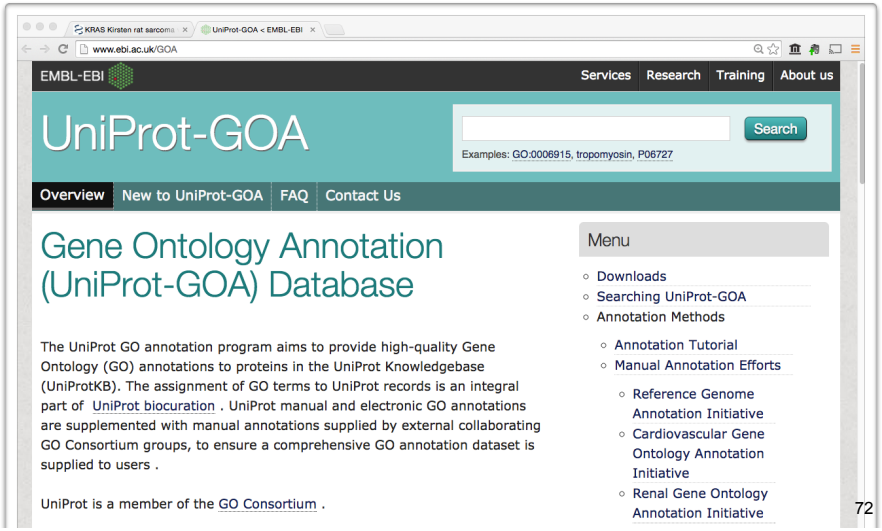


Example Questions:
What 'molecular functions', 'biological processes', and 'cellular component' information is available?



GO: Gene Ontology

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

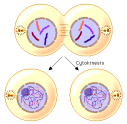


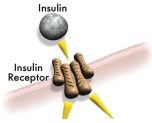
Why do we need Ontologies?

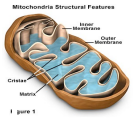
- Annotation is essential for capturing the understanding and knowledge associated with a sequence or other molecular entity
- Annotation is traditionally recorded as "free text", which is easy to read by humans, but has a number of disadvantages, including:
  - Difficult for computers to parse
  - Quality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

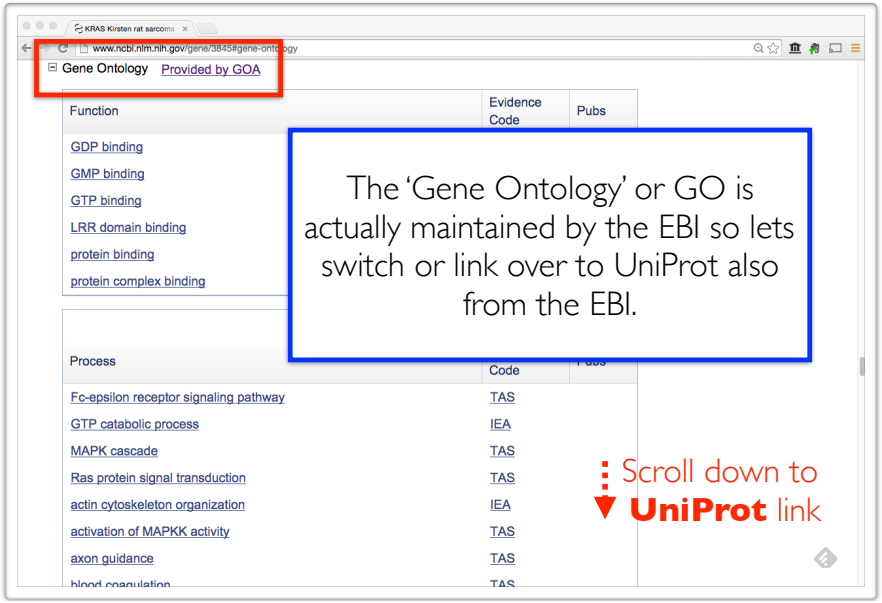
GO Ontologies

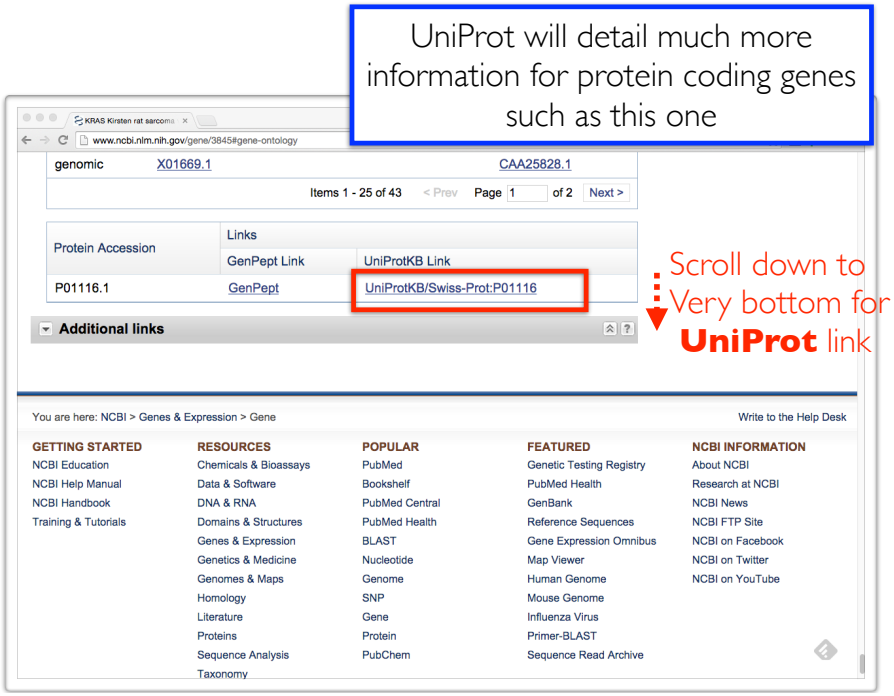
- There are three ontologies in GO:
  - **Biological <u>Process</u>**
    A commonly recognized series of events
    e.g. cell division, mitosis,
  - **Molecular <u>Function</u>**
    An elemental activity, task or job
    e.g. kinase activity, insulin binding
  - **Cellular <u>Component</u>**
    Where a gene product is located
    e.g. mitochondrion, mitochondrial membrane

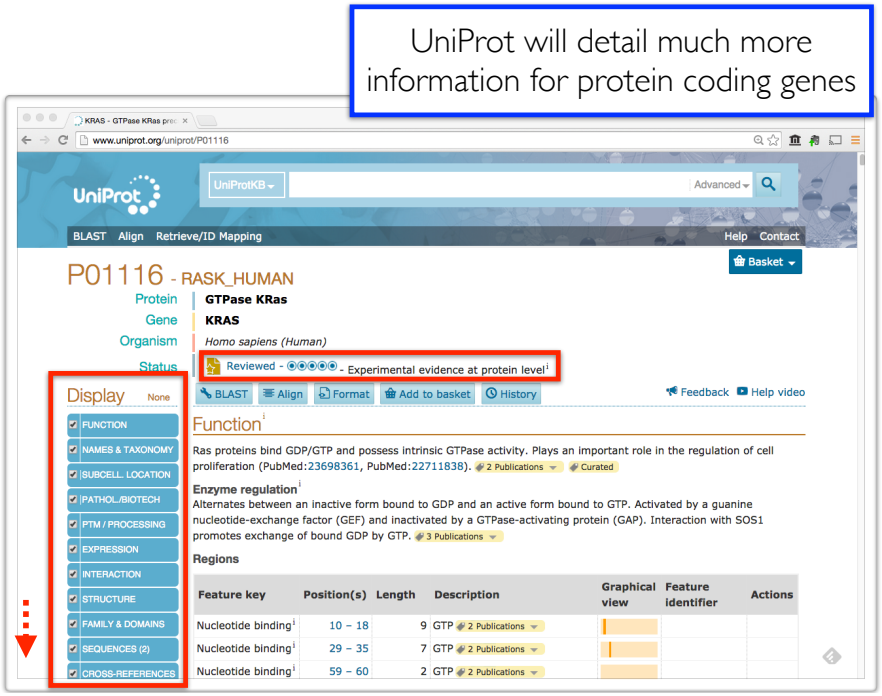












Example Questions:
What positions in the protein are responsible for GTP binding?

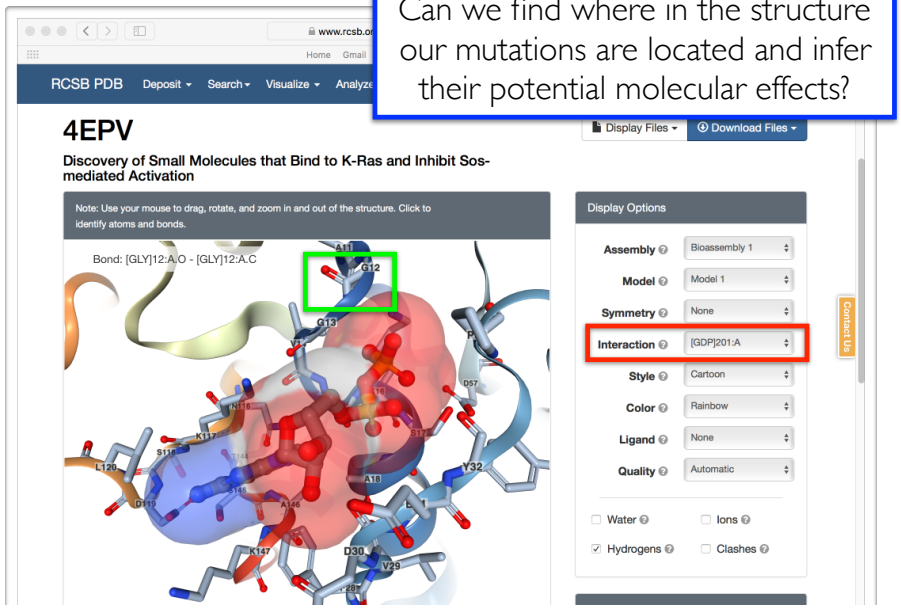
Example Questions:
What variants of this enzyme are involved in gastric cancer and other human diseases?

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

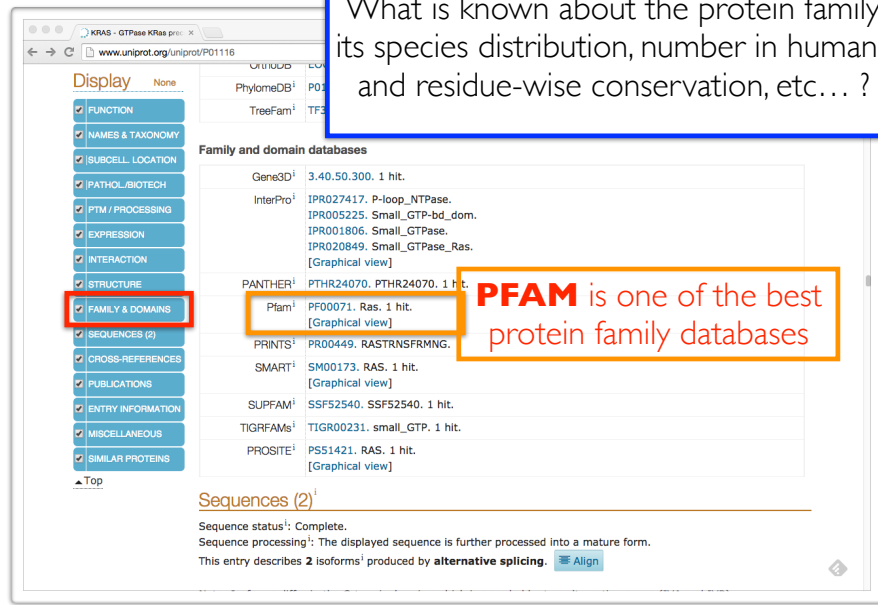
Open link in a new tab!

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

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

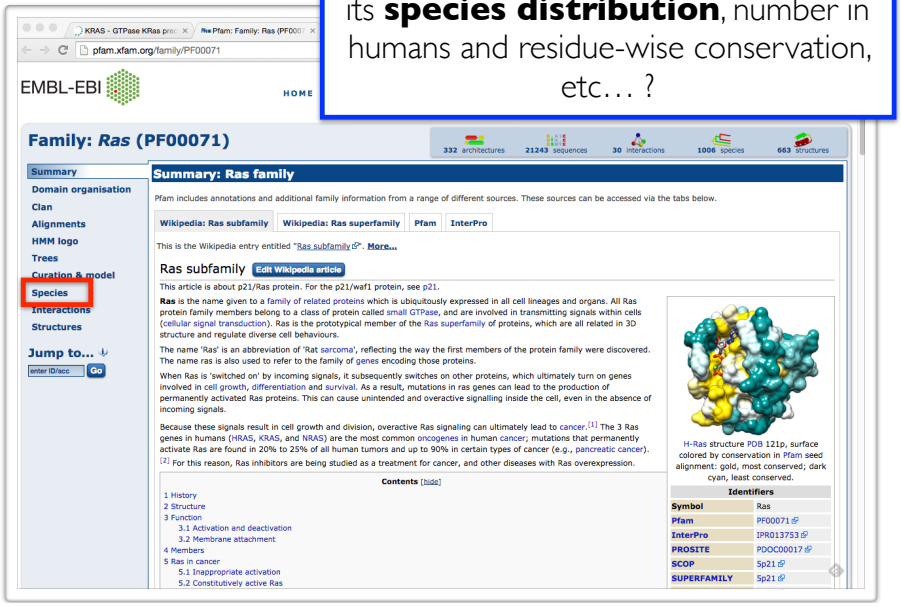


Back to UniProt:
What is known about the protein family, its species distribution, number in humans and residue-wise conservation, etc… ?

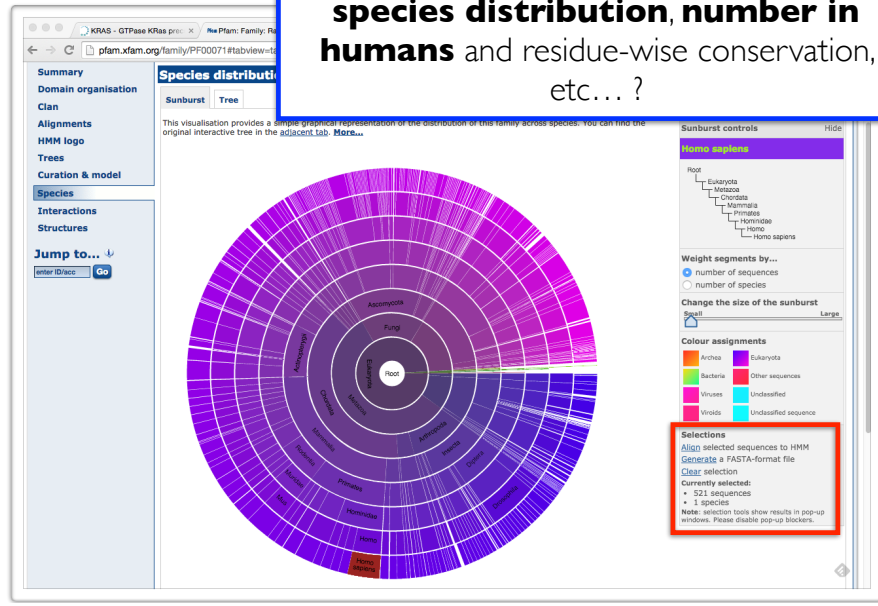


PFAM is one of the best protein family databases

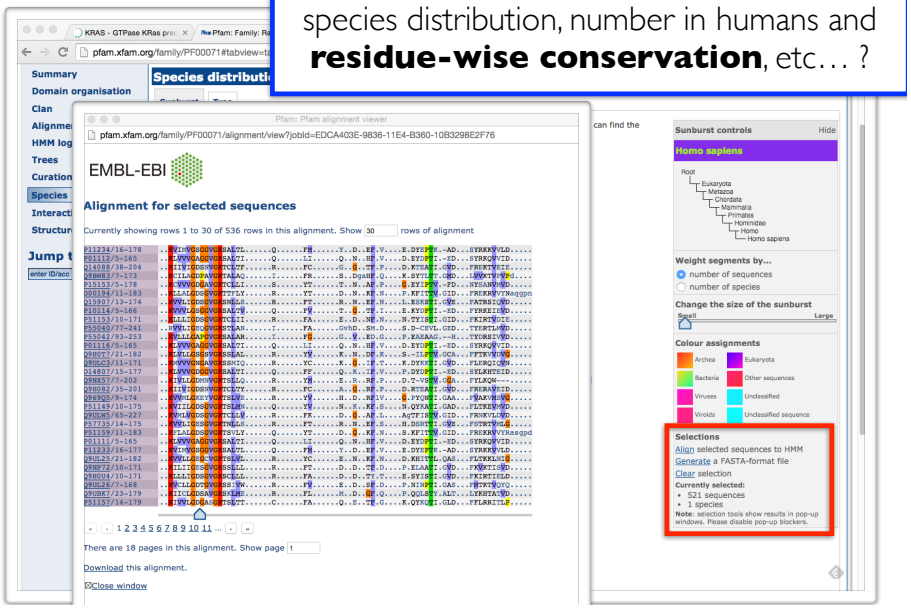
Example Questions:
What is known about the protein family, its **species distribution**, number in humans and residue-wise conservation, etc… ?



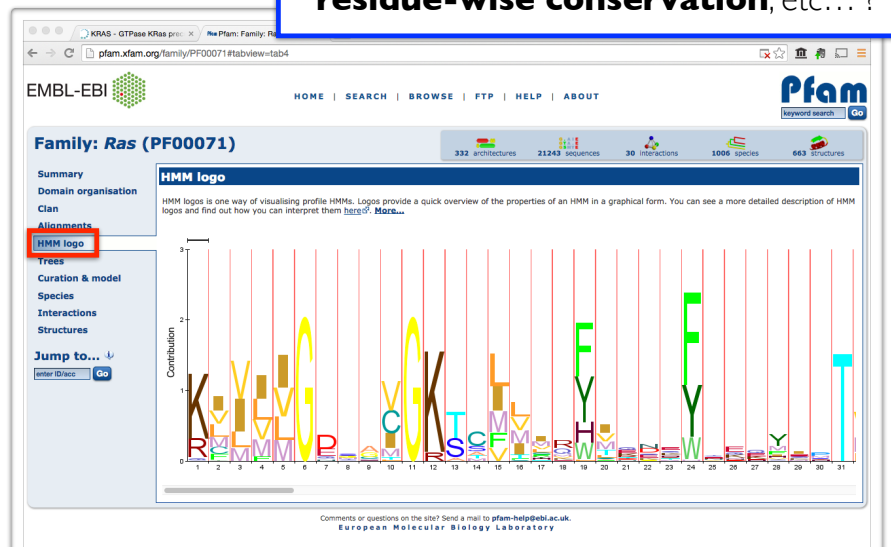
Example Questions:
What is known about the protein family, its **species distribution**, **number in humans** and residue-wise conservation, etc… ?

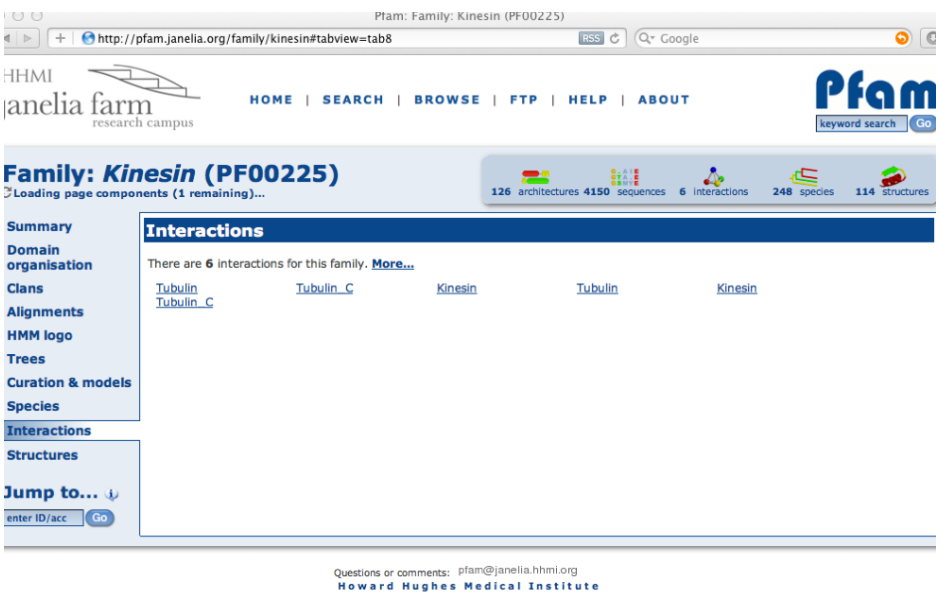


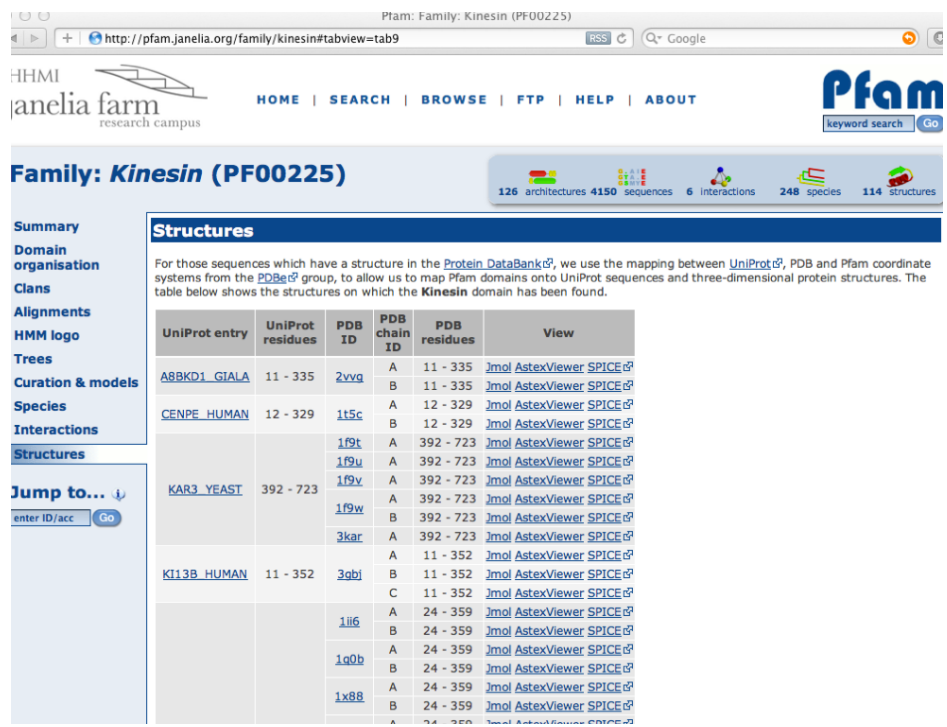
Example Questions:
What is known about the protein family, its species distribution, number in humans and **residue-wise conservation**, etc… ?

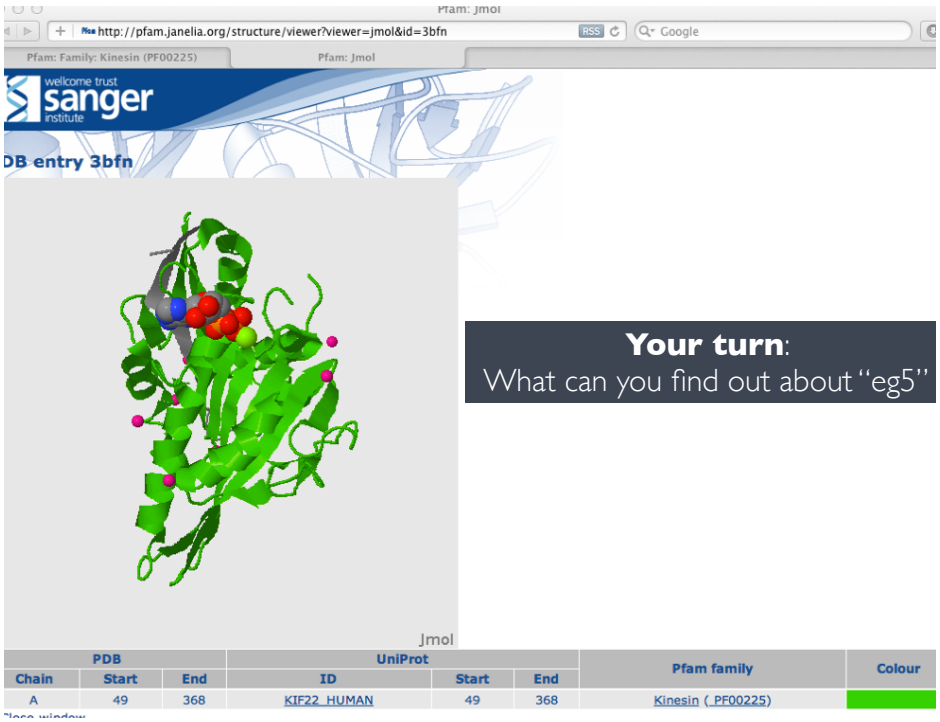


Example Questions:
What is known about the protein family, its species distribution, number in humans and **residue-wise conservation**, etc… ?









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

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

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