INTRODUCTION TO BIOINFORMATICS

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

Barry Grant
University of Michigan
www.thegrantlab.org


Barry Grant, Ph.D.
bjgrant@umich.edu


Ryan Mills, Ph.D.
remills@umich.edu


Hongyang Li (GSI)
hyangl@umich.edu

COURSE LOGISTICS

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

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

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

MODULE OVERVIEW

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

1.1. *Introduction to Bioinformatics*

1.2. *Sequence Alignment and Database Searching*

1.3 *Structural Bioinformatics*

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

TODAYS MENU

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

Bioinformatics databases and associated tools
- Primary, secondary and composite databases.
  - Nucleotide sequence databases (GenBank & RefSeq).
  - Protein sequence database (UniProt).
  - Composite databases (PFAM & OMIM).

Database usage vignette
- Searching with ENTREZ and BLAST.
- Reference slides and handout on major databases.

HOMEWORK

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

Q. What is Bioinformatics?

Q. What is Bioinformatics?

"*Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.*"
[After Orengo, 2003]

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

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

MORE DEFINITIONS

- "Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying **"informatics" techniques** (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**. Luscombe NM, *et al.* Methods Inf Med. 2001;40:346.

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

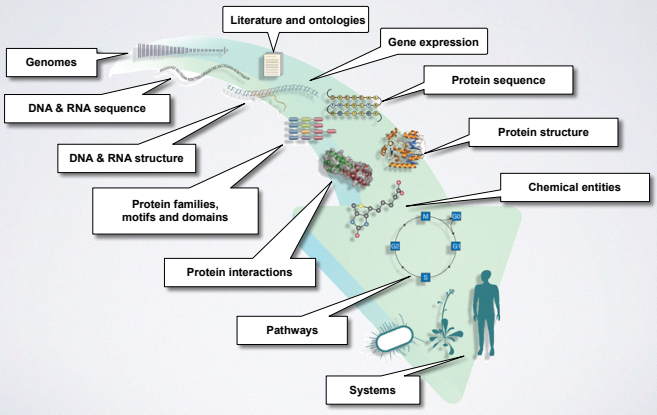
MORE DEFINITIONS

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

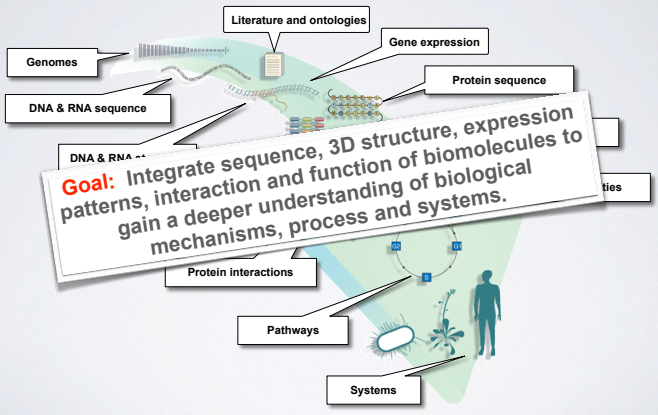
- "Bioinforma… search, development, or application of **comp… approaches** for expanding the use of **bi… medical**, **behavioral** or **health data**, including those to **a…quire**, **store**, **organize** and **analyze** such data." National Institutes of Health (NIH) (http://tinyurl.com/l3gxr6b)

Key Point: Bioinformatics is Computer Aided Biology

Major types of Bioinformatics Data

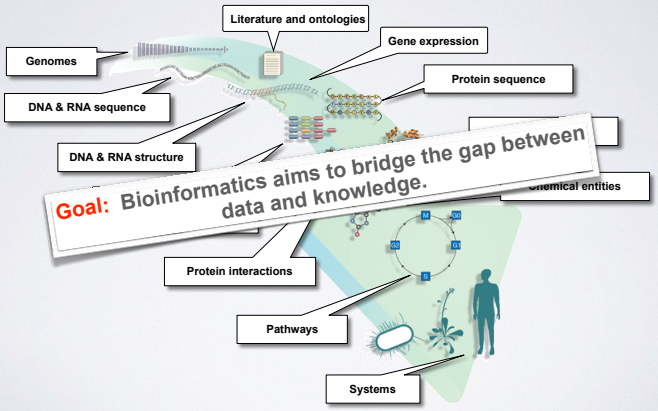


Major types of Bioinformatics Data



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 RESEARCH AREAS

Include but are not limited to:

- Organization, classification, dissemination and analysis of biological and biomedical data (particularly '-omics' data).
- Biological sequence analysis and phylogenetics.
- Genome organization and evolution.
- Regulation of gene expression and epigenetics.
- Biological pathways and networks in healthy & disease states.
- Protein structure prediction from sequence.
- Modeling and prediction of the biophysical properties of biomolecules for binding prediction and drug design.
- Design of biomolecular structure and function.

With applications to Biology, Medicine, Agriculture and Industry

Where did bioinformatics come from?

Bioinformatics arose as molecular biology began to be transformed by the emergence of molecular sequence and structural data

Recap: The key dogmas of molecular biology

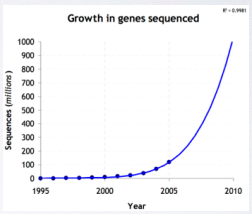
- *DNA sequence* determines *protein sequence.*
- *Protein sequence* determines *protein structure.*
- *Protein structure* determines *protein function.*
- *Regulatory mechanisms* (e.g. gene expression) determine the amount of a particular *function in space and time.*

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

Why do we need Bioinformatics?

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

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

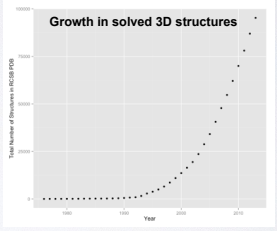


Bioinformatics is essential for the archiving, organization and analysis of data from sequencing, structural genomics, microarrays, proteomics and new high throughput assays.

Why do we need Bioinformatics?

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

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



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

How do we do Bioinformatics?

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

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

How do we actually do Bioinformatics?

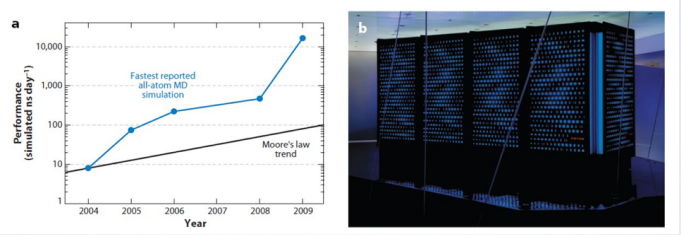
Pre-packaged tools and databases

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

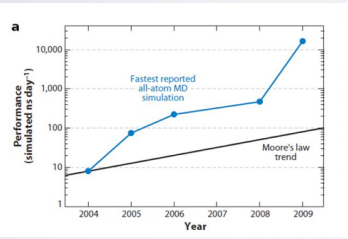
Tool development

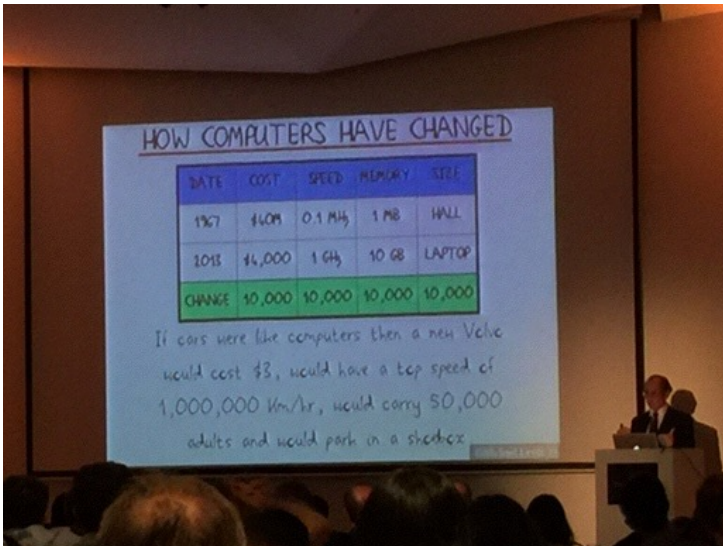
- Mostly on a UNIX environment
- Knowledge of programing languages frequently required (Python, Perl, R, C Java, Fortran)
- May require specialized or high performance computing resources…

SIDE-NOTE: SUPERCOMPUTERS AND GPUS



SIDE-NOTE: SUPERCOMPUTERS AND GPUS





Skepticism & Bioinformatics

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

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

Common problems with Bioinformatics

Confusing multitude of tools available

- Each with many options and settable parameters

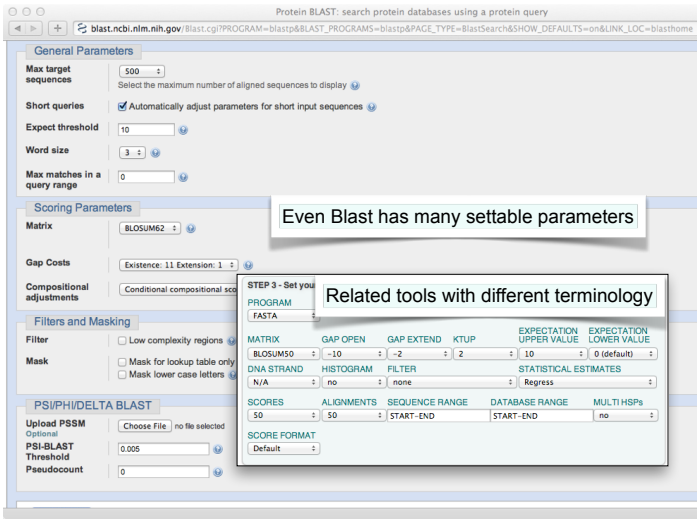
Most tools and databases are written by and for nerds

- Same is true of documentation - if any exists!

Most are developed independently

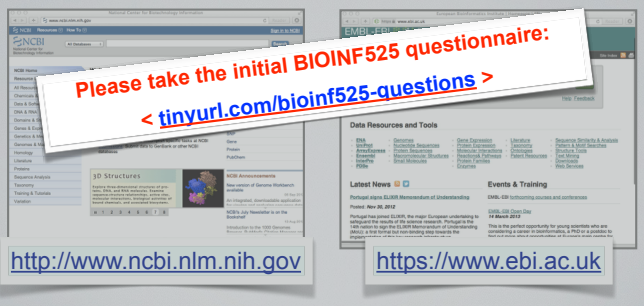
Notable exceptions are found at the:

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



Key Online Bioinformatics Resources: NCBI & EBI

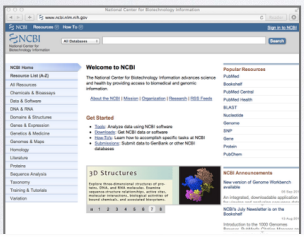
The NCBI and EBI are invaluable, publicly available resources for biomedical research

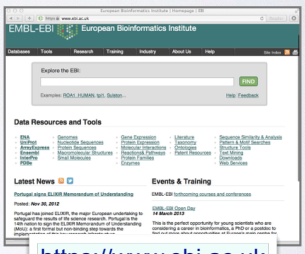


http://www.ncbi.nlm.nih.gov https://www.ebi.ac.uk

Key Online Bioinformatics Resources: NCBI & EBI

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





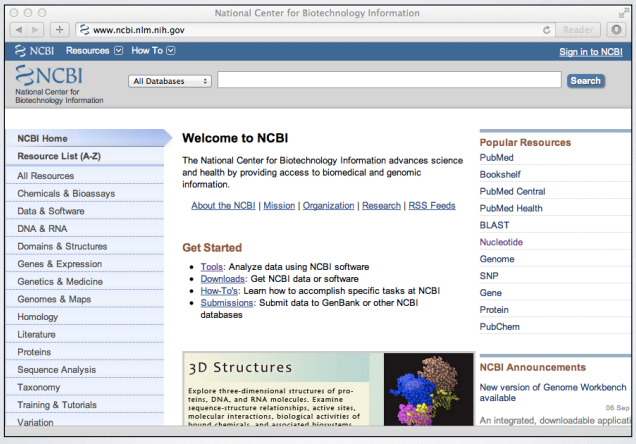
http://www.ncbi.nlm.nih.gov https://www.ebi.ac.uk

National Center for Biotechnology Information (NCBI)

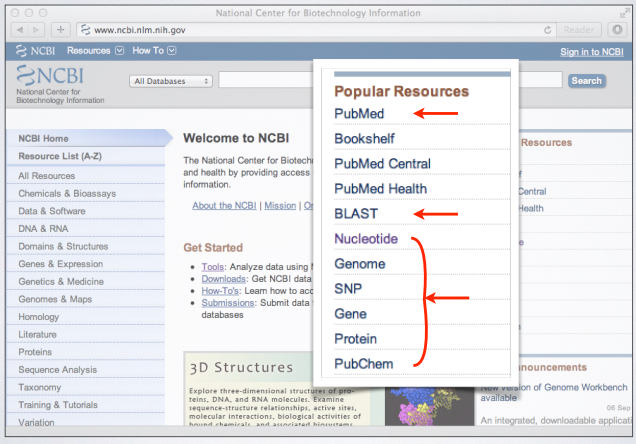
- Created in 1988 as a part of the National Library of Medicine (NLM) at the National Institutes of Health
- NCBI's mission includes:
  - Establish **public databases**
  - Develop **software tools**
  - **Education** on and dissemination of biomedical information

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

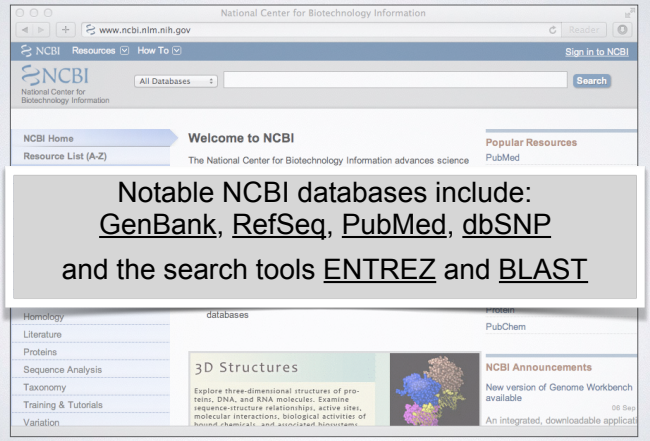
http://www.ncbi.nlm.nih.gov



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



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



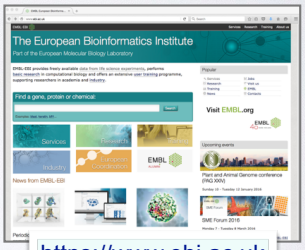
Notable NCBI databases include:
GenBank, RefSeq, PubMed, dbSNP

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



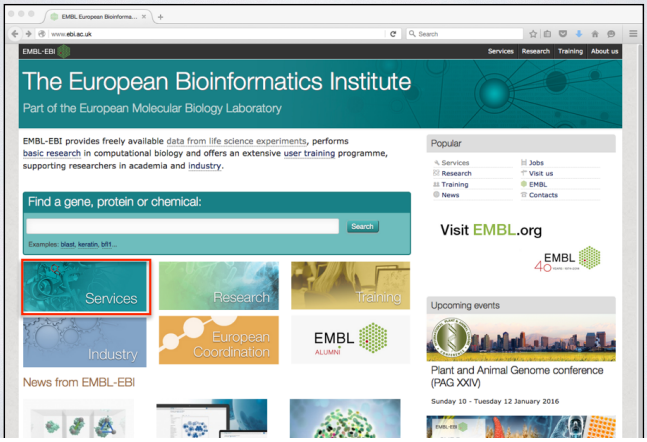
https://www.ebi.ac.uk

European Bioinformatics Institute (EBI)

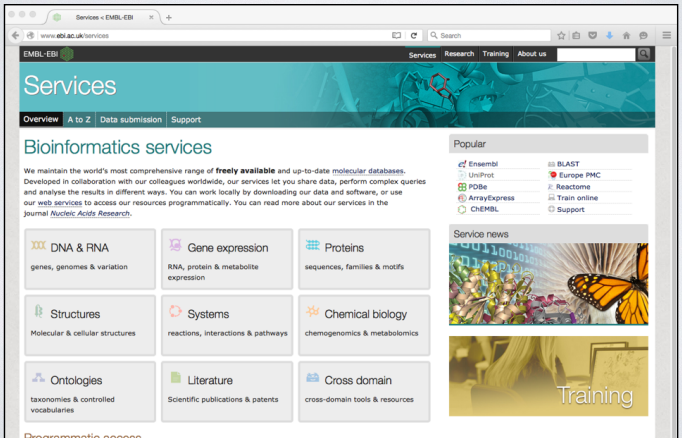
- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
  - providing freely available **data** and **bioinformatics services**
  - and providing advanced **bioinformatics training**

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

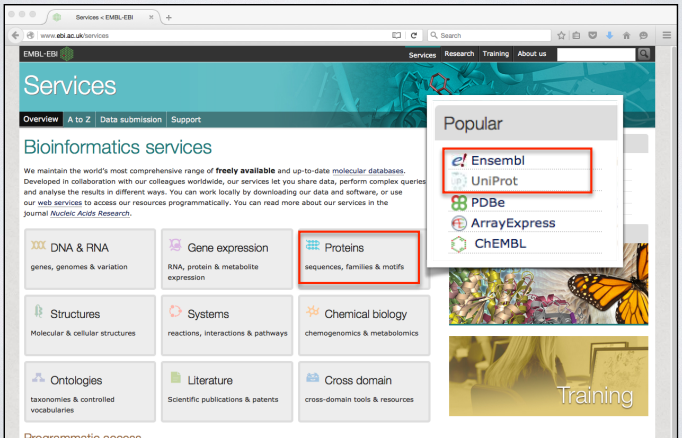
The EBI maintains a number of high quality curated **secondary databases** and associated tools



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

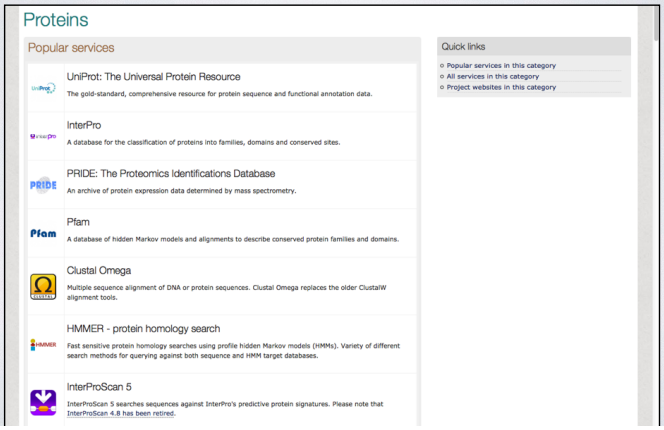


The EBI maintains a number of high quality curated **secondary databases** and associated tools

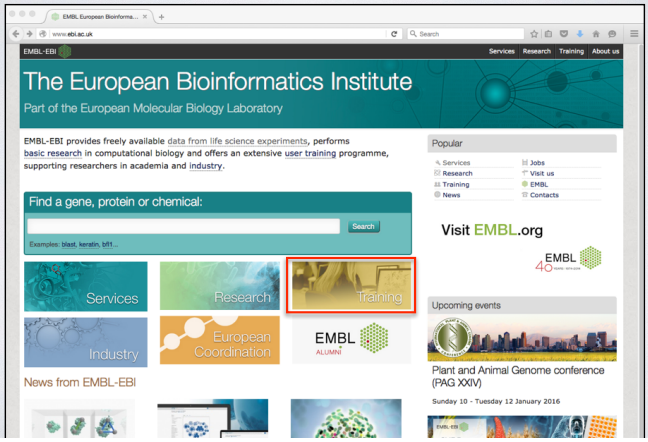


https://www.ebi.ac.uk

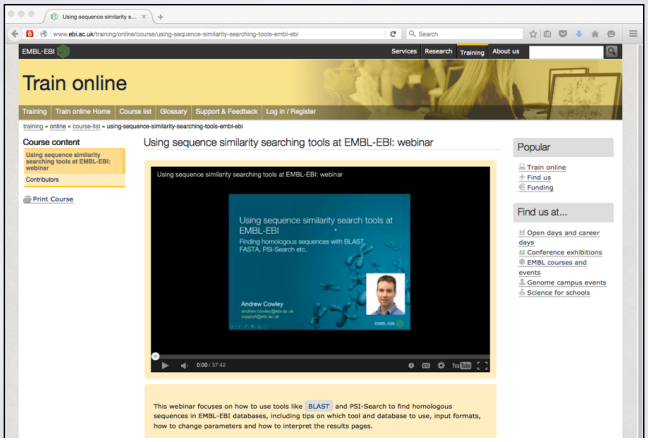
The EBI makes available a wider variety of **online tools** than NCBI



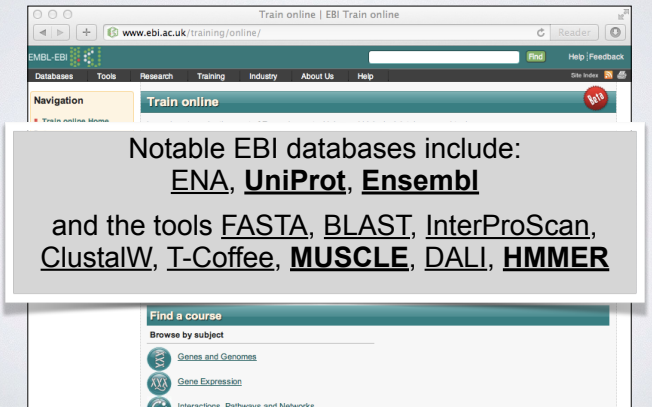
The EBI also provides a growing selection of **online tutorials** on EBI databases and tools



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, ClustalW, T-Coffee, **MUSCLE**, DALI, **HMMER**

BIOINFORMATICS DATABASES AND ASSOCIATED TOOLS

What is a database?

Computerized store of data that is organized to provide efficient retrieval.
- Uses standardized data (record) formats to enable computer handling

Key database features allow for:
- Adding, changing, removing and merging of records
- User-defined queries and extraction of specified records

Desirable features include:
- Contains the data you are interested in
- Allows fast data access
- Provides annotation and curation of entries
- Provides links to additional information (possibly in other databases)
- Allows you to make discoveries

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb,BBDB, BCGD, Beanref, 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 ..... !!!!

Bioinformatics Databases

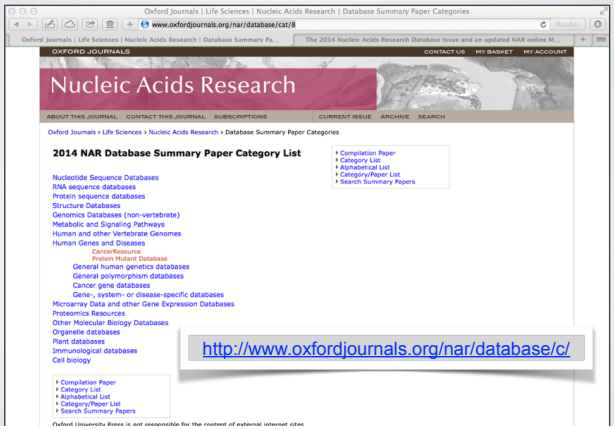
AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb,BBDB, BCGD, Beanref, BioImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, 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 ..... !!!!

There are lots of Bioinformatics Databases

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

< Handout_Major_Databases.pdf >

Finding Bioinformatics Databases



Major Molecular Databases

The most popular bioinformatics databases focus on:

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

The are also many popular "*boutique*" databases for:

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

Major Molecular Databases

The most popular bioinformatics databases focus on:

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

The are also many p... *...que*" databases for:

- Classifyin... ...ies, domains and motifs (e.g. PFAM, PROSITE)
- Spe... ...isms (e.g. WormBase, FlyBase)
- Specific proteins of biomedical importance (e.g. KinaseDB, GPCRDB)
- Specific diseases, mutations (e.g. OMIM, HGMD)
- Specific fields or methods of study (e.g. GOA, IEDB)

See Online: 'Handout_Major_Databases.pdf'

Primary, secondary & composite databases

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

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

GENBANK & REFSEQ: NCBI'S NUCLEOTIDE SEQUENCE DATABASES

What is GenBank?

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

GenBank, ENA and DDBJ Share and synchronize data

ENA ⟷ **GenBank** ⟷ **DDBJ**

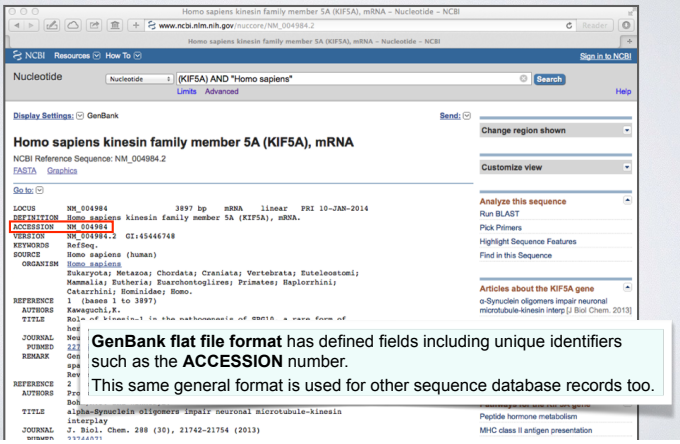
ENA — Housed at EBI — European Bioinformatics Institute

GenBank — Housed at NCBI — National Center for Biotechnology Information

DDBJ — Housed in Japan

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

GenBank sequence record



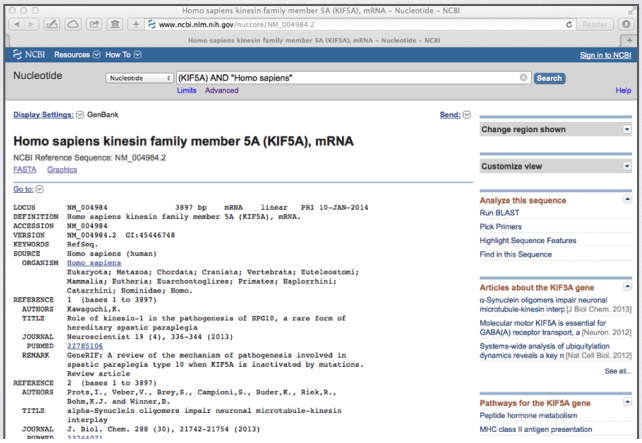
Side node: Database accession numbers

Database **accession numbers** are strings of letters and numbers used as **identifying labels** for sequences and other data within databases

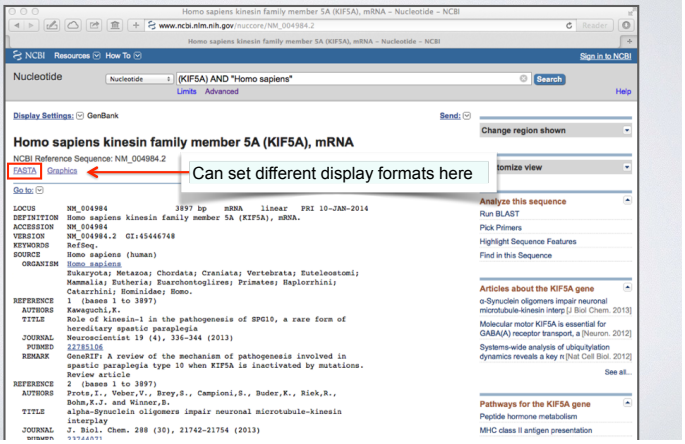
- Examples (all for retinol-binding protein, RBP4):

X02775	GenBank genomic DNA sequence	DNA
NT_030059	Genomic contig	
N91759.1	An expressed sequence tag (1 of 170)	RNA
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	Protein
AAC02945	GenBank protein	
Q28369	UniProtKB/SwissProt protein	
1KT7	Protein Data Bank structure record	
PMID: 12205585	PubMed IDs identify articles at NCBI/NIH	Literature

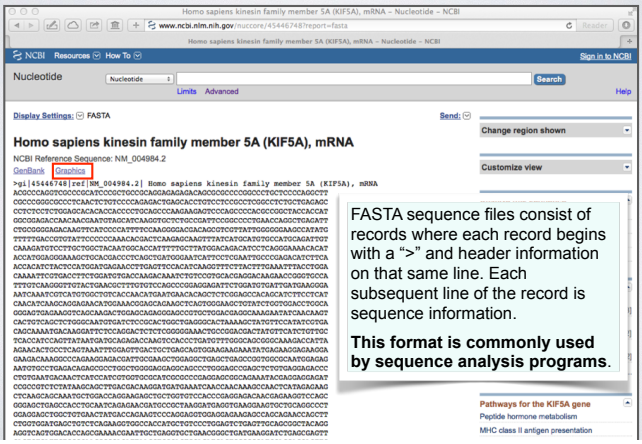
GenBank sequence record



GenBank sequence record



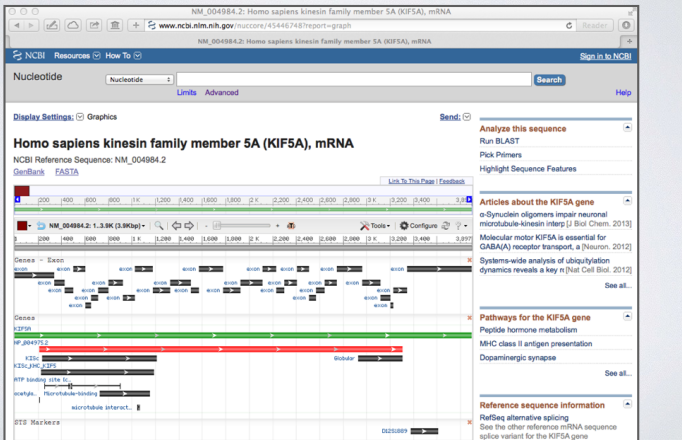
FASTA sequence record



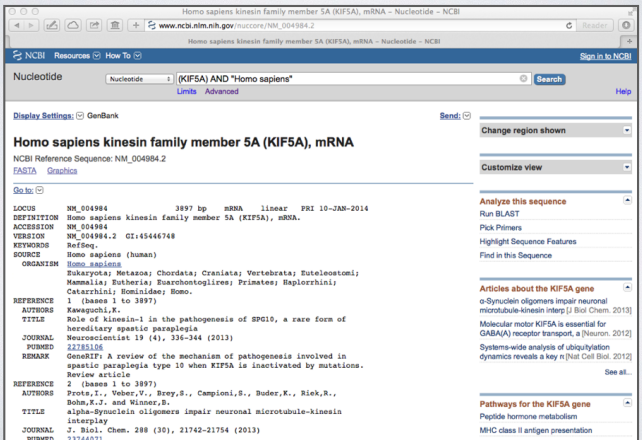
FASTA sequence files consist of records where each record begins with a “>” and header information on that same line. Each subsequent line of the record is sequence information.

This format is commonly used by sequence analysis programs.

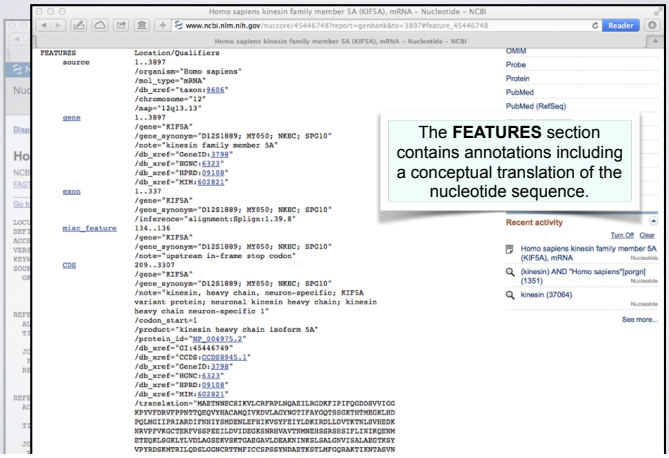
GenBank ‘graphics’ sequence record



GenBank sequence record, cont.

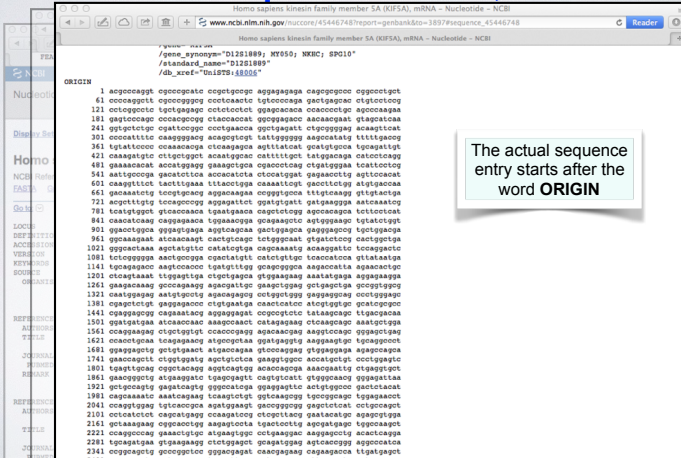


GenBank sequence record, cont.



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

GenBank sequence record, cont.



The actual sequence entry starts after the word **ORIGIN**

RefSeq: NCBI's Derivative Sequence Database

- RefSeq entries are hand curated best representation of a transcript or protein (in their judgement)
- Non-redundant for a given species although alternate transcript forms will be included if there is good evidence

- Experimentally verified transcripts and proteins
  accession numbers begin with "NM_" or "NP_"
- Model transcripts and proteins based on bioinformatics predictions with little experimental support
  accession numbers begin with "XM_" or "XP_"
- RefSeq also contains contigs and chromosome records

UNIPROT: THE PREMIER PROTEIN SEQUENCE DATABASE

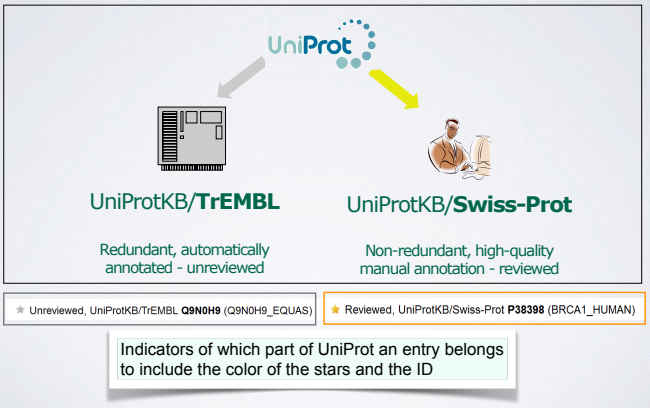
UniProt: Protein sequence database

UniProt is a comprehensive, high-quality resource of protein sequence and functional information

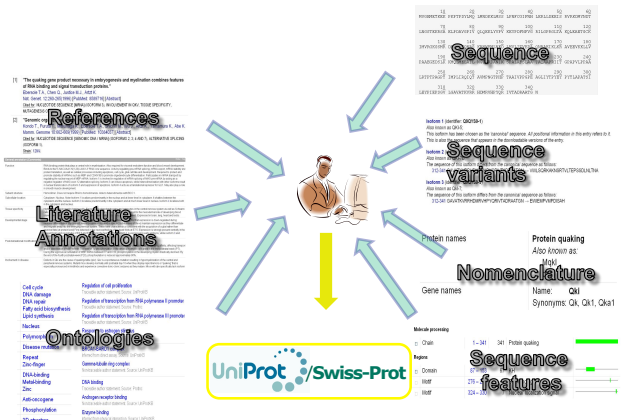
- UniProt comprises four databases:

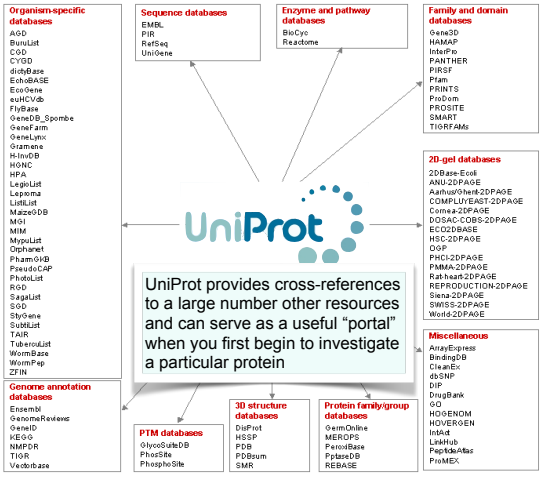
1. **UniProtKB** (Knowledgebase)
   Containing **Swiss-Prot** and **TrEMBL** components
   (these correspond to hand curated and automatically annotated entries respectively)
2. **UniRef** (Reference Clusters)
   Filtered version of UniProtKB at various levels of sequence identity
   e.g. UniRef90 contains sequences with a maximum of 90% sequence identity to each other
3. **UniParc** (Archive) with database cross-references to source.
4. **UniMES** (Metagenomic and Environmental Sequences)

The two sides of UniProtKB



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

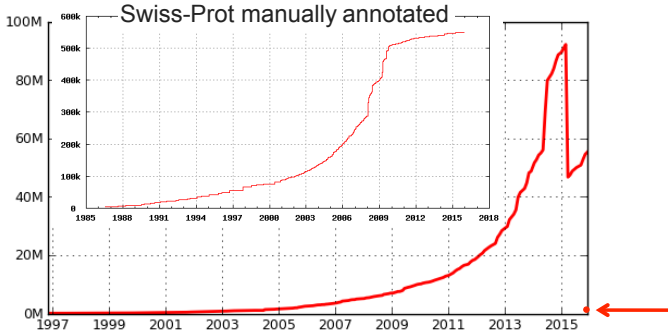




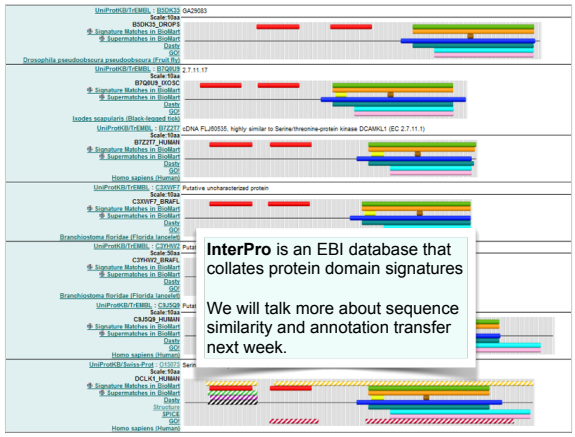
UniProt/Swiss-Prot *vs* UniProt/TrEMBL

- *UniProtKB/Swiss-Prot* is a **non-redundant** database with one entry per protein
- *UniProtKB/TrEMBL* is a **redundant** database with one entry per translated ENA entry (ENA is the EBI's equivalent of GenBank)
  - Therefore TrEMBL can contain multiple entries for the same protein
  - Multiple UniProtKB/TrEMBL entries for the same protein can arise due to:
    - Erroneous gene model predictions
    - Sequence errors (Frame shifts)
    - Polymorphisms
    - Alternative start sites
    - Isoforms
    - OR because the same sequence was submitted by different people

Side note: Automatic Annotation (sharing the wealth)



Same domain composition = same function = annotation transfer



InterPro is an EBI database that collates protein domain signatures

We will talk more about sequence similarity and annotation transfer next week.

DATABASE VIGNETTE

You have just come out a seminar about gastric cancer and one of your co-workers asks:

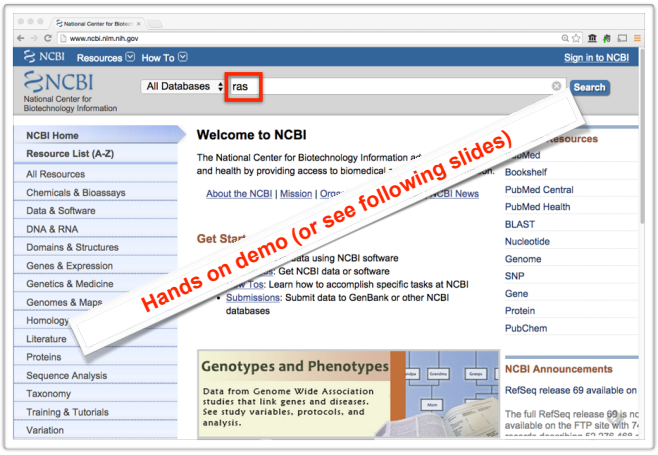
"What do you know about that 'Kras' gene the speaker kept taking about?"

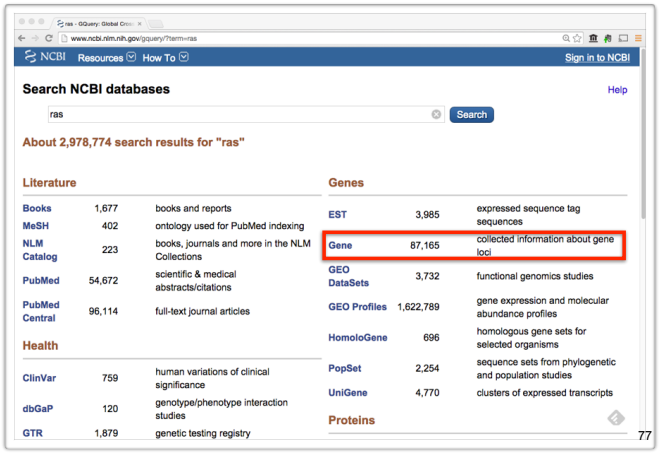
You have some recollection about hearing of 'Ras' before. How would you find out more?

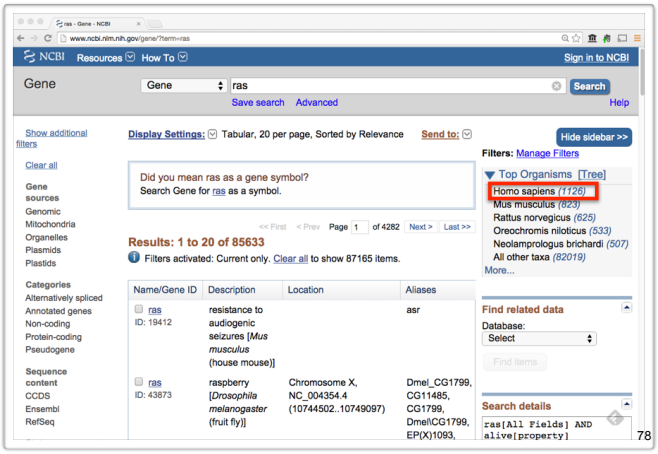
- Google?
- Library?
- **Bioinformatics databases at NCBI and EBI!**

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

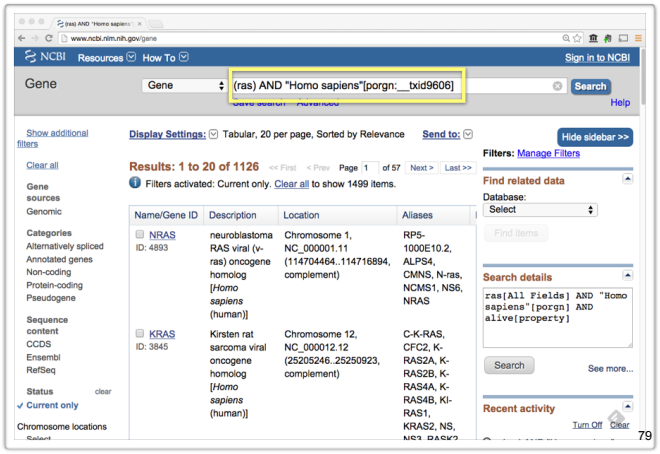
http://www.ncbi.nlm.nih.gov/



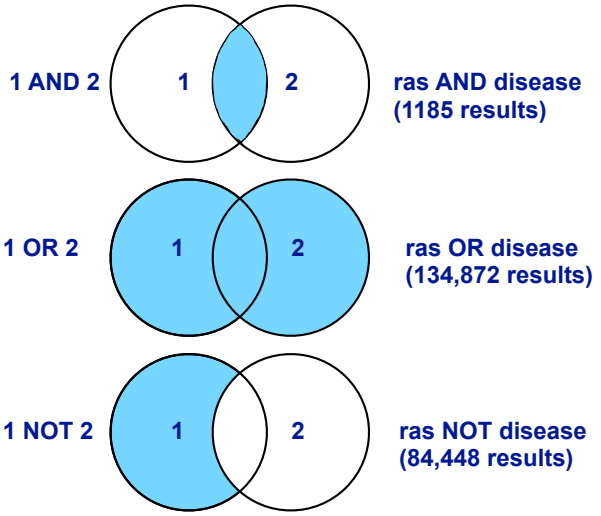




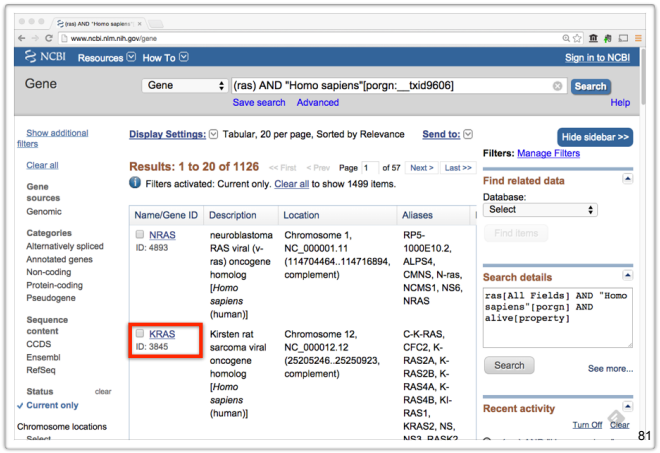
78



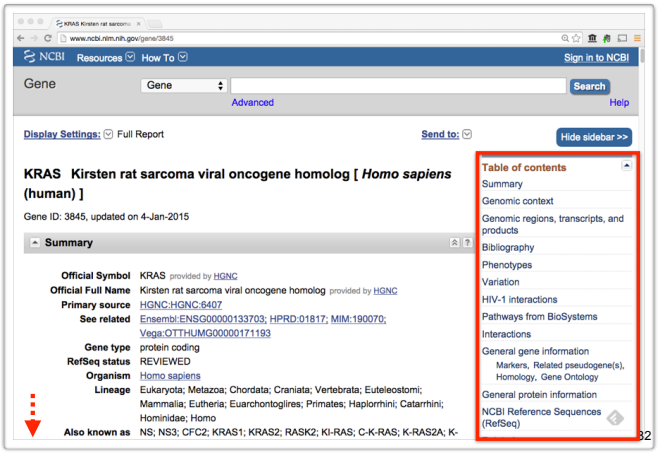
79



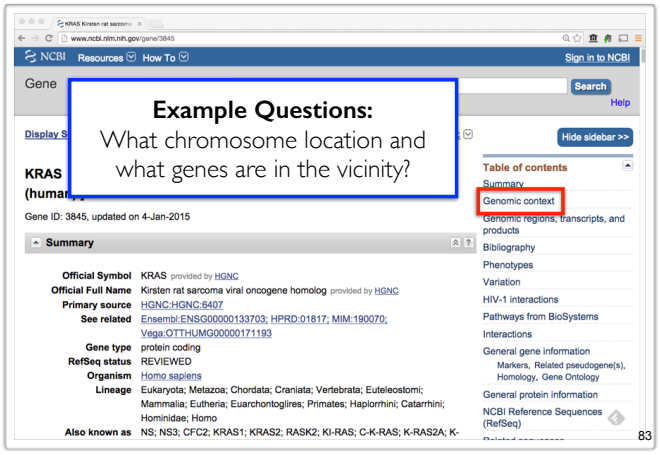
80



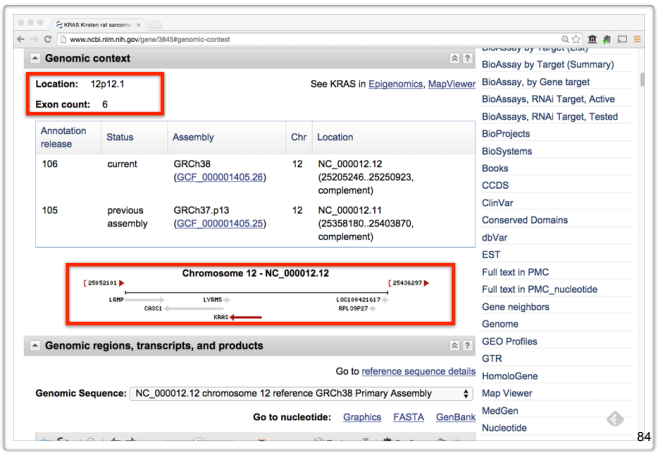
81



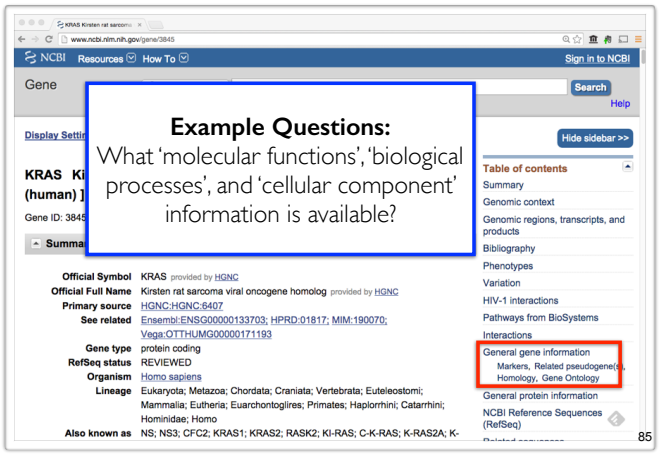
82



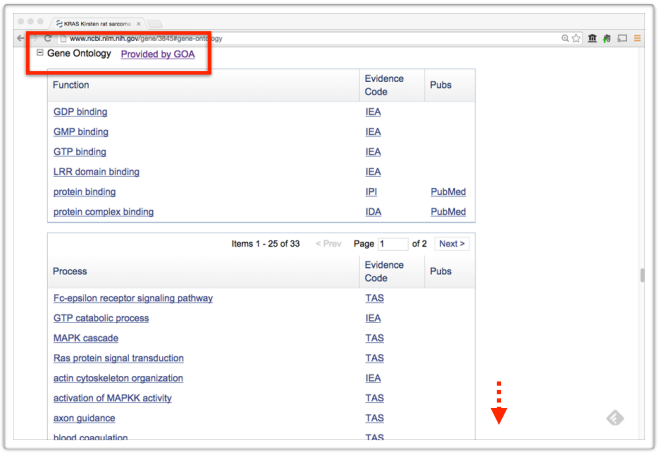
83



84

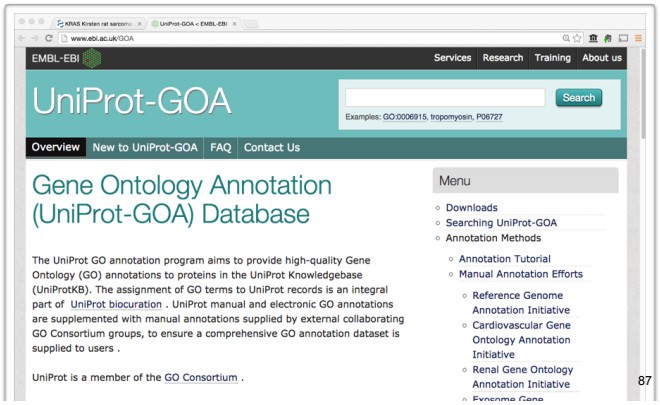


85



GO: Gene Ontology

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



87

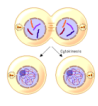
Why do we need Ontologies?

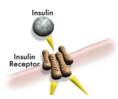
- Annotation is essential for capturing the understanding and knowledge associated with a sequence or other molecular entity
- Annotation is traditionally recorded as "free text", which is easy to read by humans, but has a number of disadvantages, including:
  - Difficult for computers to parse
  - Quality varies from database to database
  - Terminology used varies from annotator to annotator
- Ontologies are annotations using standard vocabularies that try to address these issues
- GO is integrated with UniProt and many other databases including a number at NCBI

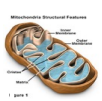
88

GO Ontologies

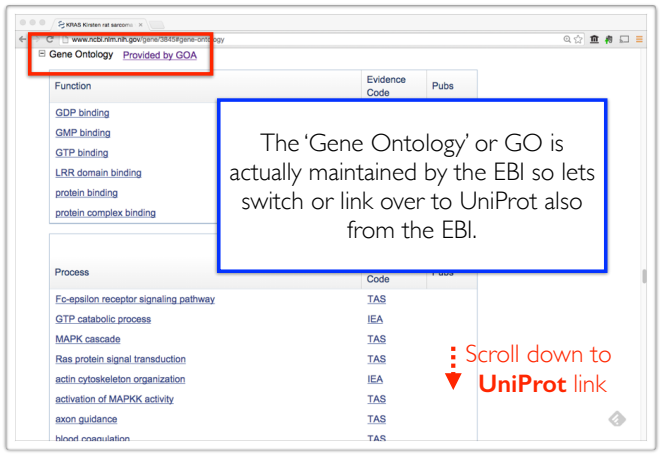
- There are three ontologies in GO:
  - **Biological Process**
    A commonly recognized series of events
    e.g. cell division, mitosis,
  - **Molecular Function**
    An elemental activity, task or job
    e.g. kinase activity, insulin binding
  - **Cellular Component**
    Where a gene product is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

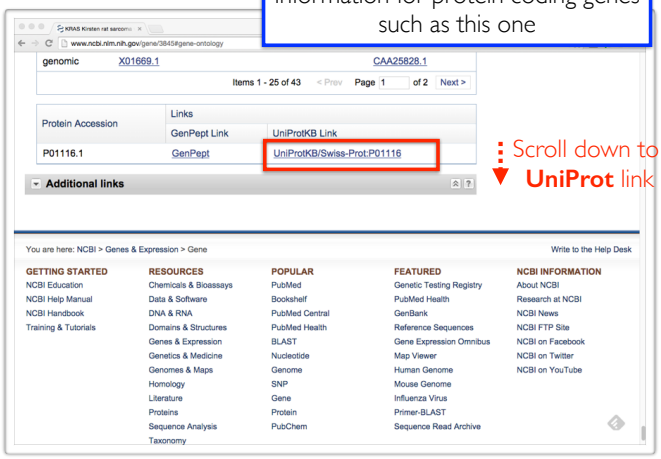


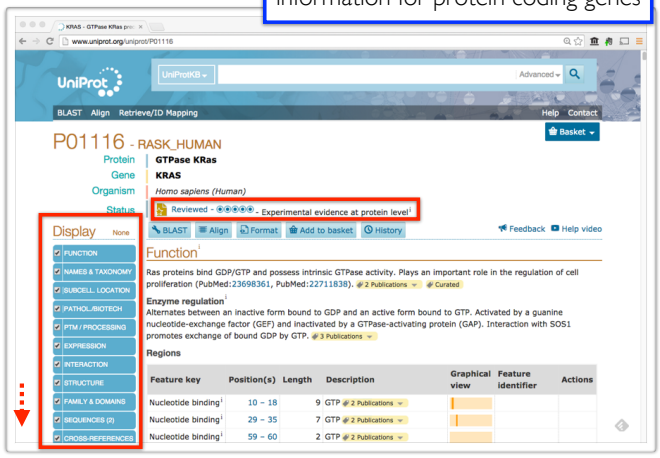


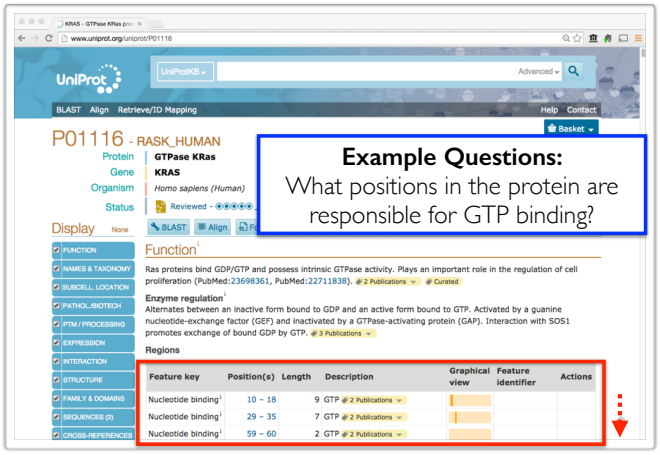


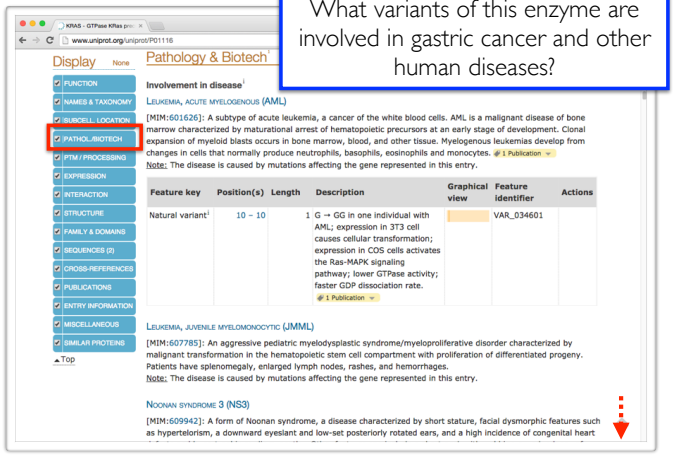
89

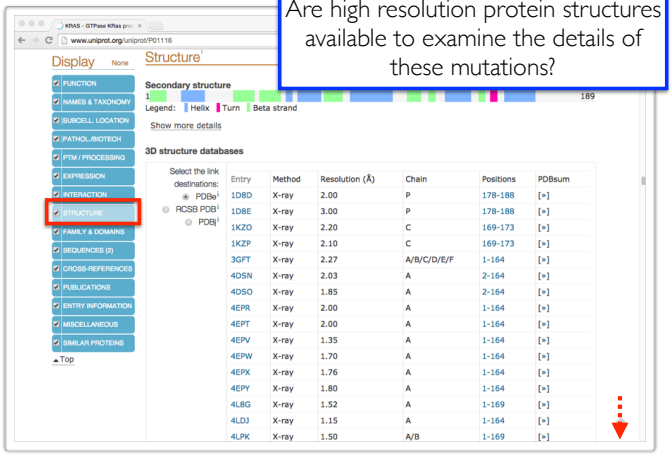

Gene Ontology Provided by GOA
Function
Evidence Code
Pubs
GDP binding
GMP binding
GTP binding
LRR domain binding
protein binding
protein complex binding
Process
Code
Fc-epsilon receptor signaling pathway
TAS
GTP catabolic process
IEA
MAPK cascade
TAS
Ras protein signal transduction
TAS
actin cytoskeleton organization
IEA
activation of MAPKK activity
TAS
axon guidance
TAS
blood coagulation
TAS
The 'Gene Ontology' or GO is actually maintained by the EBI so lets switch or link over to UniProt also from the EBI.
Scroll down to UniProt link


UniProt will detail much more information for protein coding genes such as this one
genomic
X01669.1
CAA25828.1
Items 1 - 25 of 43
Page 1 of 2
Protein Accession
Links
GenPept Link
UniProtKB Link
P01116.1
GenPept
UniProtKB/Swiss-Prot:P01116
Additional links
Scroll down to UniProt link
GETTING STARTED
RESOURCES
POPULAR
FEATURED
NCBI INFORMATION

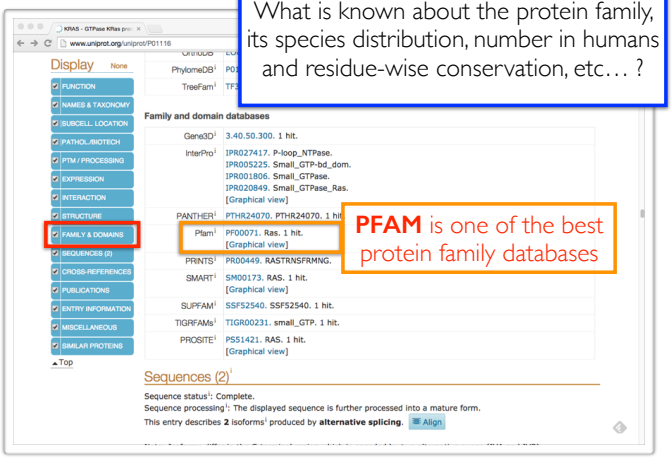

UniProt will detail much more information for protein coding genes
P01116 - RASK_HUMAN
Protein GTPase KRas
Gene KRAS
Organism Homo sapiens (Human)
Status Reviewed - Experimental evidence at protein level
Display
Function
Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23698361, PubMed:22711838).
Enzyme regulation
Regions
Feature key
Position(s)
Length
Description
Graphical view
Feature identifier
Actions
Nucleotide binding 10 – 18 9 GTP
Nucleotide binding 29 – 35 7 GTP
Nucleotide binding 59 – 60 2 GTP


Example Questions:
What positions in the protein are responsible for GTP binding?
P01116 - RASK_HUMAN
Protein GTPase KRas
Gene KRAS
Organism Homo sapiens (Human)
Function
Regions
Feature key
Position(s)
Length
Description
Nucleotide binding 10 – 18 9 GTP
Nucleotide binding 29 – 35 7 GTP
Nucleotide binding 59 – 60 2 GTP


Example Questions:
What variants of this enzyme are involved in gastric cancer and other human diseases?
Display
Pathology & Biotech
Involvement in disease
Leukemia, acute myelogenous (AML)
[MIM:601626]: A subtype of acute leukemia, a cancer of the white blood cells. AML is a malignant disease of bone marrow characterized by maturational arrest of hematopoietic precursors at an early stage of development. Clonal expansion of myeloid blasts occurs in bone marrow, blood, and other tissue. Myelogenous leukemias develop from changes in cells that normally produce neutrophils, basophils, eosinophils and monocytes.
Note: The disease is caused by mutations affecting the gene represented in this entry.
Feature key
Position(s)
Length
Description
Graphical view
Feature identifier
Actions
Natural variant 10 – 10 1 G → GG in one individual with AML; expression in 3T3 cell causes cellular transformation; expression in COS cells activates the Ras-MAPK signaling pathway; lower GTPase activity; faster GDP dissociation rate.
VAR_034601
Leukemia, juvenile myelomonocytic (JMML)
[MIM:607785]: An aggressive pediatric myelodysplastic syndrome/myeloproliferative disorder characterized by malignant transformation in the hematopoietic stem cell compartment with proliferation of differentiated progeny. Patients have splenomegaly, enlarged lymph nodes, rashes, and hemorrhages.
Note: The disease is caused by mutations affecting the gene represented in this entry.
Noonan syndrome 3 (NS3)
[MIM:609942]: A form of Noonan syndrome, a disease characterized by short stature, facial dysmorphic features such as hypertelorism, a downward eyeslant and low-set posteriorly rotated ears, and a high incidence of congenital heart

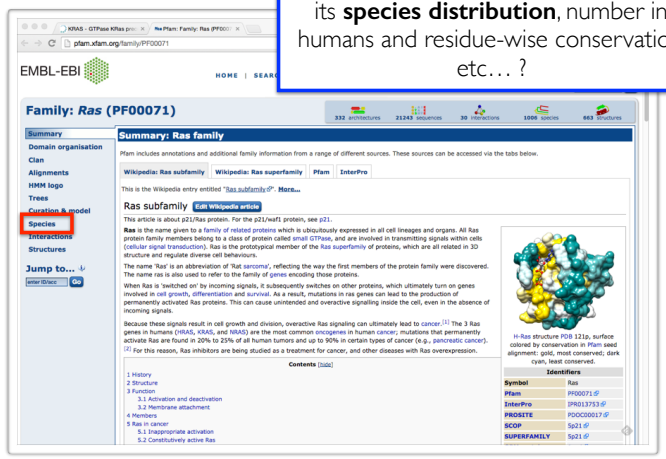

Example Questions:
Are high resolution protein structures available to examine the details of these mutations?
Display
Structure
Secondary structure
Legend: Helix Turn Beta strand
Show more details
3D structure databases
Select the link destinations:
PDBe
RCSB PDB
PDBj
Entry
Method
Resolution (Å)
Chain
Positions
PDBsum
1D8D 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 1-169
4LDJ X-ray 1.15 A 1-164
4LPK X-ray 1.50 A/B 1-169

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

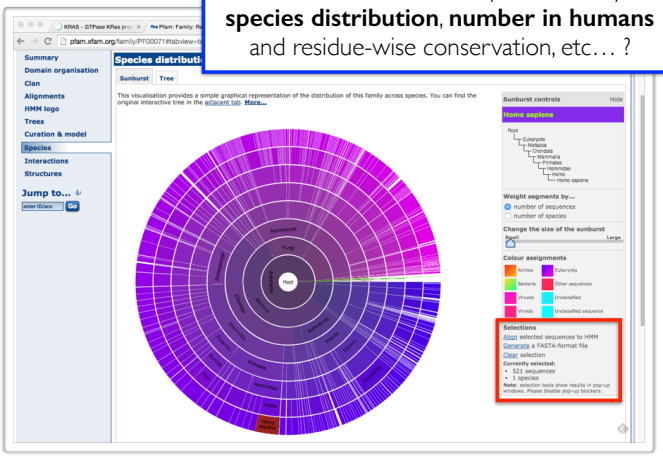


PFAM is one of the best protein family databases

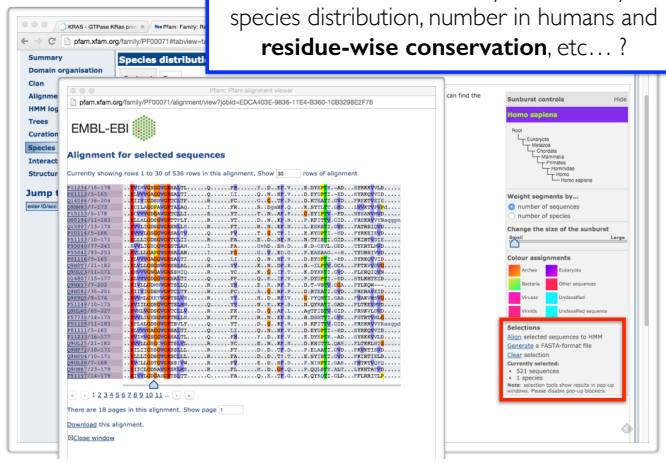
Example Questions:
What is known about the protein family, its **species distribution**, number in humans and residue-wise conservation, etc… ?



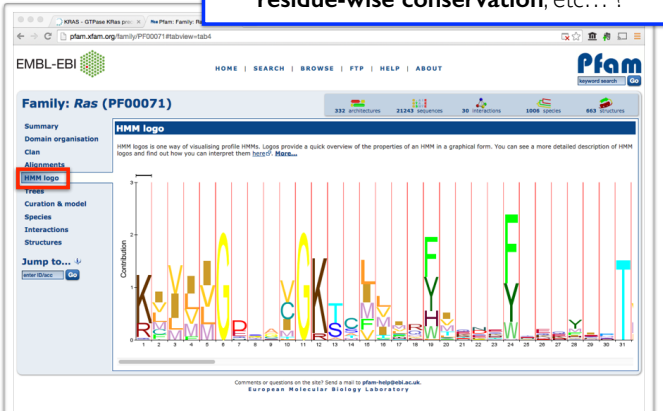
Example Questions:
What is known about the protein family, its **species distribution**, **number in humans** and residue-wise conservation, etc… ?

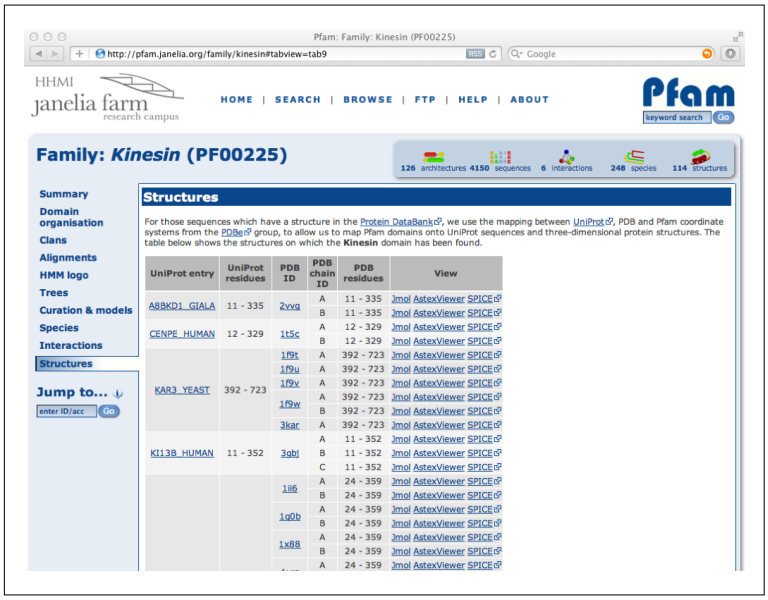


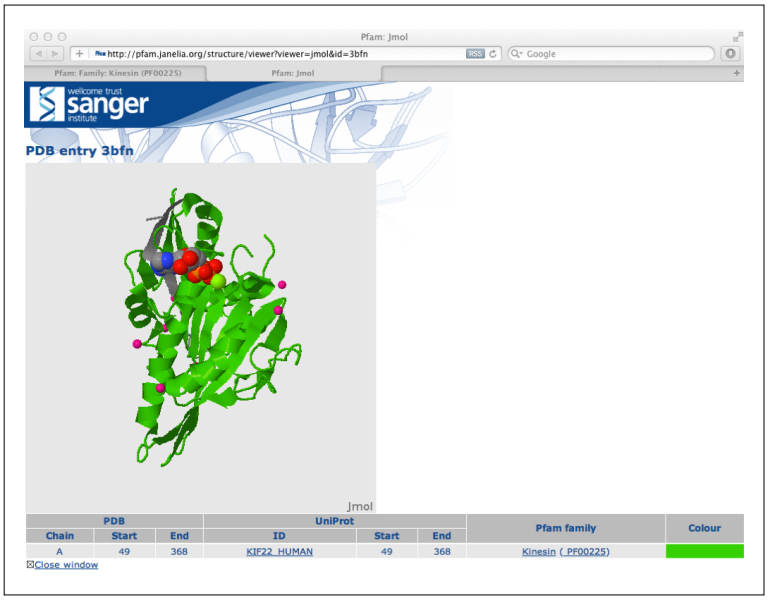
Example Questions:
What is known about the protein family, its species distribution, number in humans and **residue-wise conservation**, etc… ?



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

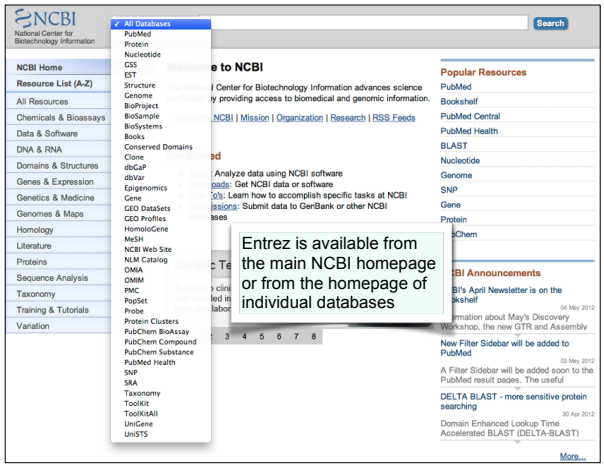




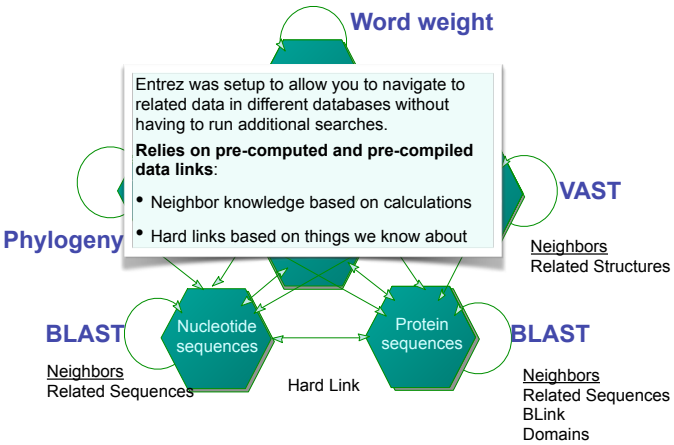


ENTREZ & BLAST:
TOOLS FOR SEARCHING AND ACCESSING MOLECULAR DATA AT NCBI

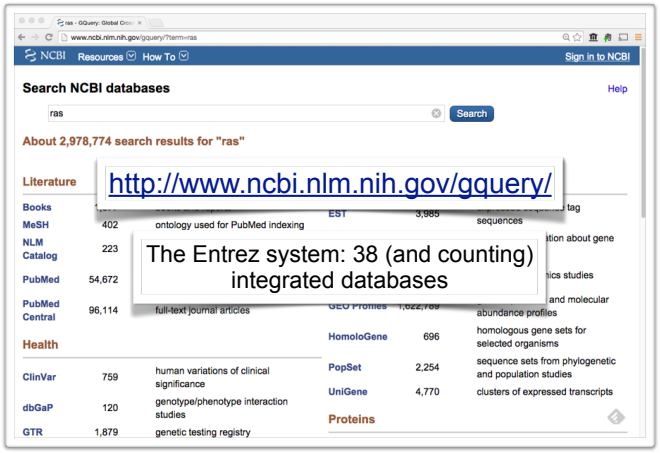
Entrez: Integrated search of NCBI databases



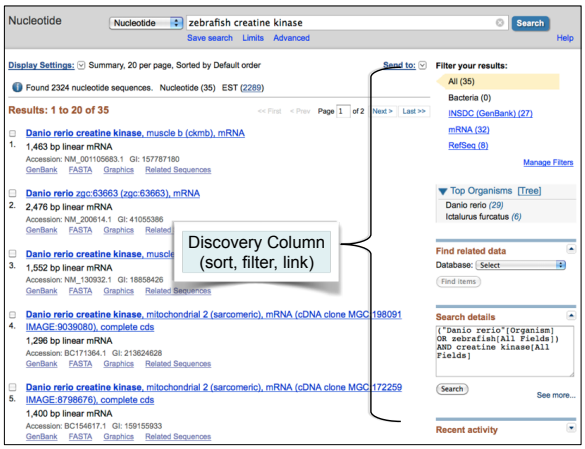
Entrez: navigating across databases



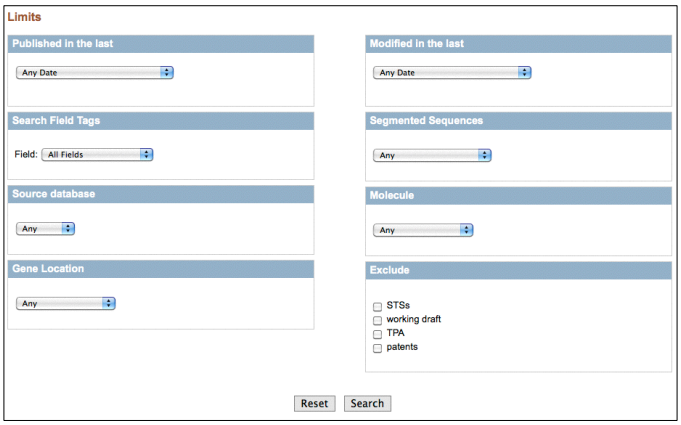
Global Entrez Query: All NCBI Databases



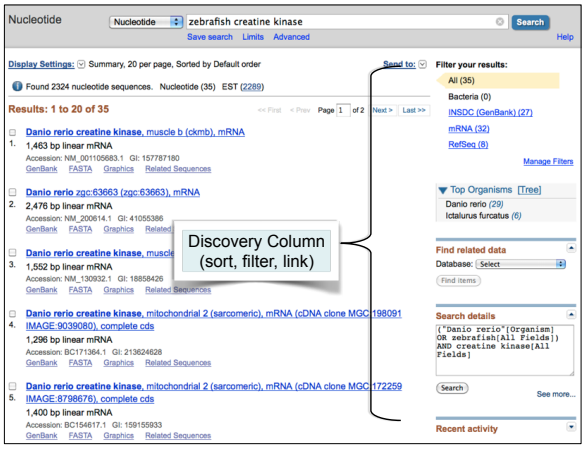
Search Results



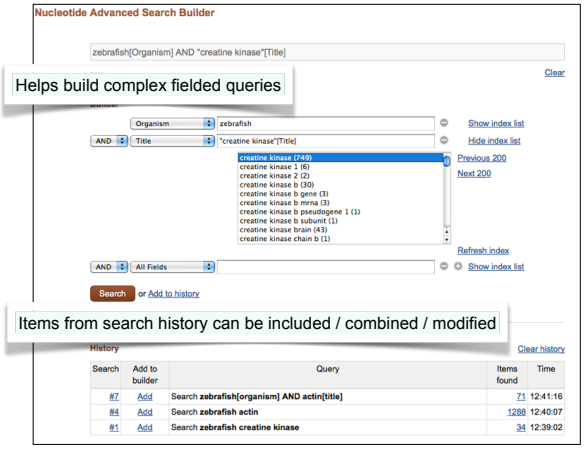
Limits



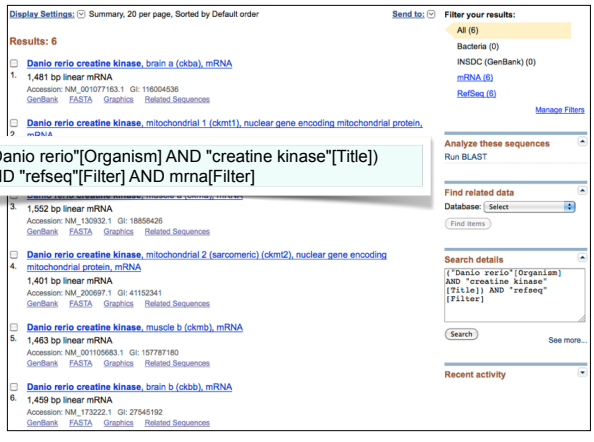
Search Results



Advanced: Search Builder

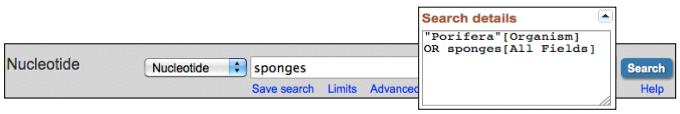


Complex Query Results

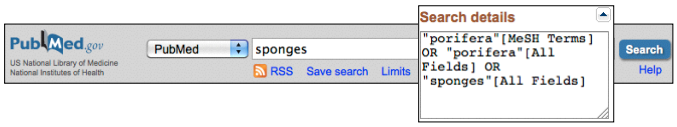


Controlled Vocabularies

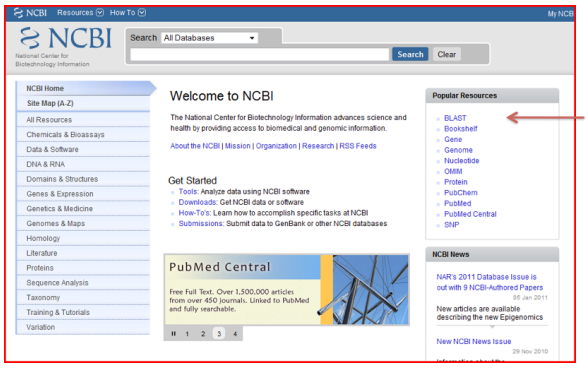
- Taxonomy primary controlled vocabulary / classification system for molecular databases at NCBI



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

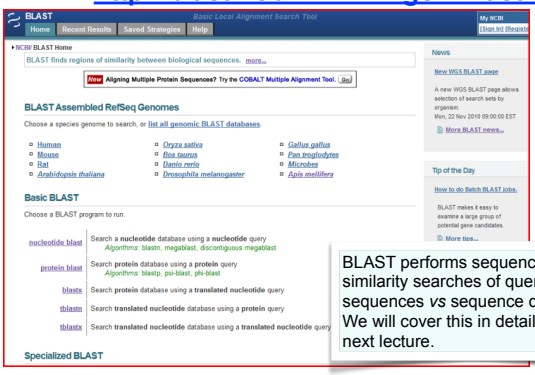


BLAST is a very important tool available from the NCBI Homepage
http://www.ncbi.nlm.nih.gov/guide/



BLAST – Basic Local Alignment Search Tool

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



BLAST performs sequence similarity searches of query sequences *vs* sequence databases. We will cover this in detail in the next lecture.

SUMMARY

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

HOMEWORK

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

ADDITIONAL DATABASES OF NOTE (SLIDES FOR YOUR REFERENCE)

NCBI Metadatabases

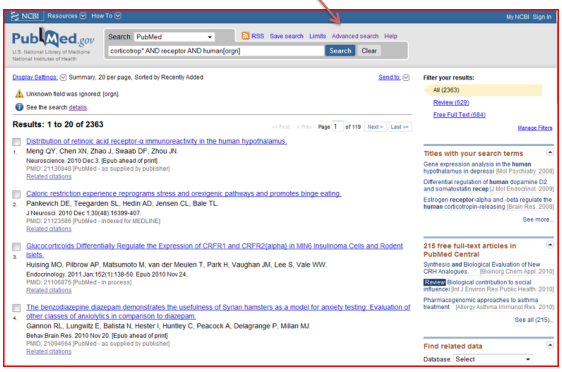
- **Gene**
  - molecular data and literature related to genes
- **HomoloGene**
  - automated collection of homologous genes from selected eukaryotes
- **Taxonomy**
  - access to NCBI data through source organism taxonomic classification
- **PubChem**
  - small organic molecules and their biological activities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

PubMed

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

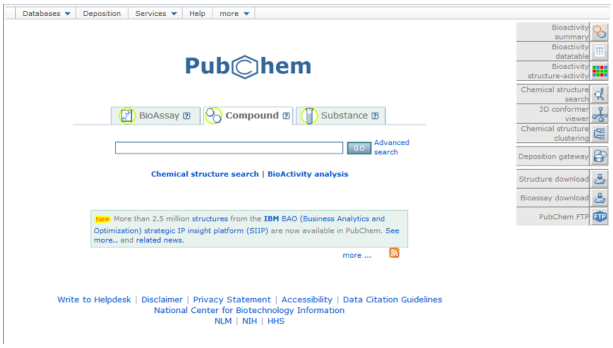
PubMed results

Limits and Advanced search can be used to refine searches

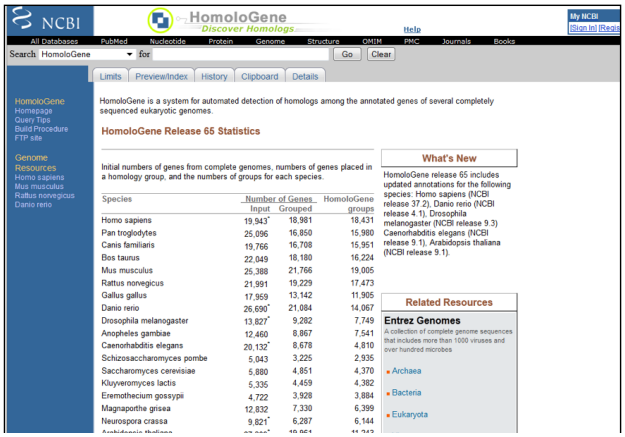


Small molecule databases have been added at NCBI

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

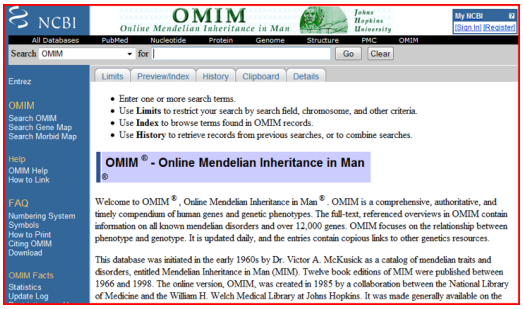


HomoloGene - Homologous genes from different organisms http://www.ncbi.nlm.nih.gov/homologene



Online Mendelian Inheritance in Man – OMIM

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



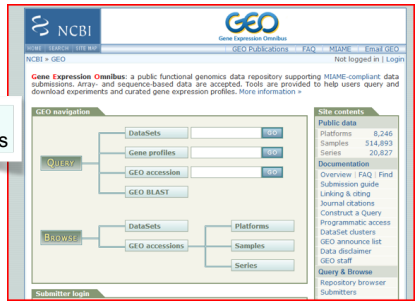
OMIM is essentially a set of reviews of human genes, gene function and phenotypes. Includes causative mutations where known.

The NCBI Bookshelf includes many well known molecular biology texts.

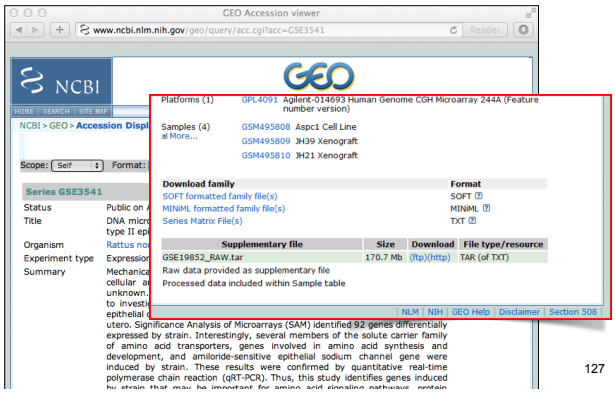
http://www.ncbi.nlm.nih.gov/books/

GEO: Gene Expression Omnibus

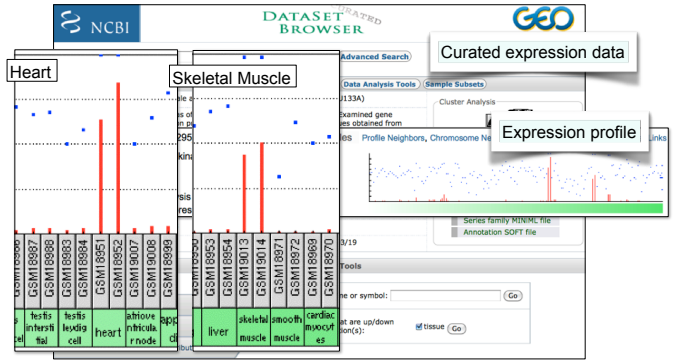
- Gene expression data (mostly from microarrays but also RNA-seq data, 2 methods for measuring RNA levels)



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



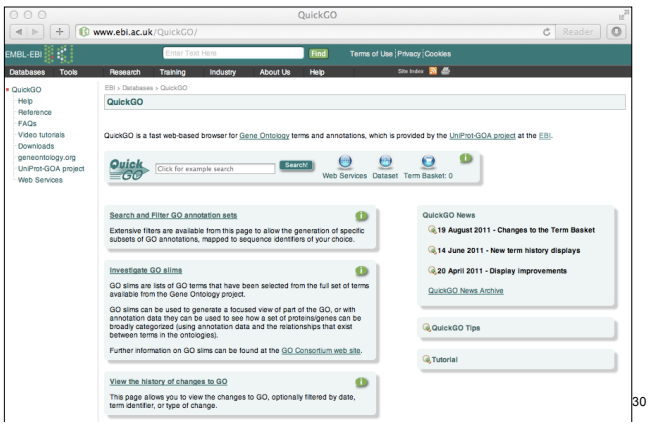
- **DataSets** - (GDSxxx) are curated collections of selected Samples that are biologically and statistically comparable



GO Ontologies

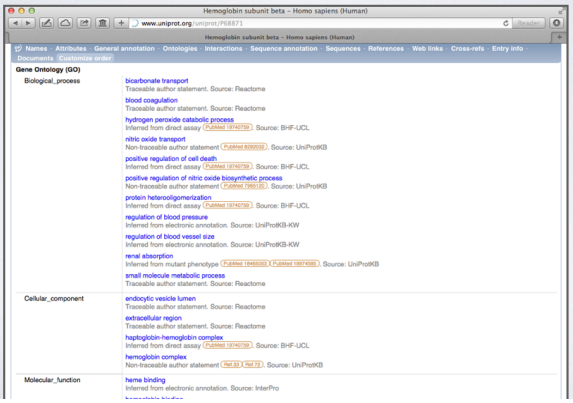
- There are three ontologies in GO:
  - **Biological Process**
    A commonly recognized series of events
    e.g. cell division, mitosis,
  - **Molecular Function**
    An elemental activity, task or job
    e.g. kinase activity, insulin binding
  - **Cellular Component**
    Where a gene product is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

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



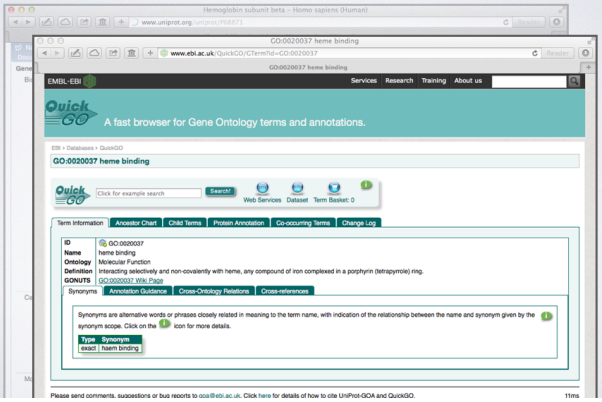
GO annotation in UniProt

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

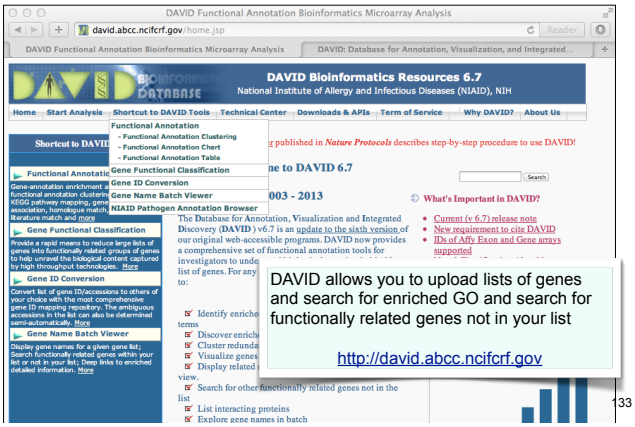


GO annotation in UniProt

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



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



Example output: enriched functions from GO