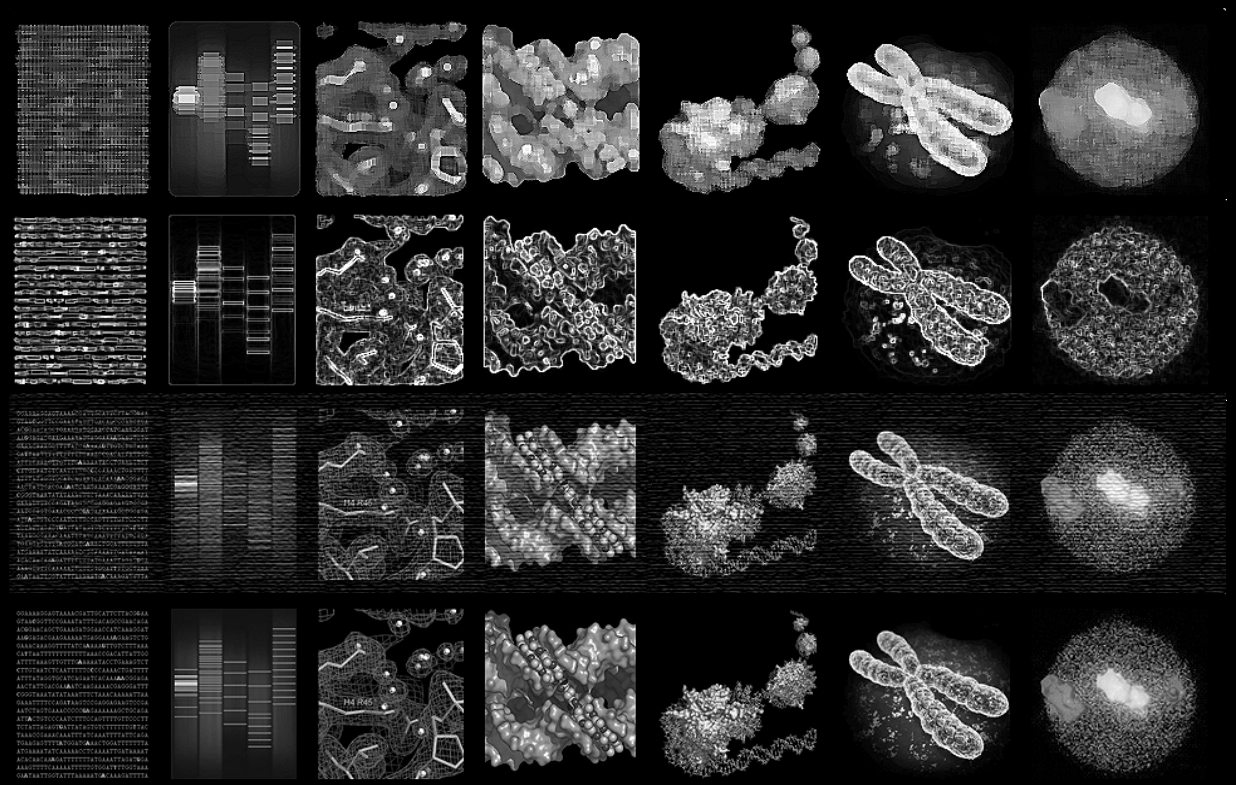
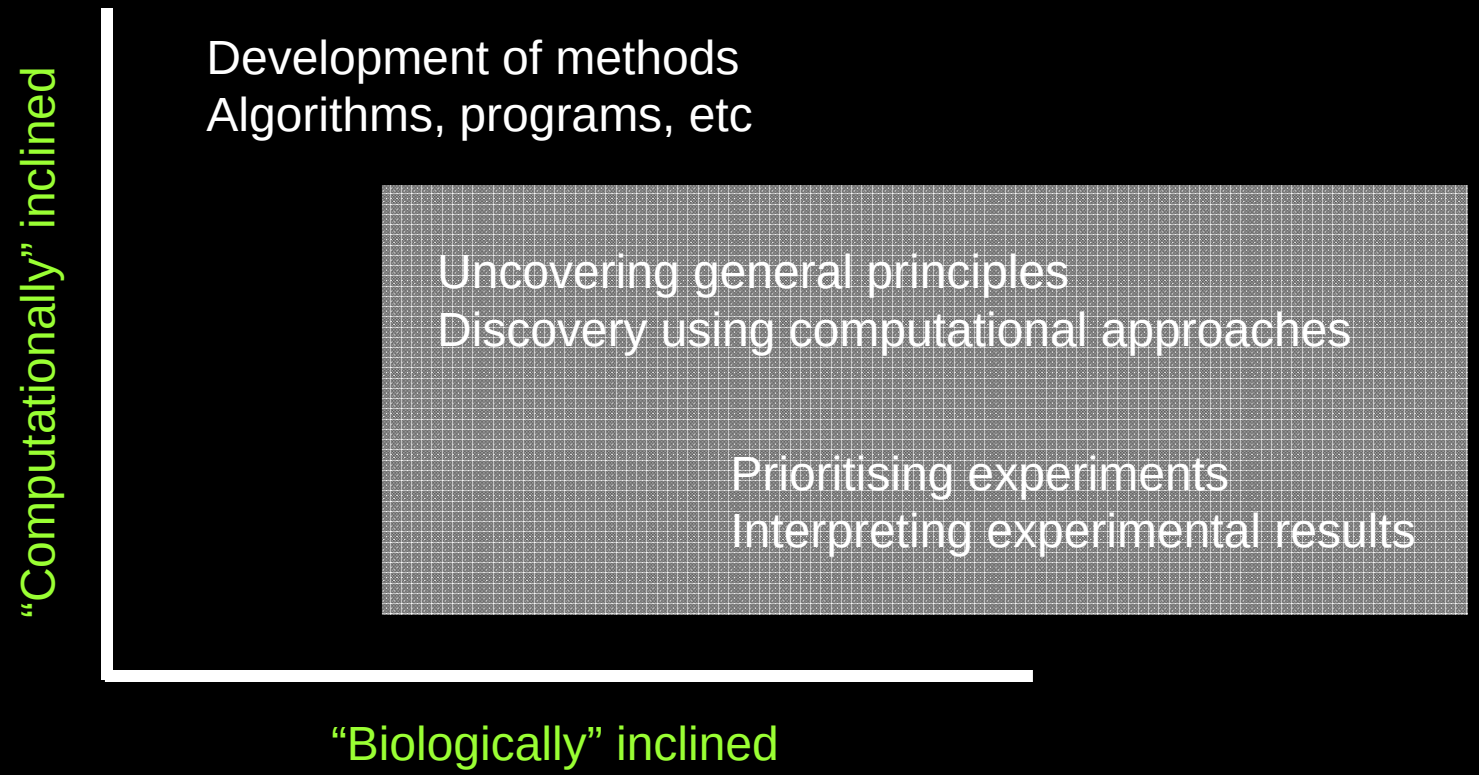
# Investigating questions in biology using computational approaches



M. Madan Babu

*Group Leader*
MRC Laboratory of Molecular Biology, Cambridge

# At what level is this talk pitched at?



# Outline

- Introduction to resources and tools (10 minutes)
- A case study to highlight data integration (10 mins)
- Specific questions (25 mins)

# How can I know more about a gene?

Knowing more about a gene is like trying to obtain all possible information about a suspect (as in a murder case) or a person (as in you are interested in someone ☺)

Treat it like investigating a case!

# Treat it like investigating a case! (suspect)

What is the suspect's name?

What does the suspect look like?

Where does the suspect live?

Where does the suspect work?

What does the suspect do?

Who does the suspect interact with?

What is the ancestry of the suspect?

# Treat it like investigating a case! (gene)

What is the sequence of the protein?

What is the structure of the protein?

Where in the genome is it encoded?

Which tissues are the genes expressed?

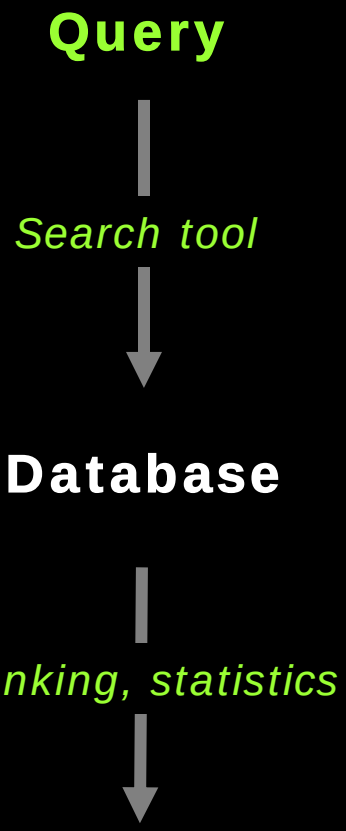
Which cellular compartment does it reside in?

What is the function of the protein?

Who are the interacting partners of the protein?

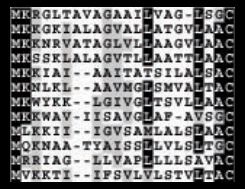
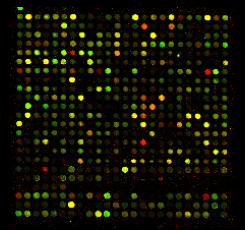
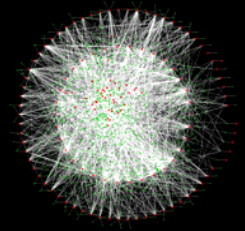
What is its evolutionary history?

# How can one extract information?



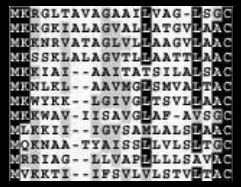

Information retrieval
(think of google)

# Explosion of information about living systems

| | | |
|---|---|---|
| Sequence | 45,000,000 sequences from<br>160,000 organisms (Genbank, NCBI) |  |
| Structure | 50,000 structures from<br>10,000 organisms (PDB, MSD) |  |
| Expression | 20,000 different conditions<br>150 organisms (SMD, GEO, ArrayExpress) |  |
| Interaction | 100,000 interactions<br>30 organisms (Bind, DIP, publications) |  |

Major challenge – How to exploit this information?

## Sequence



**Query sequence**

*BLAST, PSI-BLAST, etc*

Sequence database (NCBI, ENSEMBL)

*E-value, score, etc*

**Sequence Alignment**

## Structure

**Query structure**

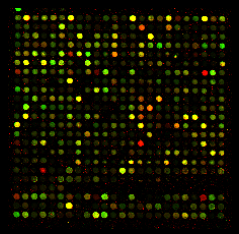
*DALI, SSM, VAST, etc*

Structure database (PDB)

*p-value, score, etc*

**Structure Alignments**

Expresion



**Query gene**

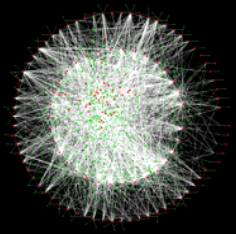
*Pearson correlation coefficient*

Expression database (GEO, BioGPS)

*Correlation coefficient*

**Other transcripts with Similar expression profile**

Network



**Gene list**

*Sub-graph extraction*

Functional networks, Protein interaction network, pathway database, etc

*P-value, enrichments*

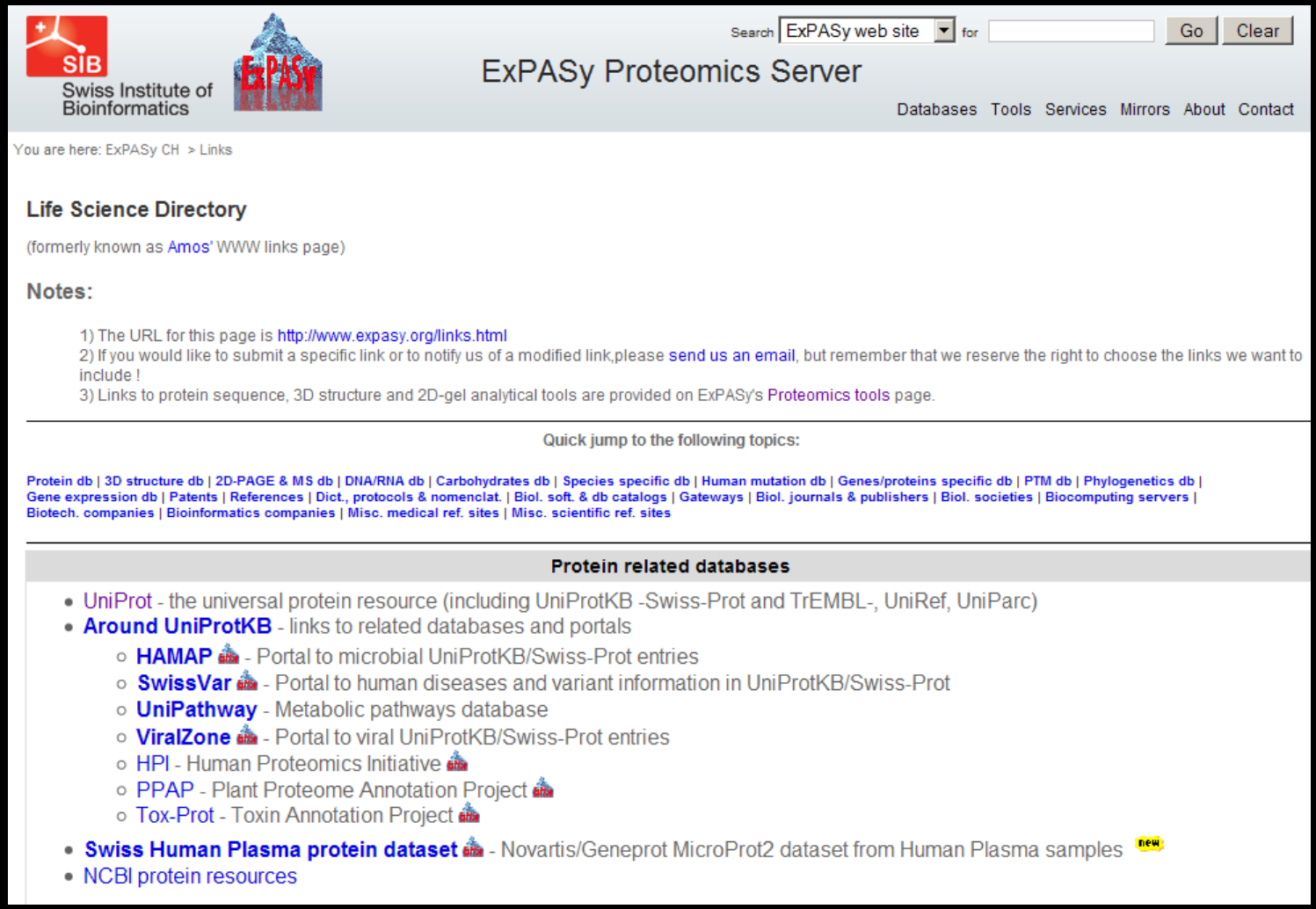
**Network of functional association Functional enrichment**

# Question #1

What is the list of databases
that is currently available?

# 1230 selected databases covering various aspects of molecular and cell biology


http://www.oxfordjournals.org/nar/database/c/




http://www.expasy.ch/links.html


http://database.oxfordjournals.org/

# Question #2

What is the list of tools, web-servers and programs that are currently available?

# 0ver 1500 selected tools covering various aspects of molecular and cell biology

bioinformatics.ca
links directory

Bioinformatics Links Directory

The Bioinformatics Links Directory features curated links to molecular resources, tools and databases. The links listed in this directory are selected on the basis of recommendations from bioinformatics experts in the field. We also rely on input from our community of bioinformatics users for suggestions. Starting in 2003, we have also started listing all links contained in the NAR Webserver issue.

Search Directory

- Main Page
- Citations
- Acknowledgements
- News
- Suggest a URL
- NAR Collaboration
- RSS Feeds
- Support

**Computer Related** (76)
This category contains links to resources relating to programming languages often used in bioinformatics. Other tools of the trade, such as web development and database resources, are also included here.

**DNA** (517)
This category contains links to useful resources for DNA sequence analyses such as tools for comparative sequence analysis and sequence assembly. Links to programs for sequence manipulation, primer design, and sequence retrieval and submission are also listed here.

**Education** (73)
Links to information about the techniques, materials, people, places, and events of the greater bioinformatics community. Included are current news headlines, literature sources, educational material and links to bioinformatics courses and workshops.

**Expression** (392)
Links to tools for predicting the expression, alternative splicing, and regulation of a gene sequence are found here. This section also contains links to databases, methods, and analysis tools for protein expression, SAGE, EST, and microarray data. Expression analysis of next-generation sequencing data sets is also covered.

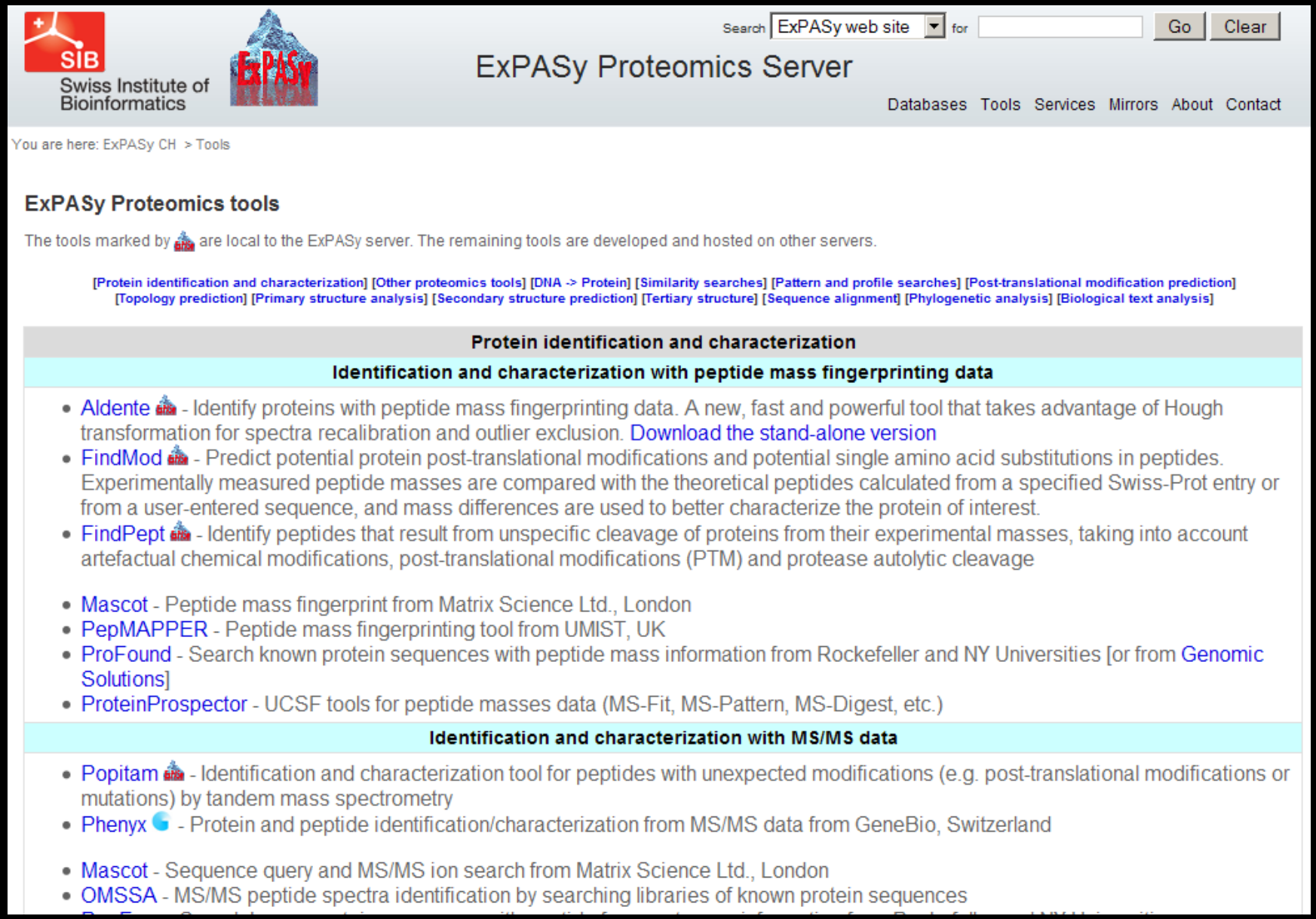
**Human Genome** (176)
This section contains links to draft annotations of the human genome in addition to resources for sequence polymorphisms and genomics. Also included are links related to ethical discussions surrounding the study of the human genome.

**Literature** (53)
Links to resources related to published literature, including tools to search for articles and through literature abstracts. Additional text mining resources, open access resources, and literature goldmines are also listed.


bioinformatics.ca

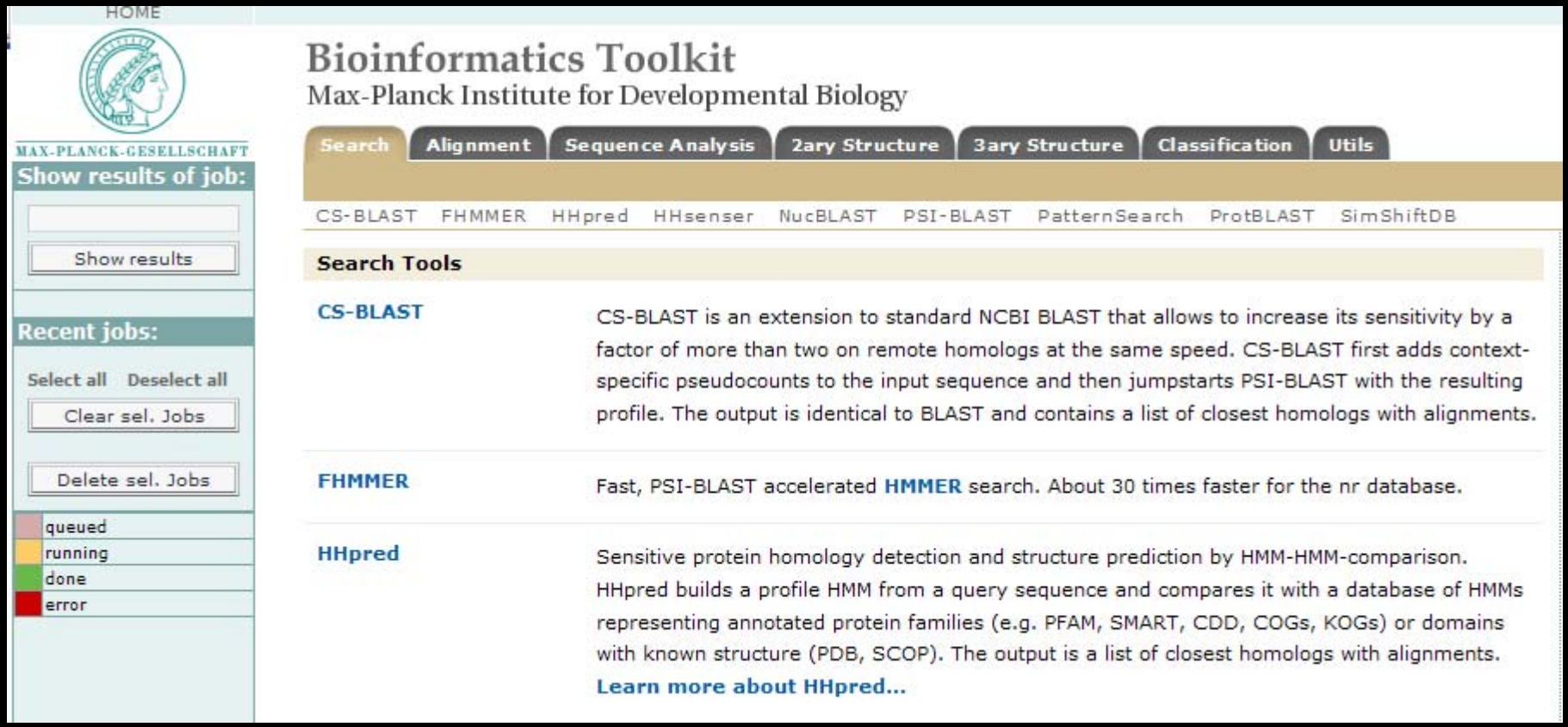
Return to BioInformatics.ca

http://bioinformatics.ca/links_directory/




http://www.expasy.ch/tools/

http://toolkit.tuebingen.mpg.de/sections/search



# Outline

- Introduction to resources and tools (10 minutes)
- A case study to highlight data integration (10 mins)
- Specific questions (25 mins)

Previous comparative genomic analysis of eukaryotes suggested lack of detectable transcription factors in Plasmodium

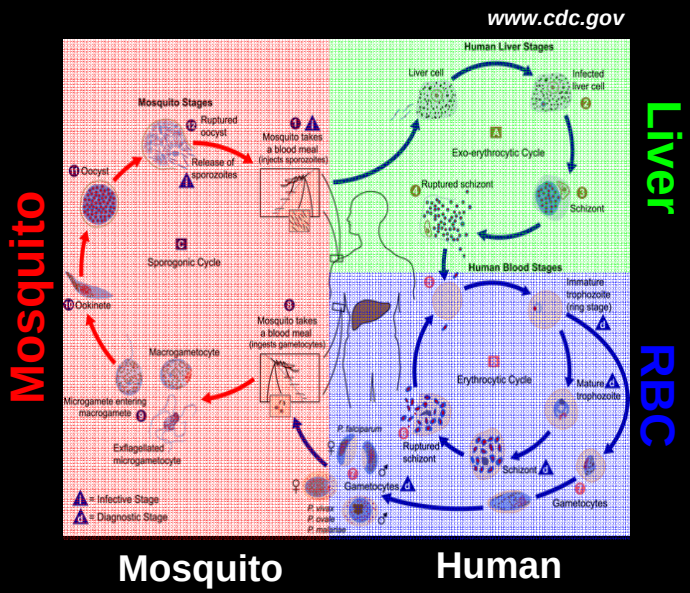
## Large number of genes


5300 genes with over 700 metabolic enzymes

Extensive complement of chromosomal regulatory proteins

Extensive complement signaling proteins (GTPases, kinases)

## Complex life cycle




# The Problem!

## How does this pathogen regulate gene expression?

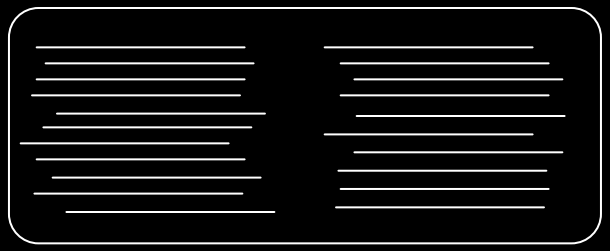
# Possible explanations for the paradoxical observation

## Alternative regulatory mechanisms

Chromatin-level regulation
Post-translational modification
RNA based regulation

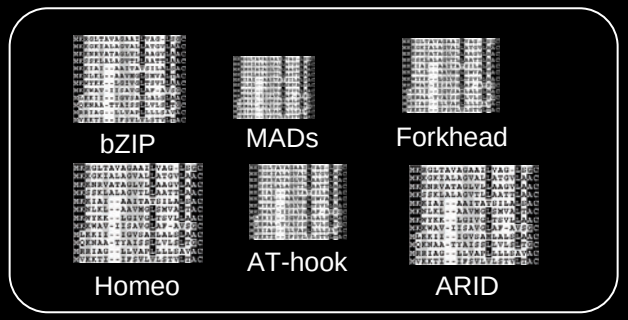
## Undetected transcription factors

Distantly related or unrelated
to known DNA binding domains

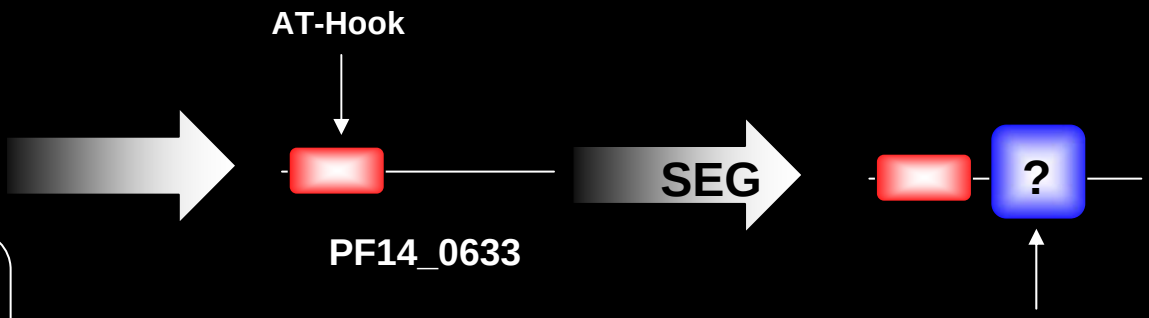


Proteome of Plasmodium

+

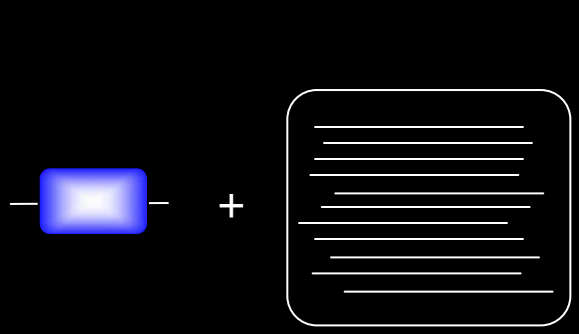



Profiles & HMMs of known DBDs

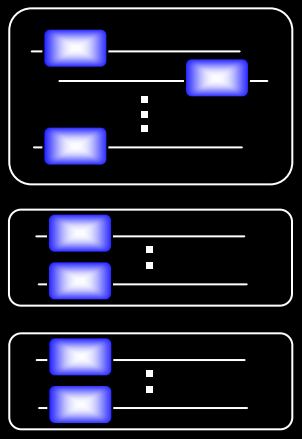



The suspect!

# Characterization of the globular domain – sequence analysis I



Non-redundant database

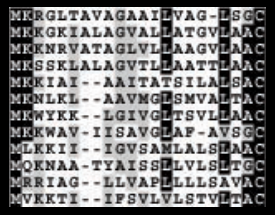


*Plasmodium falciparum*
*Plasmodium vivax*

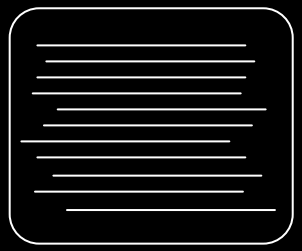
*Cryptosporidium parvum*
*Cryptosporidium hominis*

*Theileria annulata*

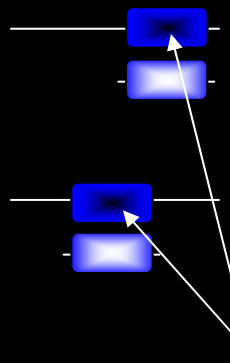
Lineage specific expansion in Apicomplexa



Profiles + HMM of this region



Non-redundant database



*Floral Homeotic protein Q (Triticum)*

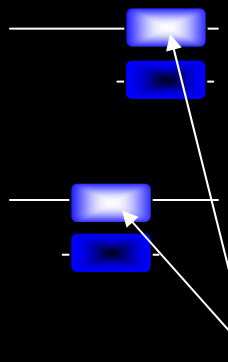
*49L, an endonuclease (X. oryzae phage Xp10)*

Globular region maps to AP2 DNA-binding domain

# Characterization of the globular domain – sequence analysis II

AP2 DNA-binding Domain from *D. Psychrophila* DP2593

Non-redundant database



*MAL6P1.287*
*(Plasmodium falciparum)*

*Cgd6_1140/Chro.60146*
*(Cryptosporidium)*

AP2 DNA-binding domain maps to the Globular region

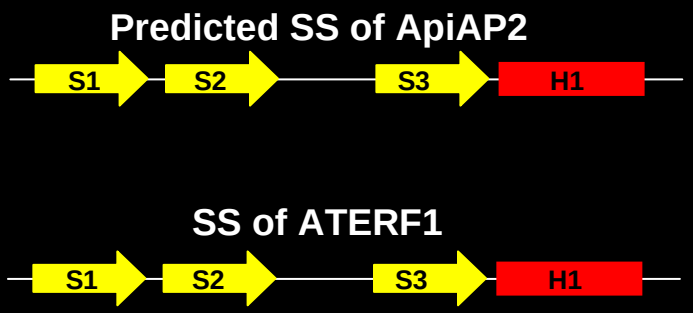
Multiple sequence alignment of all globular domains

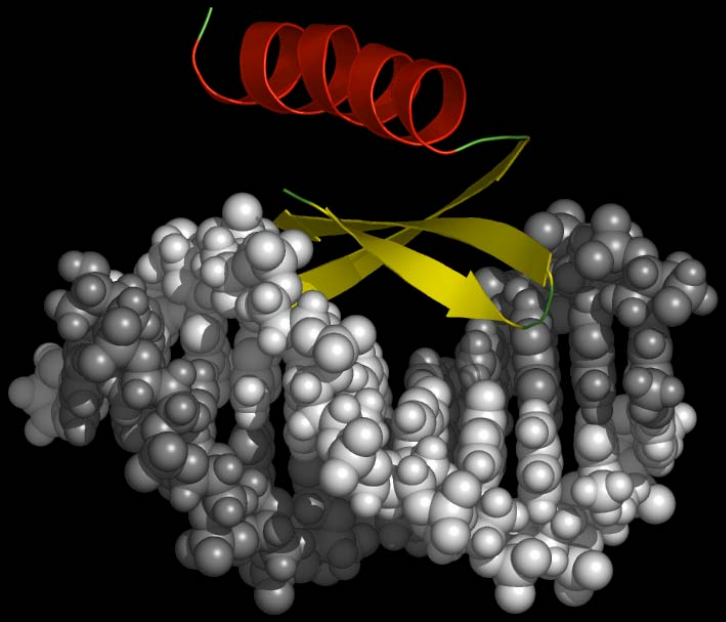
JPRED/PHD


Sequence of secondary structure is similar to the AP2 DNA-binding domain

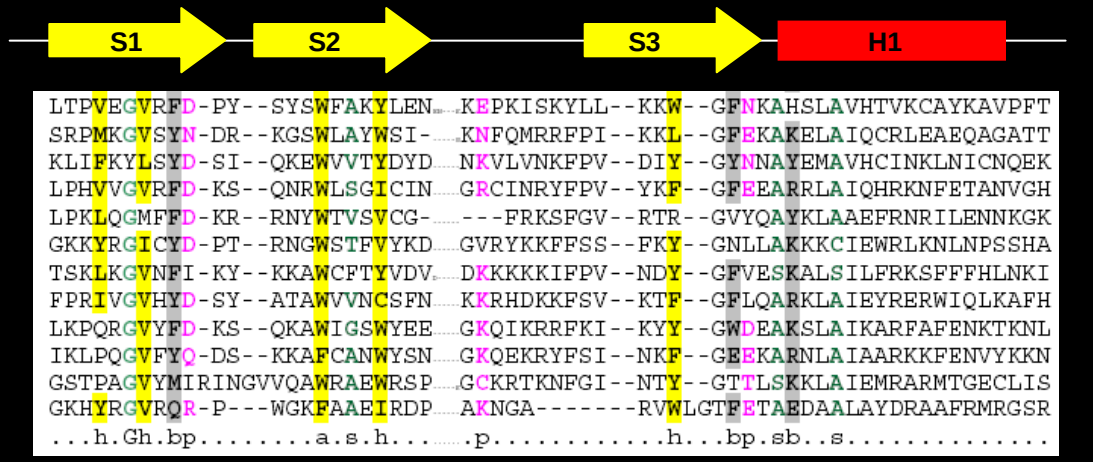
**Homologs of the conserved globular domain constitutes a novel family of the AP2 DNA-binding domain**

# Characterization of the globular domain – structural analysis I



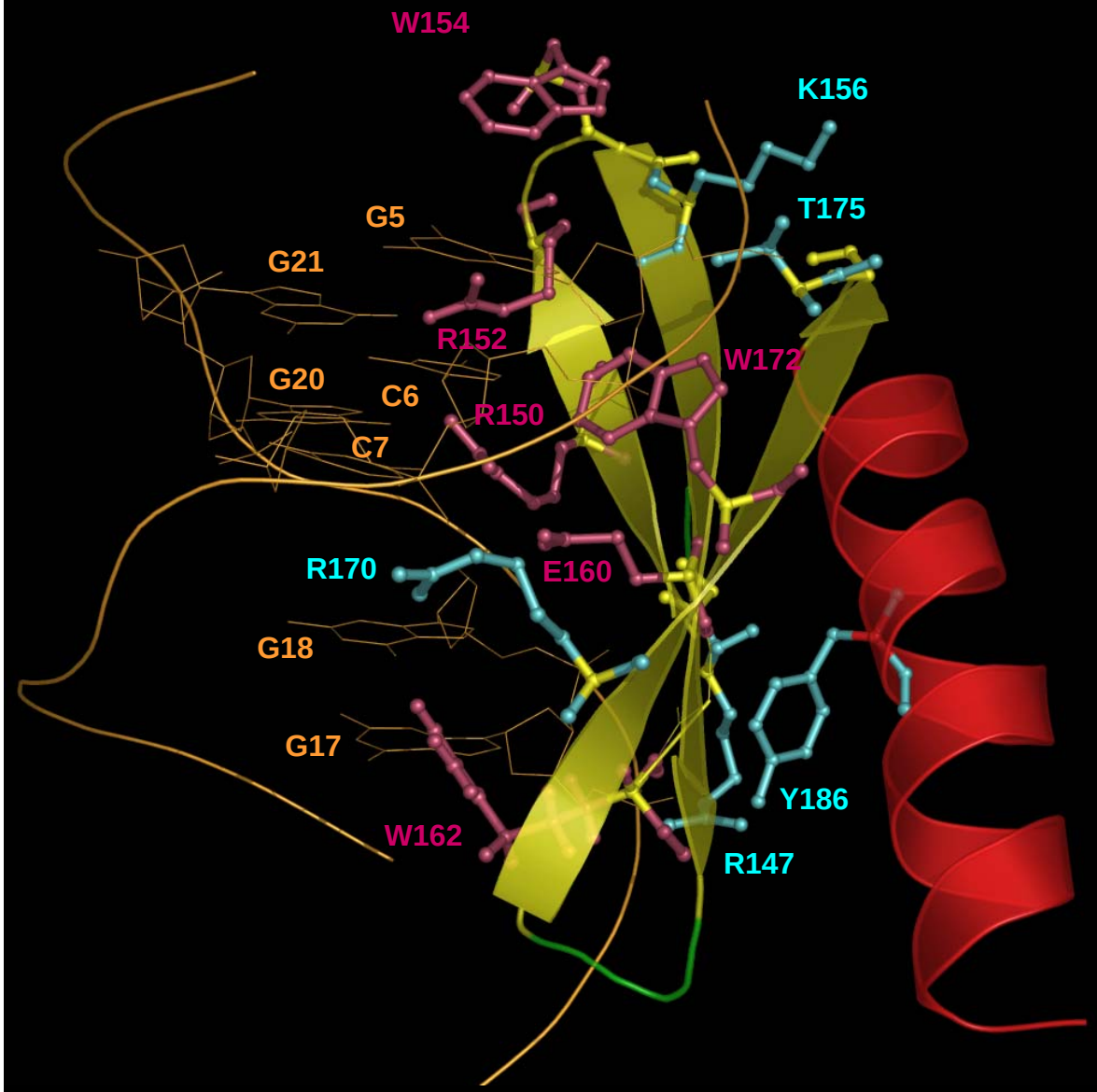
*A. thaliana* ethylene response factor
(ATERF1 - 1gcc – NMR structure)
Binds GC rich sequences




12 residues show a strong pattern of conservation and these are involved in key stabilizing hydrophobic interactions that determine the path of the backbone in the three strands and helix of the AP2 domain

**Core fold of the ApiAp2 domain will be similar to the plant AP2 DNA-binding domain**

# Characterization of the globular domain – structural analysis II

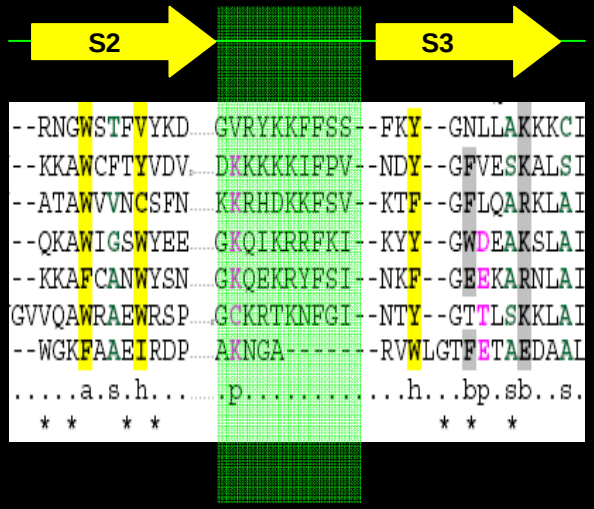



**Changes in base-contacting residues suggest binding to AT-rich sequence**

**R152 --- G5 (oxo group)**
D/N --- A (amino group)

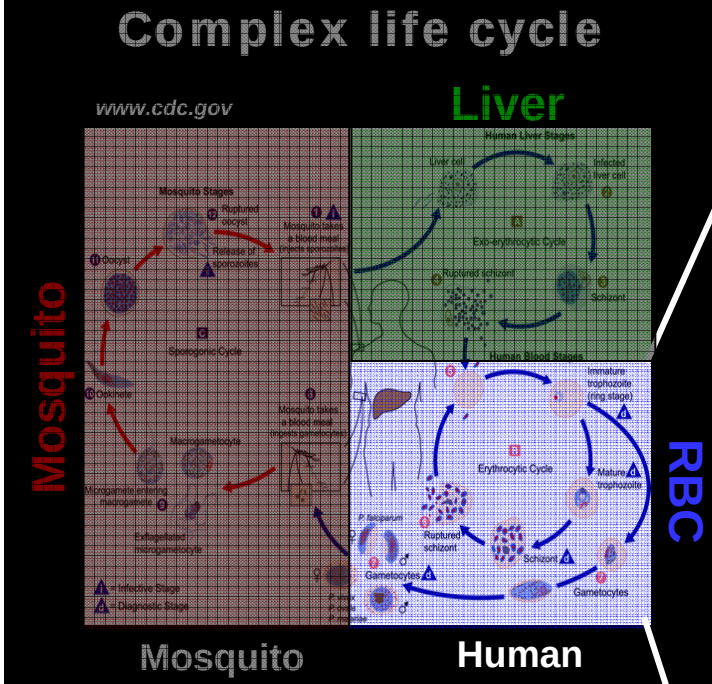
**R150 --- G20 (oxo group)**
S/T --- A (amino group)

**Charged residues in the insert may contact multiple phosphate groups to provide affinity**

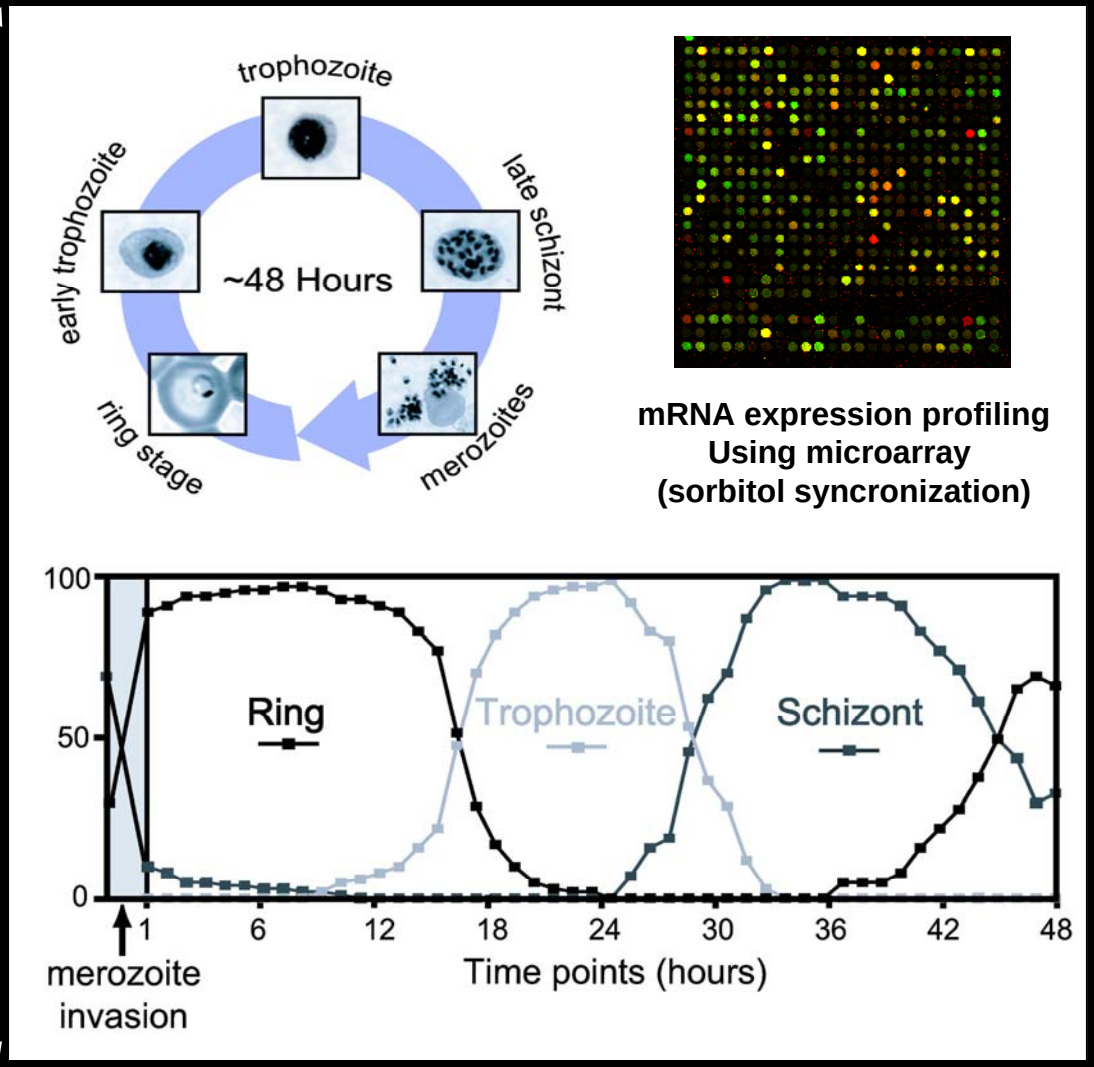



**ApiAp2 domain binds DNA in a sequence specific manner**

# Characterization of the globular domain – expression analysis I




**RBC infection & merozoite burst**



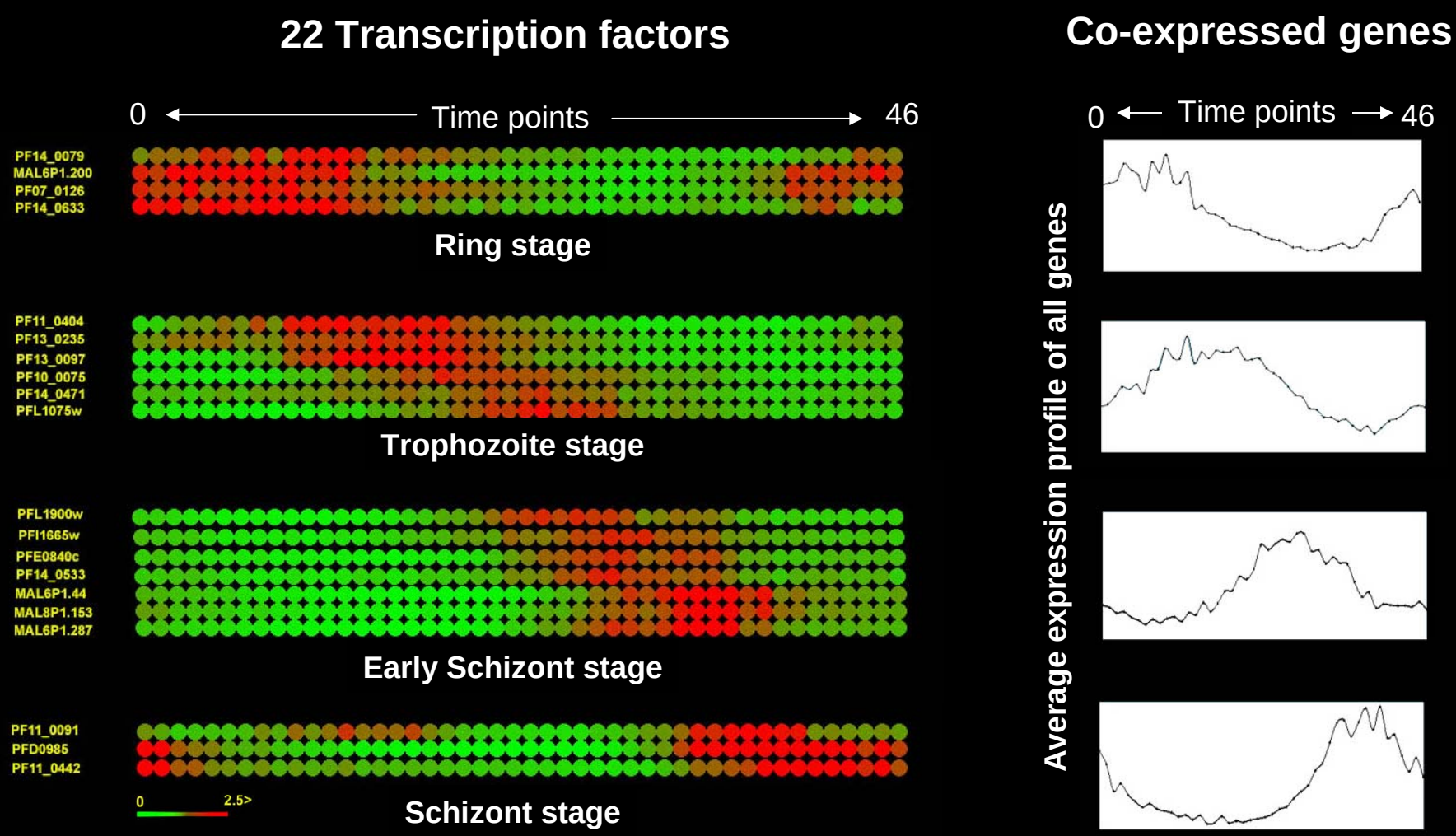



**mRNA expression profiling Using microarray (sorbitol syncronization)**

**Intra-erythrocyte developmental cycle**

DeRisi Lab

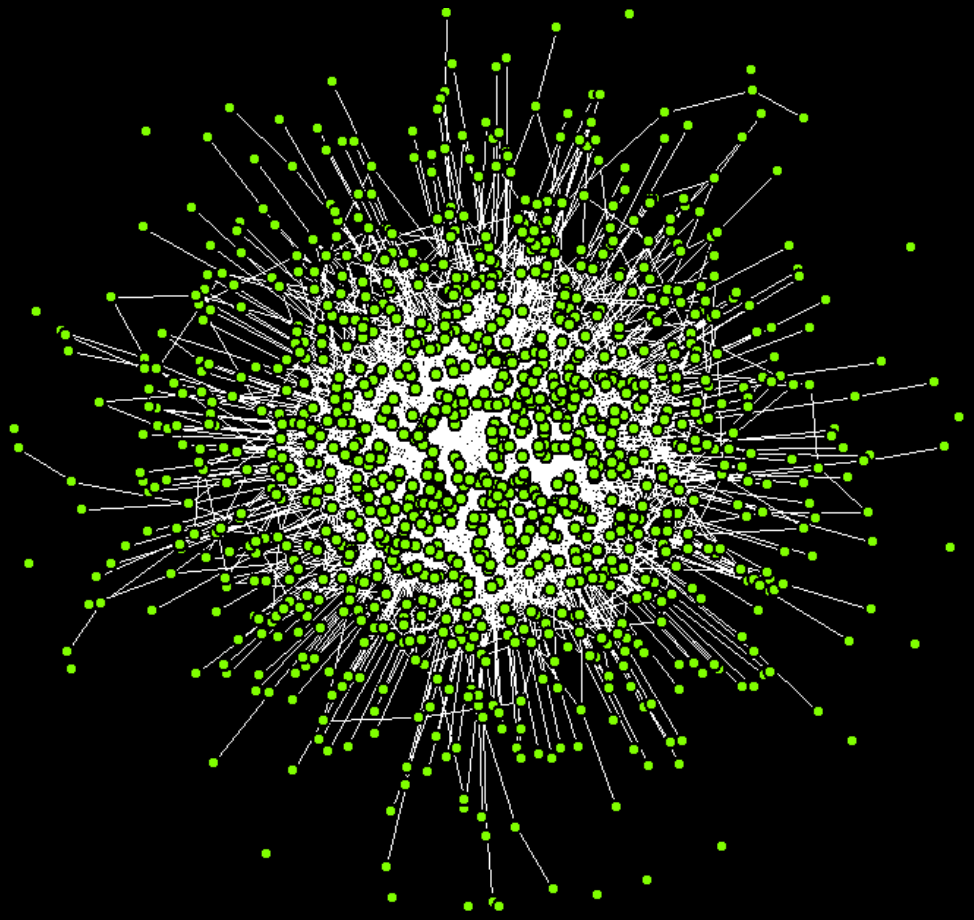
# Characterization of the globular domain – expression analysis II




**Striking expression pattern in specific developmental stages suggests that they could mediate transcriptional regulation of stage specific genes**

# Characterization of the globular domain – interaction analysis I

Protein interaction network of P. falciparum



Protein (1267)

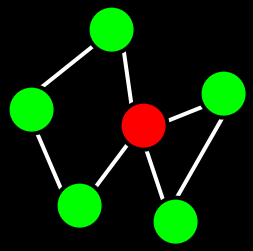
Physical interaction (2846)

LaCount et. al. Nature (2005)

Modified Y2H:
Gal4 DBD + Protein + auxotrophic gene

RNA isolated from mixed stages of
Intra-erythrocyte developmental cycle

**Guilt by association**

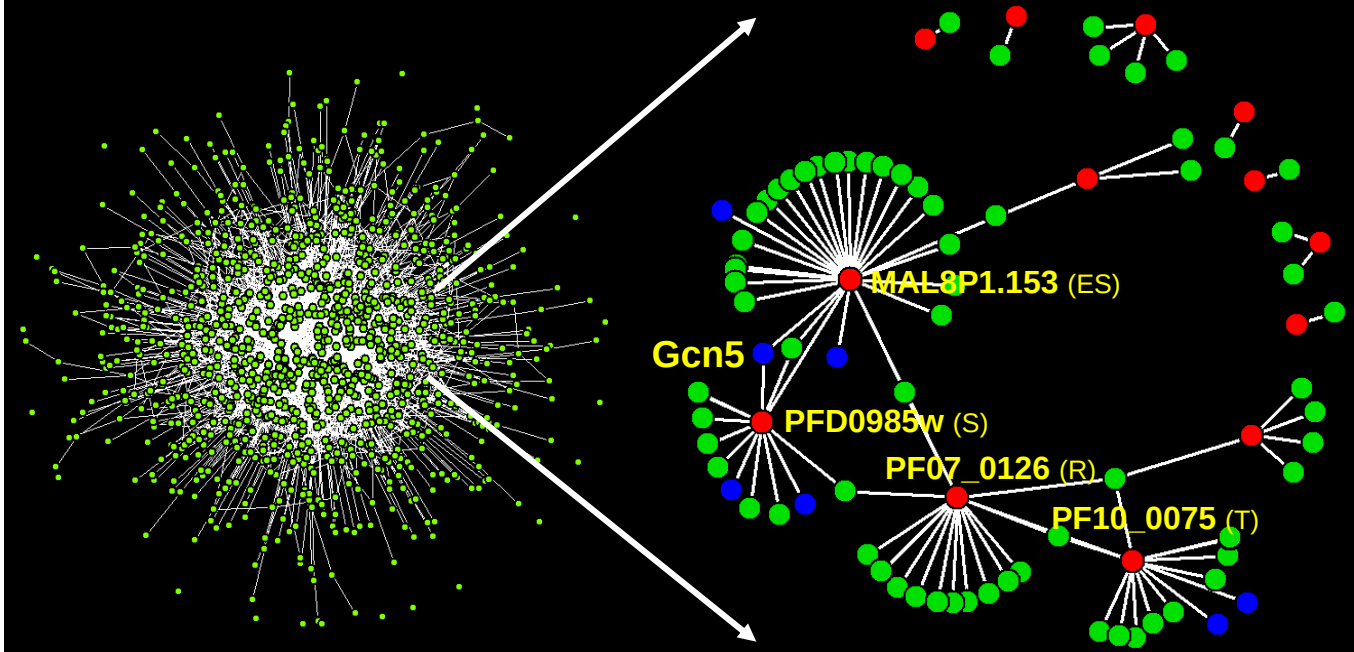


Function of interacting neighbors provides clues about function of the protein

# Characterization of the globular domain – interaction analysis II

Protein interaction network
of *P. falciparum*

Network of ApiAp2 proteins
(97 interactions, 93 proteins)



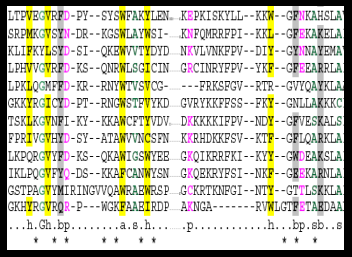

ApiAp2 proteins (13)

Chromatin proteins (8)

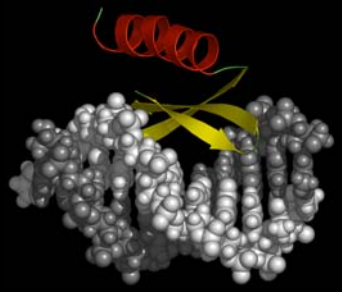
50% hypothetical
Nucleosome assembly
HMG protein
Glycolytic enzymes
Antigenic proteins
Have a PPint domain

**Guilt by association supports the role of ApiAp2 proteins to be involved in regulation of gene expression**

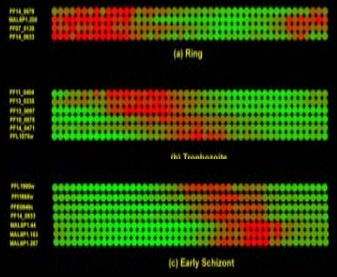
Integration of different types of experimental data allowed us to discover potential transcription factors in the *Plasmodium* genome



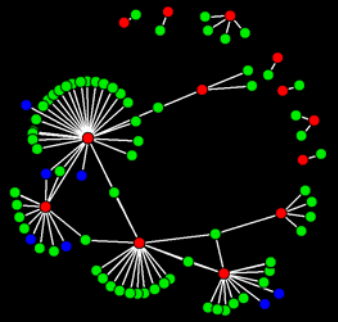
**Sequence**



**Structure**




**Expression**



**Interaction**

Integration of data can generate experimentally testable hypotheses

# The verdict!

PNAS PNAS PNAS

## Specific DNA-binding by Apicomplexan AP2 transcription factors

Erandi K. De Silva*, Andrew R. Gehrke†, Kellen Olszewski*, Ilsa León*, Jasdave S. Chahal*, Martha L. Bulyk†‡§, and Manuel Llinás*¶

*Department of Molecular Biology and Lewis–Sigler Institute for Integrative Ge
†Division of Genetics, Department of Medicine, and ‡Department of Pathology,
02115; and §Harvard/Massachusetts Institute of Technology Division of Health S

Edited by Thomas E. Wellems, National Institutes of Health, Bethesda, MD, and

Malaria remains one of the most prevalent infectious diseases worldwide, affecting more than half a billion people annually. Despite many years of research, the mechanisms underlying transcriptional regulation in the malaria-causing *Plasmodium* spp., and in Apicomplexan parasites generally, remain poorly understood. In *Plasmodium*, few regulatory elements sufficient to drive gene expression have been characterized, and their cognate DNA-binding proteins remain unknown. This study characterizes the DNA-binding specificities of two members of the recently identified Apicomplexan AP2 (ApiAP2) family of putative transcriptional regulators from *Plasmodium falciparum*. The ApiAP2 proteins contain AP2 domains homologous to the well characterized plant AP2 family of transcriptional regulators, which play key roles in development and environmental stress response pathways. We assayed ApiAP2 protein–DNA interactions using protein-binding microarrays and combined these results with computational predictions of coexpressed target genes to couple these putative *trans* factors to corresponding *cis*-regulatory motifs in *Plasmodium*. Furthermore, we show that protein–DNA sequence specificity is conserved in orthologous proteins between phylogenetically distant Apicomplexan species. The identification of the DNA-binding specificities for ApiAP2 proteins lays the foundation for the exploration of their role as transcriptional regulators during all stages of parasite development. Because of their origin in the plant lineage, ApiAP2 proteins have no homologues in the human host and may prove to be ideal antimalarial targets.

Guilty!

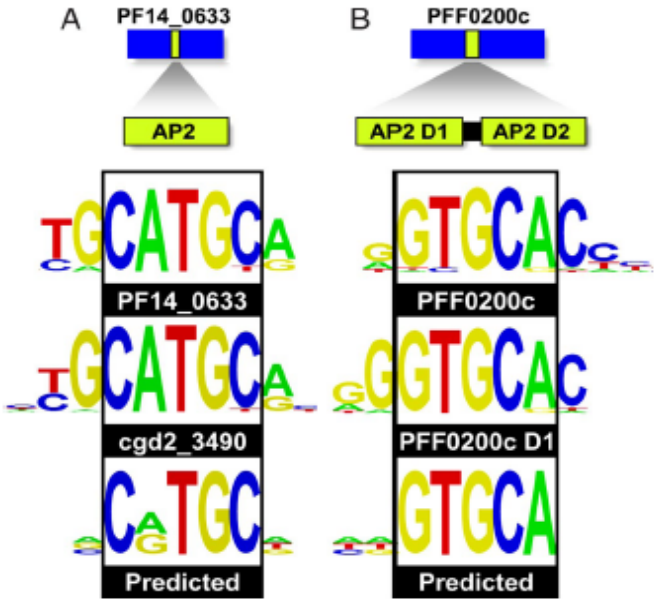



Fig. 2. DNA motifs specifically bound by AP2 domains predicted using PBMs and computational analysis (FIRE algorithm). (*A*) The core nucleotides (boxed) in the motif specifically bound by the *P. falciparum* AP2 domain of PF14_0633 are highly similar to those bound by its *C. parvum* orthologue cgd2_3490 (*Top* and *Middle*). The motifs determined from the PBM are very similar to motifs predicted using the FIRE algorithm (*Bottom*, Predicted) (24). (*B*) The PBM-derived motif bound by the tandem AP2 domains of PFF0200c (*Top*) is highly similar to the motif bound by the first domain alone (*Middle*). Domain 2 of PFF0200c did not bind a specific DNA motif (data not shown). Both PBM-derived motifs for PFF0200c match the computationally predicted motif (*Bottom*).

# Outline

- Introduction to resources and tools (10 minutes)
- A case study to highlight data integration (10 mins)
- Specific questions (25 mins)

# General approach to investigate biological questions using computational approach

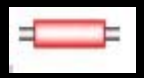
1. Formulate the big question
2. Come up with several specific questions
3. Prioritise questions and prepare a checklist
4. Identify the database
5. Identify the tools
6. Be aware of the basic statistics
7. Retrieve and integrate the information
8. Formulate hypothesis and READ A LOT!
9. Design experiments
10. Publish work & be happy ever after ☺

# Question #1

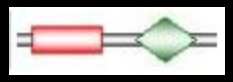
## How can I analyse the sequence of a protein?

```
>gi|75018194
MESTEDEFYTICLNLTAEDPSFGNCNYTTDFENGELLEKVVSRVVPIFFGFIGIVGLVGNALVVLVVAAN
PGMRSTTNLLIINLAVADLLFVIFCVPFTATDYVMPRWPFGDWWCKVVQYFIVVTAHASVYTLVLMSLDR
FMAVVHPIASMSIRTEKNALLAIACIWVVILTTAIPVGICHGEREYSYFNRNHSSCVFLEERGYSKLGFQ
MSFFLSSYVIPLALISVLYMCMLTRLWKSAPGGRVSAESRRGRKKVTRMVVVVVVVFAVCWCPIQIILLV
KALNKYHITYFTVTAQIVSHVLAYMNSCVNPVLYAFLSENFRVAFRKVMYCPPPYNDGFSGRPQATKTTR
TGNGNSCHDIV
```

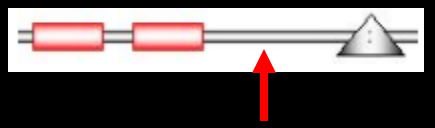
# Proteins are made of domains, which are independent evolutionary units



Single domain



Two domain



Unassigned region
New domain or unstructured region

```
B9G4I0_ORYSJ/14-218          ..............ARKRFDKASL.L.YDQARERYLS.LKK-GIRT-Dva.
Q0J0G6_ORYSJ/14-218          ..............ARKRFDKASL.L.YDQARERYLS.LKK-GIRT-Dva.
C0PDH5_MAIZE/14-218          ..............AHKRFDKASL.S.YDQIREKYLS.LKKGT----Rpdi
Q0J435_ORYSJ/14-218          ..............ARKRFDKASL.L.YDQVRDKYLS.LK--KGTRADit.
B9G206_ORYSJ/14-189          ..............ARKRFDKASL.L.YDQVRDKYLS.LK--KGTRADit.
B8B950_ORYSI/14-190          ..............ARKRFDKASL.L.YDQVRDKYLS.LK--KGTRADit.
A7QT71_VITVI/14-218          ..............ARKRFDKASL.L.YDQAREKYLS.LRK--GTKSDia.
B9RFZ3_RICCO/14-218          ..............ARKRFDKASL.L.YDQAREKFLS.LR--KGTKTDva.
AGD3_ARATH/4-218             ..............ARKRFDKASL.T.YDQAREKFLS.LRK--GTKSDva.
B9GMT1_POPTR/11-218          ..............ARKRFDKASL.L.YDQAREKFLS.LRK--GTRSDia.
B9GZ26_POPTR/12-218          ..............ARKRFDKASL.L.YDQAREKFLS.LRK-GTRS-Dva.
```

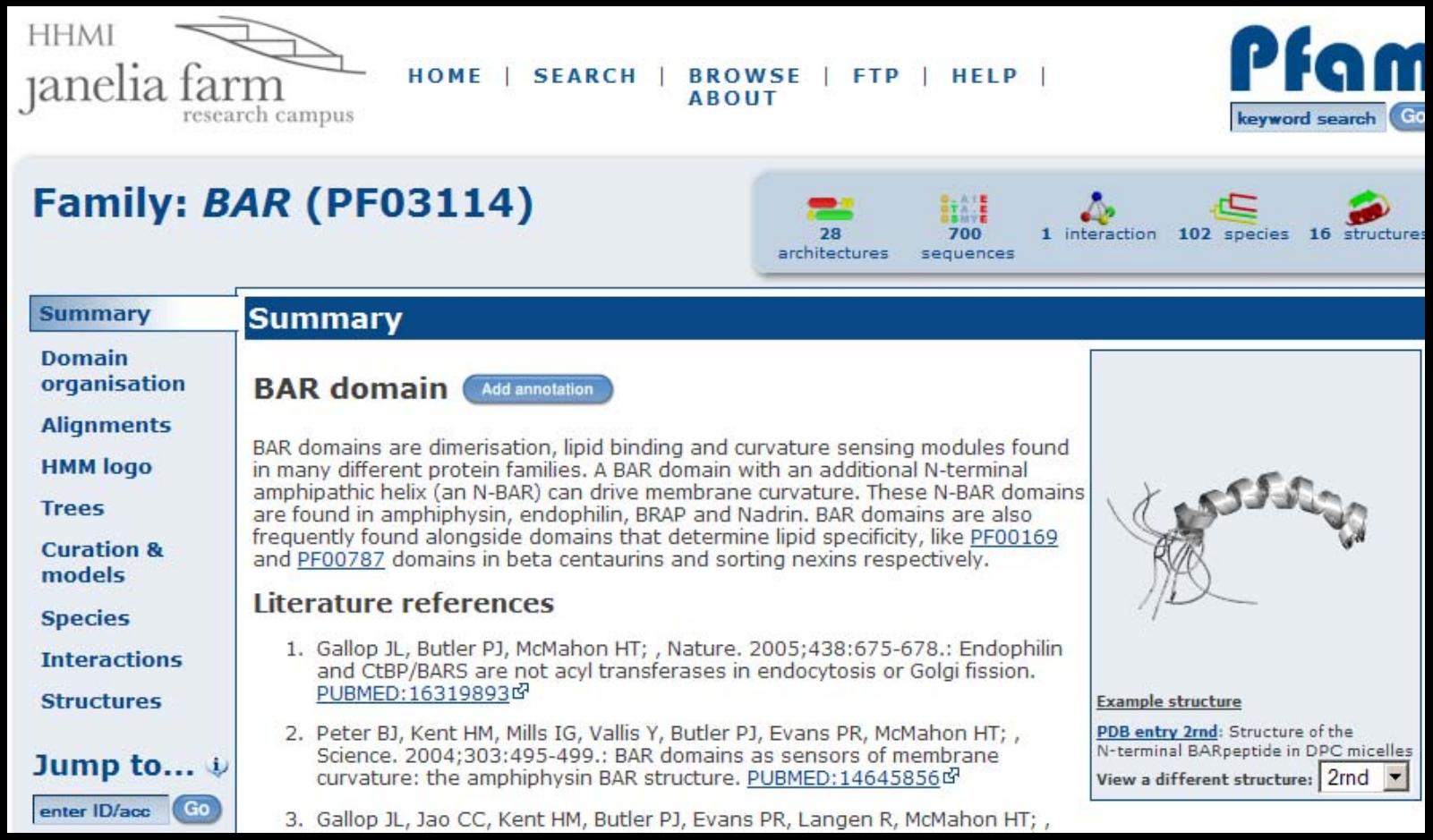
HTH domain

BAR domain

Sequence and structure based domain definition

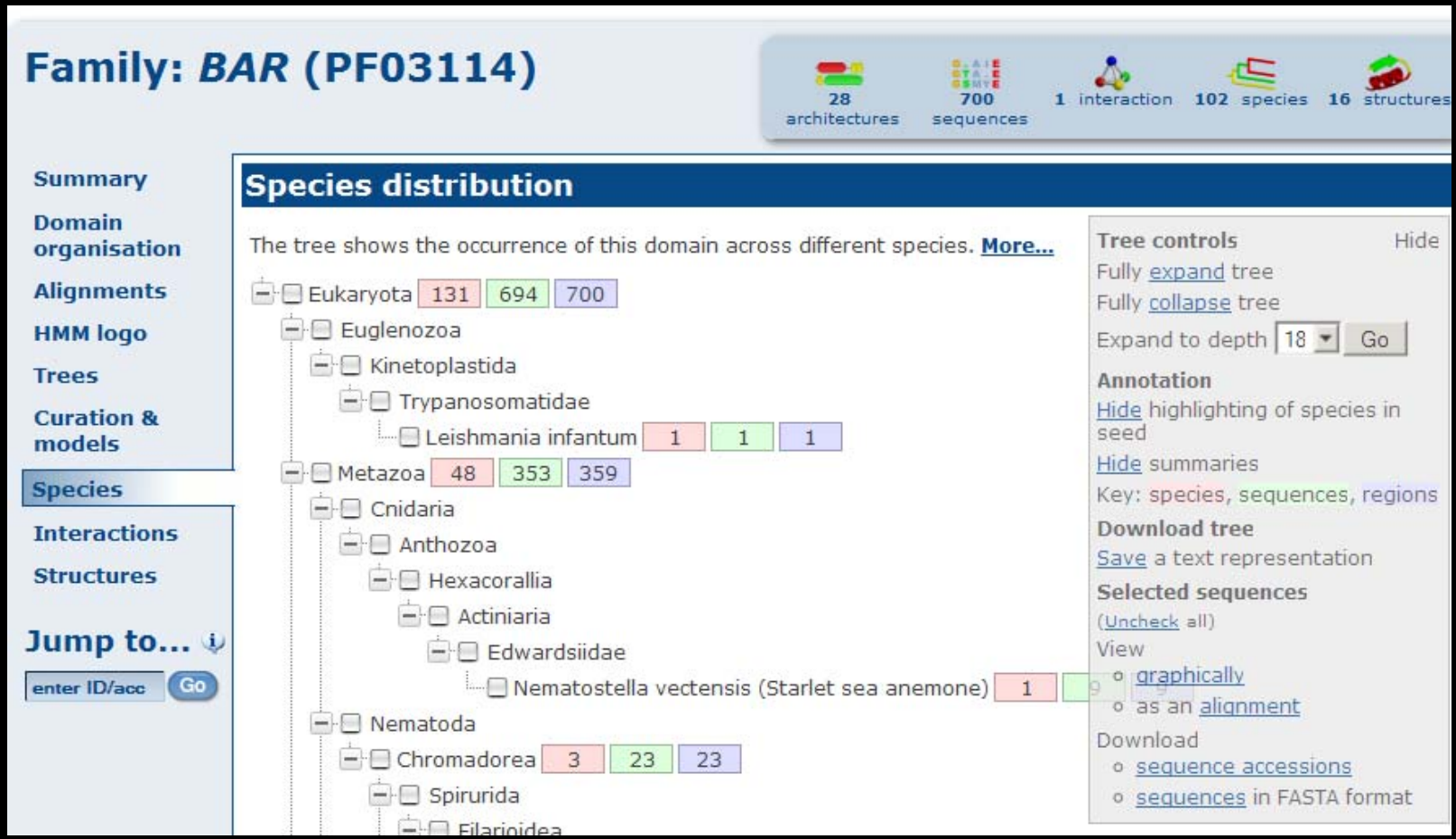
Structure based domain definition

http://pfam.janelia.org/family/bar




Database of protein domains

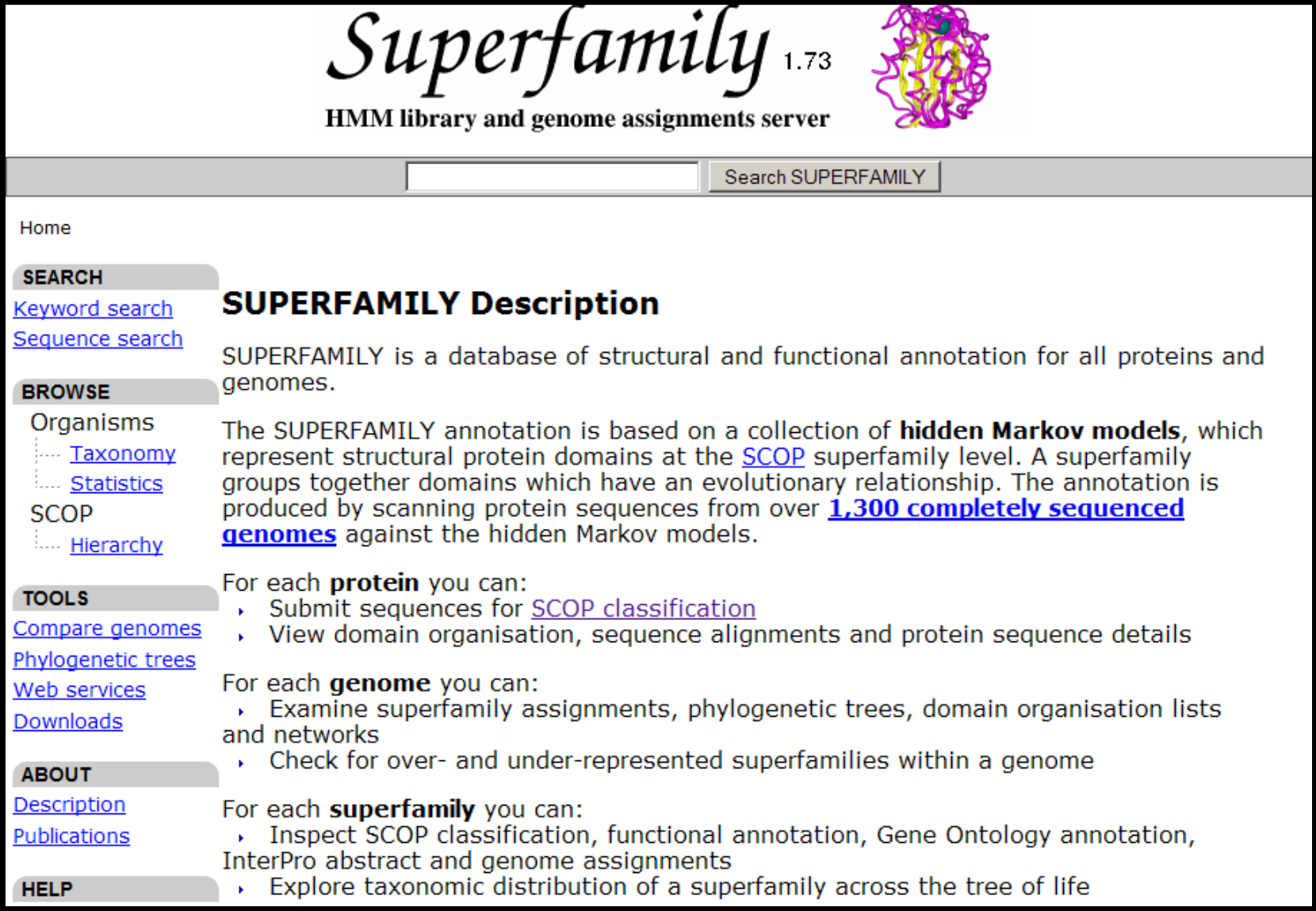
# How can I obtain all sequences containing a particular domain?




Click “Species” to obtain all proteins with a particular domain

For Marwah Hassan

http://supfam.org/SUPERFAMILY/cgi-bin/search.cgi




Structural domain assignments to completely sequenced genomes

Superfamily 1.73

HMM library and genome assignments server

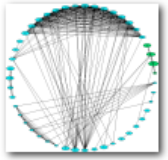
Search SUPERFAMILY

2HC zinc fingers domains

Structural Classification | Genome Assignments | Sequence Alignments | Domain Combinations | Taxonomic Distribution

## C2H2 and C2HC zinc fingers superfamily domain assignments

**Network of domain occurrence in all genomes**



**No domain assignments for these genomes.**

### Add assignments from groups of genomes

View all assignments containing a C2H2 and C2HC zinc fingers domain in each group of genomes.

- All genomes
- All eukaryote genomes (domain E)
- All bacterial genomes (domain B+A)
- All eubacterial genomes (domain B)
- All archaeal genomes (domain A)

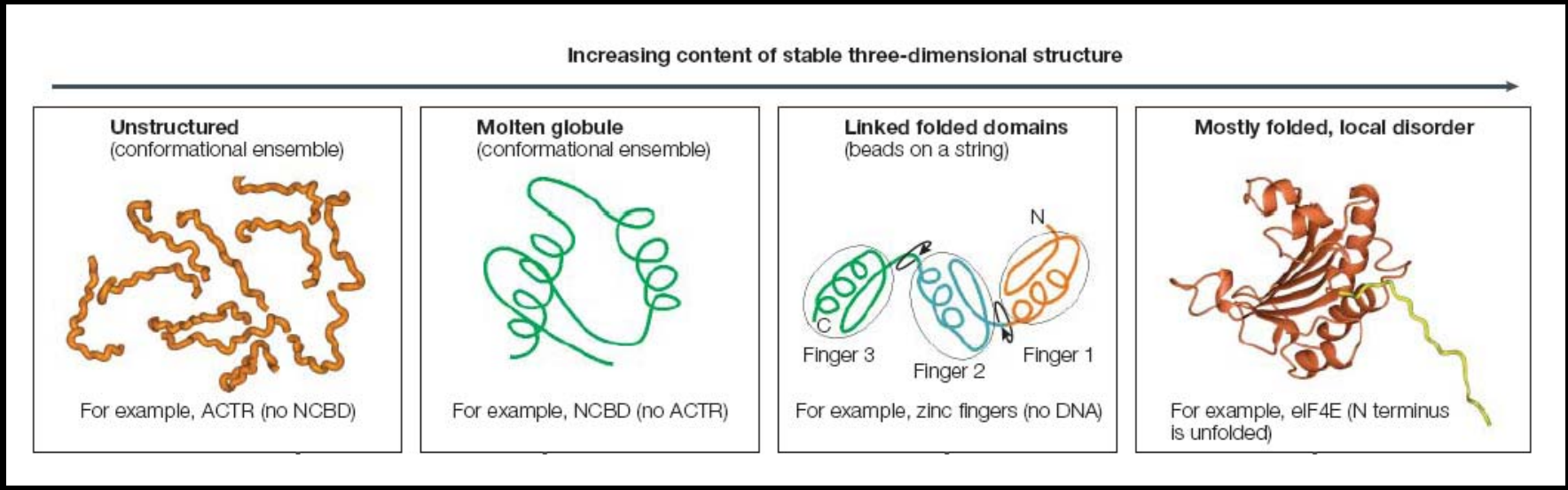
Jump to [ Top of page · Assignment details · Add assignments from groups of genomes ]

Genome assignment, sequence alignment, domain combinations and taxonomic distribution

For Marwah Hassan

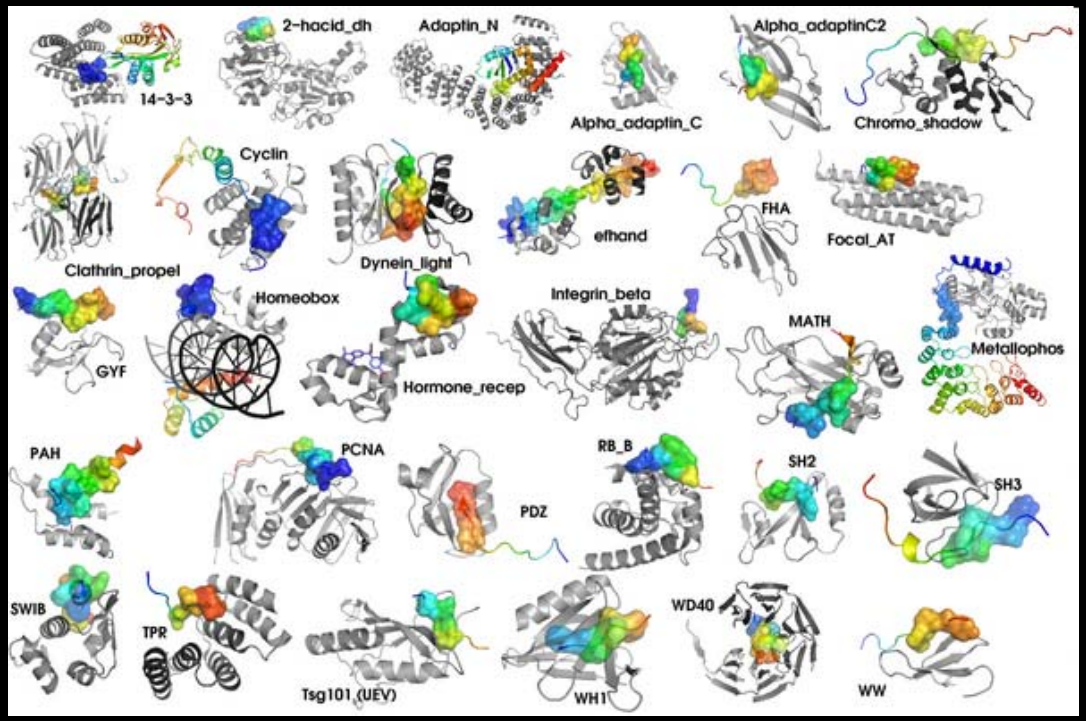
# Intrinsically Unstructured Proteins

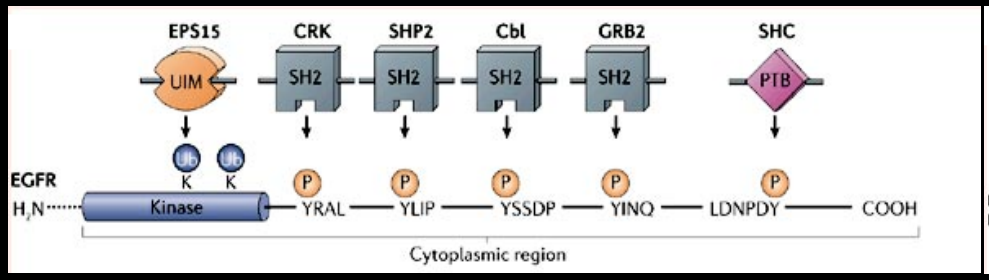
IUPs: lack an unique 3-dimensional structure, either entirely or in parts. It is assumed that they sample a variety of conformations that are in dynamic equilibrium under physiological conditions.

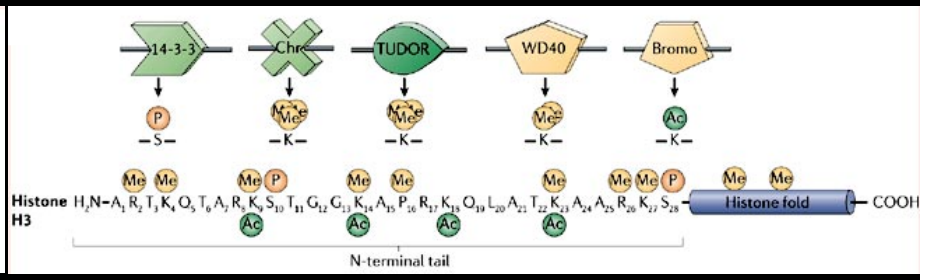



Structural continuum of intrinsically unstructured proteins

Exposed peptide may mediate protein-peptide interaction, PTM, etc



# Identification of intrinsically unstructured proteins

## Computational approaches to predict unstructured regions

### Sequence composition based on specific scales

- FoldIndex: Identifies regions of low hydrophobicity but high net charge using a sliding window
- PONDR: Local amino-acid composition, flexibility and hydropathy
- GlobPlot: amino-acid propensity to form globular domains using the russel-linding scale

HCA, IUPred, RONN, SEG, PreLink and several others

### Computationally trained on existing structures in the PDB

- DisoPred2: Support vector machine trained on structure data
- DisEMBL: Neural network trained on structure data

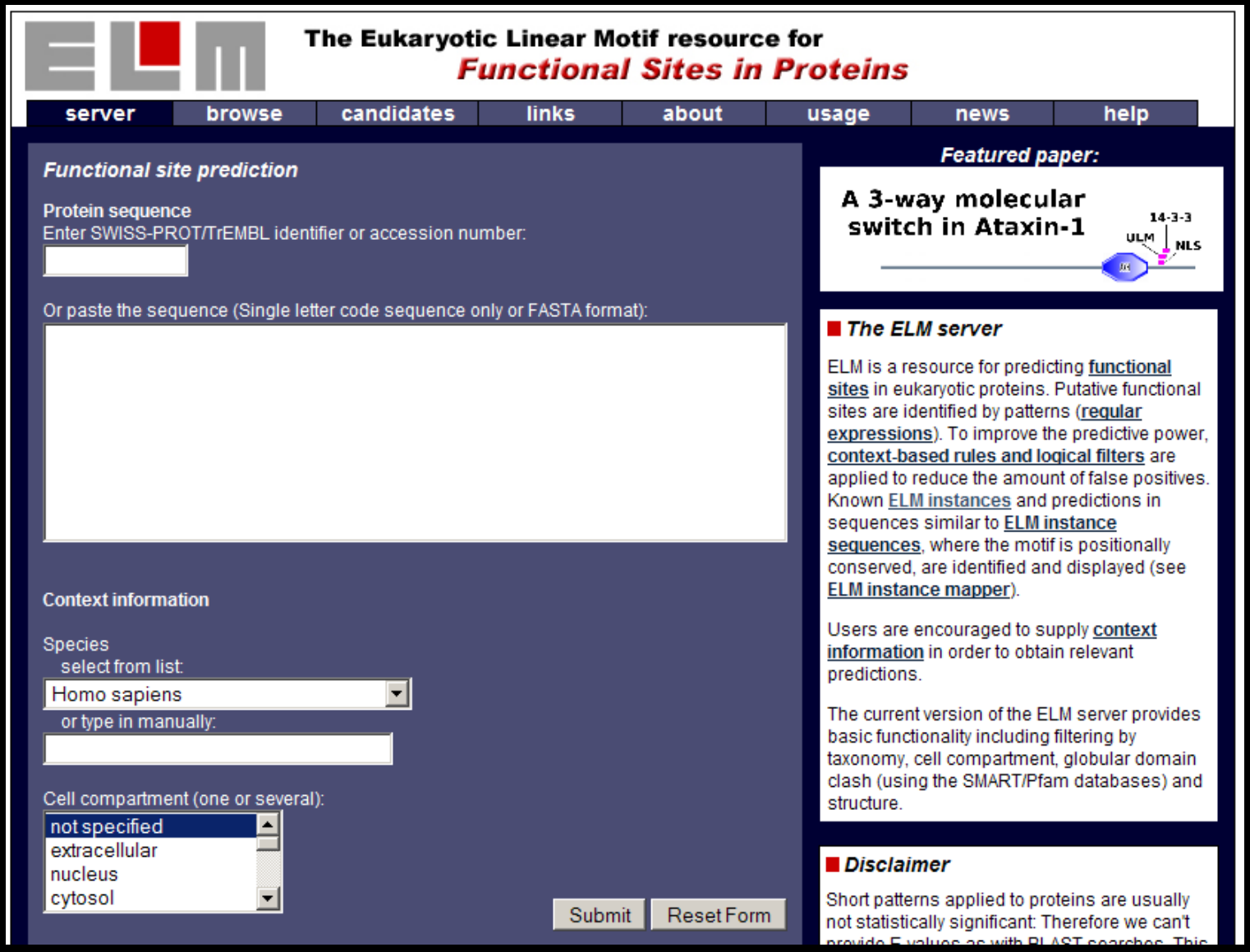
Accuracies are similar to current secondary structure prediction methods from sequence (~80%)

http://en.wikipedia.org/wiki/Intrinsically_unstructured_proteins#Database_of_Protein_Disorder

| Predictor | What is predicted | Based on | Generates and uses multiple sequence alignment? |
|---|---|---|---|
| PONDR [1] | All regions that are not rigid including random coils, partially unstructured regions, and molten globules | Local aa composition, flexibility, hydropathy, etc | No |
| SEG [2] | Low-complexity segments that is, "simple sequences" or "compositionally biased regions". | Locally optimized low-complexity segments are produced at defined levels of stringency and then refined according to the equations of Wootton and Federhen | No |
| Disopred2 [3] | Regions devoid of ordered regular secondary structure | Cascaded support vector machine classifiers trained on PSI-BLAST profiles | Yes |
| Globplot [4] | Regions with high propensity for globularity on the Russell/Linding scale (propensities for secondary structures and random coils) | Russell/Linding scale of disorder | No |
| Disembl [5] | LOOPS (regions devoid of regular secondary structure); HOT LOOPS (highly mobile loops); REMARK465 (regions lacking electron density in crystal structure) | Neural networks trained on X-ray structure data | No |
| NORSp [6] | Regions with No Ordered Regular Secondary Structure (NORS). Most, but not all, are highly flexible. | Secondary structure and solvent accessibility | Yes |
| FoldIndex [7] | Regions that have a low hydrophobicity and high net charge (either loops or unstructured regions) | Charge/hydrophaty analyzed locally using a sliding window | No |
| Charge/hydropathy method. See (Uversky et al., 2000). | Fully unstructured domains (random coils) | Global sequence composition | No |
| HCA (Hydrophobic Cluster Analysis) [8] | Hydrophobic clusters, which tend to form secondary structure elements | Helical visualization of amino acid sequence | No |
| PreLink [9] | Regions that are expected to be unstructured in all conditions, regardless of the presence of a binding partner | Compositional bias and low hydrophobic cluster content. | No |
| IUPred [10] | Regions that lack a well-defined 3D-structure under native conditions | Energy resulting from inter-residue interactions, estimated from local amino acid composition | No |
| RONN [11] | Regions that lack a well-defined 3D structure under native conditions | Bio-basis function neural network trained on disordered proteins | No |
| MD (Meta-Disorder predictor) [12] | Regions of different "types"; for example, unstructured loops and regions containing few stable intra-chain contacts | A neural-network based meta-predictor that uses different sources of information predominantly obtained from orthogonal approaches | Yes |
| GeneSilico Metadisorder [13] | Regions that lack a well-defined 3D structure under native conditions (REMARK-465) | Meta method which uses other disorder predictors (like RONN, IUPred, POODLE and many more). Based on them the consensus is calculated according method accuracy (optimized using ANN, filtering and other techniques). Currently the best available method (first 2 places in last CASP experiment (blind test)) | Yes |
| IUPforest-L [14] | Long disordered regions in a set of proteins | Moreau-Broto auto-correlation function of amino acid indices (AAIs[15]) | No |

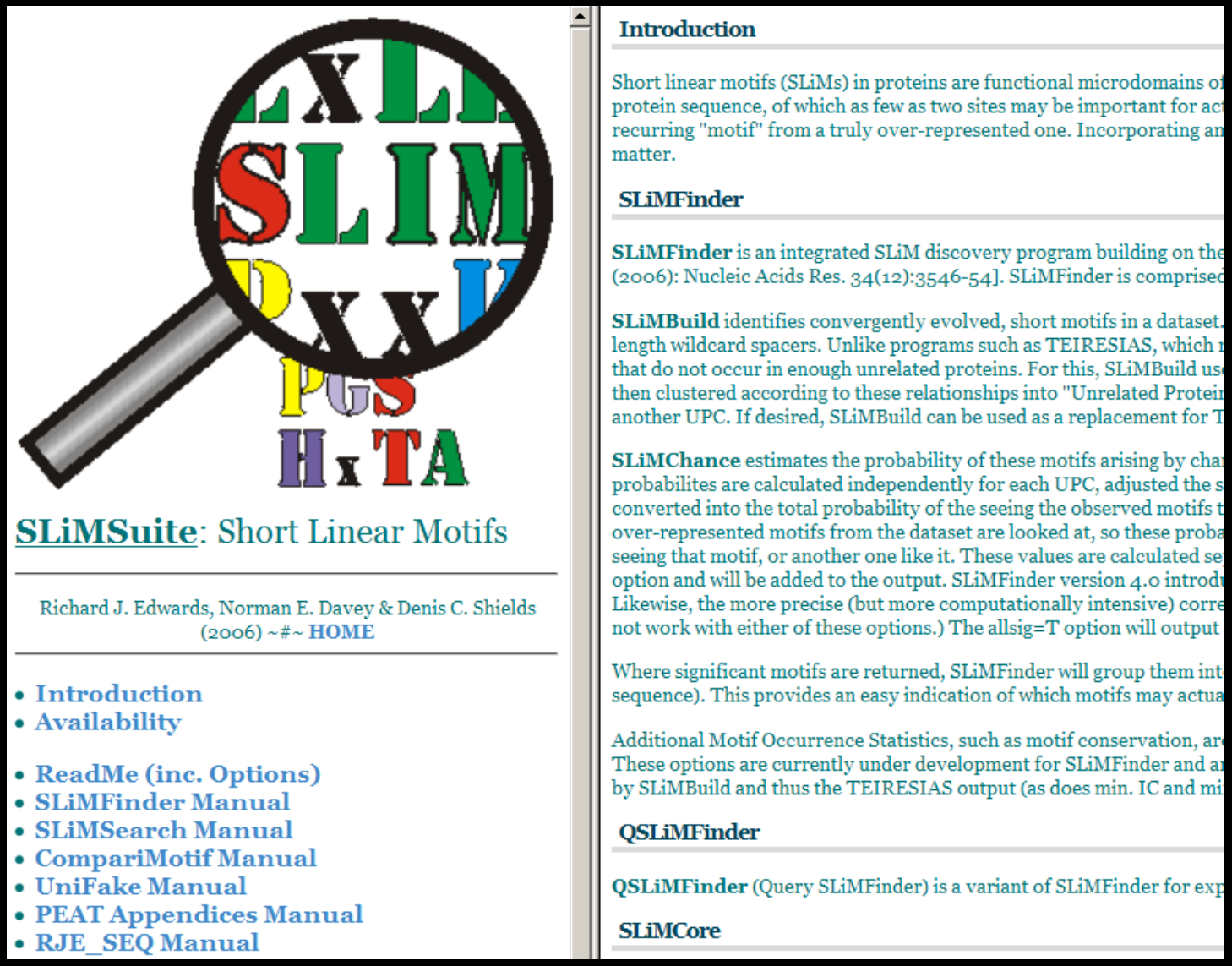
Programs for identifying unstructured regions in a protein sequence

http://elm.eu.org/




Identification of potential eukaryotic linear motif

http://www.southampton.ac.uk/~re1u06/software/slimsuite/



**SLiMSuite**: Short Linear Motifs

Richard J. Edwards, Norman E. Davey & Denis C. Shields (2006) ~#~ HOME

- Introduction
- Availability

- ReadMe (inc. Options)
- SLiMFinder Manual
- SLiMSearch Manual
- CompariMotif Manual
- UniFake Manual
- PEAT Appendices Manual
- RJE_SEQ Manual

**Introduction**

Short linear motifs (SLiMs) in proteins are functional microdomains of
protein sequence, of which as few as two sites may be important for act
recurring "motif" from a truly over-represented one. Incorporating an
matter.

**SLiMFinder**

**SLiMFinder** is an integrated SLiM discovery program building on the
(2006): Nucleic Acids Res. 34(12):3546-54]. SLiMFinder is comprised

**SLiMBuild** identifies convergently evolved, short motifs in a dataset.
length wildcard spacers. Unlike programs such as TEIRESIAS, which
that do not occur in enough unrelated proteins. For this, SLiMBuild us
then clustered according to these relationships into "Unrelated Protei
another UPC. If desired, SLiMBuild can be used as a replacement for T

**SLiMChance** estimates the probability of these motifs arising by cha
probabilites are calculated independently for each UPC, adjusted the s
converted into the total probability of the seeing the observed motifs t
over-represented motifs from the dataset are looked at, so these proba
seeing that motif, or another one like it. These values are calculated se
option and will be added to the output. SLiMFinder version 4.0 introdu
Likewise, the more precise (but more computationally intensive) corre
not work with either of these options.) The allsig=T option will output

Where significant motifs are returned, SLiMFinder will group them int
sequence). This provides an easy indication of which motifs may actua

Additional Motif Occurrence Statistics, such as motif conservation, ar
These options are currently under development for SLiMFinder and a
by SLiMBuild and thus the TEIRESIAS output (as does min. IC and mi

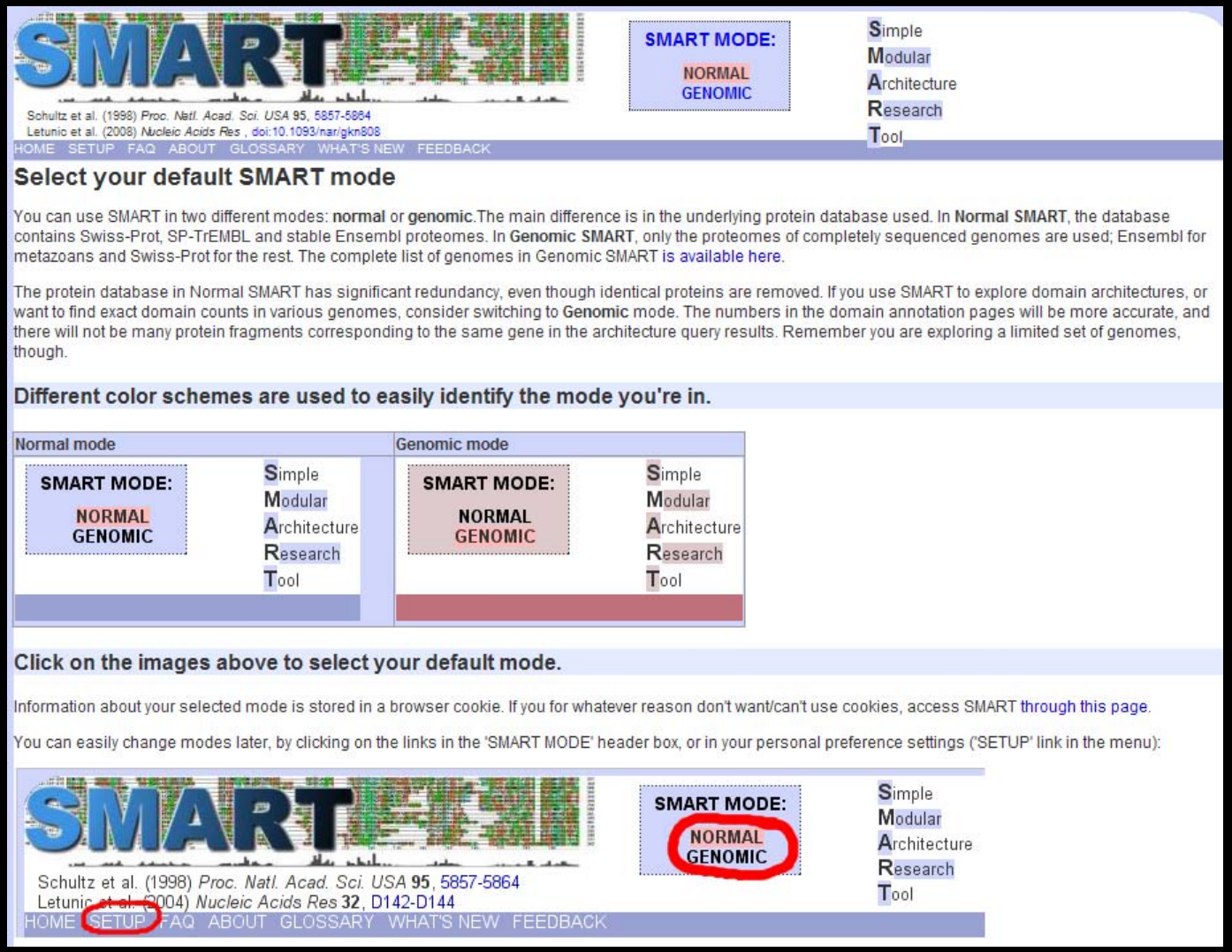
**QSLiMFinder**

**QSLiMFinder** (Query SLiMFinder) is a variant of SLiMFinder for exp

**SLiMCore**

Identification of potential eukaryotic linear motif

http://smart.embl-heidelberg.de/




SMART combines domain assignment with prediction of linear motifs in unstructured regions

SMART

SMART MODE:
NORMAL
GENOMIC

Schultz et al. (1998) *Proc. Natl. Acad. Sci. USA* 95, 5857-5864
Letunic et al. (2008) *Nucleic Acids Res* , doi:10.1093/nar/gkn808

HOME SETUP FAQ ABOUT GLOSSARY WHAT'S NEW FEEDBACK

## Sequence analysis

You may use either a Uniprot/Ensembl sequence identifier (ID) / accession number (ACC) or the protein service.

### Sequence ID or ACC

Examples: TEC_HUMAN, C1S_HUMAN

### Sequence

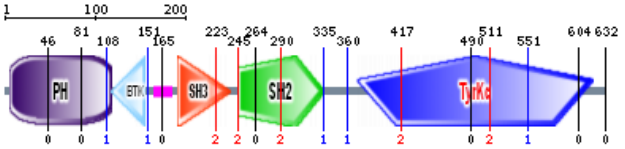
Sequence SMART | Reset

HMMER searches of the SMART database occur by default. You may also find:

- [x] Outlier homologues and homologues of known structure
- [x] PFAM domains
- [x] signal peptides
- [x] internal repeats
- [x] intrinsic protein disorder

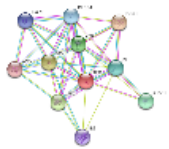
## Domains within *Homo sapiens* protein TEC_HUMAN (P42680)

Tyrosine-protein kinase Tec



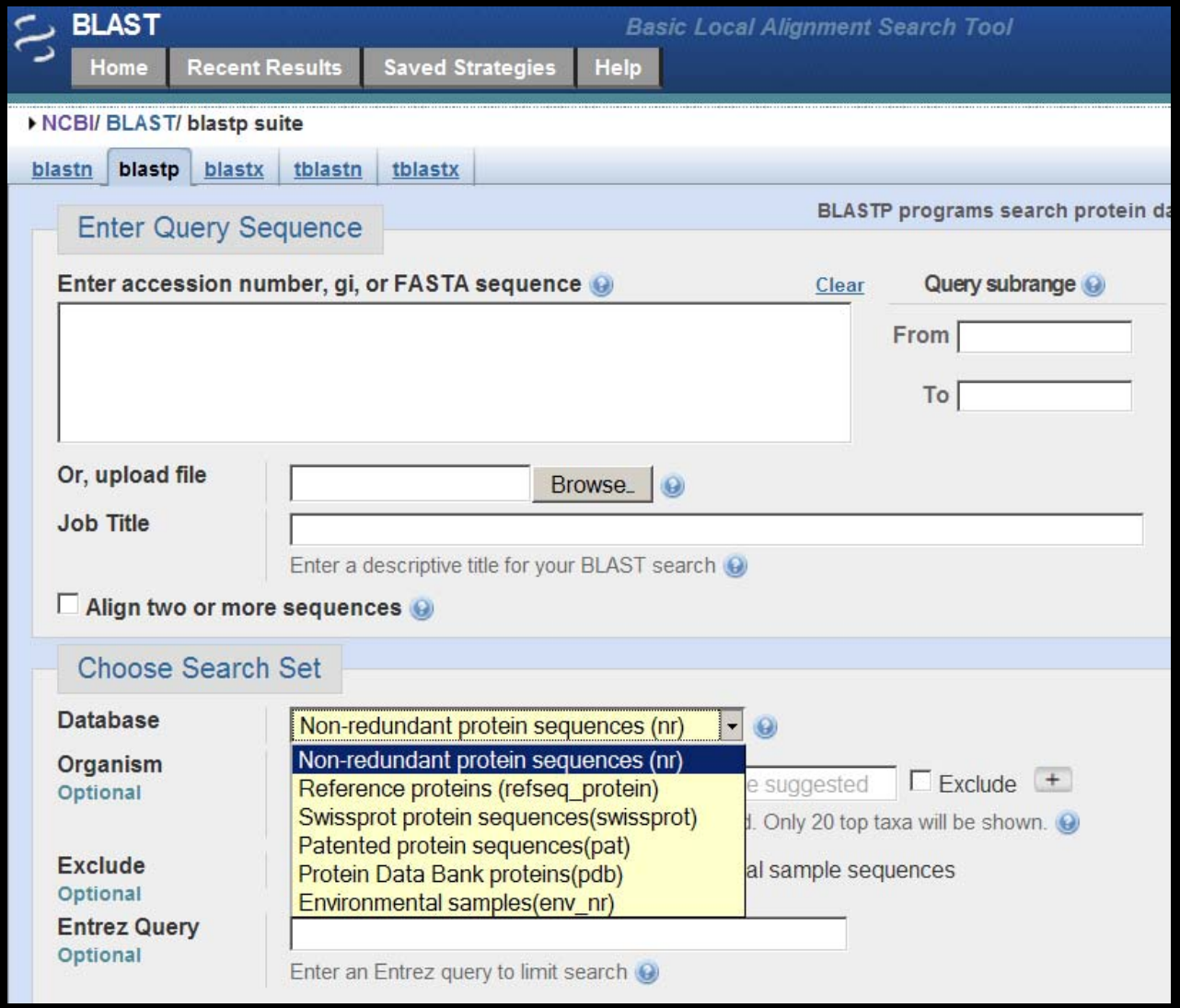

**Mouse over domain / undefined region for more info; click on it to go to detailed annotation; right-click to save wh**

Transmembrane segments as predicted by the *TMHMM2* program ( ), coiled coil regions determined by the *Coils2* are indicated by BLAST for hits in the schnipsel database and PDB for hits against PDB. Regions containing repeats phase and exact position in AA.

| Protein information | Domain architecture analysis |
|---|---|
| Display orthology and other data | Display all proteins with similar domain ORGANISATION or COMPOSITION.<br>This domain architecture was probably invented with the emergence of Metazoa. |
| **Interaction network** | **Pathway information** |
|  | TEC_HUMAN is possibly involved in the following metabolic pathways:<br>map04660: T cell receptor signaling pathway<br>These assignments are based on similarity to the following orthologous groups:<br>K07364: TEC (tec protein tyrosine kinase [EC:2.7.10.2]) |

TEC_HUMAN shown as TEC

# How to characterise an unassigned region in a protein?



# checklist

1. Get a good text editor (textpad, bbedit, nedit)
2. Run a domain analysis (pfam, superfamily, SMART)
3. Identify regions that are of low complexity
4. Run Linear Motif prediction
5. Run a post-translation modification prediction
6. Run a signal peptide prediction
7. Run a PSI-BLAST search on individual domains
8. Run a PSI-BLAST search on the whole sequence
9. Ask if predicted regions are evolutionarily conserved

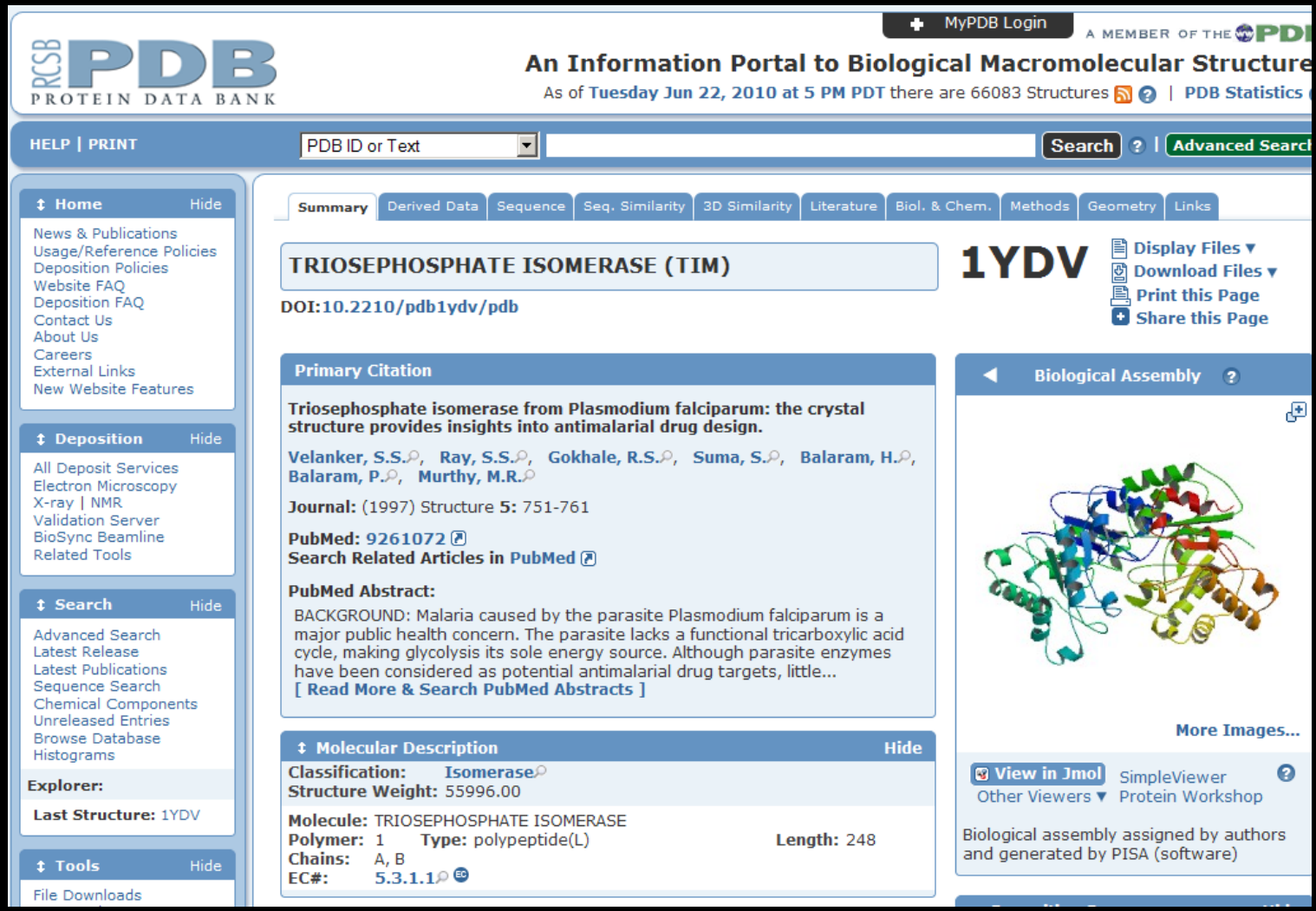
# Question #2

How can I identify structurally similar proteins?

Which amino acids contact the ligand of interest in a structure?

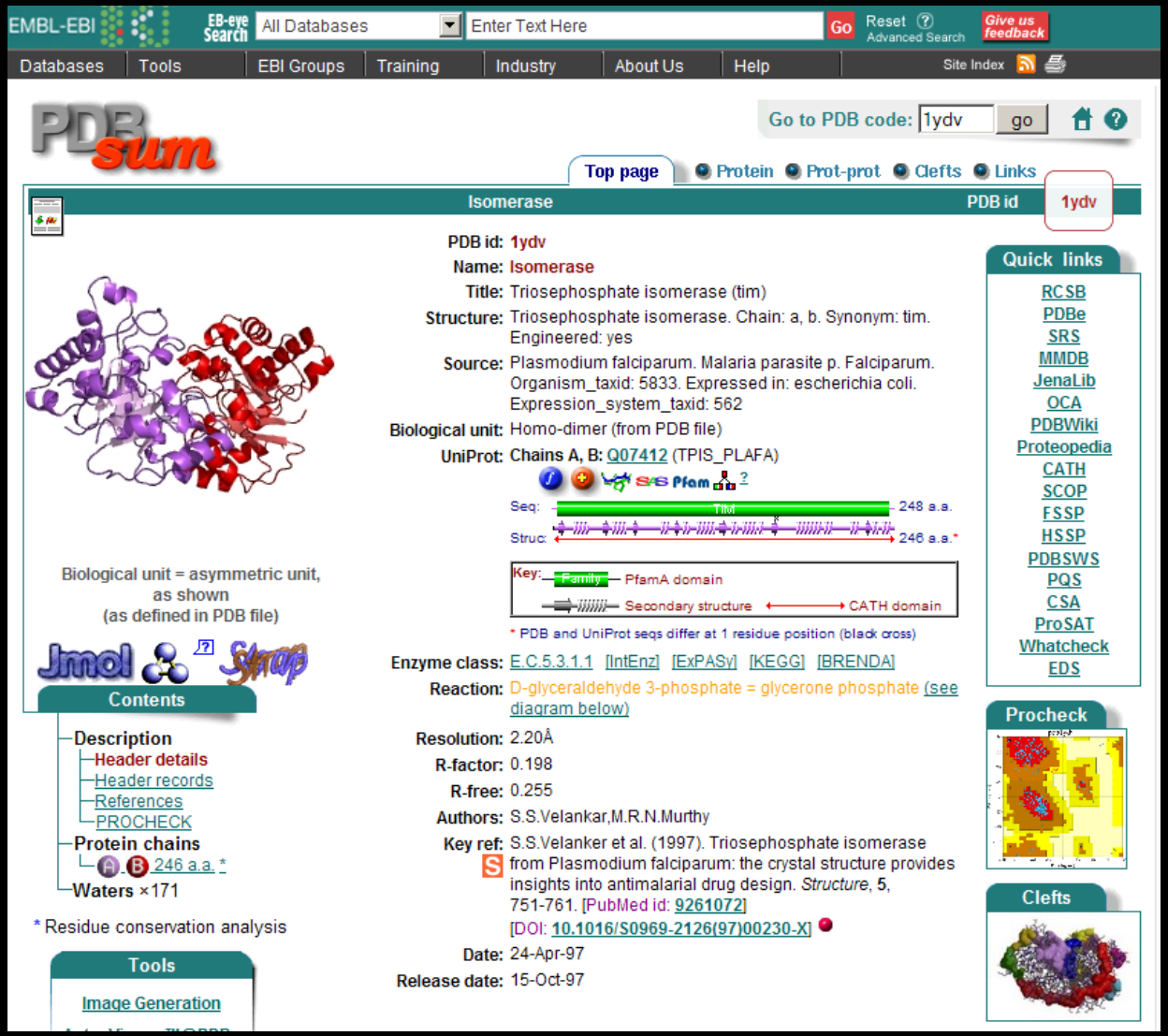
What analyses can I do with a structure?

http://www.pdb.org/pdb/explore/explore.do?structureId=1YDV




Protein Data Bank: repository for protein structures

http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl




Comprehensive summary of deposited structures

http://redpoll.pharmacy.ualberta.ca/vadar/


**Single (or Multiple) Model Protein Structure Analysis**

VADAR (Volume, Area, Dihedral Angle Reporter) is a compilation of more than 15 different algorithms and programs for analyzing and assessing peptide and protein structures from their PDB coordinate data. The results have been validated through extensive comparison to published data and careful visual inspection. The VADAR web server supports the submission of either PDB formatted files or PDB accession numbers. VADAR produces extensive tables and high quality graphs for quantitatively and qualitatively assessing protein structures determined by X-ray crystallography, NMR spectroscopy, 3D-threading or homology modelling.

Please cite the following: Leigh Willard, Anuj Ranjan,Haiyan Zhang,Hassan Monzavi, Robert F. Boyko, Brian D. Sykes, and David S. Wishart "VADAR: a web server for quantitative evaluation of protein structure quality" Nucleic Acids Res. 2003 July 1; 31 (13): 3316.3319

**Acknowledgements:**
We would like to thank PENCE and CIHR for their financial support.

**Program Options:**

1. ☐ Calculate hydrogen bonds to water
2. Values for Van der waals radii
   - ○ Chothia
   - ○ Eisenberg
   - ○ Richards
   - ◉ Shrake
3. Take definition of polar/nonpolar ASA and charged ASA from
   - ○ Chothia
   - ○ Eisenberg
   - ◉ Shrake
4. Type of volume calculation
   - ◉ Standard Voronoi procedure>
   - ○ Richards Method B
   - ○ Radical Plane procedure

**Table Output Options:**

- ☑ Main Chain Information
- ☑ Side Chain Information
- ☑ Hydrogen Bond Information
- ☑ Quality Index Information
- ☑ Statistics Information

Interactions, surface area and volume calculations

http://www.ebi.ac.uk/thornton-srv/software/LIGPLOT/

- Home
- Manual
- Reference
- FAQ


EBI > Groups > Thornton > Software > LIGPLOT

## LIGPLOT v.4.5.3 - Program for automatically plotting protein-ligand interactions

**Written by Andrew Wallace and Roman Laskowski**

LIGPLOT

Automatically generates schematic diagrams of protein-ligand interactions for a given PDB file. (Click on the example on the left).

The interactions shown are those mediated by **hydrogen bonds** and by **hydrophobic contacts**. Hydrogen bonds are indicated by dashed lines between the atoms involved, while hydrophobic contacts are represented by an arc with spokes radiating towards the ligand atoms they contact. The contacted atoms are shown with spokes radiating back.

### Availability

The official LIGPLOT download site is UCLB E-LUCID:

UCLB E-LUCID LigPlot for Unix/linux LigPlot for Windows

Alternatively, commercial users may wish to contact our distributor Ebisu.

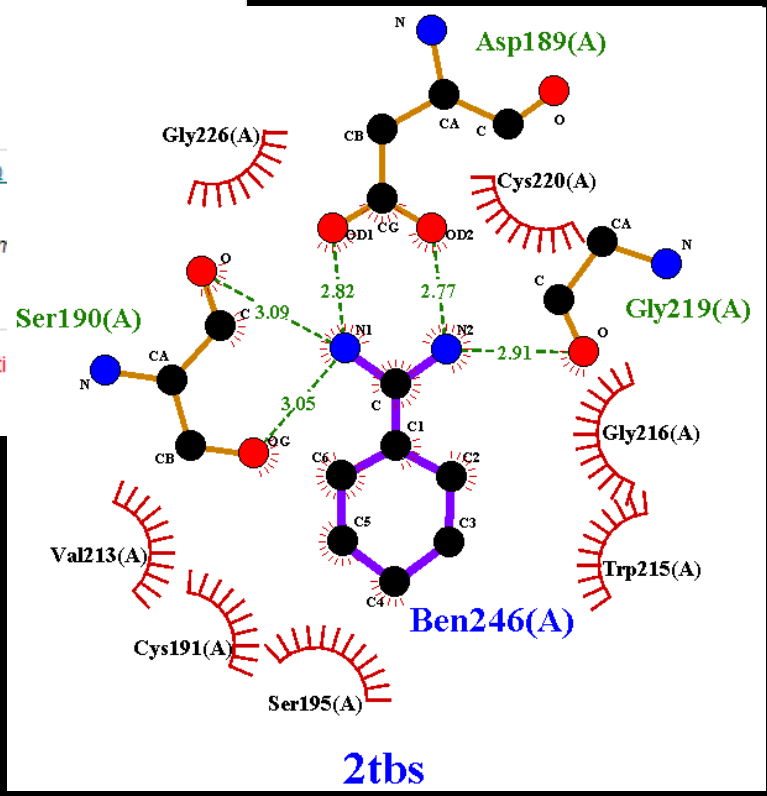
### Contact

If you have any questions about LIGPLOT or problems using it, please check out the FAQ Roman Laskowski at roman@ebi.ac.uk.

*Last n*

### Notes

A new, GUI-based version of **LIGPLOT**, called LigPlot+, is now available for beta-testi




Ligand contacting residues

http://scop.mrc-lmb.cam.ac.uk/scop/

*Structural Classification of Proteins*

Welcome to **SCOP**: Structural Classification of Proteins.
**1.75 release** (June 2009)

38221 PDB Entries. 1 Literature Reference. 110800 Domains. (excluding nucleic acids and theoretical models).
Folds, superfamilies, and families statistics here.
New folds superfamilies families.
List of obsolete entries and their replacements.

**Authors**. Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte
Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia
**Reference:** Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995)
database for the investigation of sequences and structures. *J. Mol. Biol.* 24
**Recent changes** are described in: Lo Conte L., Brenner S. E., Hubba
SCOP database in 2002: refinements accommodate structural genomics.

1. Root: scop
2. Class: All alpha proteins [46456]
3. Fold: Globin-like [46457]
   *core: 6 helices; folded leaf, partly opened*
4. Superfamily: Globin-like [46458]
   *Superfamily*
5. Family: Globins [46463]
   *Heme-binding protein*

**Protein Domains:**

1. Hemoglobin I [46464]
   1. Ark clam (Scapharca inaequivalvis) [TaxId: 6561] [46465] (19)
   2. Clam (Lucina pectinata) [TaxId: 29163] [46466] (4)
2. Trematode hemoglobin/myoglobin [63438]
   1. Paramphistomum epiclitum [TaxId: 54403] [63439] (2)
3. Glycera globin [46467]

Domain definition and evolutionary relationship of domains

CATH
PROTEIN STRUCTURE CLASSIFICATION

Home | Search | Documentation | Tools | Download

Search Clear

Home

**Welcome to CATH**

CATH is a manually curated classification of protein domain structures. Each protein has been chopped into structural domains and assigned into homologous superfamilies (groups of domains that are related by evolution). This classification procedure uses a combination of automated and manual techniques which include computational algorithms, empirical and statistical evidence, literature review and expert analysis.

Find out more about CATH >>

**New in CATH v3.3**

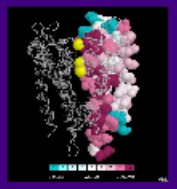
CATH v3.3 is built from 97,625 PDB chains. We have added the following data since v3.2:

- 124 folds (total 1,288)
- 226 superfamilies (total 2,593)
- 1,148 sequence families (total 10,019)
- 14,473 domains (total 128,688)

Download CATH data >>

http://www.cathdb.info/

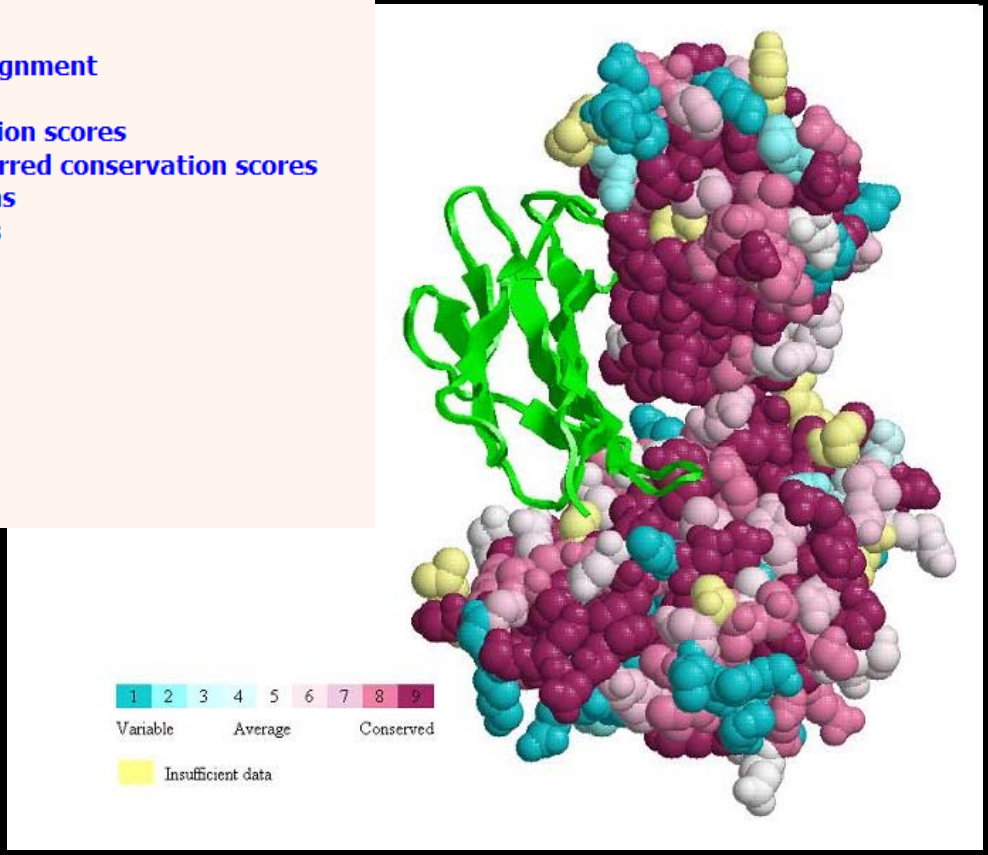
http://consurf.tau.ac.il/overview.html



HOME
GALLERY
OVERVIEW
QUICK HELP
FAQ
BENTAL GROUP
PUPKO GROUP
CITING & CREDITS

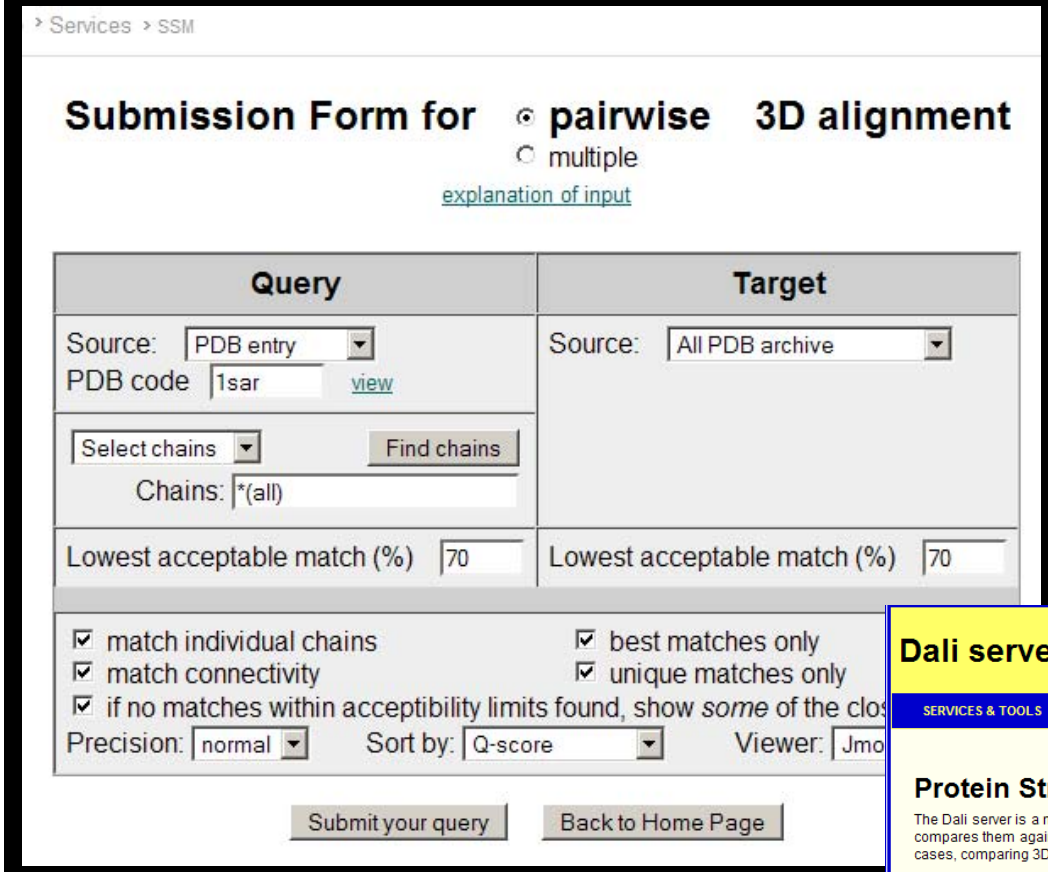
## ConSurf Overview

- Introduction
- Methodology
  - PDB 3D-structure
  - Searching for homologous sequences
  - Generating the multiple sequence alignment
  - Generating the phylogenetic tree
  - Calculating the amino acid conservation scores
    - A confidence interval of the inferred conservation scores
    - Model of substitution for proteins
    - Normalized Conservation Scores
    - Coloring scheme
- Outputs
  - Graphic visualization
- Comparison with other servers
- References

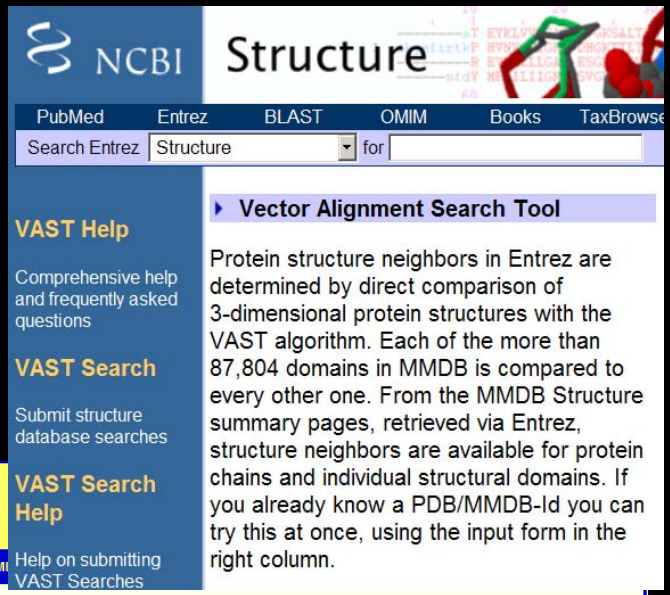


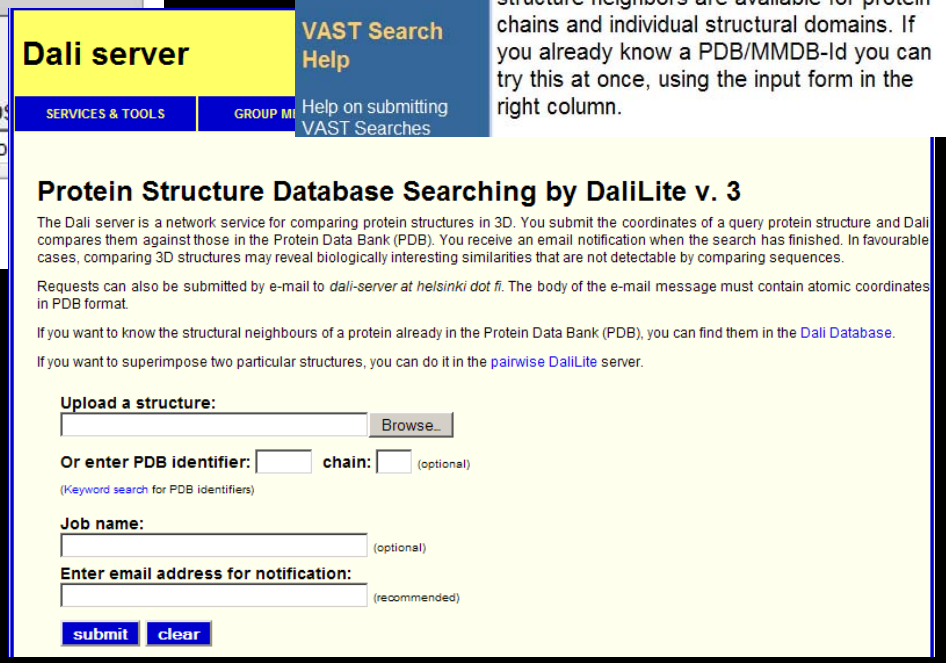

Evolutionary analysis of individual amino acids on the structure

http://www.ebi.ac.uk/msd-srv/ssm/cgi-bin/ssmserver




http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml




Programs to retrieve similar structures

http://ekhidna.biocenter.helsinki.fi/dali_server/

http://supfam.mrc-lmb.cam.ac.uk/elevy/3dcomplex/About.cgi

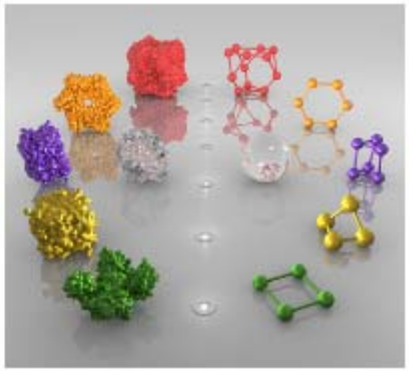
# 3D Complex.org v2.0

A Web Server to browse Protein Complexes of known 3D Structure

Home | Browse | Search & Custom Hierarchy | Download | Errors | About | Links

## Outline

Most proteins act in concert with other proteins, forming permanent or transient complexes. Understanding these interactions at an atomic level is only possible through analysis of protein structures. Over the past few years, there have been efforts to infer the quaternary structures of X-ray crystal structures (Henrick and Thornton 1998; Valdar and Thornton 2001; Ponstingl, Kabir et al. 2003; Bahadur, Chakrabarti et al. 2004), which support the prediction of the Biological Unit to the structure in PDB.



Based on these predictions, we present a visualization and comparison strategy, and construct a hierarchical classification of complexes to integrate and organise the structures. Our strategy is organized in two main steps that we illustrate below:

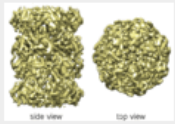
1. The representation of proteins complexes as graphs.
2. The comparison of the graphs.

We also discuss about the symmetry types that are found in protein complexes, and brief explanations can be found here

Obtain related proteins with similar topology to investigate how interfaces evolve in homologous proteins

http://www.ebi.ac.uk/pdbe-srv/emsearch/atlas/1740_visualization.html

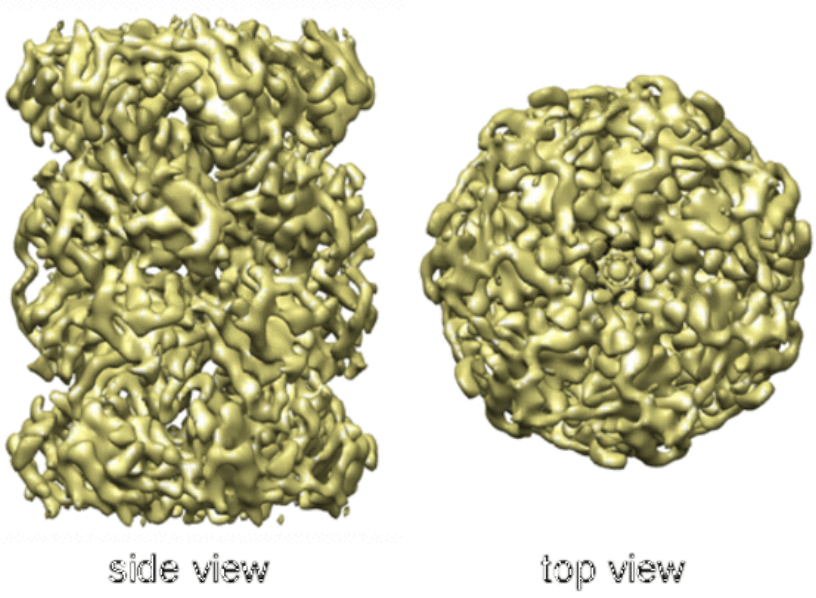
EMDB Home | EMDB Entry EMD-1740 | Contact EMDB



**Title**: Mechanism of Gate Opening in the 20S Proteasome by the Proteasomal ATPases
**Authors**: Rabl J, Smith DM, Yu Y, Chang SC, Goldberg AL, and Cheng Y
**Source**: Thermoplasma acidophilum 20S proteasome
**Aggregation State**: singleParticle, (resolution 6.8 Angstroms)

- Summary
- Visualisation
- Sample
- Experiment
- Processing
- Map Information
- Downloads

## Visualisation

| BIOLOGICAL CONTEXT | EMDB SNAPSHOT |
|---|---|
| Archaeal 20S proteasome with a closed gate |  side view top view |
| **Suggested contour level for viewing the map: 2.5** | Launch EmViewer |

The EMDB images on this site were either supplied by the depositor or generated using CHIMERA.

EMDB: repository for electron microscopy structures

# Databases and tools for structure analysis

| Database | Description |
|---|---|
| CSA | The Catalytic Site Atlas (CSA) is a resource of catalytic sites and residues identified in enzymes using structural data. |
| DSSP | The DSSP database is a database of secondary structure assignments (and much more) for all of the entries in the Protein Data Bank (PDB). |
| EMDB | The Electron Microscopy Data Bank (EMDB) is a public repository for electron microscopy density maps of macromolecular complexes and subcellular structures. |
| FSSP | Fold classification based on structure-structure assignments. |
| HSSP | HSSP (homology-derived structures of proteins) is a derived database merging structural (2-D and 3-D) and sequence information (1-D). |
| PDBe | The Protein Databank in Europe (PDBe). This includes PDBe search tools |
| PDBeChem | The ligand library, a complete chemical description of all the distinct chemical components found within the PDB. |
| PDBeFold | Secondary Structure Matching (SSM) is an interactive service for comparing protein structures in 3D. |
| PDBePisa | A tool for the exploration of macromolecular (protein, DNA/RNA and ligand) interfaces, prediction of probable quaternary structures (assemblies), database searches of structurally similar interfaces and assemblies, as well as searches on various assembly and PDB entry parameters. |
| PDBeMotif | PDBeMotif is an extremely fast and powerful search tool that facilitates exploration of the Protein Data Bank (PDB). |
| PDBe NMR | A respository of NMR structures in PDBe. |
| PDBeView | Provides easy access to the PDB. |
| PDBsum | PDBsum is a pictorial database providing an at-a-glance overview of every macromolecular structure deposited in the Protein Data Bank (PDB). |
| ProFunc | The ProFunc server had been developed to help identify the likely biochemical function of a protein from its three-dimensional structure. |

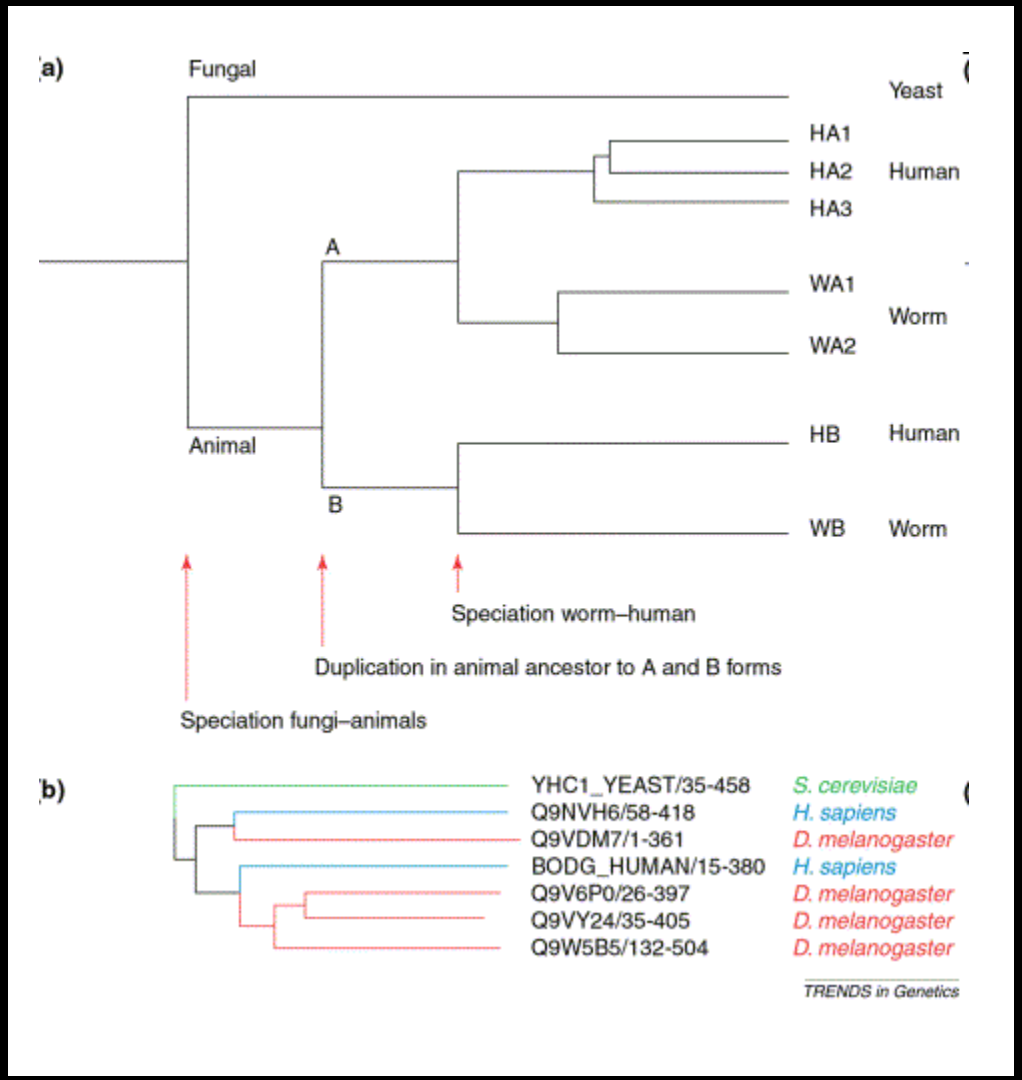
http://www.ebi.ac.uk/Databases/structure.html

# Question set #3

How can I find the ortholog of my protein in other genomes?

Which residues are functionally important?

# Ortholog detection is not trivial




Orthologs are genes in different species that evolved from a common ancestral gene by speciation.

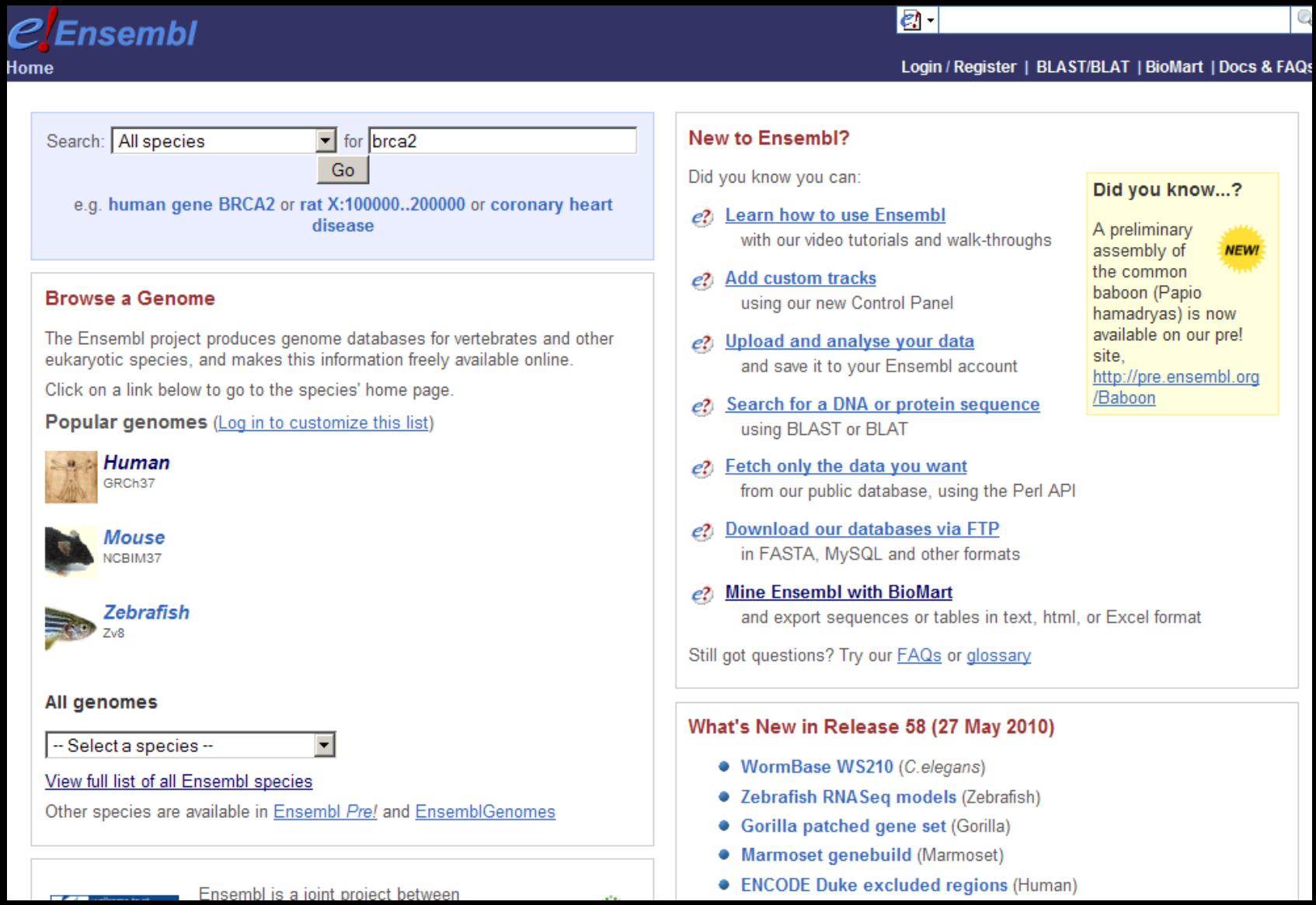
Paralogs are genes related by duplication within a genome.

Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

Need to infer evolutionary history of the gene family

Obtain the tree

Investigate synteny




ENSEMBL: a repository for genome sequence

## Ensembl Species


**Alpaca**
*Vicugna pacos*


**Anole Lizard**
*Anolis carolinensis*


**Armadillo**
*Dasypus novemcinctus*


**Bushbaby**
*Otolemur garnettii*


**Caenorhabditis elegans**


**Ciona intestinalis**


**Ciona savignyi**


**Cat**
*Felis catus*


**Chicken**
*Gallus gallus*


**Chimpanzee**
*Pan troglodytes*


**Cow**
*Bos taurus*


**Dog**
*Canis familiaris*


**Dolphin**
*Tursiops truncatus*


**Elephant**
*Loxodonta africana*


**Guinea Pig**
*Cavia porcellus*


**Hedgehog**
*Erinaceus europaeus*


**Horse**
*Equus caballus*


**Human**
*Homo sapiens*


**Hyrax**
*Procavia capensis*


**Kangaroo rat**
*Dipodomys ordii*


**Lamprey** (preview - assembly only)
*Petromyzon marinus*


**Lesser hedgehog tenrec**
*Echinops telfairi*


**Macaque**
*Macaca mulatta*


**Marmoset**
*Callithrix jacchus*


**Medaka**
*Oryzias latipes*


**Megabat**
*Pteropus vampyrus*


**Microbat**
*Myotis lucifugus*


**Mouse**
*Mus musculus*


**Pig**
*Sus scrofa*


**Pika**
*Ochotona princeps*


**Platypus**
*Ornithorhynchus anatinus*


**Rabbit**
*Oryctolagus cuniculus*


**Rat**
*Rattus norvegicus*


**Saccharomyces cerevisiae**


**Shrew**
*Sorex araneus*


**Sloth**
*Choloepus hoffmanni*


**Squirrel**
*Spermophilus tridecemlineatus*

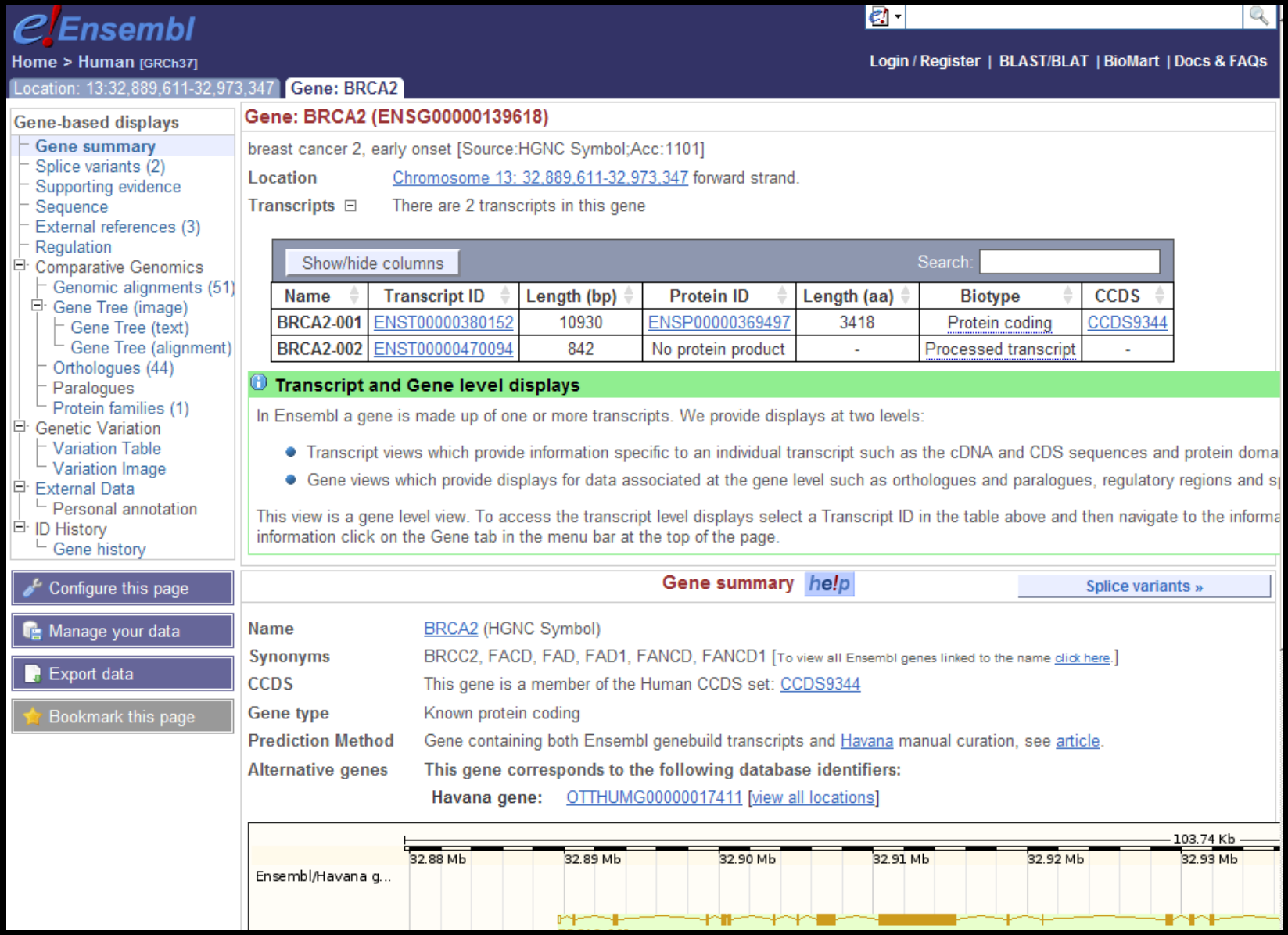

**Stickleback**
*Gasterosteus aculeatus*


**Tarsier**
*Tarsius syrichta*


**Tetraodon**
*Tetraodon nigroviridis*

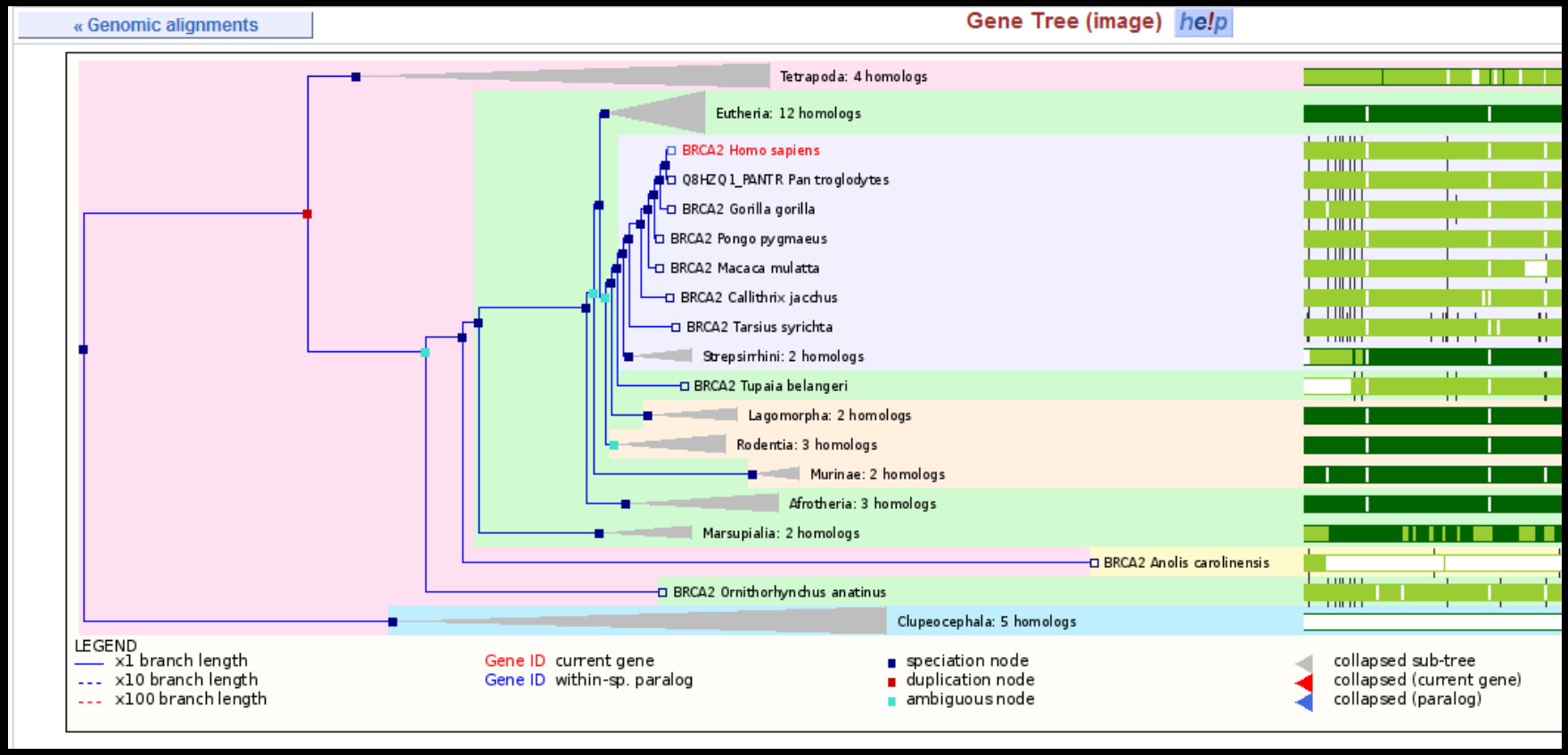

**Tree Shrew**
*Tupaia belangeri*


**Wallaby**
*Macropus eugenii*



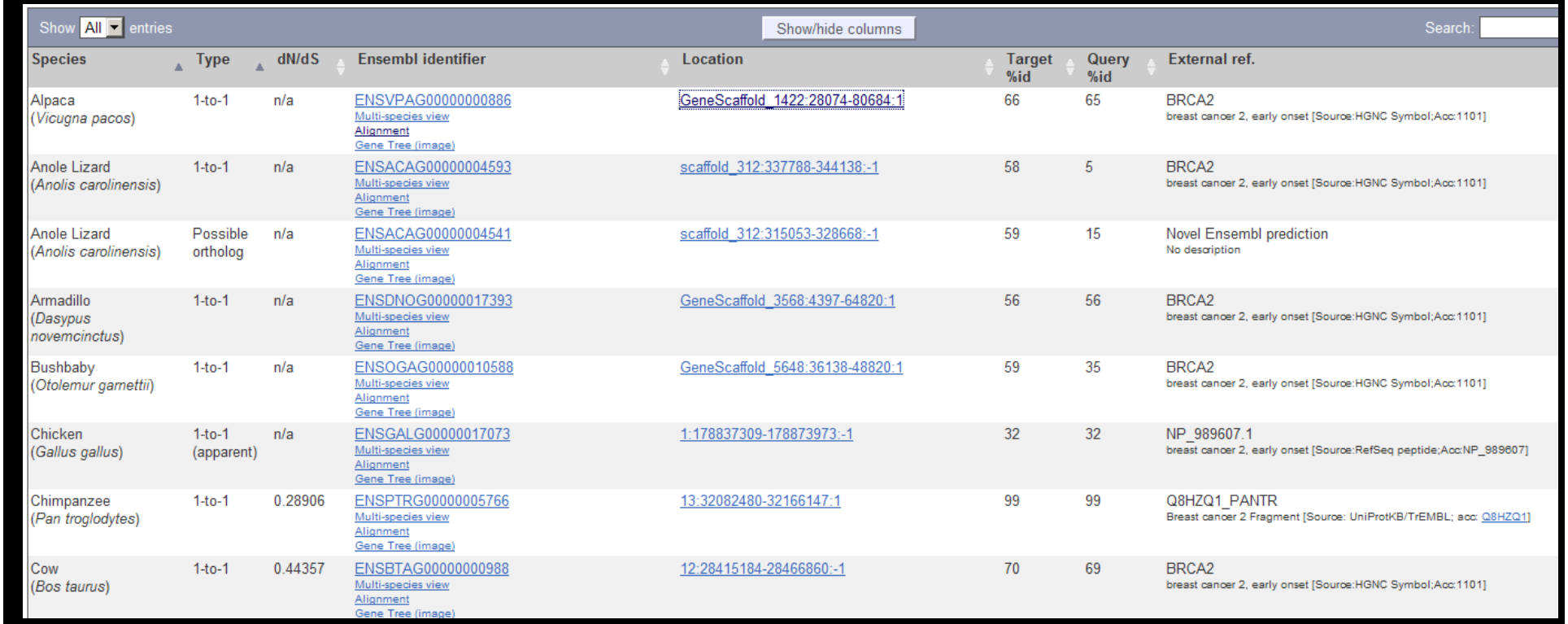

Go to the entry page of a gene of interest

Tree of the evolutionary history of a gene




Click “Gene Tree” to get the tree for further detailed analysis

# Orthology relationship table

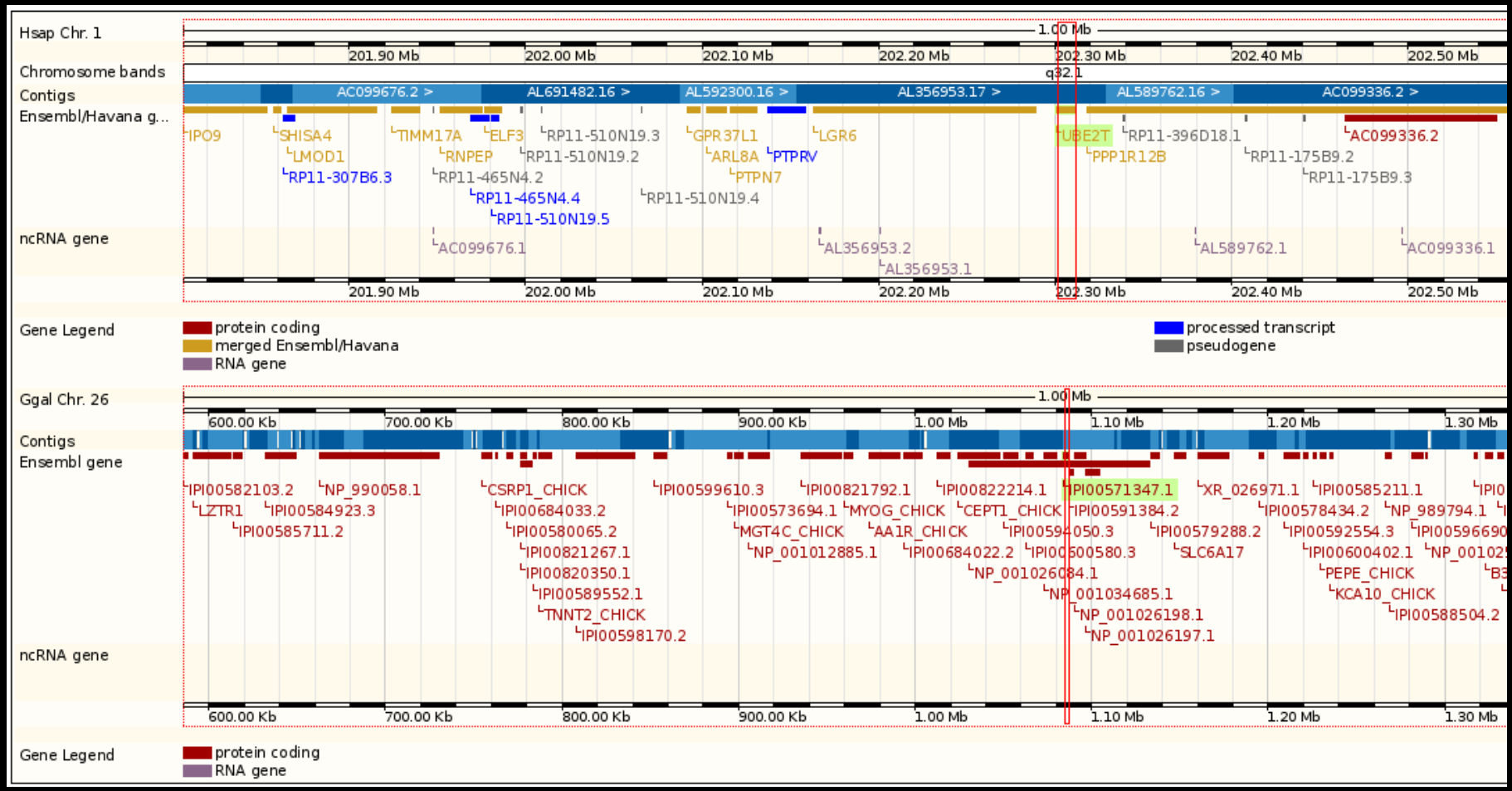


Show All entries | Show/hide columns | Search:

| Species | Type | dN/dS | Ensembl identifier | Location | Target %id | Query %id | External ref. |
|---|---|---|---|---|---|---|---|
| Alpaca (*Vicugna pacos*) | 1-to-1 | n/a | ENSVPAG00000000886<br>Multi-species view<br>Alignment<br>Gene Tree (image) | GeneScaffold_1422:28074-80684:1 | 66 | 65 | BRCA2<br>breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101] |
| Anole Lizard (*Anolis carolinensis*) | 1-to-1 | n/a | ENSACAG00000004593<br>Multi-species view<br>Alignment<br>Gene Tree (image) | scaffold_312:337788-344138:-1 | 58 | 5 | BRCA2<br>breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101] |
| Anole Lizard (*Anolis carolinensis*) | Possible ortholog | n/a | ENSACAG00000004541<br>Multi-species view<br>Alignment<br>Gene Tree (image) | scaffold_312:315053-328668:-1 | 59 | 15 | Novel Ensembl prediction<br>No description |
| Armadillo (*Dasypus novemcinctus*) | 1-to-1 | n/a | ENSDNOG00000017393<br>Multi-species view<br>Alignment<br>Gene Tree (image) | GeneScaffold_3568:4397-64820:1 | 56 | 56 | BRCA2<br>breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101] |
| Bushbaby (*Otolemur garnettii*) | 1-to-1 | n/a | ENSOGAG00000010588<br>Multi-species view<br>Alignment<br>Gene Tree (image) | GeneScaffold_5648:36138-48820:1 | 59 | 35 | BRCA2<br>breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101] |
| Chicken (*Gallus gallus*) | 1-to-1 (apparent) | n/a | ENSGALG00000017073<br>Multi-species view<br>Alignment<br>Gene Tree (image) | 1:178837309-178873973:-1 | 32 | 32 | NP_989607.1<br>breast cancer 2, early onset [Source:RefSeq peptide;Acc:NP_989607] |
| Chimpanzee (*Pan troglodytes*) | 1-to-1 | 0.28906 | ENSPTRG00000005766<br>Multi-species view<br>Alignment<br>Gene Tree (image) | 13:32082480-32166147:1 | 99 | 99 | Q8HZQ1_PANTR<br>Breast cancer 2 Fragment [Source: UniProtKB/TrEMBL; acc: Q8HZQ1] |
| Cow (*Bos taurus*) | 1-to-1 | 0.44357 | ENSBTAG00000000988<br>Multi-species view<br>Alignment<br>Gene Tree (image) | 12:28415184-28466860:-1 | 70 | 69 | BRCA2<br>breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101] |

Click “orthologs” to obtain pre-computed orthology relationship table

# Synteny




Click “Genomic Alignments” to obtain details about synteny

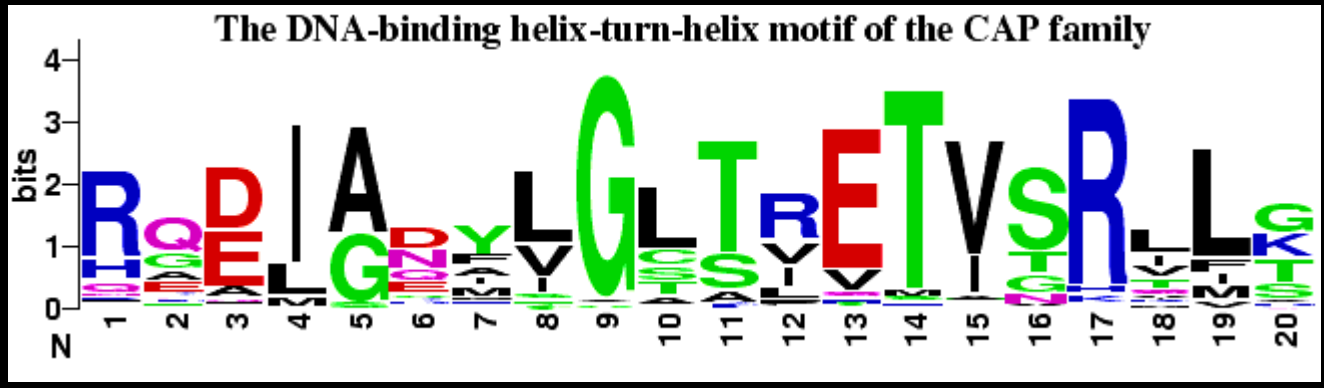
# Alignment

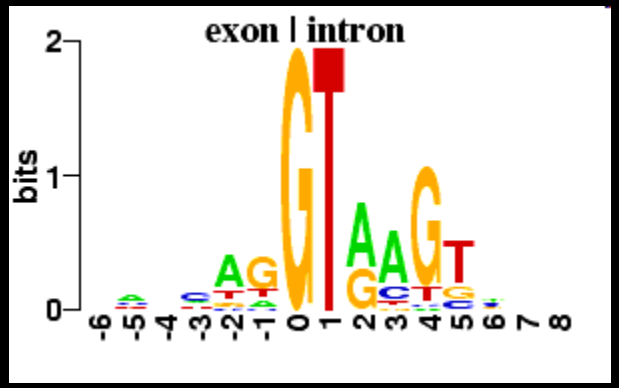
```
ENSMICP00000010933/1-3908  ---------- ---------- ---------- LGPISLNWFD ELSSEAPPYN
ENSSTOP00000004979/1-3908  M------PIG SKERPTFFEI FKTRCSGA-D LGPISLNWFE ELSSEASLYN
ENSOCUP00000014514/1-3908  M------PIG SKERPTLFDI FKMRCNTTVD LGPISLNWFE ELSSEAPLYN
ENSRNOP00000001475/1-3908  M------TVE YKRRPTFWEI FKARCSTA-D LGPISLNWFE ELSSEAPPYN
ENSTGUP00000012130/1-3908  M------ACK PVERPTFFEI FKAHCSES-D LGPISLNWFE ELSAEAPPYE
ENSTRUP00000015030/1-3908  ---------- ---------- ---------- ---------- ----------
ENSECAP00000013146/1-3908  M------PIG CKERPSFFDI FKTRCNKA-D LGPISLNWFE ELSSEAPPYN
ENSDARP00000099674/1-3908  M--------- -------FEN FFHQIDR--E LGPLNPDWFE ELNERAS--R
ENSPCAP00000000440/1-3908  M------PVG FKERPTFFEI FKARCSKA-D LGPISLNWFE ELSSKAPLYQ
ENSMODP00000033276/1-3908  M------SAG YQGKTTFFEV FKTRCSES-D LGPISLNWFE ELTSEAPPYN
ENSGGOP00000015446/1-3908  M------PIG SKERPTFFEI FKTRCNKAVD LGPISLNWFE ELSSEAPPYN
ENSDNOP00000013476/1-3908  ---------- ---------- ---------- LGPISLNWFE ELSSEAPPYI
ENSTBEP00000013856/1-3908  ---------- ---------- ---------- ---------- ----------
ENSGACP00000015199/1-3908  ---------- ---------- ---------- ---------- ----------
ENSPVAP00000000225/1-3908  ---------- ---------- ---------- LGPISLNWFE ELSSEAPLYN
ENSCHOP00000007822/1-3908  M------PVR AKKRPTFFEV FKMRCSKA-X XXXXXXXXXX XXXXXXXXXX
ENSCPOP00000004635/1-3908  M------PIG SKERPTFFEI FKRRCDKA-D LGPISLNWFE ELSSEAPPFN
ENSTSYP00000000441/1-3908  M------PIG SKERPTFFEI FKTRCNKA-D LGPISLNWFE ELSSEAPLYN
ENSEEUP00000008968/1-3908  M------PVG CKERPTFFEI FKRRCNEA-D LGPISLNWFK ELSSEAQPYN
ENSMUSP00000038576/1-3908  M------PVE YKRRPTFWEI FKARCSTA-D LGPISLNWFE ELSSEAPPYN
ENSOANP00000024376/1-3908  M------PVE PKERPTFFEI FKARCSNSGD LGPISLNWFE ELSLEAPPYN
ENSOPRP00000014082/1-3908  M------PTG TKERPTLFEI FKMRCNVT-D LGPISLNWFE ELSSEAPPYN
ENSLAFP00000002234/1-3908  M------PVG FKERPTFFEI FKAQCSKA-D LGPVSLNWFE ELSSEAPPYK
ENSSSCP00000009970/1-3908  M------PIG CKERPTFFEI FKTRCNEA-D LGPVSLNWFE ELSLEAPPYN
ENSDORP00000006609/1-3908  M------PIE SPGRPTFFEI FKTQCSKA-D LGPISLNWFE ELSSEAPPYN
ENSBTAP00000001311/1-3908  M------PIG CKERPTFFDI FKARCNKA-D LGPISLNWFE ELSSEAPLCN
ENSMLUP00000012516/1-3908  M------PIG CKERPTFFDI FET-CSQ--- LGPISINWFE ELSSEAPPYN
ENSMEUP00000009812/1-3908  ---------- ---------- ---------- ---------- ----------
ENSACAP00000004460/1-3908  MAVMVKENDA PTKKPSFFEI YKARCDDS-D LGPISLNWFE ELSSEAPPYD
ENSTNIP00000002435/1-3908  ---------- ---------- ---------- ---------- ----------
ENSORLP00000004773/1-3908  ---------- ---------- ---------- ---------- ----------
ENSSARP00000002541/1-3908  M------PVG CKERPSFWEI FQTRCNQA-D LGPISLNWFE ELSLEAPPYN
ENSCAFP00000009557/1-3908  M------PVG CKERPTFFEI FKTRCNQA-D LGPISLNWFE ELSLEAPPYN
ENSOGAP00000009477/1-3908  ---------- ---------- ---------- ---------- ----------
ENSTTRP00000010004/1-3908  M------PIG CKERPTFFEI FRTRCNKA-D LGPISLNWFE ELSSEAPPYN
ENSMMUP00000009432/1-3908  M------SIG SKERPTFFEI FKTRCNKA-D LGPISLNWFE ELSSEAAPYN
ENSPTRP00000009810/1-3908  M------PIG SKERPTFFEI FKTRCNKA-D LGPISLNWFE ELSSEAPPYN
ENSCJAP00000034250/1-3908  M------PIG SKERPTFFEI FKTRCNKAVD IGPISLNWFE ELSSEASPCN
ENSXETP00000037057/1-3908  ---------- ---------- ---------- ---------- ----------
ENSVPAP00000000821/1-3908  M------PIG CKERPTFFEI FRTRCNKA-D LGPISLNWFE ELSSEAPPYN
ENSPPYP00000005997/1-3908  M------PVG SKERPTFFEI FKTRCNKA-D LGPISLHWFE ELSSEAPPYN
ENSETEP00000003277/1-3908  L-------VG TKERPTFFEI FKARCSKA-D LGPISLNWFE ELSLEAPPYR
ENSGALP00000027524/1-3908  M------AYK SGKRPTFFEV FKAHCSDS-D LGPVSLDWFE ELSSEAPPYE
```

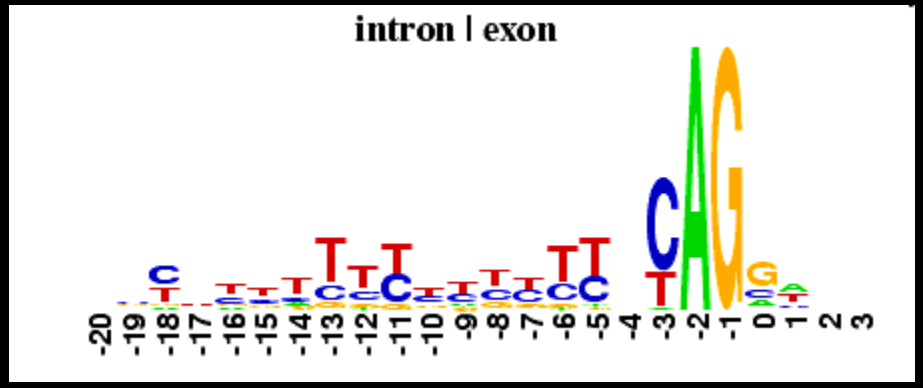
Click “Alignment” to obtain alignments

Weblogo for position specific analysis of an alignment of orthologs

http://weblogo.berkeley.edu/examples.html

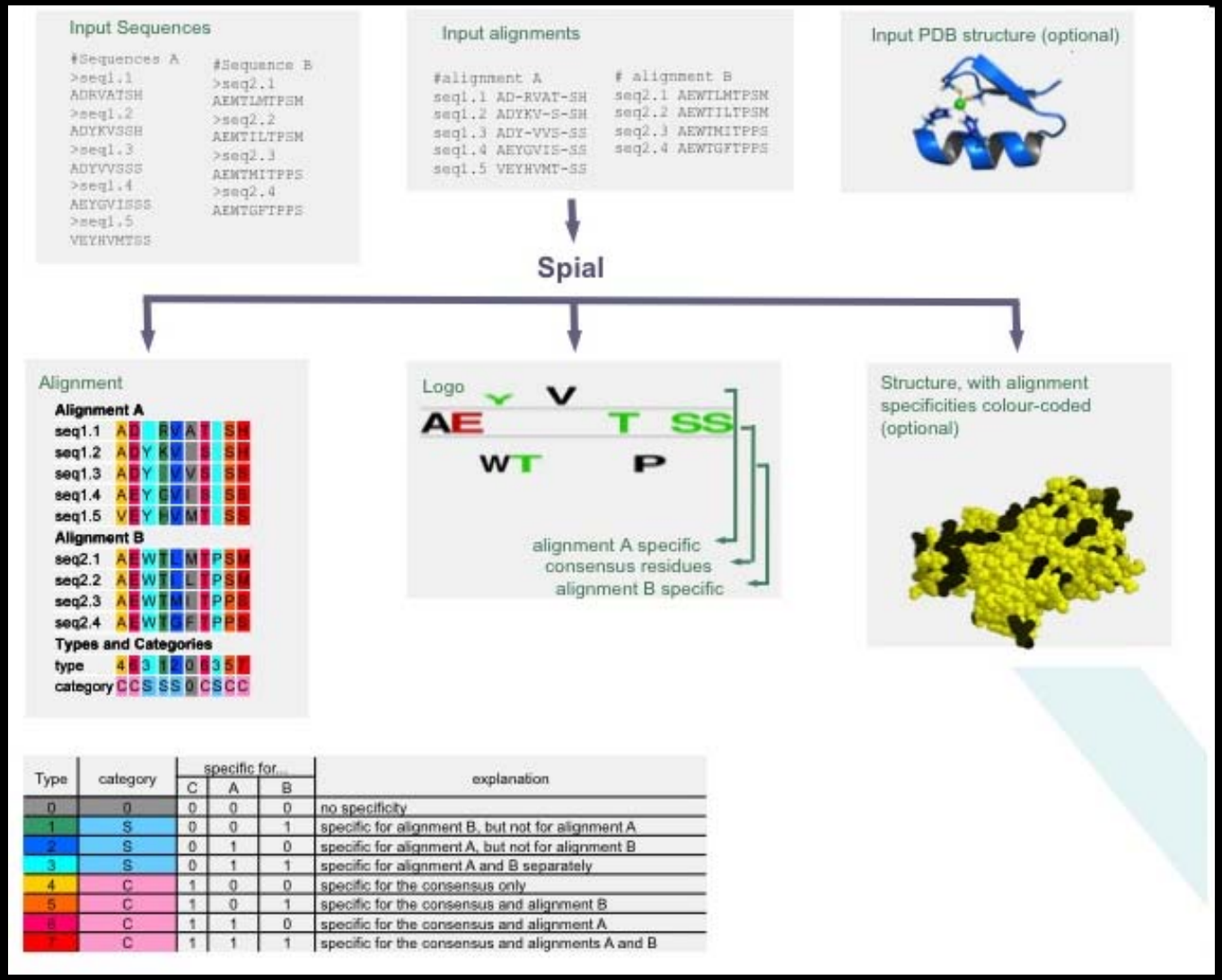



Evolutionary conservation reflects structural or a functional constrain on a position

http://www.mrc-lmb.cam.ac.uk/genomes/spial/

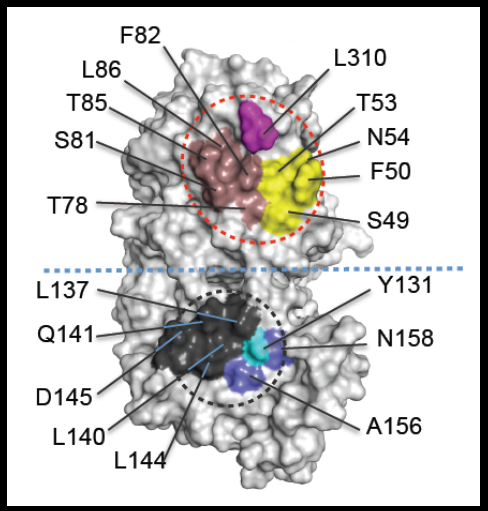


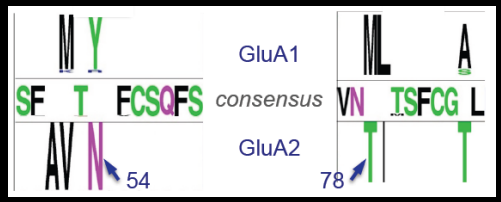

| Type | category | specific for... | | | explanation |
|---|---|---|---|---|---|
| | | C | A | B | |
| 0 | 0 | 0 | 0 | 0 | no specificity |
| 1 | S | 0 | 0 | 1 | specific for alignment B, but not for alignment A |
| 2 | S | 0 | 1 | 0 | specific for alignment A, but not for alignment B |
| 3 | S | 0 | 1 | 1 | specific for alignment A and B separately |
| 4 | C | 1 | 0 | 0 | specific for the consensus only |
| 5 | C | 1 | 0 | 1 | specific for the consensus and alignment B |
| 6 | C | 1 | 1 | 0 | specific for the consensus and alignment A |
| 7 | C | 1 | 1 | 1 | specific for the consensus and alignments A and B |

SPIAL: Specificity conferring residues in paralogous families

Interface residues between stat4 and stat5

Interface residues in glutamate receptor N-terminal domain




Residues that are different between paralogs but conserved among orthologs are specificity conferring positions

# Pre-computed ortholog databases


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

http://eggnog.embl.de/

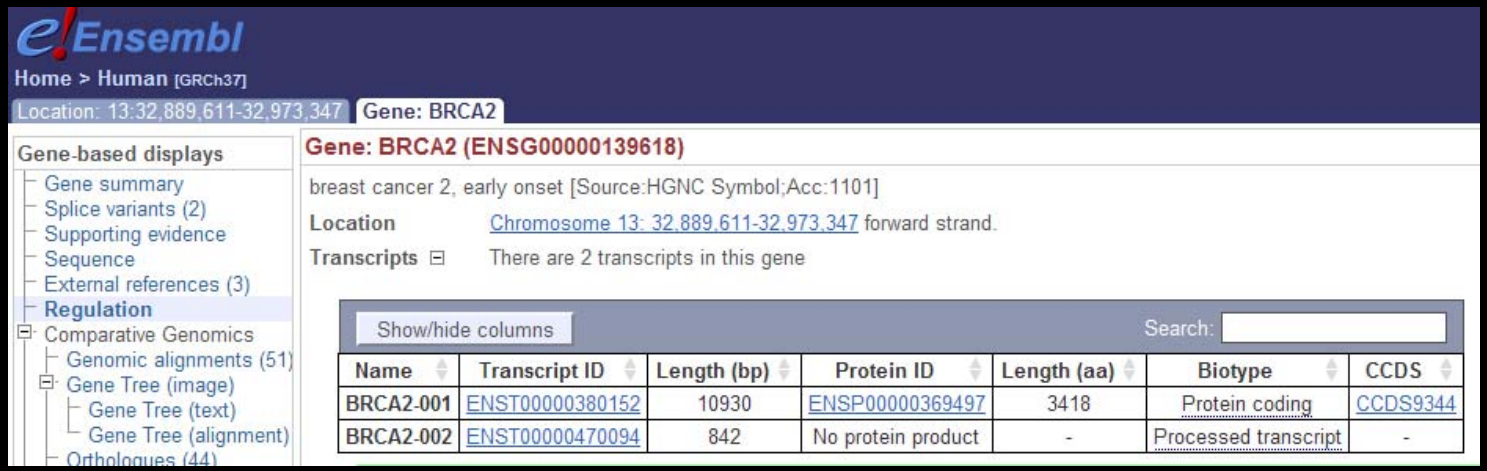
http://inparanoid.sbc.su.se/cgi-bin/index.cgi

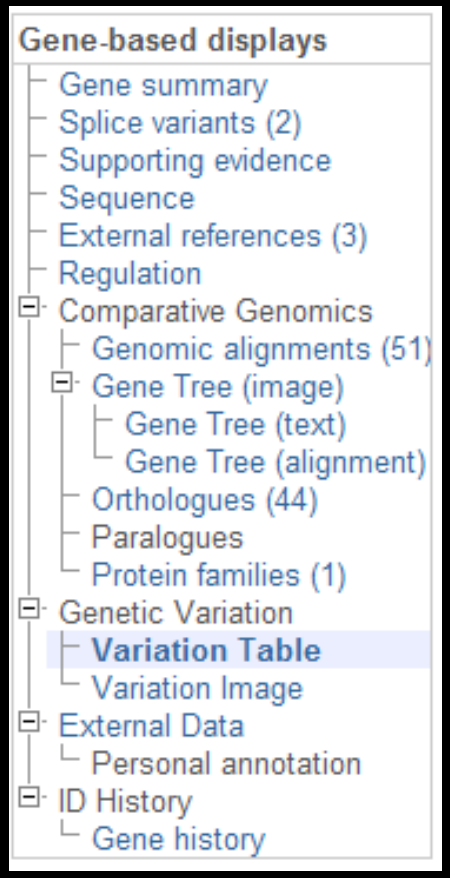
http://mbgd.genome.ad.jp/

# Question Set #4

How can I know more about my favourite gene?

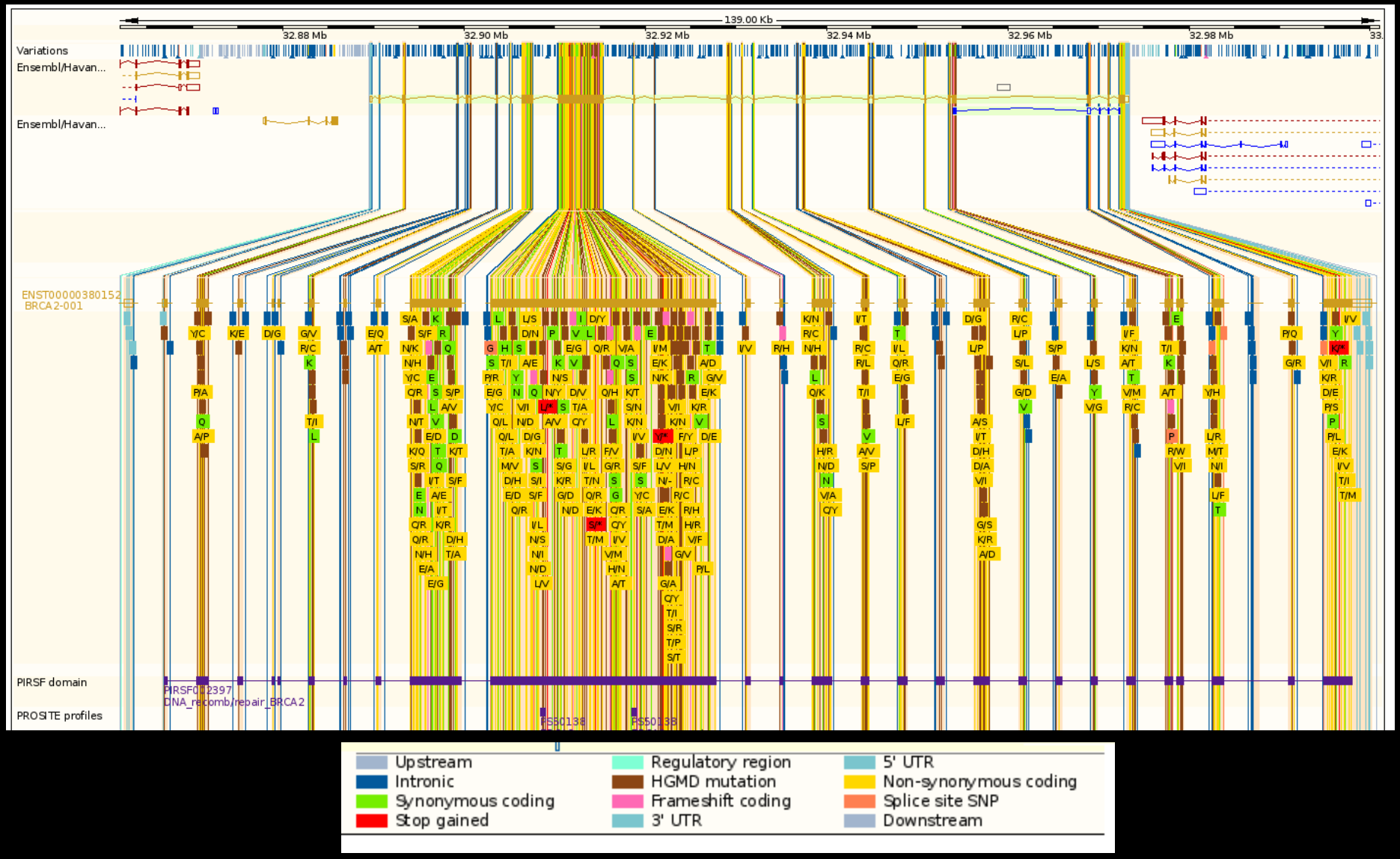
Splice forms, regulators, promoter structure, protein domains, SNPs, paralogs, genomic position, etc..



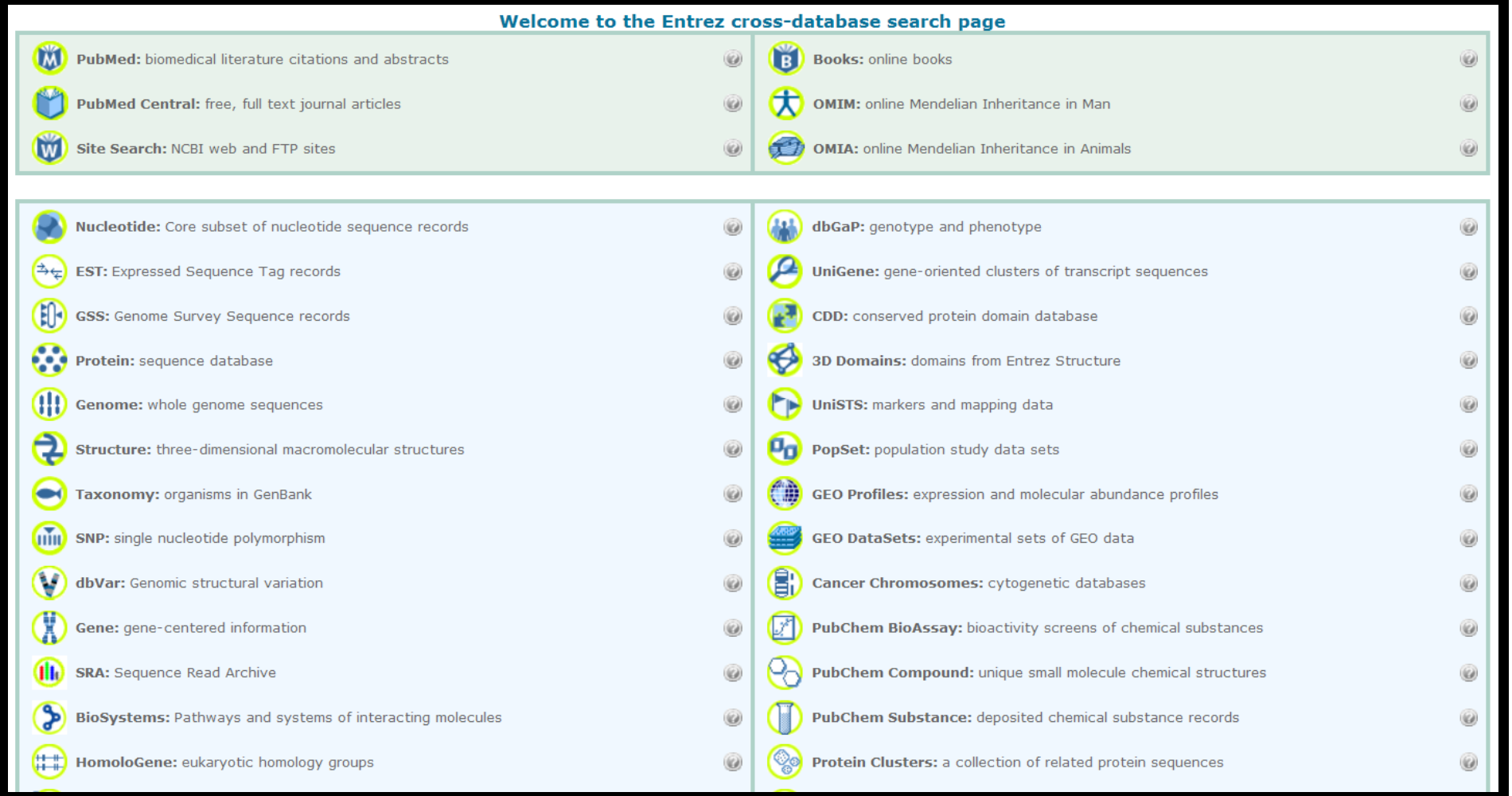

ENSEMBL entry for a gene has links to a lot of relevant information!

# What are the SNPs in my gene (or region) of interest?



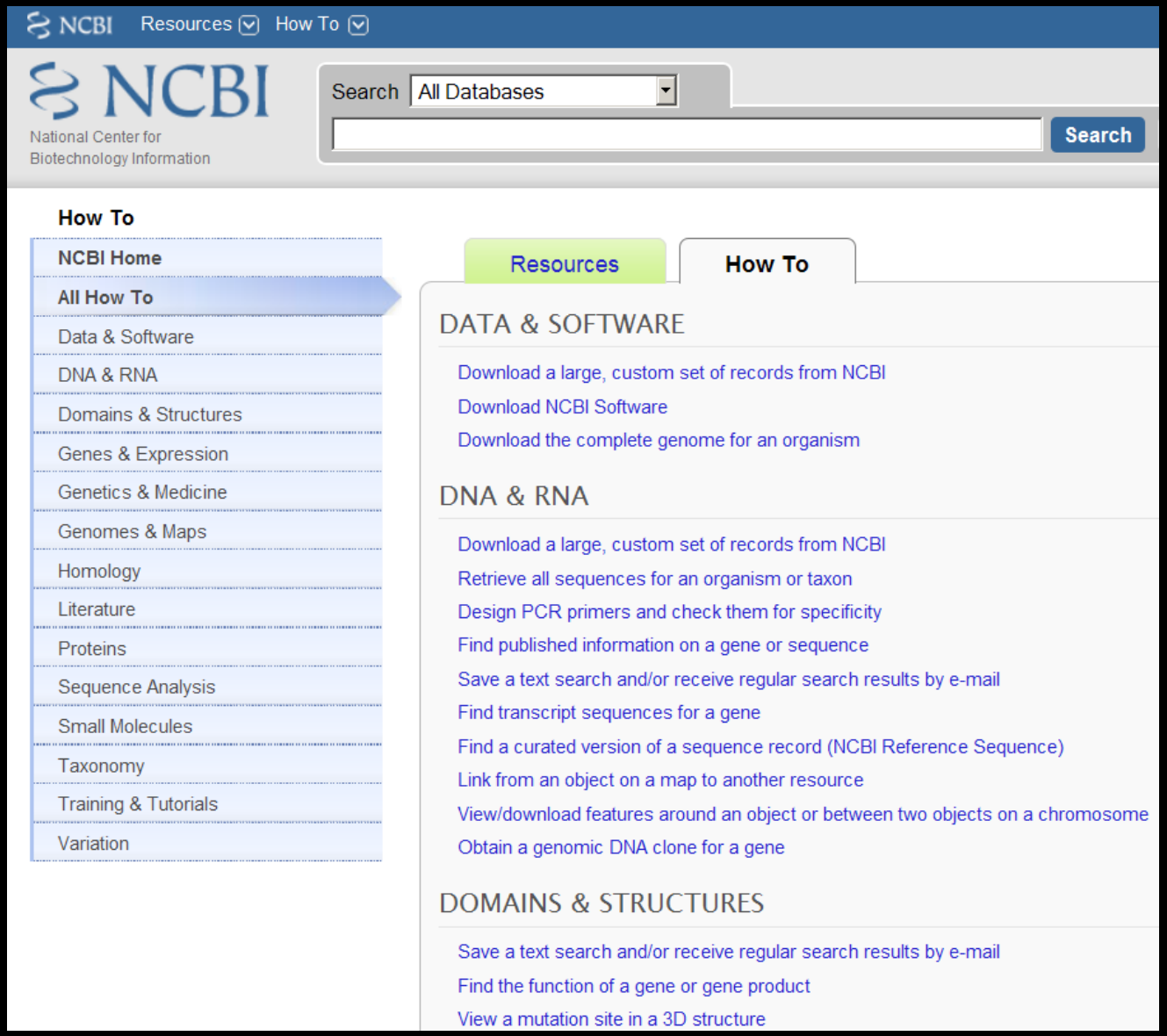
# NCBI

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



# How to use NCBI website effectively

http://www.ncbi.nlm.nih.gov/guide/all/howto/

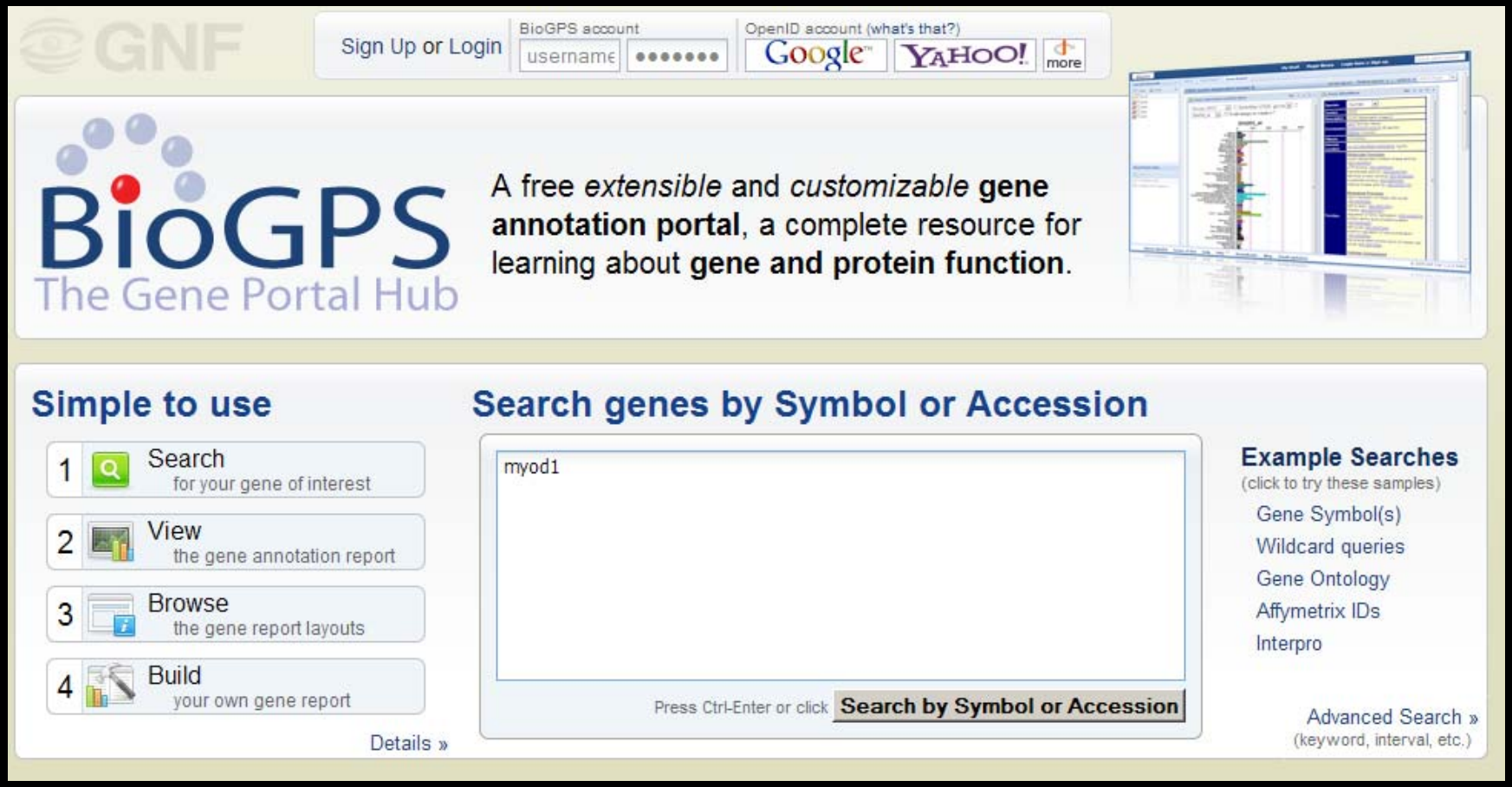


# Question Set #5

In which tissues, cell lines, and conditions are my genes expressed?

What are the transcript levels and protein levels for my gene?

http://biogps.gnf.org/#goto=welcome



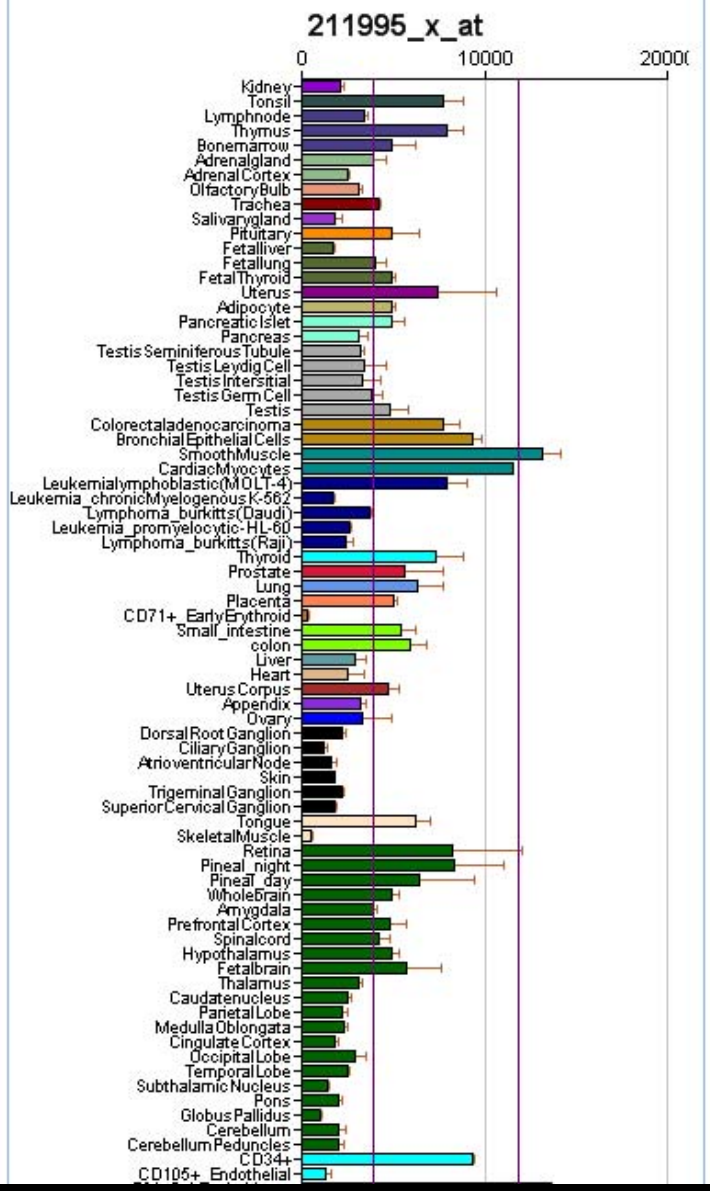

A web-portal for data on expression levels of transcripts in 70 different human and mouse tissues and cancer cell lines

## Act5C (Actin 5C)

### Gene expression/activity chart

Species: Hs

Image | Correlation | Downloads

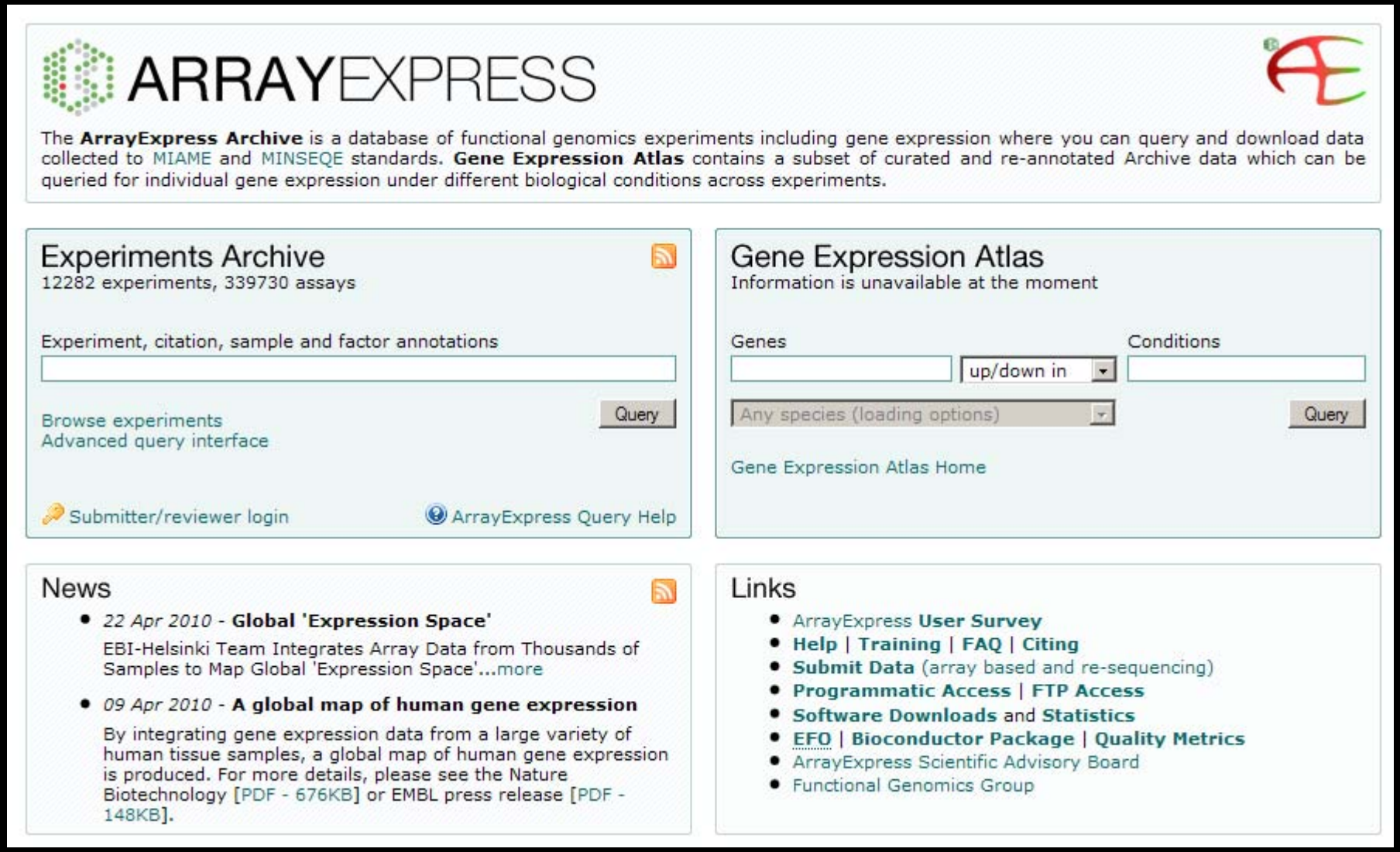



### Gene Identifiers

Species: Dm

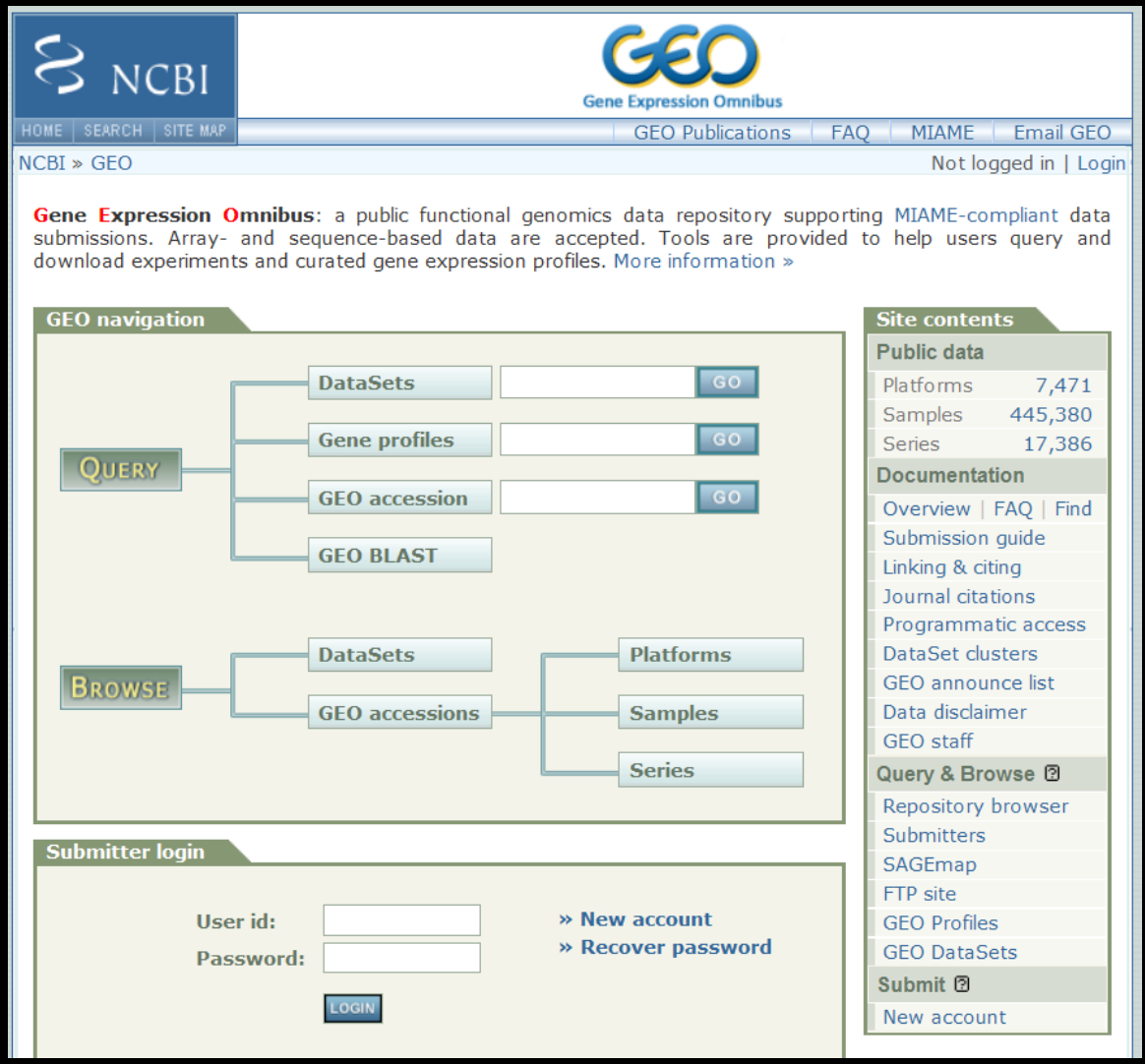
| | |
|---|---|
| Symbol: | Act5C |
| Description: | Actin 5C |
| Accessions: | 31521 (Entrez Gene)<br>FBgn0000042 (Ensembl) |
| Aliases: | A, Ac5C, Act, Act-5C, Act5c, Actin, Actin/BAP47, Actin5C, BAP47, Bap47, CG4027, Dmel\CG4027, M32055, T11, act 5C, act5C, actin, actin-5C, actin5C, anon-EST:fe2D2, beta-actin, beta-actin/Bap47, cyt5C, l(1)G0009, l(1)G0010, l(1)G0025, l(1)G0079, l(1)G0117, l(1)G0177, l(1)G0245, l(1)G0330, l(1)G0420, l(1)G0486 |
| Genome Location: | chrX:5794897-5798823 (dm3) |
| Function: | **Molecular Function**<br>structural constituent of cytoskeleton (GO:0005200)<br>protein binding (GO:0005515)<br>ATP binding (GO:0005524)<br><br>**Biological Process**<br>cytokinesis (GO:0000910)<br>inter-male aggressive behavior (GO:0002121)<br>phagocytosis, engulfment (GO:0006911)<br>cytoskeleton organization (GO:0007010)<br>sperm individualization (GO:0007291)<br>mushroom body development (GO:0016319)<br>maintenance of protein location in cell (GO:0032507)<br><br>**Cellular Component**<br>lipid particle (GO:0005811)<br>actin filament (GO:0005884)<br>Ino80 complex (GO:0031011) |
| Interpro: | Actin/actin-like (IPR004000)<br>Actin, conserved site (IPR004001) |
| Transcripts: | NM_001014725.1<br>NM_001014726.1<br>NM_078497.2<br>NM_167053.1 |
| Proteins: | NP_001014725.1<br>NP_001014726.1<br>NP_511052.1<br>NP_727048.1 |
| Reporters: | |

http://www.ebi.ac.uk/microarray-as/ae/



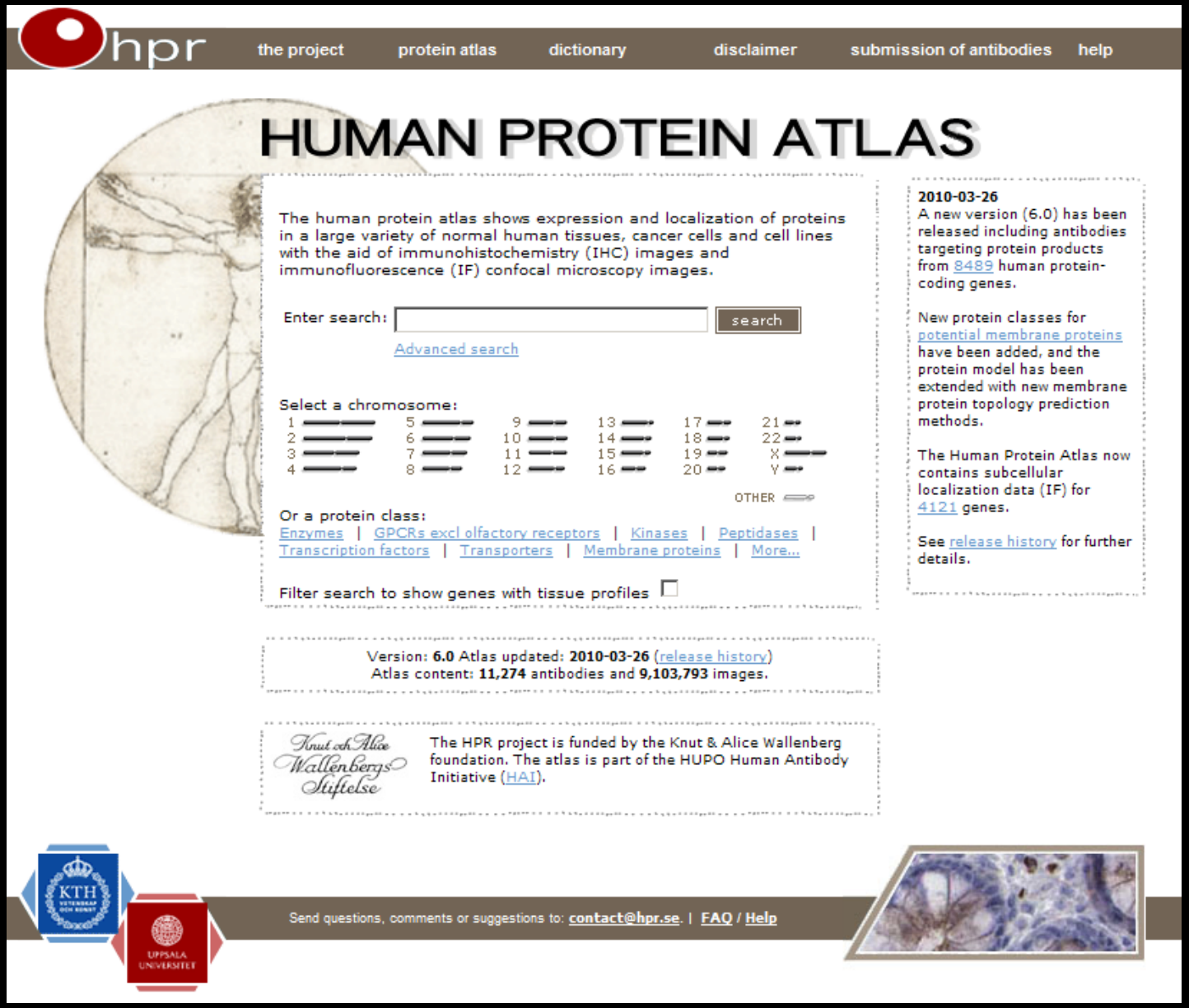

Public repository for gene expression datasets at the EBI

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



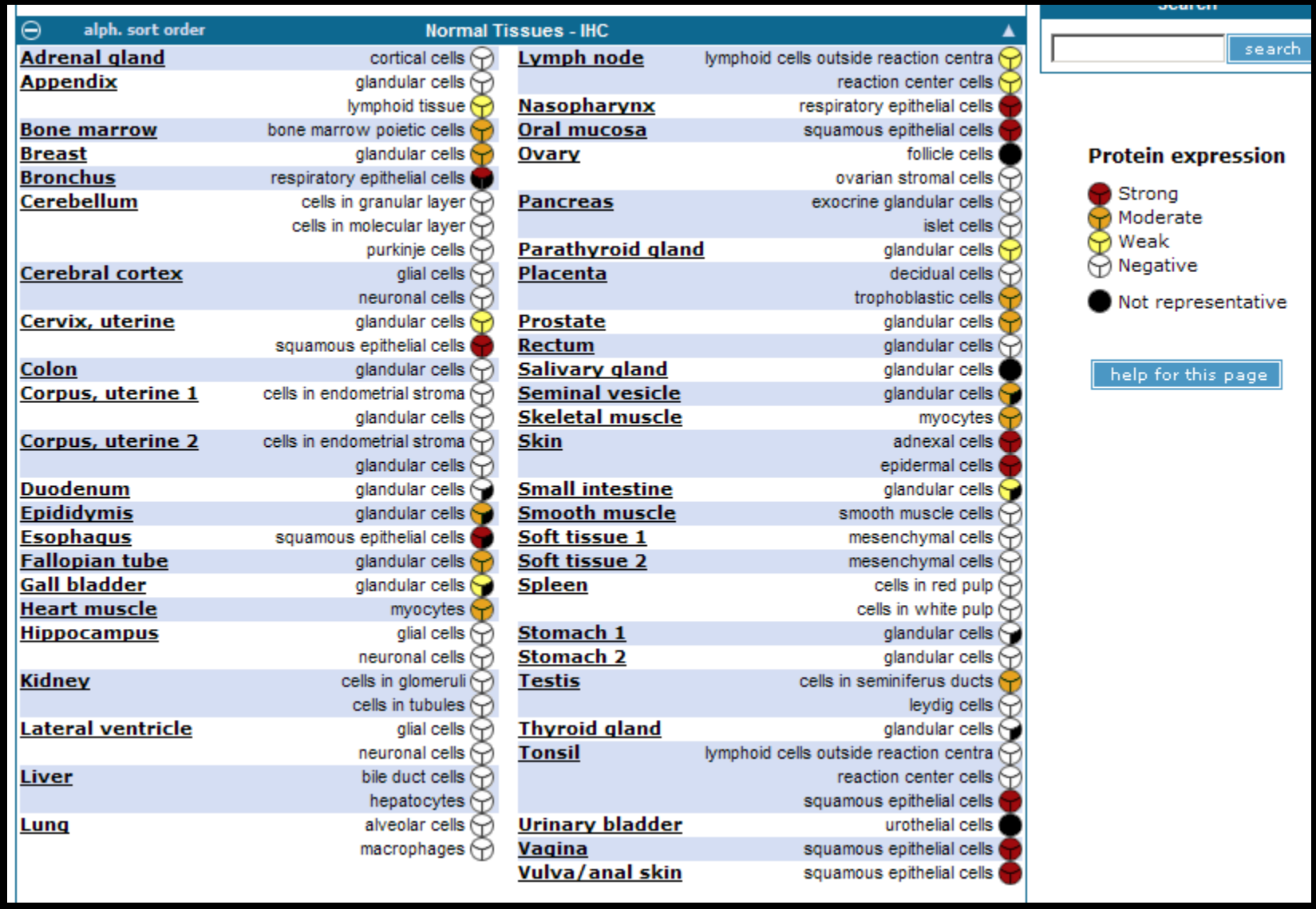

Public repository for gene expression datasets at the NCBI

http://www.proteinatlas.org/index.php




Human protein expression levels (Immunohistochemistry)

# Immunohistochemistry in normal cells



alph. sort order

Normal Tissues - IHC

| | | | |
|---|---|---|---|
| Adrenal gland | cortical cells | Lymph node | lymphoid cells outside reaction centra |
| Appendix | glandular cells | | reaction center cells |
| | lymphoid tissue | Nasopharynx | respiratory epithelial cells |
| Bone marrow | bone marrow poietic cells | Oral mucosa | squamous epithelial cells |
| Breast | glandular cells | Ovary | follicle cells |
| Bronchus | respiratory epithelial cells | | ovarian stromal cells |
| Cerebellum | cells in granular layer | Pancreas | exocrine glandular cells |
| | cells in molecular layer | | islet cells |
| | purkinje cells | Parathyroid gland | glandular cells |
| Cerebral cortex | glial cells | Placenta | decidual cells |
| | neuronal cells | | trophoblastic cells |
| Cervix, uterine | glandular cells | Prostate | glandular cells |
| | squamous epithelial cells | Rectum | glandular cells |
| Colon | glandular cells | Salivary gland | glandular cells |
| Corpus, uterine 1 | cells in endometrial stroma | Seminal vesicle | glandular cells |
| | glandular cells | Skeletal muscle | myocytes |
| Corpus, uterine 2 | cells in endometrial stroma | Skin | adnexal cells |
| | glandular cells | | epidermal cells |
| Duodenum | glandular cells | Small intestine | glandular cells |
| Epididymis | glandular cells | Smooth muscle | smooth muscle cells |
| Esophagus | squamous epithelial cells | Soft tissue 1 | mesenchymal cells |
| Fallopian tube | glandular cells | Soft tissue 2 | mesenchymal cells |
| Gall bladder | glandular cells | Spleen | cells in red pulp |
| Heart muscle | myocytes | | cells in white pulp |
| Hippocampus | glial cells | Stomach 1 | glandular cells |
| | neuronal cells | Stomach 2 | glandular cells |
| Kidney | cells in glomeruli | Testis | cells in seminiferus ducts |
| | cells in tubules | | leydig cells |
| Lateral ventricle | glial cells | Thyroid gland | glandular cells |
| | neuronal cells | Tonsil | lymphoid cells outside reaction centra |
| Liver | bile duct cells | | reaction center cells |
| | hepatocytes | | squamous epithelial cells |
| Lung | alveolar cells | Urinary bladder | urothelial cells |
| | macrophages | Vagina | squamous epithelial cells |
| | | Vulva/anal skin | squamous epithelial cells |

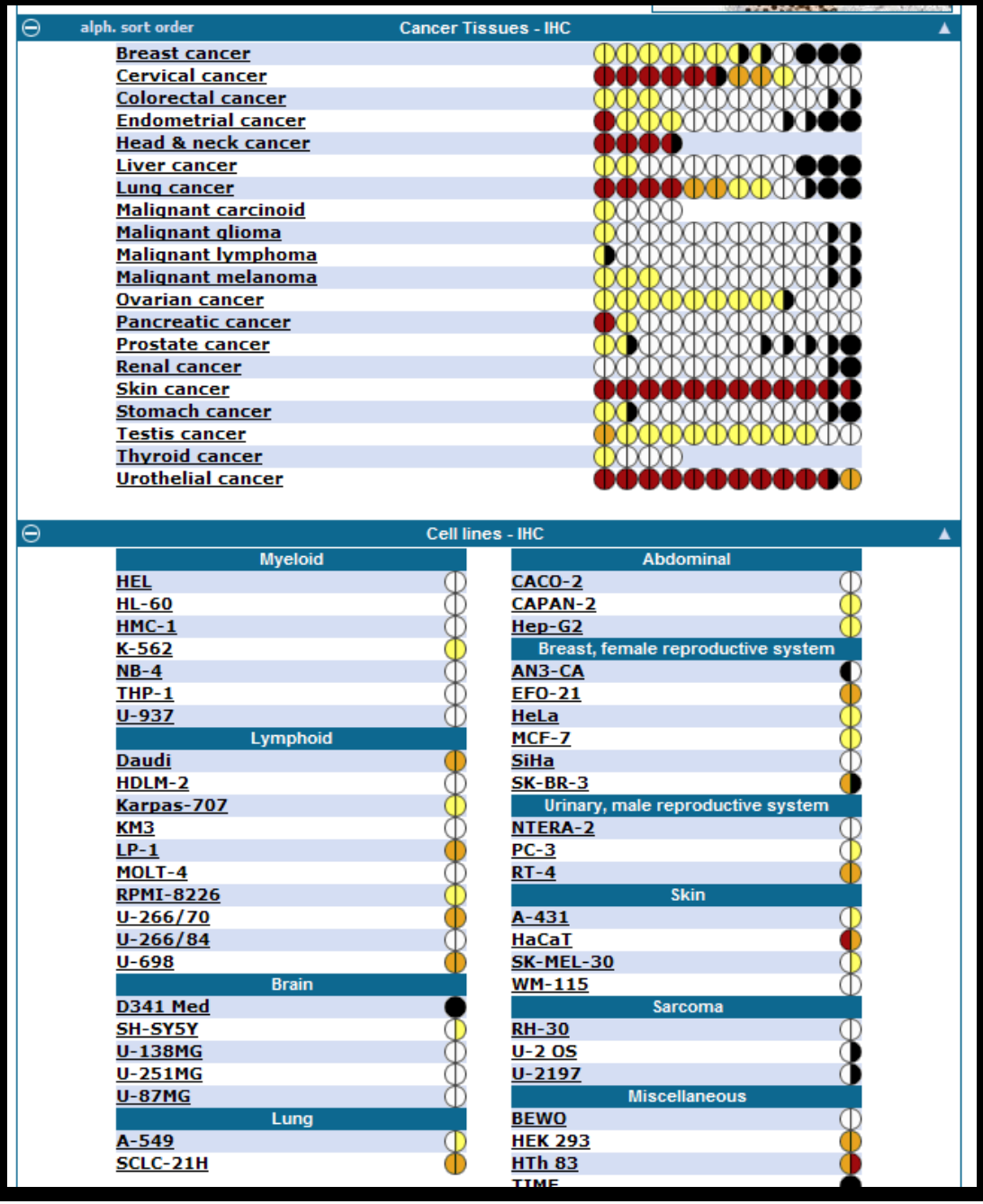
search

**Protein expression**

- Strong
- Moderate
- Weak
- Negative
- Not representative

help for this page

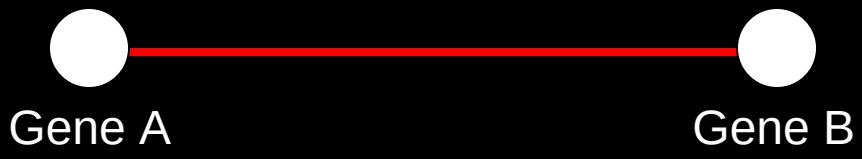
# Immunohistochemistry in cancer cells and cell lines



# Question Set #6

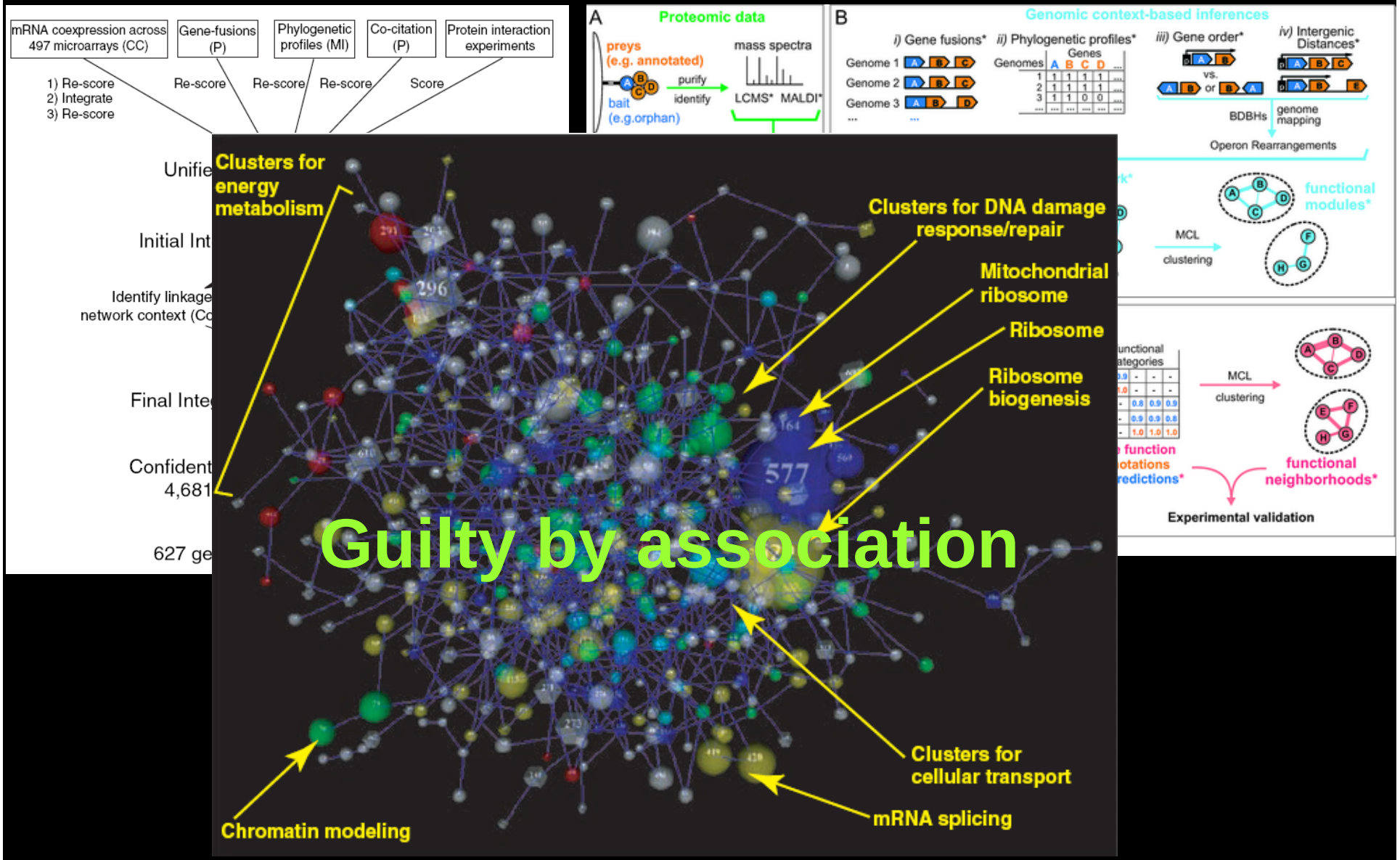
Where can I obtain functional networks for my gene(s)?

# What is a functional interaction network?

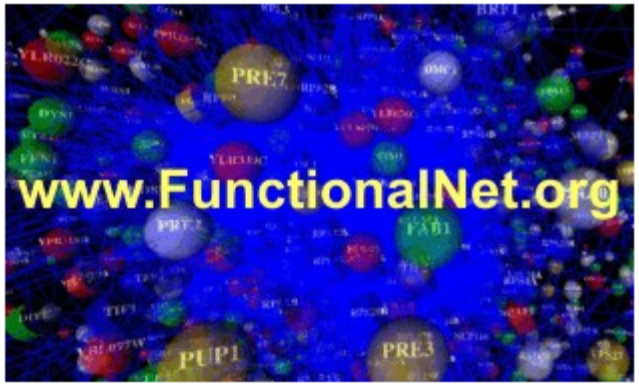



Edge and its weight is a function of:
Co-regulation
Co-expression across several conditions
Physical interaction
Co-localisation
Co-evolution (present/absent together in several genomes)
Gene fusion in another organism
Similar phenotype when knocked out
Toxic when over expressed
Genetically interact
Epistatic interaction
Synthetic lethality
Co-citation

# Functional interaction network



Functional interaction network for eukaryotes



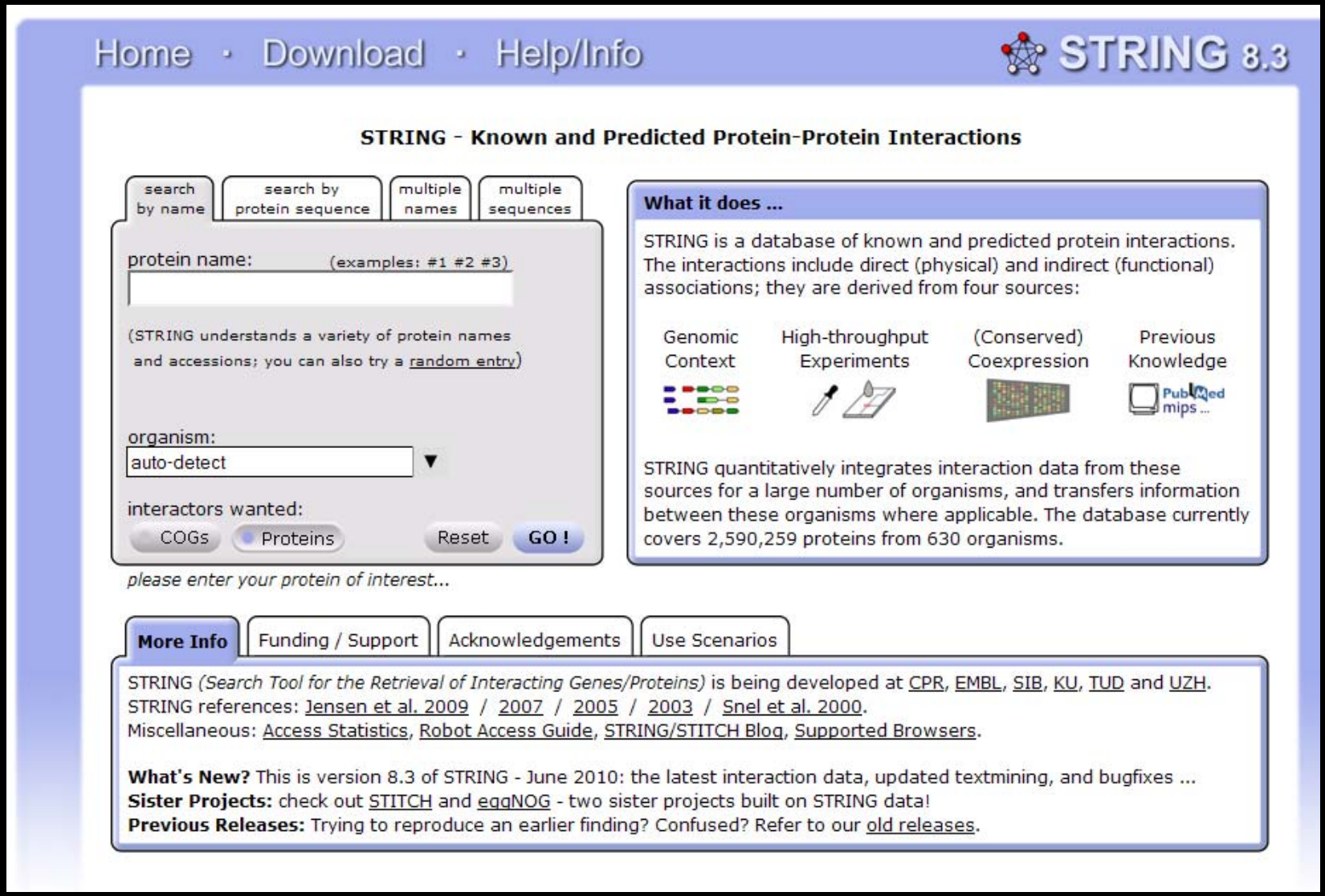

**Welcome to the www.FunctionalNet.org server**

**Networks:** *C. elegans* *S. cerevisiae* *M. musculus* *A. thaliana*

Questions/Comments: Email marcotte AT icmb dot utexas dot edu

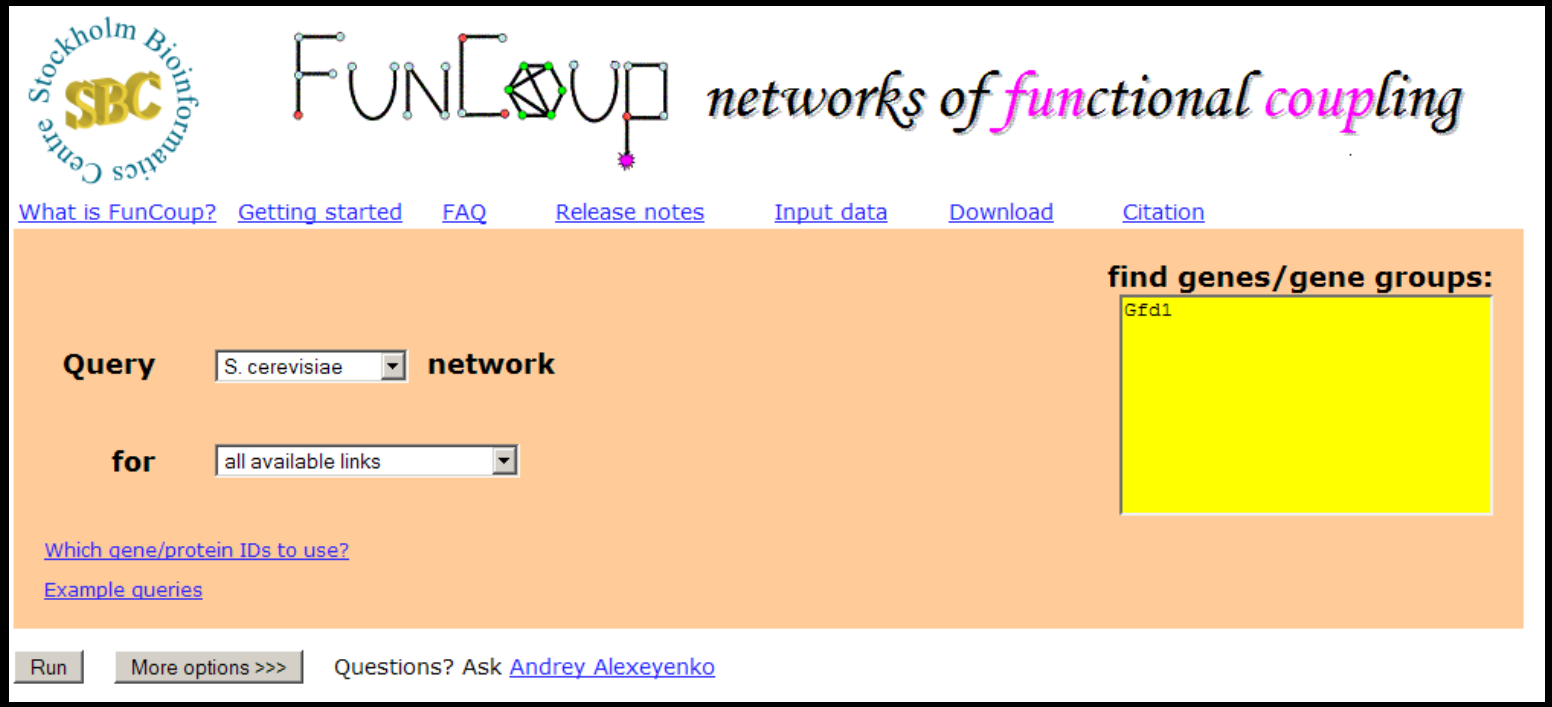
http://www.functionalnet.org/

# Functional interaction network for prokaryotes and eukaryotes




http://string-db.org/

http://funcoup.sbc.su.se/



http://sonorus.princeton.edu/hefalmp/

**HEFalMp** Download | Help | About

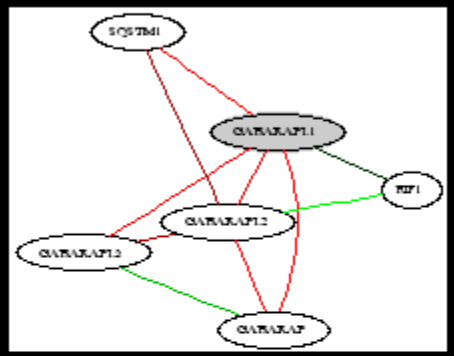
Exploring GABARAPL1 in relation to all genes in the context of all biological processes Update

| Gene ? | Score ? | Description ? |
|---|---|---|
| GABARAPL1 | 1 | GABA(A) receptor-associated protein like 1 |
| GABARAP | 0.9992 | GABA(A) receptor-associated protein |
| GABARAPL3 | 0.9971 | GABA(A) receptors associated protein like 3 |
| SQSTM1 | 0.9947 | sequestosome 1 |
| GABARAPL2 | 0.978 | GABA(A) receptor-associated protein-like 2 |
| PNRC1 | 0.7895 | proline-rich nuclear receptor coactivator 1 |
| EIF1 | 0.7723 | eukaryotic translation initiation factor 1 |
| FOXO3 | 0.7338 | forkhead box O3 |
| C1R | 0.7235 | complement component 1, r subcomponent |
| EIF1B | 0.7115 | eukaryotic translation initiation factor 1B |
| SERINC1 | 0.7078 | serine incorporator 1 |
| HRASLS3 | 0.697 | HRAS-like suppressor 3 |
| CDKN1C | 0.6862 | cyclin-dependent kinase inhibitor 1C (p57, Kip2) |
| WBP2 | 0.6774 | WW domain binding protein 2 |
| SAT1 | 0.6746 | spermidine/spermine N1-acetyltransferase 1 |
| TOM1 | 0.6697 | target of myb1 (chicken) |
| NINJ1 | 0.6461 | ninjurin 1 |
| DNAJB9 | 0.6293 | DnaJ (Hsp40) homolog, subfamily B, member 9 |
| ZFAND5 | 0.6119 | zinc finger, AN1-type domain 5 |
| KLF9 | 0.6116 | Kruppel-like factor 9 |

Download

## GABARAPL1

GABA(A) receptor-associated protein like 1

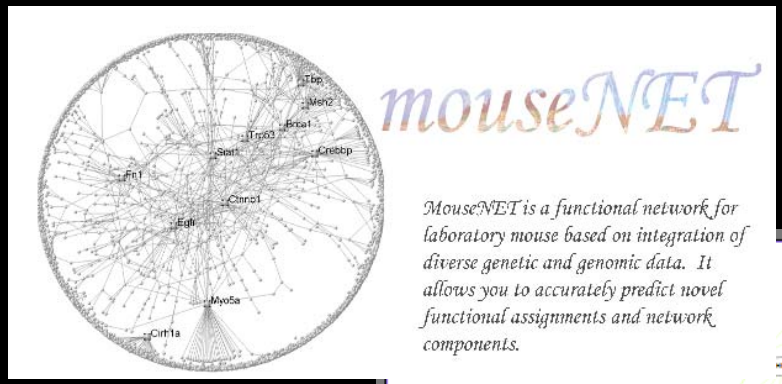


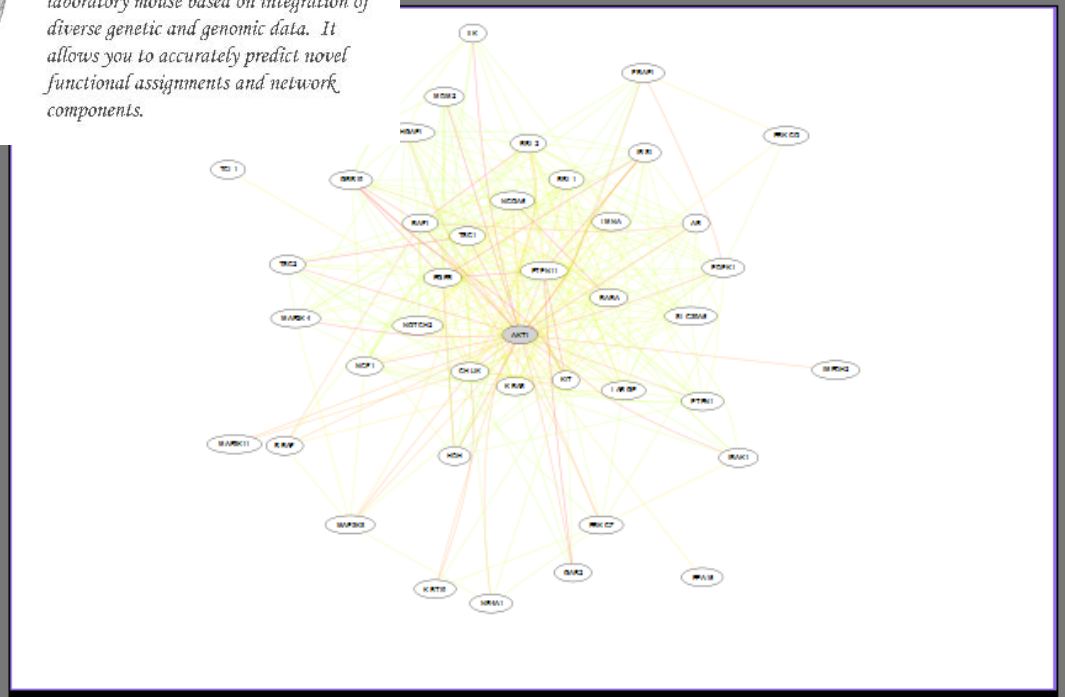
Ensembl: ENSG00000139112
Entrez: 23710
GeneCard: GABARAPL1
Uniprot: Q6FIE6, Q9BY60, Q9H0R8

**Legend**

Relationship confidence

0.0 0.5 1.0



**Legend**

| | |
|---|---|
| PHO85 | **Characterized Gene** |
| BNI1 | **Input Gene** |
| PWP1 | **Uncharacterized Gene** |
| 1 | **Single gene graph**- Click to view graph around that gene |
| L | **Single gene list**- Click to view list of all interactions for that gene |
| Click on a gene to get MGI info | |
| | **Interaction** - Click highlighted edge to view evidence. Edge strength from low to high below. |

0 0.5 1.0

Maximize Graph | Download Ranked list of Genes In Graph | View All Interactions | View Disease-related genes

Mouse over nodes to highlight connecting edges, and click on highlighted edges to view interaction evidence. Click "Maximize Graph" to view the graph in a larger window. Can't see the graphs? Download the Adobe Scalable Vector Graphic (SVG) plugin. You will need to enable pop-ups to view evidence.

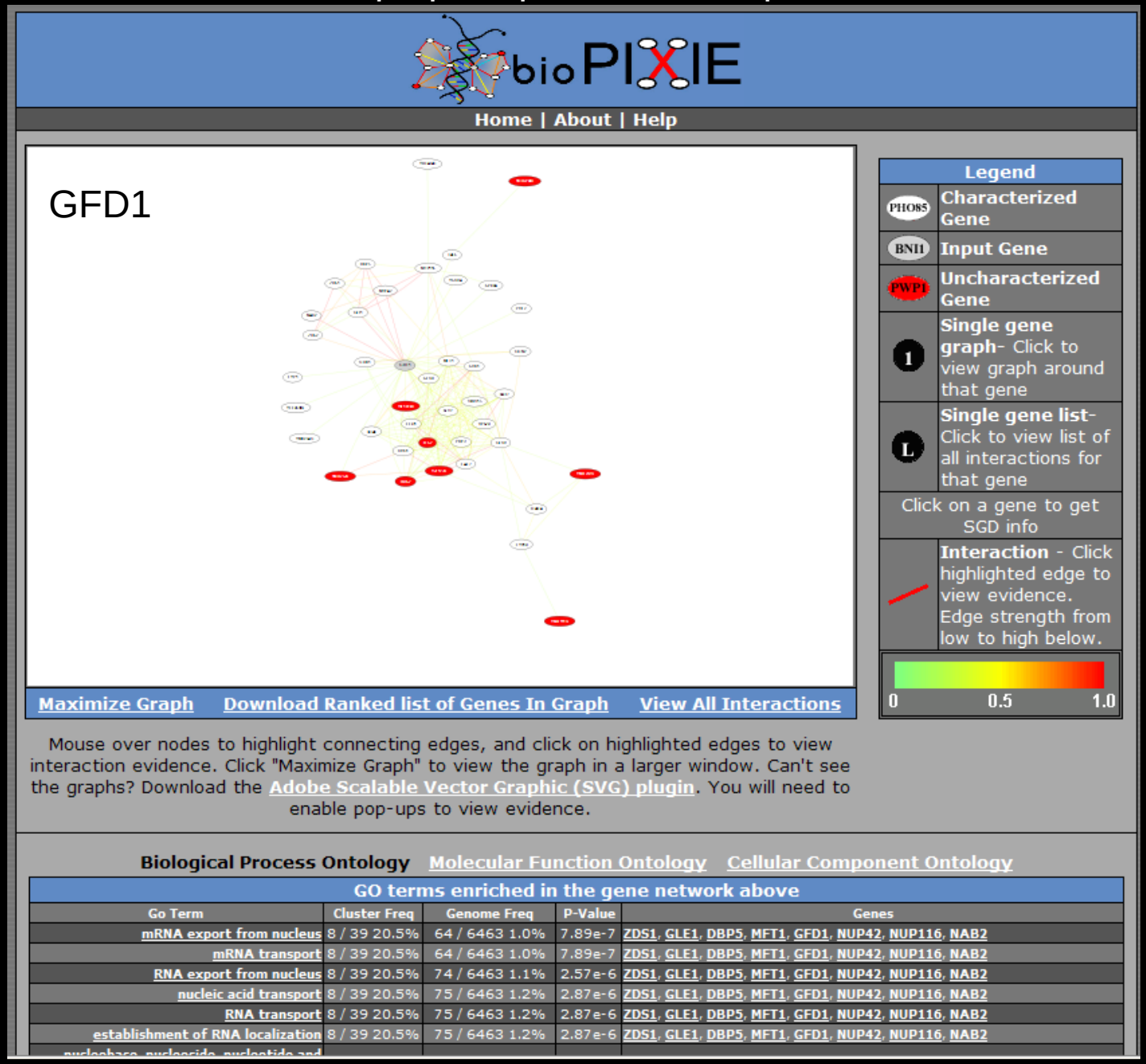
Biological Process Ontology | Molecular Function Ontology | Cellular Component Ontology

**GO terms enriched in the gene network above**

| Go Term | Cluster Freq | Genome Freq | P-Value | Genes |
|---|---|---|---|---|
| phosphorus metabolic process | 17 / 40 42.5% | 887 / 15817 5.6% | 6.19e-9 | RAF1, PRKCQ, MAP3K11, EGFR, PRKCZ, MAP2K4, KIT, PDPK1, FRAP1, CHUK, MAP3K8, PTPN11, PTPN1, IRAK1, ILK, BRAF, AKT1 |
| phosphate metabolic process | 17 / 40 | 887 / 15817 | 6.19e-9 | RAF1, PRKCQ, MAP3K11, EGFR, PRKCZ, MAP2K4, KIT, PDPK1, FRAP1, |

http://avis.princeton.edu/mouseNET/viewgraph.php?graphID=42

http://pixie.princeton.edu/pixie/




For Jack Dean

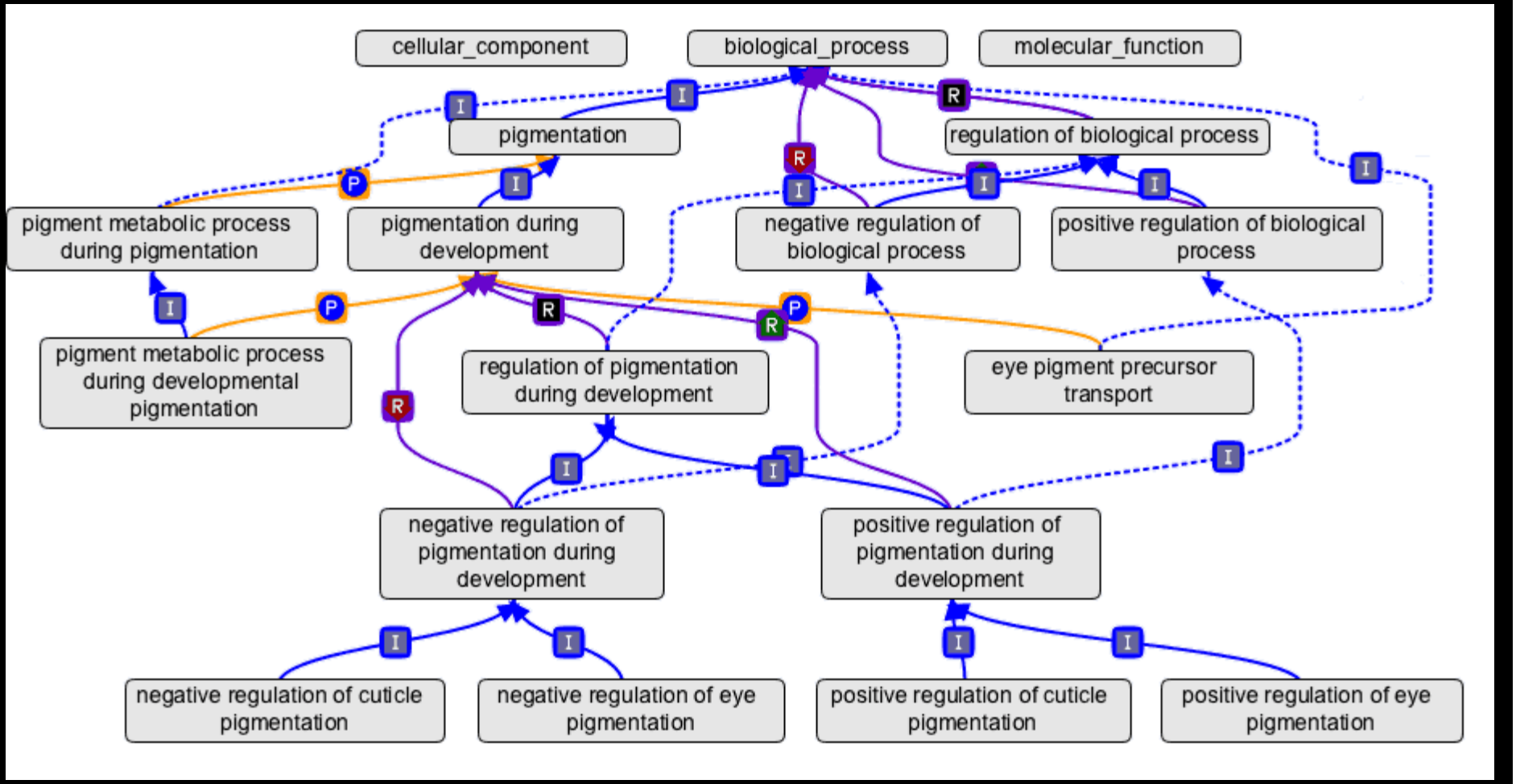
# Question Set #7

I have a set of genes from a Y2H experiment, what can I make of it?

I have a list of differentially expressed genes, what can I do with that list?

For Alex Hamilton

# Gene Ontology



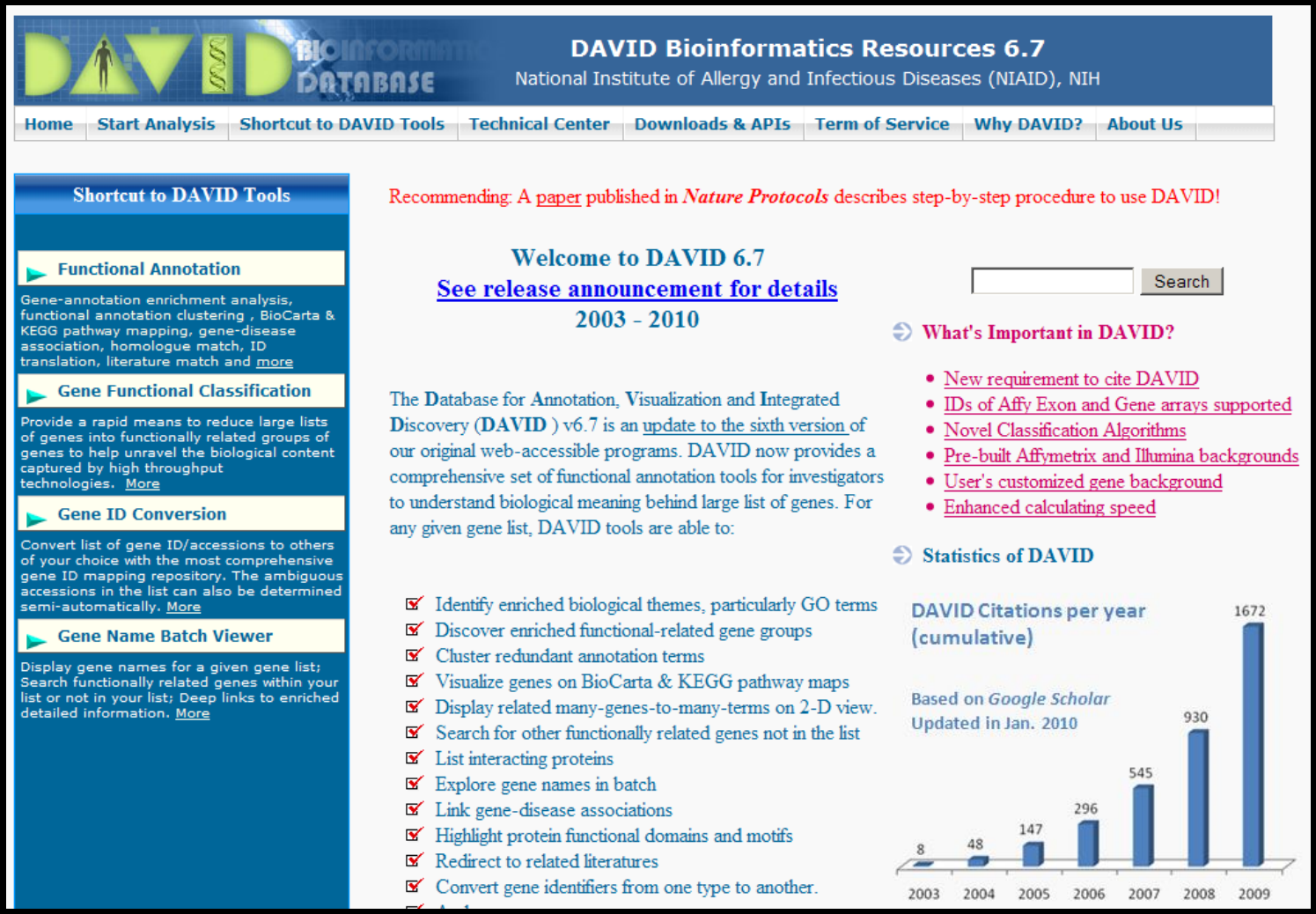

The Gene Ontology is a *controlled vocabulary*, a set of standard terms—words and phrases—used for indexing and retrieving information. In addition to defining terms, GO also defines the *relationships* between the terms, making it a *structured* vocabulary

The Gene Ontology is a a set of standard terms

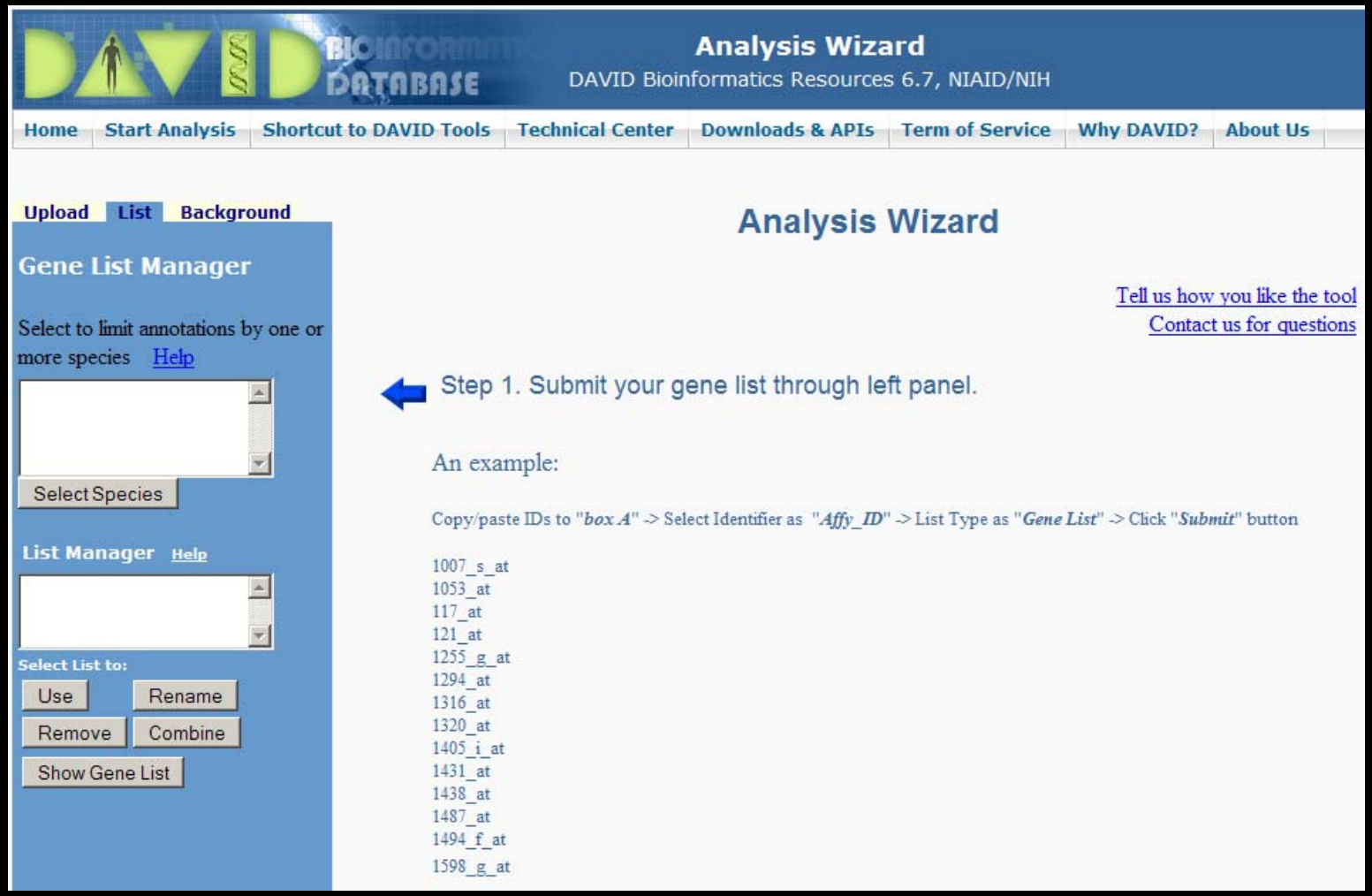
Every gene has a set of gene ontology terms associated with it

Given a set of genes, the objective is to ask if particular functions (or pathways) are enriched

# Functional Enrichment

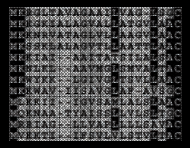



http://david.abcc.ncifcrf.gov/tools.jsp


Analysis Wizard
DAVID Bioinformatics Resources 6.7, NIAID/NIH
Home
Start Analysis
Shortcut to DAVID Tools
Technical Center
Downloads & APIs
Term of Service
Why DAVID?
About Us
Upload
List
Background
Gene List Manager
Select to limit annotations by one or more species Help
Select Species
List Manager Help
Select List to:
Use
Rename
Remove
Combine
Show Gene List
Analysis Wizard
Tell us how you like the tool
Contact us for questions
Step 1. Submit your gene list through left panel.
An example:
Copy/paste IDs to "box A" -> Select Identifier as "Affy_ID" -> List Type as "Gene List" -> Click "Submit" button
1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at

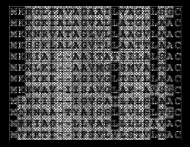
# Approach

- *Are they evolutionarily conserved?*



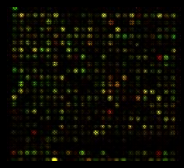
*Presence/absence in*
*Humans, worm, fly*

- *What are their molecular functions?*



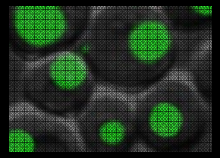

*Sequence &*
*structure*

- *What are their expression patterns?*



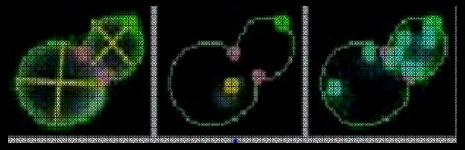
*Expression pattern &*
*co-expression analysis*

- *What are their localization patterns?*



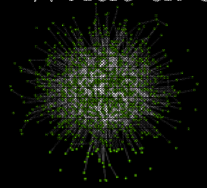
*Membrane*
*Nucleus*

- *What are the morphological changes?*



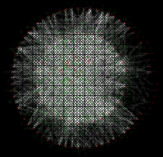
*No. Nucleus*
*Vesicle size*

- *What are their interaction partners?*



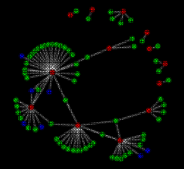
*Protein complexes*
*Evolutionary conservation of partners*

- *Which TFs regulate their expression?*



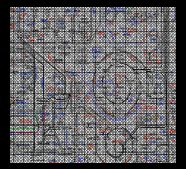
*Transcription factors involved*
*Evolutionary conservation of regulators*

- *Which signaling proteins are involved?*



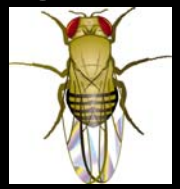
*Signaling proteins involved*
*Evolutionary conservation*

- *Which metabolic pathways are involved?*



*Metabolic pathways involved*
*Evolutionary conservation*

- *Do orthologs behave similarly?*


# General approach to investigate biological questions using computational approach

1. Formulate the big question
2. Come up with several specific questions
3. Prioritise questions and prepare a checklist
4. Identify the database
5. Identify the tools
6. Be aware of the basic statistics
7. Retrieve and integrate the information
8. Formulate hypothesis and READ A LOT!
9. Design experiments
10. Publish work & be happy ever after ☺

THANK YOU!

questions welcome