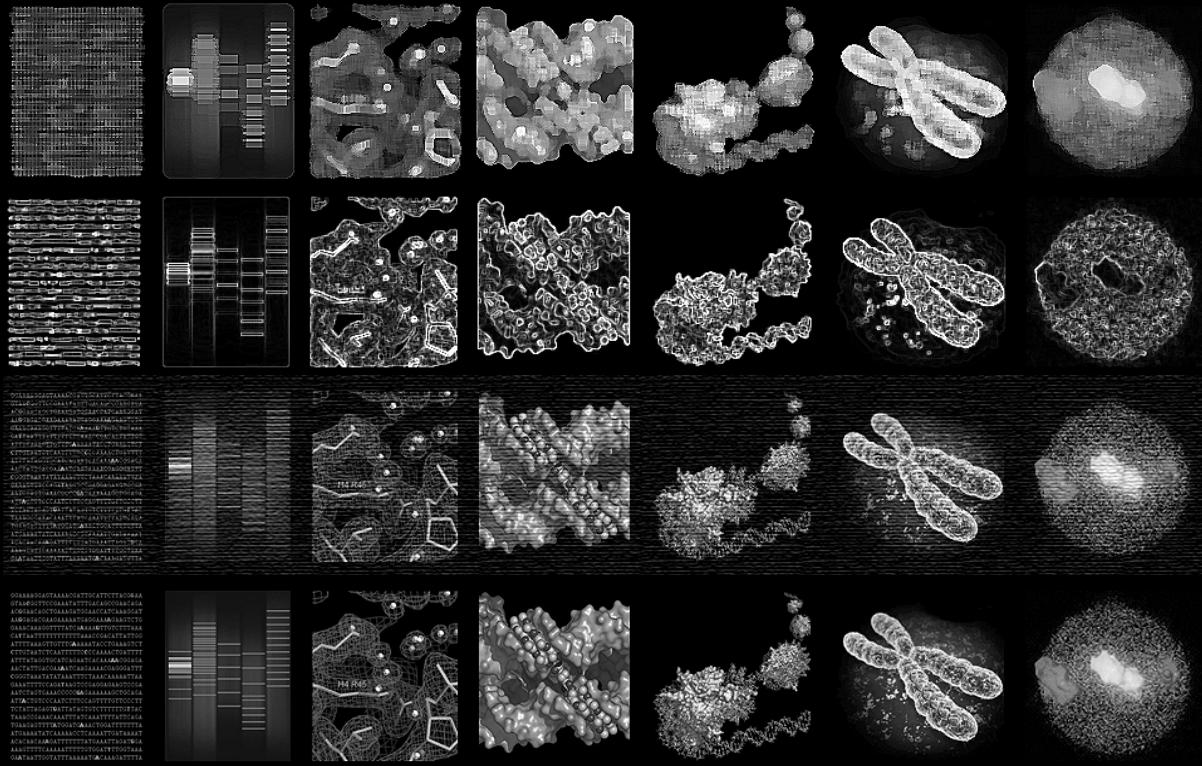


Investigating questions in biology using computational approaches



M. Madan Babu

Group Leader

MRC Laboratory of Molecular Biology, Cambridge

At what level is this talk pitched at?

“Computationally” inclined

Development of methods
Algorithms, programs, etc

Uncovering general principles
Discovery using computational approaches

Prioritising experiments
Interpreting experimental results

“Biologically” inclined

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?

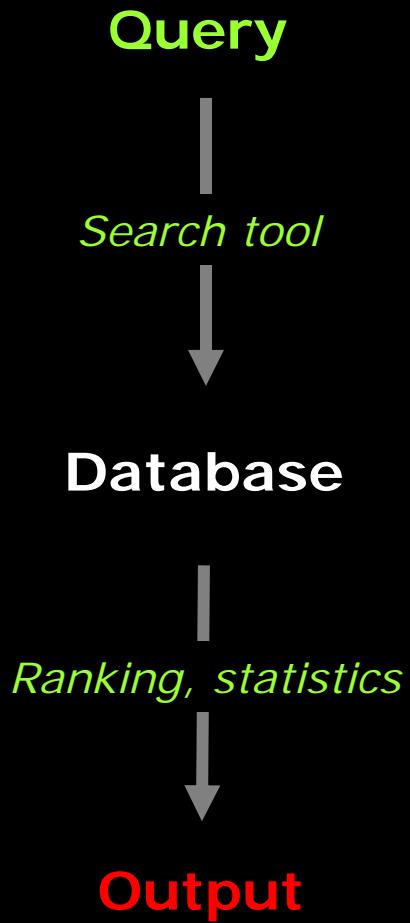
Which cellular compartment does it reside in?

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
160,000 organisms (Genbank, NCBI)

```
MERGLDTAVAGAAIIVAG-EQGC  
MKKGKIALAGVALLATGVEVAC  
MKKNRVTAGLVLLAAGVETAC  
MSSKLALAGVTLAAATTAC  
MKKIAI--AKATATSLAESAC  
MKNLKL--AAVMGEISMVAVAC  
MKWYKK--LGIVGLTSVLLAAC  
MKKWA--IXSAGVLAFAVSGC  
MLKKII--IGVSAMLALSQAC  
MQKNAA-TYAISSLLVLSLQGC  
MRRIAG--ILVAPLLLSSAVAC  
MKVKTI--IFPSVLVLSTVQDQC
```

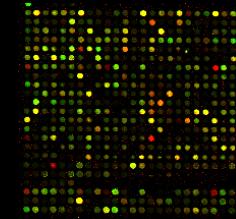
Structure

50,000 structures from
10,000 organisms (PDB, MSD)



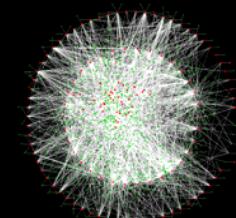
Expression

20,000 different conditions
150 organisms (SMD, GEO, ArrayExpress)



Interaction

100,000 interactions
30 organisms (Bind, DIP, publications)



Major challenge – How to exploit this information?

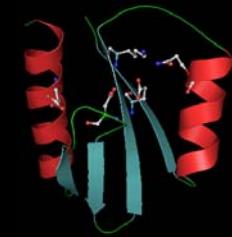
Sequence

Query sequence

```
MKRGLTAVAGAAIIVAG-LSGC  
MKKGKIALAGVALATGVLAAC  
MKNRVAAGLVLAAGVLAAC  
MSSKLALAGVTLLAATTLLAAC  
MKKIAI--AAITATSIALLSAC  
MNLLEL--AAVMGESSMVLLTAC  
MWYKK--LGIVGTSVLLAAC  
MKWAV--IISAVGQAF-AVSGC  
MLKXII--IGVSAMIALSPLAAC  
PQNAAA-TYAISSELVLSPLTGCG  
PRRIAG--LLVAPLLLSAVAC  
MVKTIE--IFPSVLVLSTVLLTAC
```

Structure

Query structure



BLAST, PSI-BLAST, etc

Sequence database (NCBI,
ENSEMBL)

E-value, score, etc

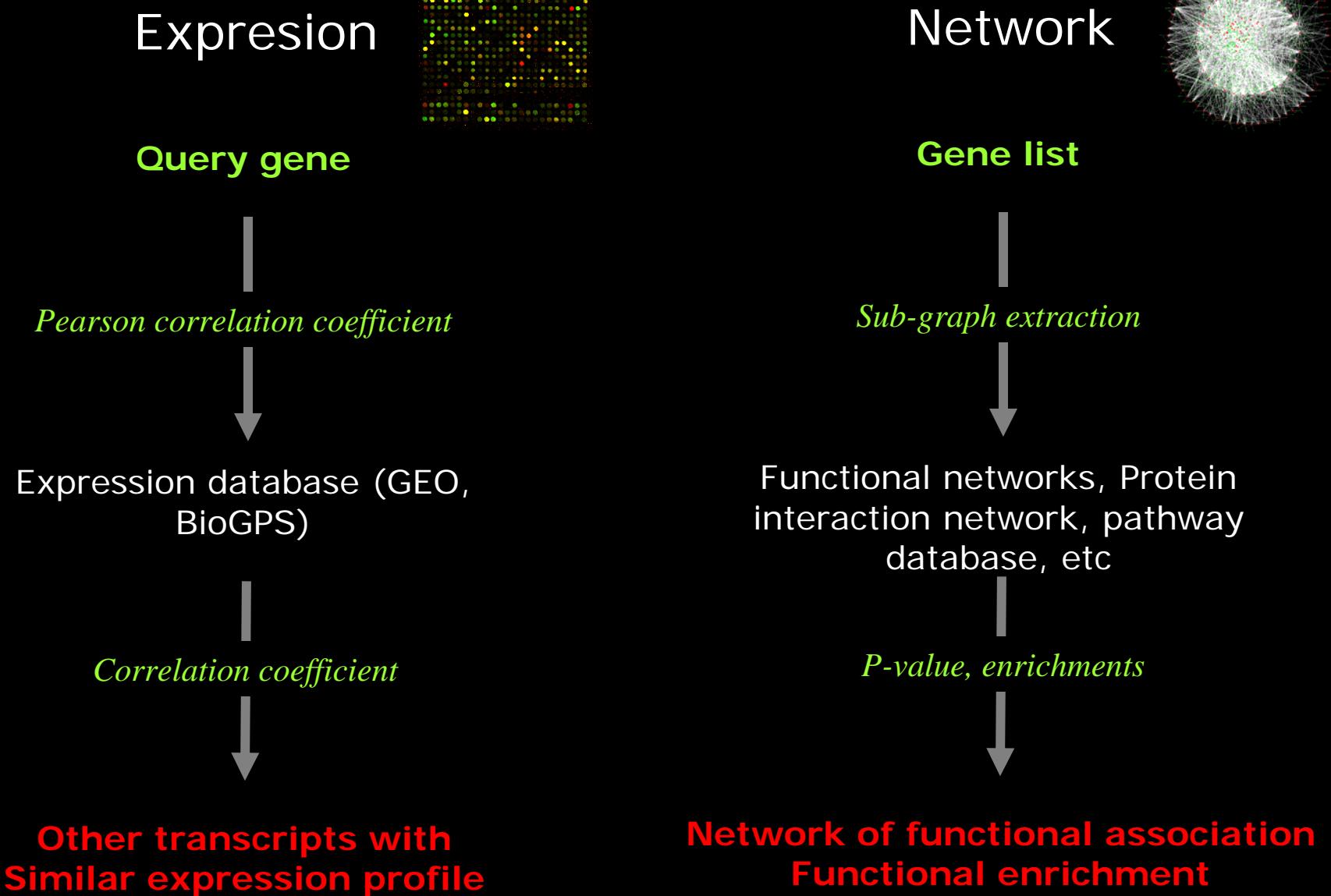
**Sequence
Alignment**

DALI, SSM, VAST, etc

Structure database (PDB)

p-value, score, etc

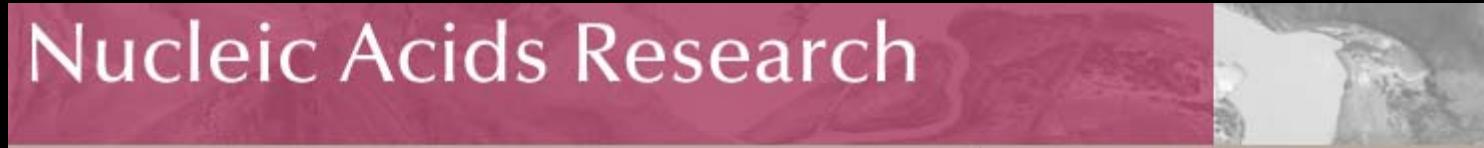
**Structure
Alignments**



Question #1

What is the list of databases
that is currently available?

1230 selected databases covering various aspects of molecular and cell biology



Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS | CURRENT ISSUE ARCHIVE SEARCH

[Oxford Journals](#) > [Life Sciences](#) > [Nucleic Acids Research](#) > Database Summary Paper Categories

2010 NAR Database Summary Paper Category List

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases

- ▶ Compilation Paper
- ▶ Category List
- ▶ Alphabetical List
- ▶ Category/Paper List
- ▶ Search Summary Papers

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

 Swiss Institute of Bioinformatics



Search ExPASy web site for Go Clear

ExPASy Proteomics Server

Databases Tools Services Mirrors About Contact

You are here: ExPASy CH > Links

Life Science Directory

(formerly known as Amos' WWW links page)

Notes:

- 1) The URL for this page is <http://www.expasy.org/links.html>
- 2) If you would like to submit a specific link or to notify us of a modified link, please [send us an email](#), but remember that we reserve the right to choose the links we want to include !
- 3) Links to protein sequence, 3D structure and 2D-gel analytical tools are provided on ExPASy's [Proteomics tools](#) page.

Quick jump to the following topics:

[Protein db](#) | [3D structure db](#) | [2D-PAGE & MS db](#) | [DNA/RNA db](#) | [Carbohydrates db](#) | [Species specific db](#) | [Human mutation db](#) | [Genes/proteins specific db](#) | [PTM db](#) | [Phylogenetics db](#) | [Gene expression db](#) | [Patents](#) | [References](#) | [Dict., protocols & nomenclat.](#) | [Biol. soft. & db catalogs](#) | [Gateways](#) | [Biol. journals & publishers](#) | [Biol. societies](#) | [Biocomputing servers](#) | [Biotech. companies](#) | [Bioinformatics companies](#) | [Misc. medical ref. sites](#) | [Misc. scientific ref. sites](#)

Protein related databases

- [UniProt](#) - the universal protein resource (including UniProtKB -Swiss-Prot and TrEMBL-, UniRef, UniParc)
- [Around UniProtKB](#) - links to related databases and portals
 - [HAMAP](#) - Portal to microbial UniProtKB/Swiss-Prot entries
 - [SwissVar](#) - Portal to human diseases and variant information in UniProtKB/Swiss-Prot
 - [UniPathway](#) - Metabolic pathways database
 - [ViralZone](#) - Portal to viral UniProtKB/Swiss-Prot entries
 - [HPI](#) - Human Proteomics Initiative
 - [PPAP](#) - Plant Proteome Annotation Project
 - [Tox-Prot](#) - Toxin Annotation Project
- [Swiss Human Plasma protein dataset](#) - Novartis/Geneprot MicroProt2 dataset from Human Plasma samples new
- [NCBI protein resources](#)

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

DATABASE

The Journal of Biological
Databases and Curation

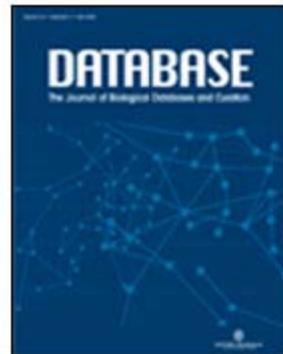
ABOUT THIS JOURNAL CONTACT THIS JOURNAL

CURRENT ISSUE ARCHIVE SEARCH

Institution: University of Cambridge [Sign In as Personal Subscriber](#)

Oxford Journals > Life Sciences > Database

READ THIS JOURNAL



Welcome to *Database: The Journal of Biological Databases and Curation*

A Fully Open Access Journal

[View Current Content](#)

[Browse the Archive](#)

[Biocuration Virtual Issue](#)

[Now Indexed in PubMed Central](#)

Huge volumes of primary data are archived in numerous open-access databases, and with new generation technologies becoming more common in laboratories, large datasets will become even more prevalent. The archiving, curation, analysis and interpretation of all of these data are a challenge. Database development and biocuration are at the forefront of the endeavor to make sense of this mounting deluge of data.

Database: The Journal of Biological Databases and Curation provides an open access platform for the presentation of novel ideas in database research and biocuration, and aims to help strengthen the bridge between database developers, curators, and users.

THE JOURNAL

- [About the journal](#)
- [Rights & permissions](#)
- [Recent Comments](#)

BIOCURATION VIRTUAL ISSUE

Editor-in-Chief

David Landsman

[View full editorial board](#)

FOR AUTHORS

- [Instructions to authors](#)
- [Services for authors](#)
- [Submit Now!](#)

<http://database.oxfordjournals.org/>

Question #2

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

Over 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.

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.

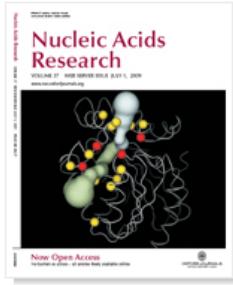
Education (73)
Links to information about the techniques, materials, people, places, and events of the greater bioinformatics community. Included are current news headlines, literature sources, educational material and links to bioinformatics courses and workshops.

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.

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




bioinformatics.ca

[Return to Bioinformatics.ca](#)

http://bioinformatics.ca/links_directory/



Swiss Institute of
Bioinformatics



Search ExPASy web site for Go Clear

ExPASy Proteomics Server

Databases Tools Services Mirrors About Contact

You are here: ExPASy CH > Tools

ExPASy Proteomics tools

The tools marked by are local to the ExPASy server. The remaining tools are developed and hosted on other servers.

[Protein identification and characterization] [Other proteomics tools] [DNA > Protein] [Similarity searches] [Pattern and profile searches] [Post-translational modification prediction]
[Topology prediction] [Primary structure analysis] [Secondary structure prediction] [Tertiary structure] [Sequence alignment] [Phylogenetic analysis] [Biological text analysis]

Protein identification and characterization

Identification and characterization with peptide mass fingerprinting data

- [Aldente](#) - Identify proteins with peptide mass fingerprinting data. A new, fast and powerful tool that takes advantage of Hough transformation for spectra recalibration and outlier exclusion. [Download the stand-alone version](#)
- [FindMod](#) - Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot entry or from a user-entered sequence, and mass differences are used to better characterize the protein of interest.
- [FindPept](#) - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications, post-translational modifications (PTM) and protease autolytic cleavage
- [Mascot](#) - Peptide mass fingerprint from Matrix Science Ltd., London
- [PepMAPPER](#) - Peptide mass fingerprinting tool from UMIST, UK
- [ProFound](#) - Search known protein sequences with peptide mass information from Rockefeller and NY Universities [or from [Genomic Solutions](#)]
- [ProteinProspector](#) - UCSF tools for peptide masses data (MS-Fit, MS-Pattern, MS-Digest, etc.)

Identification and characterization with MS/MS data

- [Popitam](#) - Identification and characterization tool for peptides with unexpected modifications (e.g. post-translational modifications or mutations) by tandem mass spectrometry
- [Phenyx](#) - Protein and peptide identification/characterization from MS/MS data from GeneBio, Switzerland
- [Mascot](#) - Sequence query and MS/MS ion search from Matrix Science Ltd., London
- [OMSSA](#) - MS/MS peptide spectra identification by searching libraries of known protein sequences

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

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

HOME

MAX-PLANCK-GESELLSCHAFT

Show results of job:

Recent jobs:
Select all Deselect all

queued
running
done
error

Bioinformatics Toolkit

Max-Planck Institute for Developmental Biology

Search Alignment Sequence Analysis 2ary Structure 3ary Structure Classification Utils

CS-BLAST FHMMER HHpred HHsenser NucBLAST PSI-BLAST PatternSearch ProtBLAST SimShiftDB

Search Tools

CS-BLAST CS-BLAST is an extension to standard NCBI BLAST that allows to increase its sensitivity by a factor of more than two on remote homologs at the same speed. CS-BLAST first adds context-specific pseudocounts to the input sequence and then jumpstarts PSI-BLAST with the resulting profile. The output is identical to BLAST and contains a list of closest homologs with alignments.

FHMMER Fast, PSI-BLAST accelerated [HHMMER](#) search. About 30 times faster for the nr database.

HHpred Sensitive protein homology detection and structure prediction by HMM-HMM-comparison. HHpred builds a profile HMM from a query sequence and compares it with a database of HMMs representing annotated protein families (e.g. PFAM, SMART, CDD, COGs, KOGs) or domains with known structure (PDB, SCOP). The output is a list of closest homologs with alignments.
[Learn more about HHpred...](#)

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

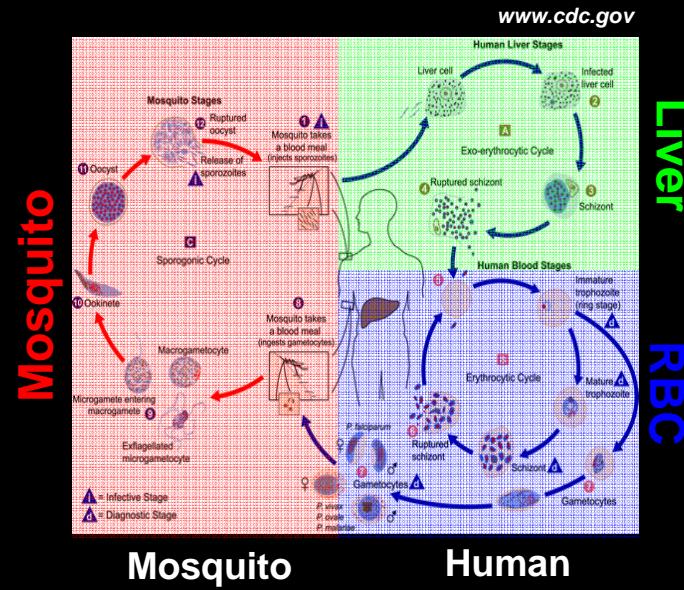


5300 genes with over 700 metabolic enzymes

Extensive complement of chromosomal regulatory proteins

Extensive complement signaling proteins (GTPases, kinases)

Complex life cycle

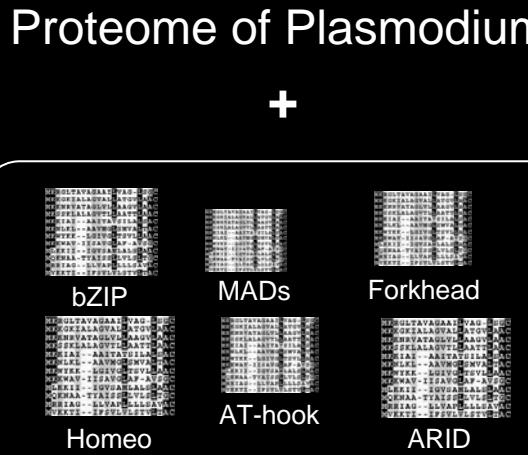
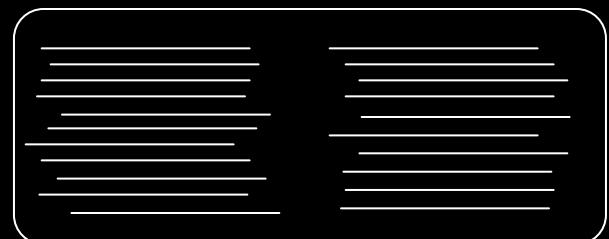


The Problem!
How does this pathogen regulate gene expression?

Possible explanations for the paradoxical observation

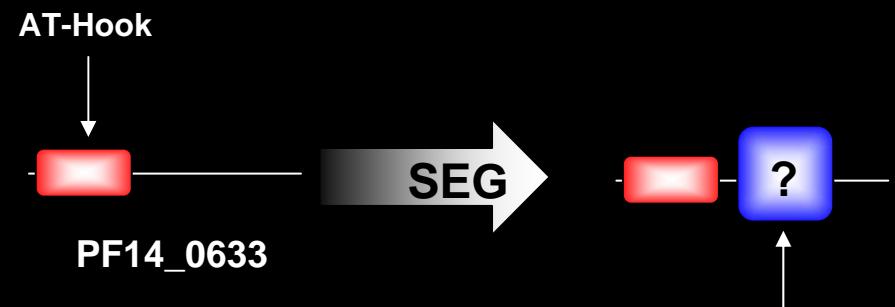
Alternative regulatory mechanisms

- Chromatin-level regulation
- Post-translational modification
- RNA based regulation



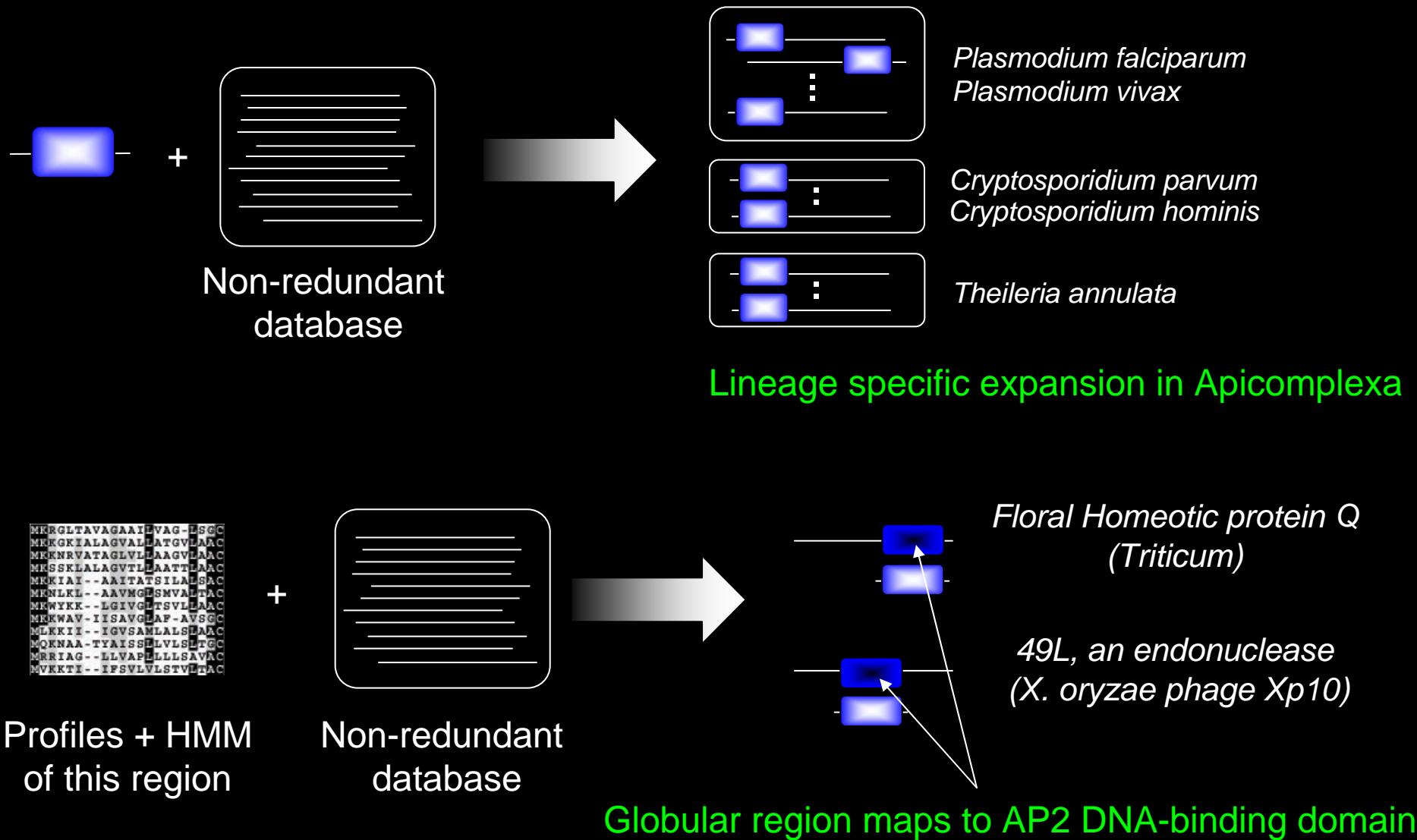
Undetected transcription factors

- Distantly related or unrelated to known DNA binding domains

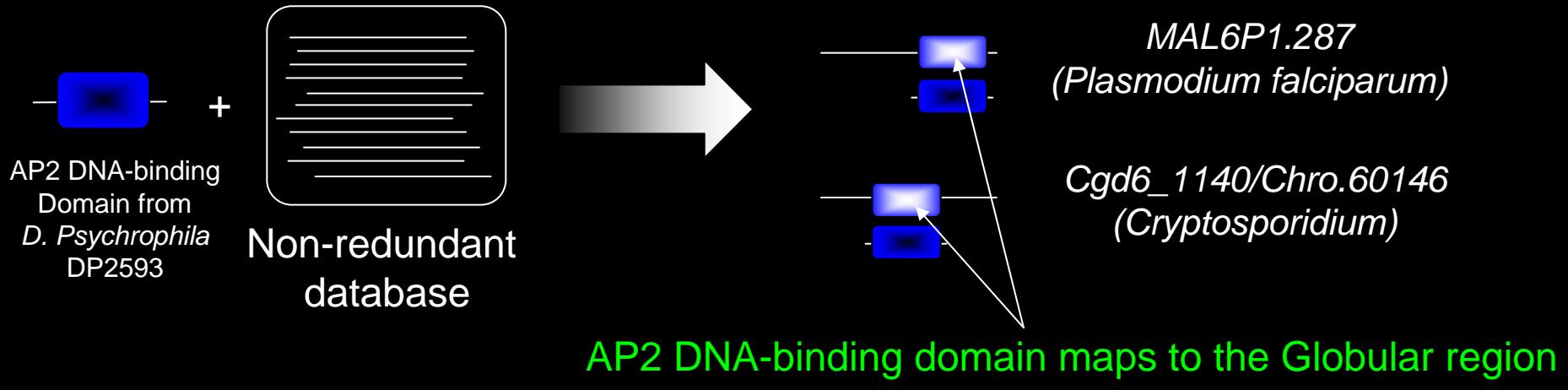


The suspect!

Characterization of the globular domain – sequence analysis I



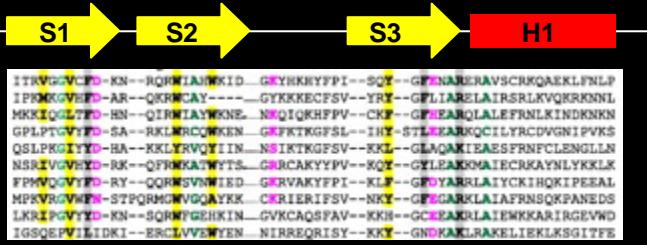
Characterization of the globular domain – sequence analysis II



MKRLGLTAVAGAAILVAG-ESGC
MKKGKIALLAGVALLATGVLAAC
MKKNRVTAGLVLLAAGVLAAC
MKSSKLALAGVTLLAATTGAAAC
MKKIAI- -AAITATSILALSAAC
MKNLKL- -AAVMGLSMVALTAC
MKWYKK- -LGIVGLTSVLLAAC
MKKWAV- -ISAVGLAF-AVSGC
MLKKII- -IGVSAMLALSAAC
MOKNAA- -TYAIISSLVLSLTGC
MRRIAG- -LLVAPLLLSSAVAC
MVKKTI- -IFSVLVLSTVLTAC

Multiple sequence alignment of all globular domains

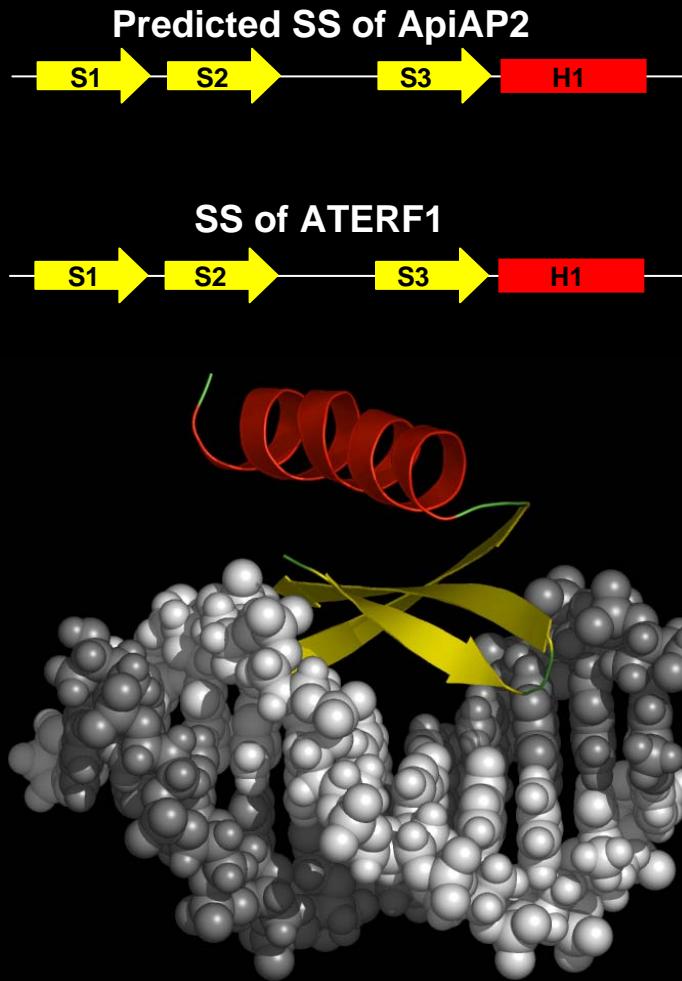
JPRED/PHD



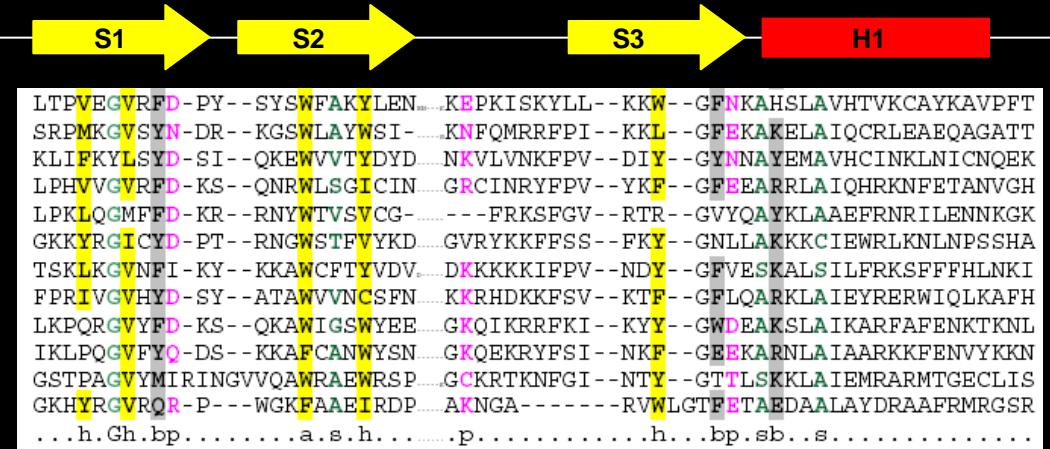
Sequence of secondary structure is similar to the AP2 DNA-binding domain

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

Characterization of the globular domain – structural analysis I



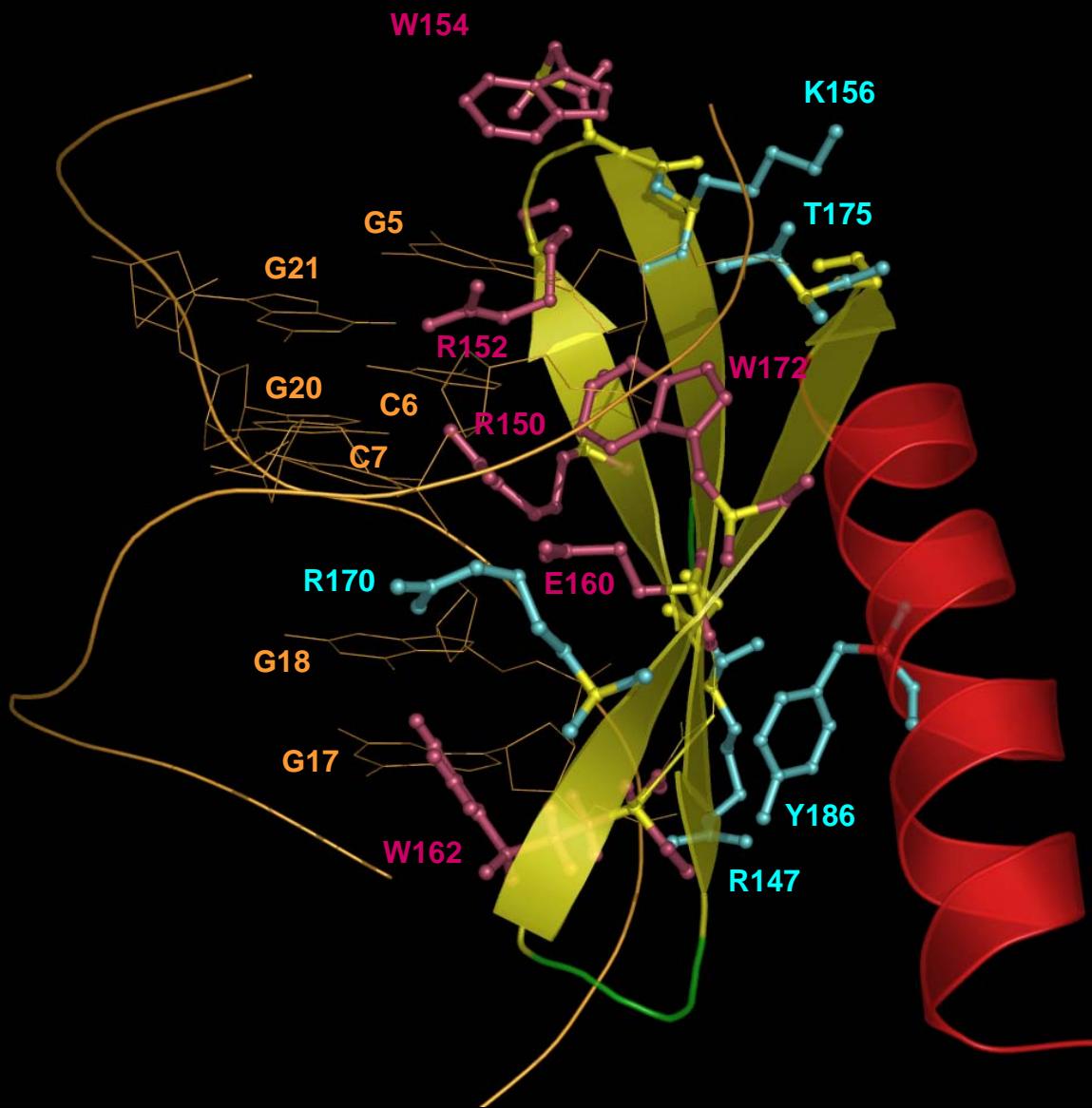
A. thaliana ethylene response factor
(ATERF1 - 1gcc – NMR structure)
Binds GC rich sequences



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

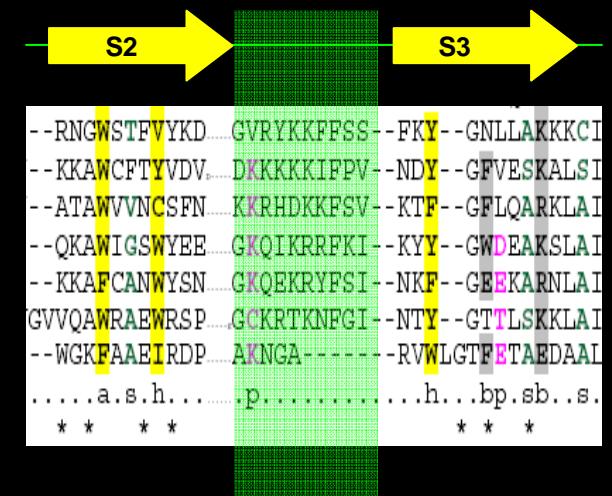
R152 --- G5 (oxo group)

D/N --- A (amino group)

R150 --- G20 (oxo group)

S/T --- A (amino group)

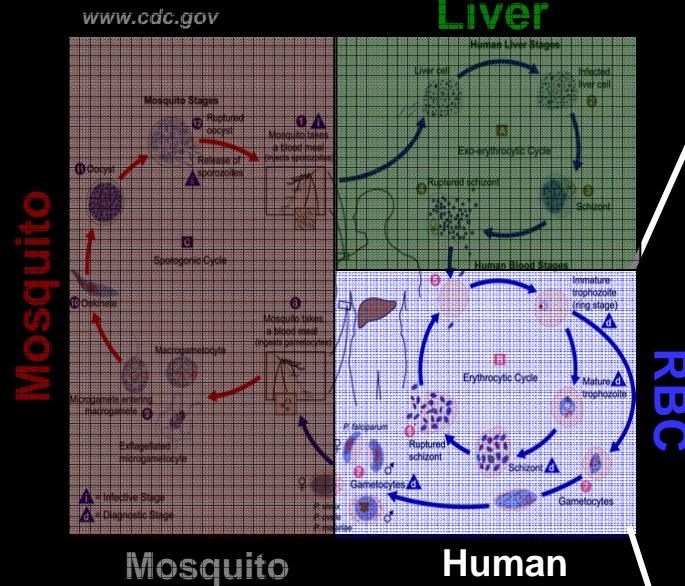
Charged residues in the insert may contact multiple phosphate groups to provide affinity



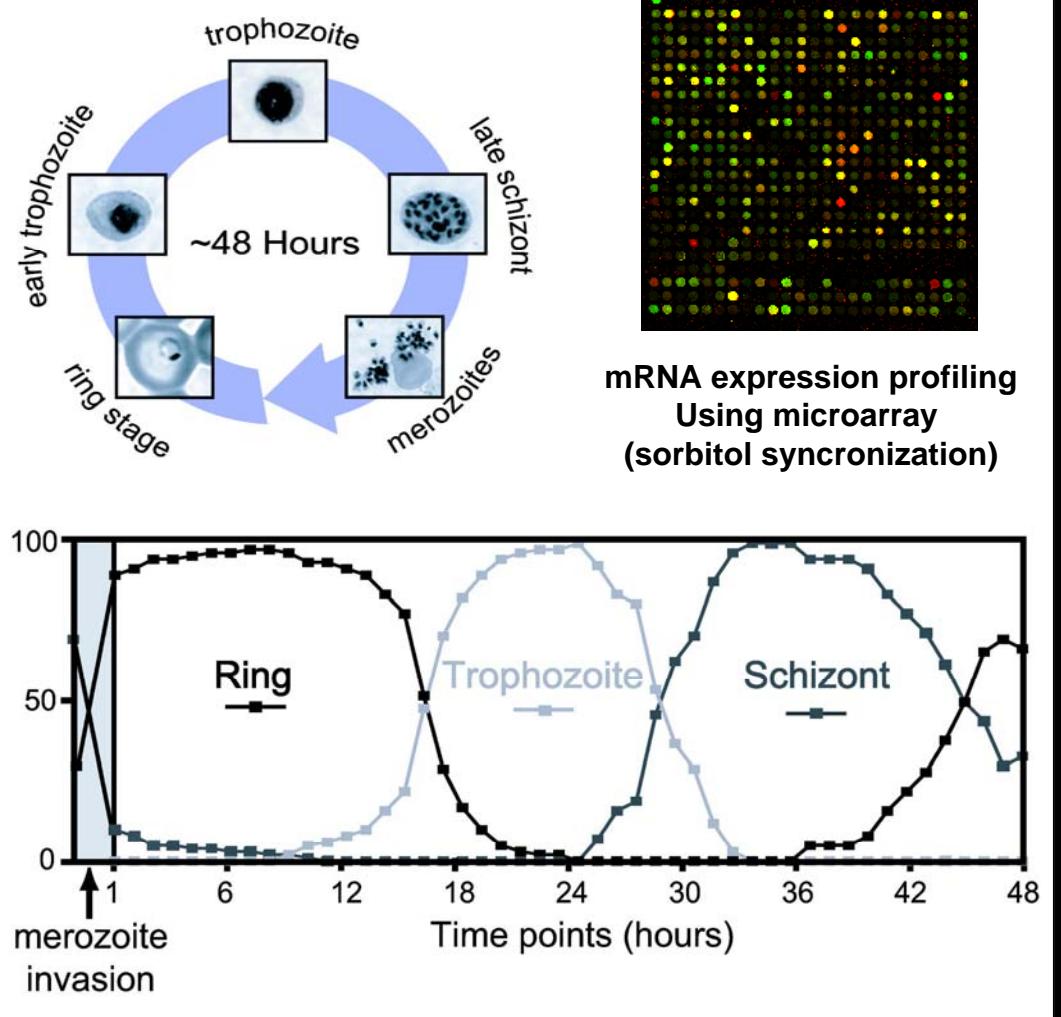
ApiAp2 domain binds DNA in a sequence specific manner

Characterization of the globular domain – expression analysis I

Complex life cycle



RBC infection & merozoite burst

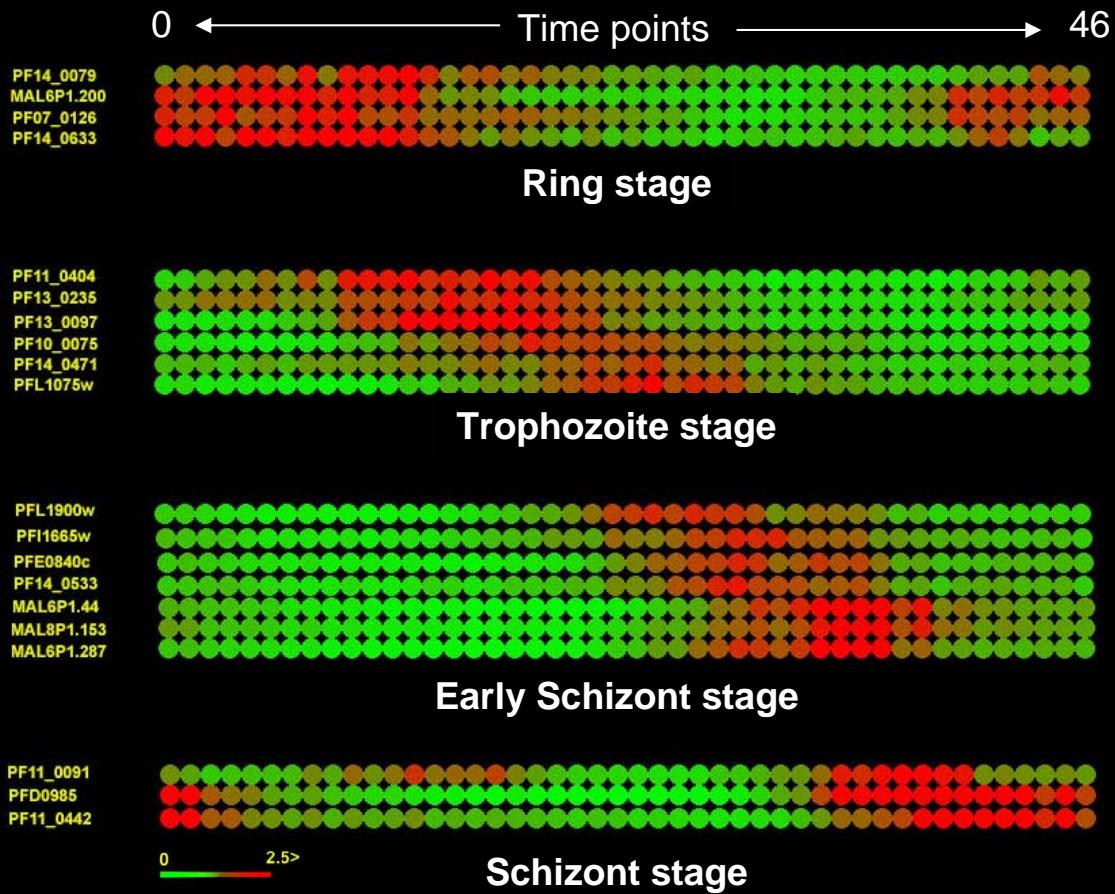


Intra-erythrocyte developmental cycle

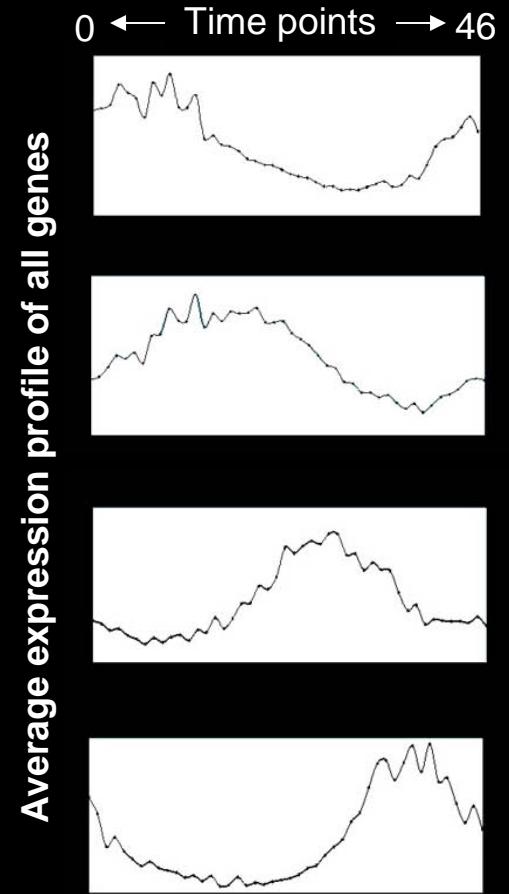
DeRisi Lab

Characterization of the globular domain – expression analysis II

22 Transcription factors



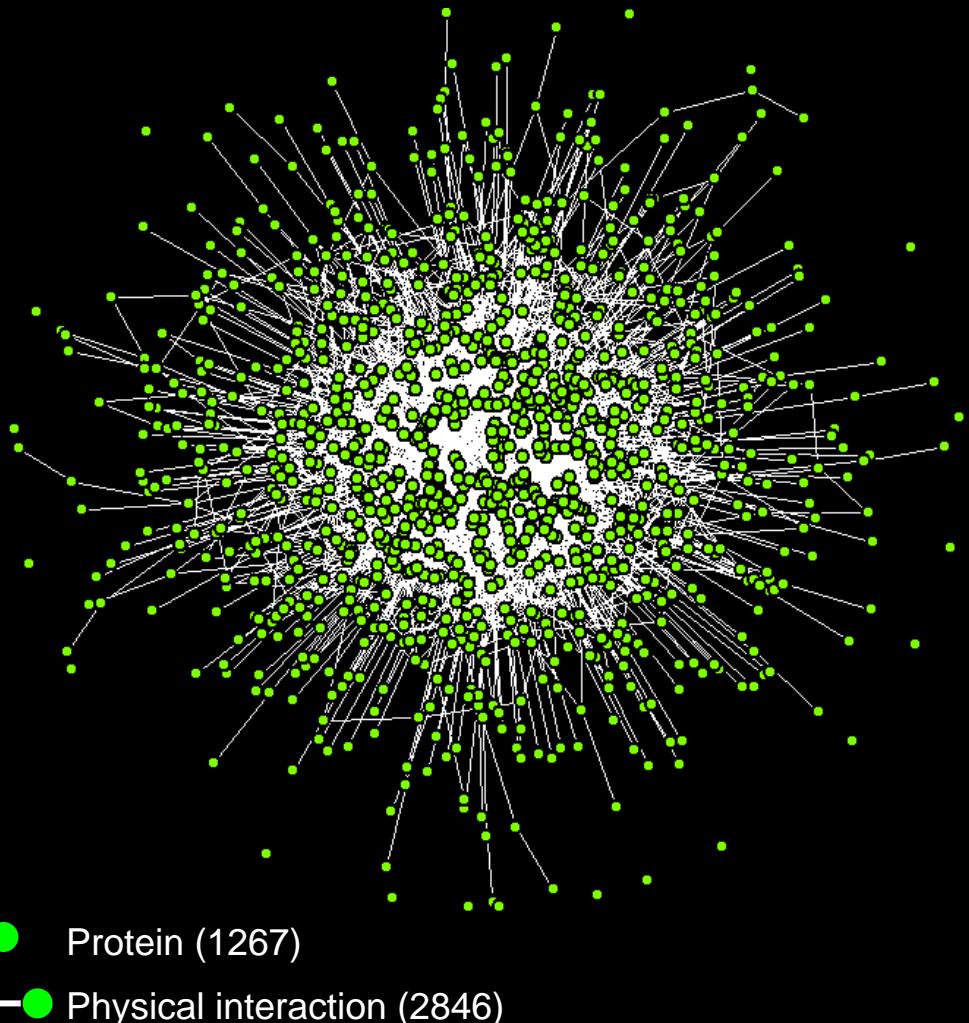
Co-expressed genes



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*

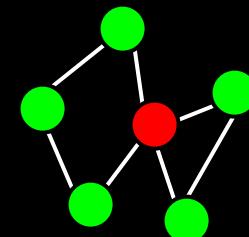


LaCount et. al. Nature (2005)

Modified Y2H:
Gal4 DBD + Protein + auxotrophic gene

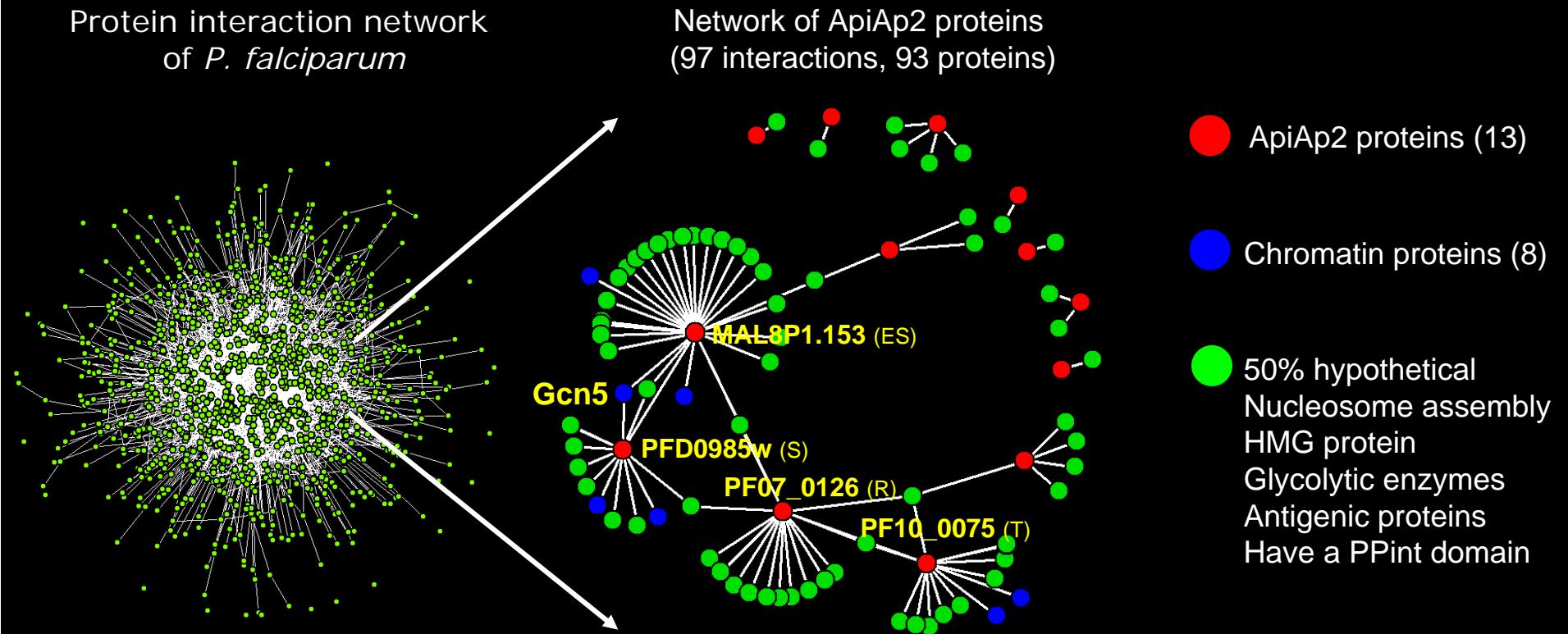
RNA isolated from mixed stages of
Intra-erythrocyte developmental cycle

Guilt by association



Function of interacting neighbors provides clues
about function of the protein

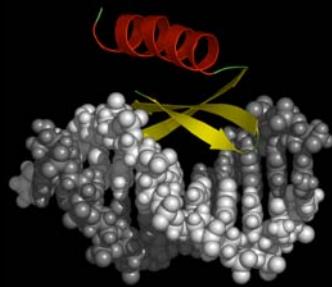
Characterization of the globular domain – interaction analysis II



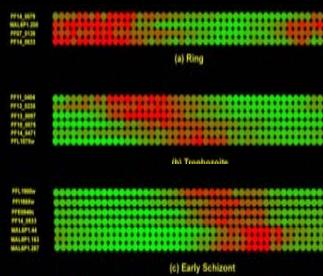
Guilt by association supports the role of ApiAp2 proteins to be involved in regulation of gene expression

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

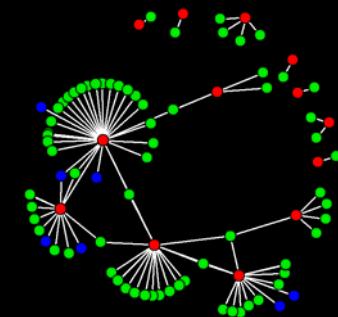
LTPVW-NRDP-PY -SISWFAIRYLEN KEPFISKYL -KIN -GFWKAHSLA
KPKVSKWNS -DR -KQSKLAWT -ENKFQPRPE -K -PGEKAHLK
KLKFLYLSD -SI -QEWNWVTTWD NWWLNVKPXY -DIY -GYWNAEMNA
LPHWVTRD -KS -QNPWNLSCICIN GRCINRYTFV -YK -GPBEARLA
LQKQKQKED -KR -XNWNTWSTCQ - -FRKSGV -TR -GQWTAQK
GKXKRJCDT -PT -ENHGTSFTVK GYRKVWFKPS -FKN -GLNLLAKK
TSKLKQUNI -KY -KKAWACPTYDV DIXKKKKIPPI -NDV -GPWEAKSL
PPRVWHD -SY -ATAWNCVWSN KHRDKKKSFS -KTV -GPQLKARL
QKLRQWVFR -KS -QKAWGMSWEE GQKQIKRPF -KYY -GWEAKSL
IKLPGQWVY -DS -KEAFCAWVSN GQKQEKRSFI -NTR -GKBRKARL
GSTPAWTMIRUNGQVQANRABW GCKRTKNG -NT -GTLSKWL
GKBRKWRQD -P -WCKGAAEIIDP ANGKA - -RWLWTGTEAHIA
h.Gh.bp.....a.s.h.....p.....h..bp..sb.....
* * * * *



Sequence



Structure



Expression

Interaction

Integration of data **can** generate experimentally testable hypotheses

The verdict!

Specific DNA-binding by Apicomplexan AP2 transcription factors

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

*Department of Molecular Biology and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544; †Division of Genetics, Department of Medicine, and §Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115; and ¶Harvard/Massachusetts Institute of Technology Division of Health Sciences and Technology, Cambridge, MA 02139

Edited by Thomas E. Wellem, 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!

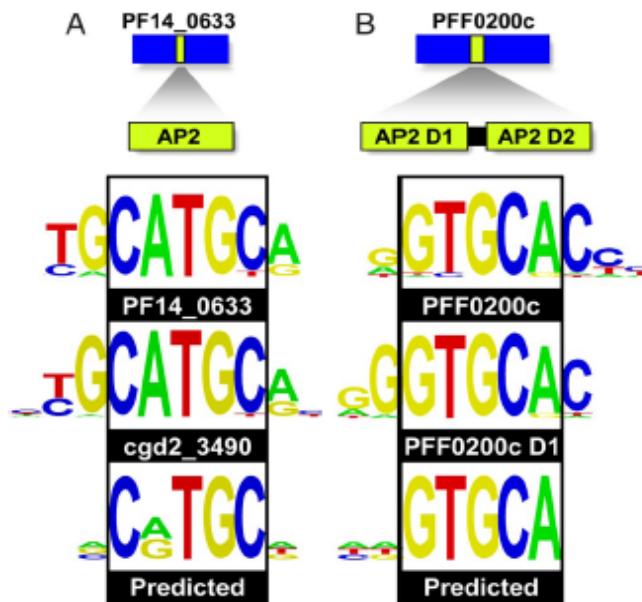


Fig. 2. DNA motifs specifically bound by AP2 domains predicted using PBMs and computational analysis (FIRE algorithm). (A) The core nucleotides (boxed) in the motif specifically bound by the *P. falciparum* AP2 domain of PF14_0633 are highly similar to those bound by its *C. parvum* orthologue cgd2_3490 (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?

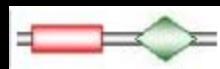
>gi | 75018194

```
MESTEDEFYTICLNLTAEDEPSFGNCNYTTDFENGELLEKVVSRRVPIFFGFIGIVGLVGNALVVLVVAAN
PGMRSTTNLLIINLAVADLLFVIFCVPFTATDYVMPRWPFGDWCKVQYFIVVTAHASVYTLVLMSDLR
FMAVVHPIASMSIRTEKNALLAIACIWVVILTTAIPVGICHGEREYSYFNRNHSSCVFLEERGYSKLGFQ
MSFFLSSYVIPLALISVLYMCMTRLWKSAPGRVSAESRRGRKKVTRMVVVVVFAVCWCPIQIILLV
KALNKYHITYFTVTAQIVSHVLAYMNCSVNPVLYAFLSENFRVAFRKVMYCPPYNDGFSGRPQATKTTR
TGNGNSCHDIV
```

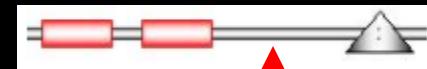
Proteins are made of domains, which are independent evolutionary units



Single domain



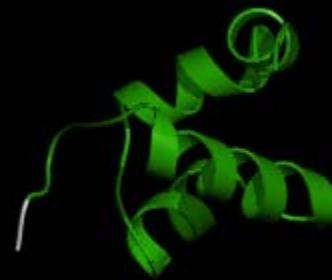
Two domain



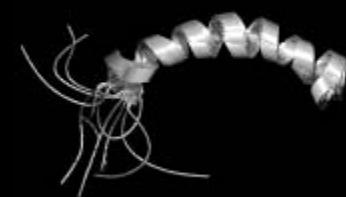
Unassigned region
New domain or unstructured region

B9G4I0_ORYSJ/14-218	ARKRFDKASI	L	YDQAREYLIS	LKK	GTRI	Dva.
Q0J0G6_ORYSJ/14-218	ARKRFDKASI	L	YDQAREYLIS	LKK	GTRI	Dva.
C0PDH5_MAIZE/14-218	AHKRFDKASI	S	YDQIREKILS	LKKGT	-----	Rpdi
Q0J435_ORYSJ/14-218	ARKRFDKASI	L	YDQVRDKVLS	LK	KGTRADit.	
B9G206_ORYSJ/14-189	ARKRFDKASI	L	YDQVRDKVLS	LK	KGTRADit.	
B8B950_ORYSI/14-190	ARKRFDKASI	L	YDQVRDKVLS	LK	KGTRADit.	
A7QT71_VITVI/14-218	ARKRFDKASI	L	YDQAREKYLIS	LRK	-----	GTKSDia.
B9RFZ3_RICCO/14-218	ARKRFDKASI	L	YDQAREKFLS	LR	KGTKIDva.	
AGD3_ARATH/4-218	ARKRFDKASI	L	YDQAREKFLS	LRK	-----	GTKSDva.
B9GMT1_POPTR/11-218	ARKRFDKASI	L	YDQAREKFLS	IRK	-----	GTRSDia.
B9GZ26_POPTR/12-218	ARKRFDKASI	L	YDQAREKFLS	LRK	GTRS	Dva.

Sequence and structure based
domain definition



HTH domain



BAR domain

Structure based
domain definition

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

HHMI
janelia farm
research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam
keyword search Go

Family: BAR (PF03114)

28 architectures 700 sequences 1 interaction 102 species 16 structures

Summary

Domain organisation

Alignments

HMM logo

Trees

Curation & models

Species

Interactions

Structures

Jump to... ↴

enter ID/acc

Summary

BAR domain

BAR domains are dimerisation, lipid binding and curvature sensing modules found in many different protein families. A BAR domain with an additional N-terminal amphipathic helix (an N-BAR) can drive membrane curvature. These N-BAR domains are found in amphiphysin, endophilin, BRAP and Nadrin. BAR domains are also frequently found alongside domains that determine lipid specificity, like [PF00169](#) and [PF00787](#) domains in beta centaurins and sorting nexins respectively.

Literature references

1. Gallop JL, Butler PJ, McMahon HT; , *Nature*. 2005;438:675-678.: Endophilin and CtBP/BARS are not acyl transferases in endocytosis or Golgi fission. [PUBMED:16319893](#)
2. Peter BJ, Kent HM, Mills IG, Vallis Y, Butler PJ, Evans PR, McMahon HT; , *Science*. 2004;303:495-499.: BAR domains as sensors of membrane curvature: the amphiphysin BAR structure. [PUBMED:14645856](#)
3. Gallop JL, Jao CC, Kent HM, Butler PJ, Evans PR, Langen R, McMahon HT; ,

Example structure
PDB entry 2rnd: Structure of the N-terminal BAR peptide in DPC micelles
View a different structure: 2rnd ▾

Database of protein domains

How can I obtain all sequences containing a particular domain?

Family: BAR (PF03114)

28 architectures 700 sequences 1 interaction 102 species 16 structures

Summary **Domain organisation** **Alignments** **HMM logo** **Trees** **Curation & models** **Species** **Interactions** **Structures** **Jump to...**

Species distribution

The tree shows the occurrence of this domain across different species. [More...](#)

Eukaryota [131] [694] [700]

- Euglenozoa
 - Kinetoplastida
 - Trypanosomatidae
 - Leishmania infantum [1] [1] [1]

Metazoa [48] [353] [359]

 - Cnidaria
 - Anthozoa
 - Hexacorallia
 - Actiniaria
 - Edwardsiidae
 - Nematostella vectensis (Starlet sea anemone) [1] [9]

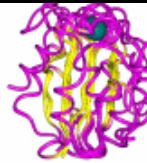
Nematoda
 - Chromadorea [3] [23] [23]
 - Spirurida
 - Filarioidea

Tree controls Hide
Fully expand tree
Fully collapse tree
Expand to depth 18 Go
Annotation
Hide highlighting of species in seed
Hide summaries
Key: species, sequences, regions
Download tree
Save a text representation
Selected sequences
([Uncheck all](#))
View
 - graphically
 - as an alignment**Download**
 - [sequence accessions](#)
 - [sequences in FASTA format](#)

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

For Marwah Hassan

Superfamily 1.73
HMM library and genome assignments server



Search SUPERFAMILY

Home

SEARCH

[Keyword search](#)

[Sequence search](#)

BROWSE

Organisms

[Taxonomy](#)

[Statistics](#)

SCOP

[Hierarchy](#)

TOOLS

[Compare genomes](#)

[Phylogenetic trees](#)

[Web services](#)

[Downloads](#)

ABOUT

[Description](#)

[Publications](#)

HELP

SUPERFAMILY Description

SUPERFAMILY is a database of structural and functional annotation for all proteins and genomes.

The SUPERFAMILY annotation is based on a collection of **hidden Markov models**, which represent structural protein domains at the [SCOP](#) superfamily level. A superfamily groups together domains which have an evolutionary relationship. The annotation is produced by scanning protein sequences from over [1,300 completely sequenced genomes](#) against the hidden Markov models.

For each **protein** you can:

- Submit sequences for [SCOP classification](#)
- View domain organisation, sequence alignments and protein sequence details

For each **genome** you can:

- Examine superfamily assignments, phylogenetic trees, domain organisation lists and networks
- Check for over- and under-represented superfamilies within a genome

For each **superfamily** you can:

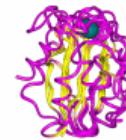
- Inspect SCOP classification, functional annotation, Gene Ontology annotation, InterPro abstract and genome assignments
- Explore taxonomic distribution of a superfamily across the tree of life

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](#)

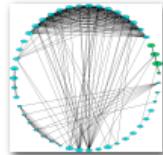
[Sequence
Alignments](#)

[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)

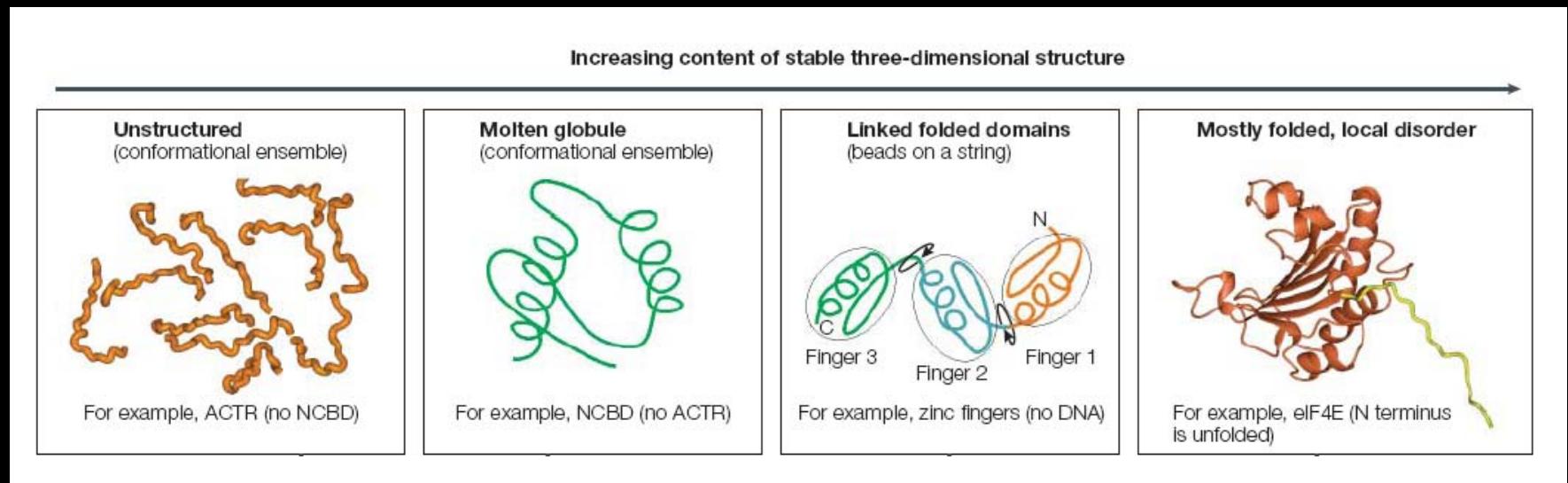
Jump to [[Top of page](#) · [Assignment details](#) · [Add assignments from groups of genomes](#)]

Genome assignment, sequence alignment, domain combinations
and taxonomic distribution

For Marwah Hassan

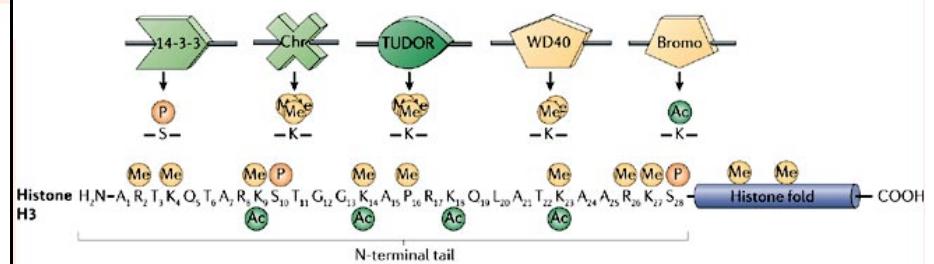
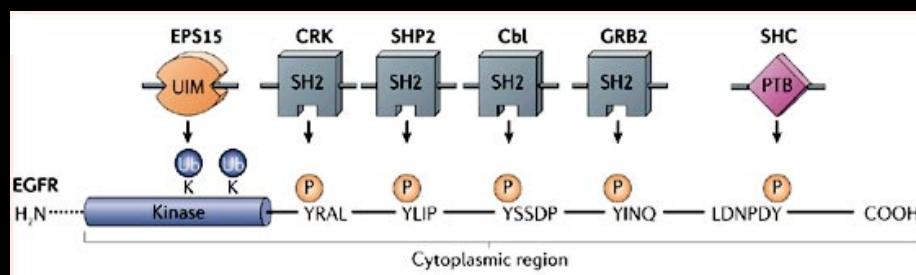
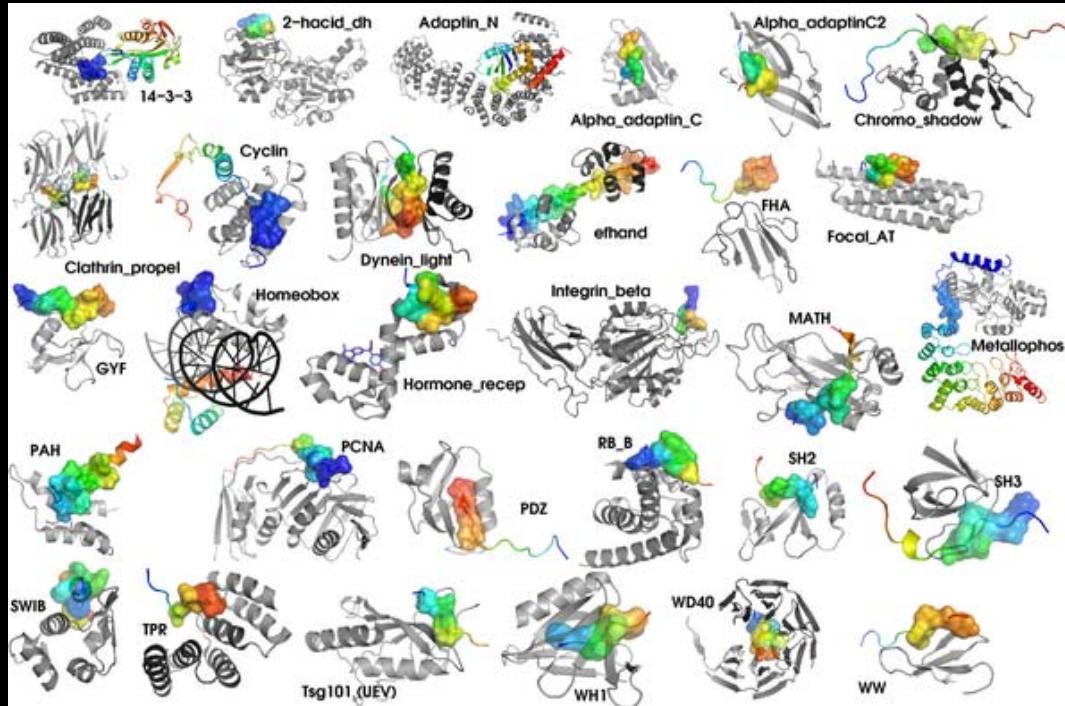
Intrinsically Unstructured Proteins

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/hydrophathy 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

**The Eukaryotic Linear Motif resource for
Functional Sites in Proteins**

server browse candidates links about usage news help

Functional site prediction

Protein sequence
Enter SWISS-PROT/TrEMBL identifier or accession number:

Or paste the sequence (Single letter code sequence only or FASTA format):

Context information

Species
select from list:
 or type in manually:

Cell compartment (one or several):

Featured paper:
A 3-way molecular switch in Ataxin-1


The ELM server
ELM is a resource for predicting functional sites in eukaryotic proteins. Putative functional sites are identified by patterns (regular expressions). To improve the predictive power, context-based rules and logical filters are applied to reduce the amount of false positives. Known ELM instances and predictions in sequences similar to ELM instance sequences, where the motif is positionally conserved, are identified and displayed (see ELM instance mapper).
Users are encouraged to supply context information in order to obtain relevant predictions.
The current version of the ELM server provides basic functionality including filtering by taxonomy, cell compartment, globular domain clash (using the SMART/Pfam databases) and structure.

Disclaimer
Short patterns applied to proteins are usually not statistically significant. Therefore we can't provide E-values as with BLAST searches. This

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 activating a recurring "motif" from a truly over-represented one. Incorporating an

[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 identify motifs that do not occur in enough unrelated proteins. For this, SLiMBuild uses then clustered according to these relationships into "Unrelated Protein Clusters" (UPCs). If desired, SLiMBuild can be used as a replacement for T

SLiMChance estimates the probability of these motifs arising by chance. Probabilities are calculated independently for each UPC, adjusted the size converted into the total probability of the seeing the observed motifs to over-represented motifs from the dataset are looked at, so these probabilities seeing that motif, or another one like it. These values are calculated separately for each motif, and will be added to the output. SLiMFinder version 4.0 introduces a new option, `sig=`, which will calculate the probability of seeing a motif in another UPC. Likewise, the more precise (but more computationally intensive) correction option will be added to the output. SLiMFinder version 4.0 introduces a new option, `sig=`, which will calculate the probability of seeing a motif in another UPC. Likewise, the more precise (but more computationally intensive) correction not work with either of these options.) The `allsig=T` option will output

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

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

[QSLiMFinder](#)

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

[SLiMCore](#)

Identification of potential eukaryotic linear motif

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

The screenshot shows the SMART homepage with the title "SMART" in large blue letters at the top left. Below it is a banner with the text "Schultz et al. (1998) Proc. Natl. Acad. Sci. USA 95, 5857-5864" and "Letunic et al. (2008) Nucleic Acids Res., doi:10.1093/nar/gkn808". A navigation bar below the banner includes links for HOME, SETUP, FAQ, ABOUT, GLOSSARY, WHAT'S NEW, and FEEDBACK.

A central section titled "Select your default SMART mode" explains the difference between Normal and Genomic modes. It states that Normal SMART uses Swiss-Prot, SP-TrEMBL, and Ensembl proteomes, while Genomic SMART uses only completely sequenced genomes like Ensembl for metazoans and Swiss-Prot for others. It also notes that Genomic mode provides more accurate domain counts.

Below this, a table compares the color schemes for Normal mode (blue background) and Genomic mode (red background). Both tables show the "SMART MODE" dropdown menu with options: NORMAL (highlighted in red) and GENOMIC.

Instructions below the table advise users to click on the images to select their default mode. It also mentions that mode selection is stored in browser cookies and can be changed later via the SETUP link in the menu.

At the bottom of the screenshot, the "SETUP" link in the footer is circled in red.

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



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:

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

SMART MODE:

NORMAL
GENOMIC

Domains within *Homo sapiens* protein TEC_HUMAN (P42680)



Mouse over domain / undefined region for more info; click on it to go to detailed annotation; right-click to save w/ Transmembrane segments as predicted by the TMHMM2 program (■), coiled coil regions determined by the Coils2 program (○) 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.

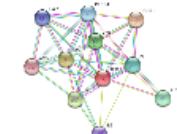
Protein information

Display orthology and other data

Domain architecture analysis

Display all proteins with similar domain ORGANISATION or COMPOSITION.
This domain architecture was probably invented with the emergence of Metazoa.

Interaction network



TEC_HUMAN shown as TEC

Pathway information

TEC_HUMAN is possibly involved in the following metabolic pathways:

map04660: T cell receptor signaling pathway

These assignments are based on similarity to the following orthologous groups:

K07364: TEC (tec protein tyrosine kinase [EC:2.7.10.2])

How to characterise an unassigned region in a protein?

The screenshot shows the NCBI BLAST Basic Local Alignment Search Tool (BLAST) interface. The top navigation bar includes links for Home, Recent Results, Saved Strategies, and Help. Below this, a sub-navigation bar for the NCBI/BLAST/blastp suite is shown, with 'blastp' selected. The main search area is titled 'Enter Query Sequence' and contains fields for entering a sequence or uploading a file, a 'Job Title' input, and a checkbox for aligning two or more sequences. A 'Query subrange' section allows specifying a range within the query sequence. The 'Choose Search Set' section on the left lists options for Database (Non-redundant protein sequences (nr)), Organism (Optional), Exclude (Optional), and Entrez Query (Optional). The 'Database' dropdown menu is open, showing several options: Non-redundant protein sequences (nr) (selected), Non-redundant protein sequences (nr), Reference proteins (refseq_protein), Swissprot protein sequences(swissprot), Patented protein sequences(pat), Protein Data Bank proteins(pdb), and Environmental samples(env_nr). The 'nr' option is highlighted in blue, while the others are yellow.

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>

PDB
PROTEIN DATA BANK

An Information Portal to Biological Macromolecular Structure
As of Tuesday Jun 22, 2010 at 5 PM PDT there are 66083 Structures | PDB Statistics

HELP | PRINT PDB ID or Text Search | Advanced Search

Home Hide
News & Publications
Usage/Reference Policies
Deposition Policies
Website FAQ
Deposition FAQ
Contact Us
About Us
Careers
External Links
New Website Features

Deposition Hide
All Deposit Services
Electron Microscopy
X-ray | NMR
Validation Server
BioSync Beamlne
Related Tools

Search Hide
Advanced Search
Latest Release
Latest Publications
Sequence Search
Chemical Components
Unreleased Entries
Browse Database
Histograms

Explorer:
Last Structure: 1YDV

Tools Hide
File Downloads

Summary Derived Data Sequence Seq. Similarity 3D Similarity Literature Biol. & Chem. Methods Geometry Links

TRIOSEPHOSPHATE ISOMERASE (TIM)
DOI:10.2210/pdb1ydv/pdb

1YDV Display Files ▾
Download Files ▾
Print this Page
Share this Page

Primary Citation
Triosephosphate isomerase from *Plasmodium falciparum*: the crystal structure provides insights into antimalarial drug design.
Velanker, S.S., Ray, S.S., Gokhale, R.S., Suma, S., Balaram, H., Balaram, P., Murthy, M.R.

Journal: (1997) Structure 5: 751-761
PubMed: 9261072 | Search Related Articles in PubMed

PubMed Abstract:
BACKGROUND: Malaria caused by the parasite *Plasmodium falciparum* is a major public health concern. The parasite lacks a functional tricarboxylic acid cycle, making glycolysis its sole energy source. Although parasite enzymes have been considered as potential antimalarial drug targets, little...
[Read More & Search PubMed Abstracts]

Molecular Description
Classification: Isomerase
Structure Weight: 55996.00
Molecule: TRIOSEPHOSPHATE ISOMERASE
Polymer: 1 Type: polypeptide(L)
Chains: A, B
EC#: 5.3.1.1 |

Length: 248

Biological Assembly More Images...

View in Jmol SimpleViewer Other Viewers ▾ Protein Workshop
Biological assembly assigned by authors and generated by PISA (software)

Protein Data Bank: repository for protein structures

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

EMBL-EBI EB-eye Search All Databases Enter Text Here Go Reset ? Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index RSS

PDB sum

Go to PDB code: 1ydv go

Top page Protein Prot-prot Clefts Links

Isomerase PDB id 1ydv

PDB id: 1ydv
Name: Isomerase
Title: Triosephosphate isomerase (tim)
Structure: Triosephosphate isomerase. Chain: a, b. Synonym: tim.
Engineered: yes
Source: Plasmodium falciparum. Malaria parasite p. Falciparum.
Organism_taxid: 5833. Expressed in: escherichia coli.
Expression_system_taxid: 562

Biological unit: Homo-dimer (from PDB file)

UniProt: Chains A, B: Q07412 (TPIS_PLAFA)

Seq: TIM 248 a.a.
Struc: 246 a.a.*

Key: Family PfamA domain
Secondary structure CATH domain

* PDB and UniProt seqs differ at 1 residue position (black cross)

Enzyme class: E.C.5.3.1.1 [IntEnz] [ExPASy] [KEGG] [BRENDA]
Reaction: D-glyceraldehyde 3-phosphate = glyceralone phosphate (see diagram below)

Resolution: 2.20Å
R-factor: 0.198
R-free: 0.255
Authors: S.S.Velankar,M.R.N.Murthy
Key ref: S.S.Velankar et al. (1997). Triosephosphate isomerase from Plasmodium falciparum: the crystal structure provides insights into antimalarial drug design. *Structure*, 5, 751-761. [PubMed id: 9261072] [DOI: 10.1016/S0969-2126(97)00230-X]

Date: 24-Apr-97
Release date: 15-Oct-97

Quick links

- RCSB
- PDBe
- SRS
- MMDB
- JenaLib
- OCA
- PDBWiki
- Proteopedia
- CATH
- SCOP
- FSSP
- HSSP
- PDBSWS
- PQS
- CSA
- ProSAT
- Whatcheck
- EDS

Jmol Snap Contents

- Description
- Header details
- Header records
- References
- PROCHECK
- Protein chains
- 246 a.a.*
- Waters ×171

* Residue conservation analysis

Tools

- Image Generation

Procheck

Clefts

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.

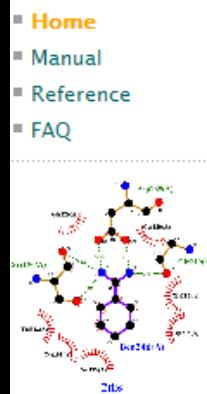
Program Options:

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

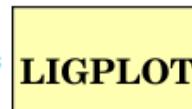


EBI > Groups > Thornton > Software > LIGPLOT

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

Written by Andrew Wallace and Roman Laskowski

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:



[LigPlot for Unix/linux](#)

[LigPlot for Windows](#)

Alternatively, commercial users may wish to contact our distributor [Ebisu](#).

Contact

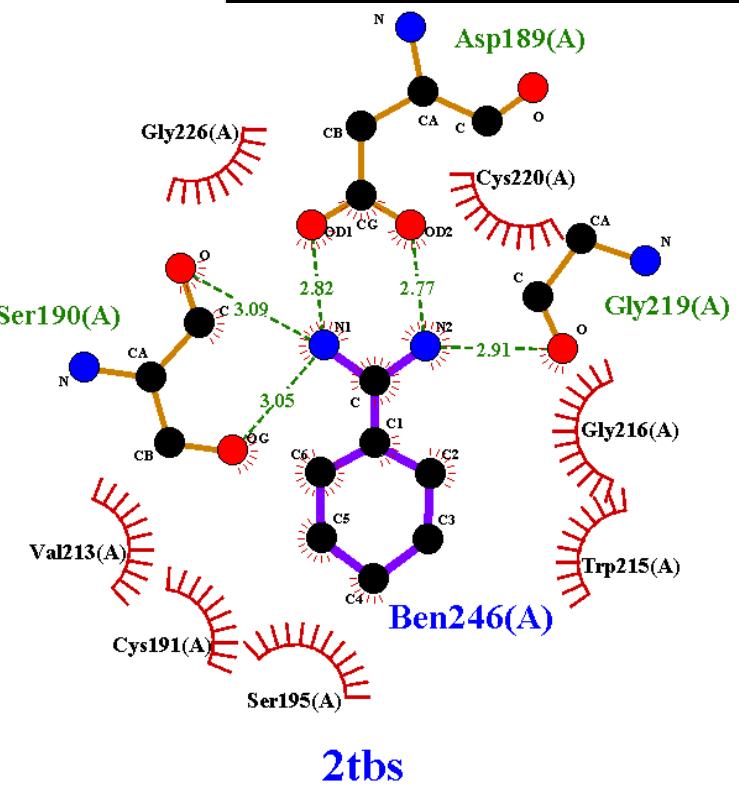
If you have any questions about LIGPLOT or problems using it, please check out the [FAQ](#) Roman Laskowski at roman@ebi.ac.uk.

Last n

Notes

► A new, GUI-based version of LIGPLOT, called [LiqPlot⁺](#), is now available for beta-test!

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 www-lab@cam.ac.uk

Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995)

database for the investigation of sequences and structures. *J. Mol. Biol.* 247: 537-546.

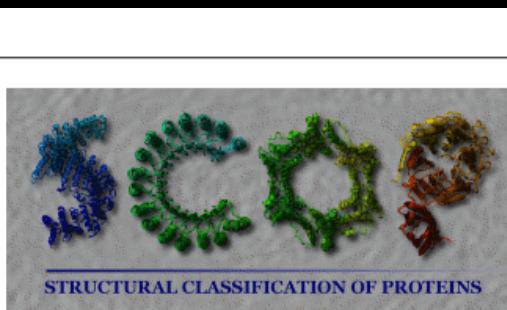
Recent changes are described in: Lo Conte L., Brenner S. E., Hubbard T., Chothia C. (2002)

SCOP database in 2002: refinements accommodate structural genomics. [Nucleic Acids Res.](#) 30: 260-263.

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

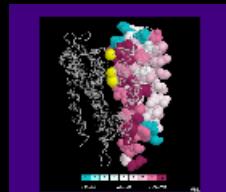
CATH v3.3 is built from 97,625 PDB chains. We have added the following data since v3.2:

- 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>

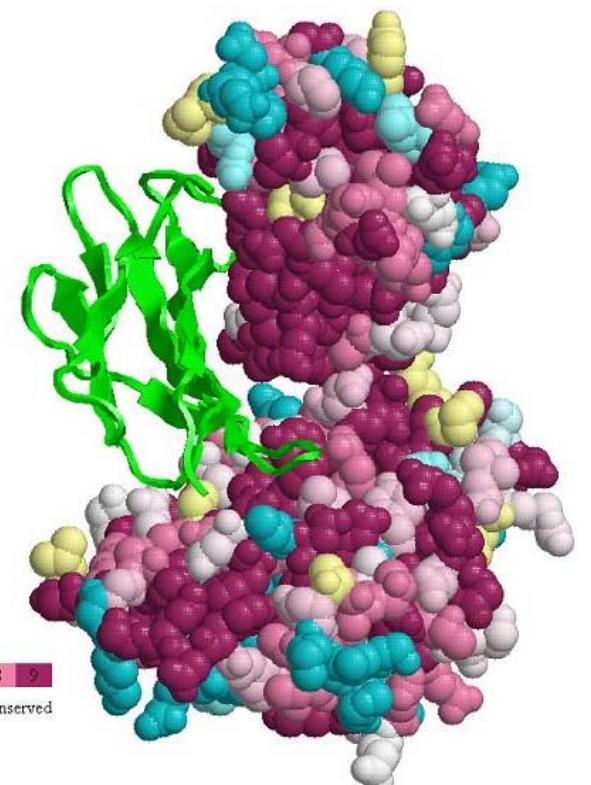


HOME
GALLERY
OVERVIEW
QUICK HELP
FAQ
BENTAL
GROUP
PUPKO
GROUP
CITING &
CREDITS

Evolutionary analysis
of individual amino
acids on the structure

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](#)



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

> Services > SSM

Submission Form for 3D alignment

pairwise multiple
[explanation of input](#)

Query	Target
Source: PDB entry PDB code 1sar view	Source: All PDB archive
Select chains Find chains Chains: * (all)	
Lowest acceptable match (%) 70	Lowest acceptable match (%) 70
<input checked="" type="checkbox"/> match individual chains <input checked="" type="checkbox"/> best matches only <input checked="" type="checkbox"/> match connectivity <input checked="" type="checkbox"/> unique matches only <input checked="" type="checkbox"/> if no matches within acceptability limits found, show some of the closest	
Precision: normal Sort by: Q-score	Viewer: Jmol
Submit your query Back to Home Page	

Programs to retrieve similar structures

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

NCBI Structure

PubMed Entrez BLAST OMIM Books TaxBrowser

Search Entrez Structure for

VAST Help Comprehensive help and frequently asked questions

VAST Search Submit structure database searches

VAST Search Help Help on submitting VAST Searches

Protein structure neighbors in Entrez are determined by direct comparison of 3-dimensional protein structures with the VAST algorithm. Each of the more than 87,804 domains in MMDB is compared to every other one. From the MMDB Structure summary pages, retrieved via Entrez, structure neighbors are available for protein chains and individual structural domains. If you already know a PDB/MMDB-ID you can try this at once, using the input form in the right column.

Dali server

SERVICES & TOOLS

GROUP MI

Help on submitting VAST Searches

Protein Structure Database Searching by DaliLite v. 3

The Dali server is a network service for comparing protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the Protein Data Bank (PDB). You receive an email notification when the search has finished. In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences.

Requests can also be submitted by e-mail to [dali-server at helsinki dot fi](mailto:dali-server@helsinki.fi). The body of the e-mail message must contain atomic coordinates in PDB format.

If you want to know the structural neighbours of a protein already in the Protein Data Bank (PDB), you can find them in the [Dali Database](#).

If you want to superimpose two particular structures, you can do it in the [pairwise DaliLite](#) server.

Upload a structure:

[Browse...](#)

Or enter PDB identifier: chain: (optional)

([Keyword search for PDB identifiers](#))

Job name:

(optional)

Enter email address for notification:

(recommended)

[submit](#) [clear](#)

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

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

The screenshot shows the homepage of the 3D Complex.org v2.0 web server. The header features a banner with protein structures and the text "3D Complex.org v2.0" and "A Web Server to browse Protein Complexes of known 3D Structure". Below the banner is a navigation menu with links: Home, Browse, Search & Custom Hierarchy, Download, Errors, About, and Links. The "Search & Custom Hierarchy" link is highlighted. A sub-menu titled "Outline" is open, showing a text block about protein complexes and a visualization of various protein complexes and their subunits.

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:

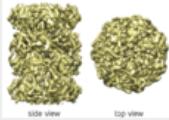
1. The representation of proteins complexes as graphs.
2. The comparison of the graphs.

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

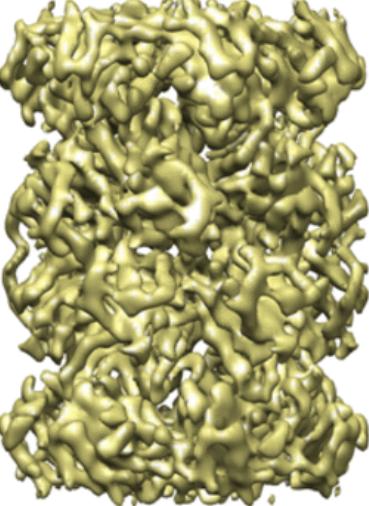
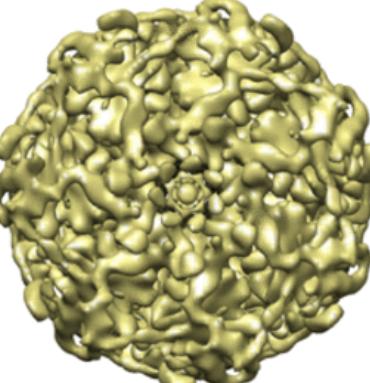
http://www.ebi.ac.uk/pdbe-srv/emsearch/atlas/1740_visualization.html

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)

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	<input type="button" value="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

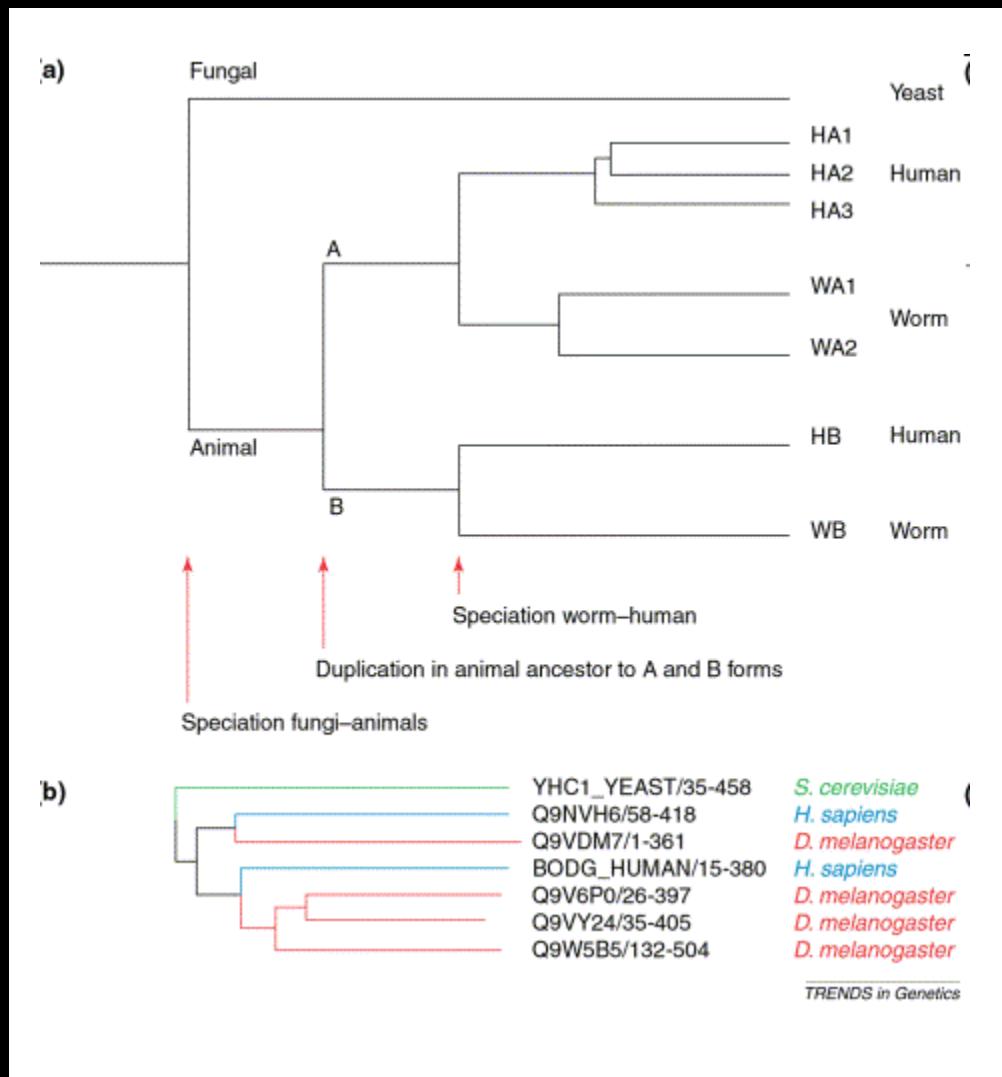
Database	Description
CSA	The Catalytic Site Atlas (CSA) is a resource of catalytic sites and residues identified in enzymes using structural data.
DSSP	The DSSP database is a database of secondary structure assignments (and much more) for all of the entries in the Protein Data Bank (PDB).
EMDB	The Electron 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 repository 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.

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

Search: for

Go

e.g. [human gene BRCA2](#) or [rat X:100000..200000](#) or [coronary heart disease](#)

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes ([Log in to customize this list](#))

**Human**

GRCh37

**Mouse**

NCBIM37

**Zebrafish**

Zv8

All genomes

[View full list of all Ensembl species](#)

Other species are available in [Ensembl Pre!](#) and [EnsemblGenomes](#)

Ensembl is a joint project between

New to Ensembl?

Did you know you can:

[Learn how to use Ensembl](#)

with our video tutorials and walk-throughs

[Add custom tracks](#)

using our new Control Panel

[Upload and analyse your data](#)

and save it to your Ensembl account

[Search for a DNA or protein sequence](#)

using BLAST or BLAT

[Fetch only the data you want](#)

from our public database, using the Perl API

[Download our databases via FTP](#)

in FASTA, MySQL and other formats

[Mine Ensembl with BioMart](#)

and export sequences or tables in text, html, or Excel format

Still got questions? Try our [FAQs](#) or [glossary](#)

Did you know...?

A preliminary assembly of the common baboon (*Papio hamadryas*) is now available on our pre! site,
<http://pre.ensembl.org/Baboon>



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

Gene-based displays

- Gene summary
- Splice variants (2)
- Supporting evidence
- Sequence
- External references (3)
- Regulation
- Comparative Genomics
 - Genomic alignments (51)
 - Gene Tree (image)
 - Gene Tree (text)
 - Gene Tree (alignment)
 - Orthologues (44)
 - Paralogues
 - Protein families (1)
- Genetic Variation
 - Variation Table
 - Variation Image
- External Data
 - Personal annotation
- ID History
 - Gene history

 Configure this page Manage your data Export data Bookmark this page**Gene: BRCA2 (ENSG00000139618)**

breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]

Location Chromosome 13: 32,889,611-32,973,347 forward strand.

Transcripts □ There are 2 transcripts in this gene

Show/hide columns

Search:

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
BRCA2-001	ENST00000380152	10930	ENSP00000369497	3418	Protein coding	CCDS9344
BRCA2-002	ENST00000470094	842	No protein product	-	Processed transcript	-

Transcript and Gene level displays

In Ensembl a gene is made up of one or more transcripts. We provide displays at two levels:

- Transcript views which provide information specific to an individual transcript such as the cDNA and CDS sequences and protein domains
- Gene views which provide displays for data associated at the gene level such as orthologues and paralogues, regulatory regions and splicing

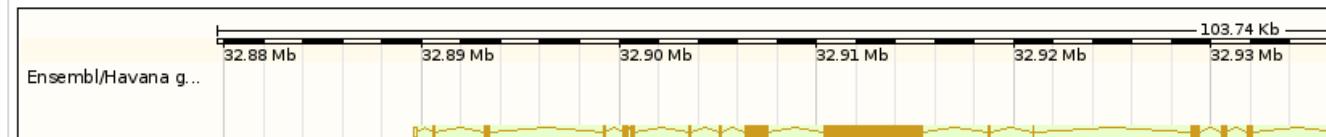
This view is a gene level view. To access the transcript level displays select a Transcript ID in the table above and then navigate to the information click on the Gene tab in the menu bar at the top of the page.

Gene summary [help](#)[Splice variants »](#)Name [BRCA2 \(HGNC Symbol\)](#)Synonyms BRCC2, FACD, FAD, FAD1, FANCD, FANCD1 [To view all Ensembl genes linked to the name [click here](#).]CCDS This gene is a member of the Human CCDS set: [CCDS9344](#)

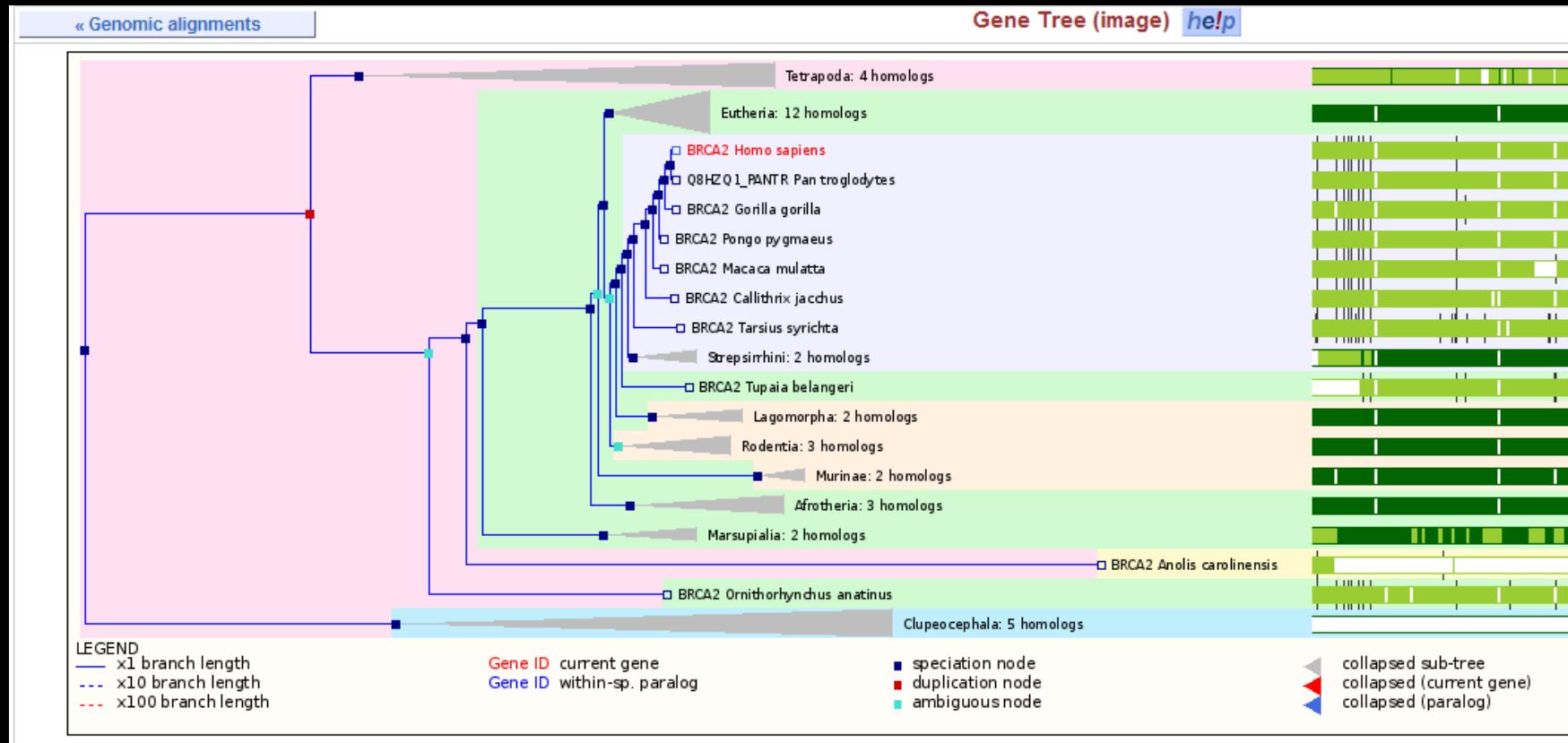
Gene type Known protein coding

Prediction Method Gene containing both Ensembl genebuild transcripts and [Havana](#) manual curation, see [article](#).

Alternative genes This gene corresponds to the following database identifiers:

Havana gene: [OTTHUMG0000017411](#) [view all locations]**Go to the entry page of a gene of interest**

Tree of the evolutionary history of a gene



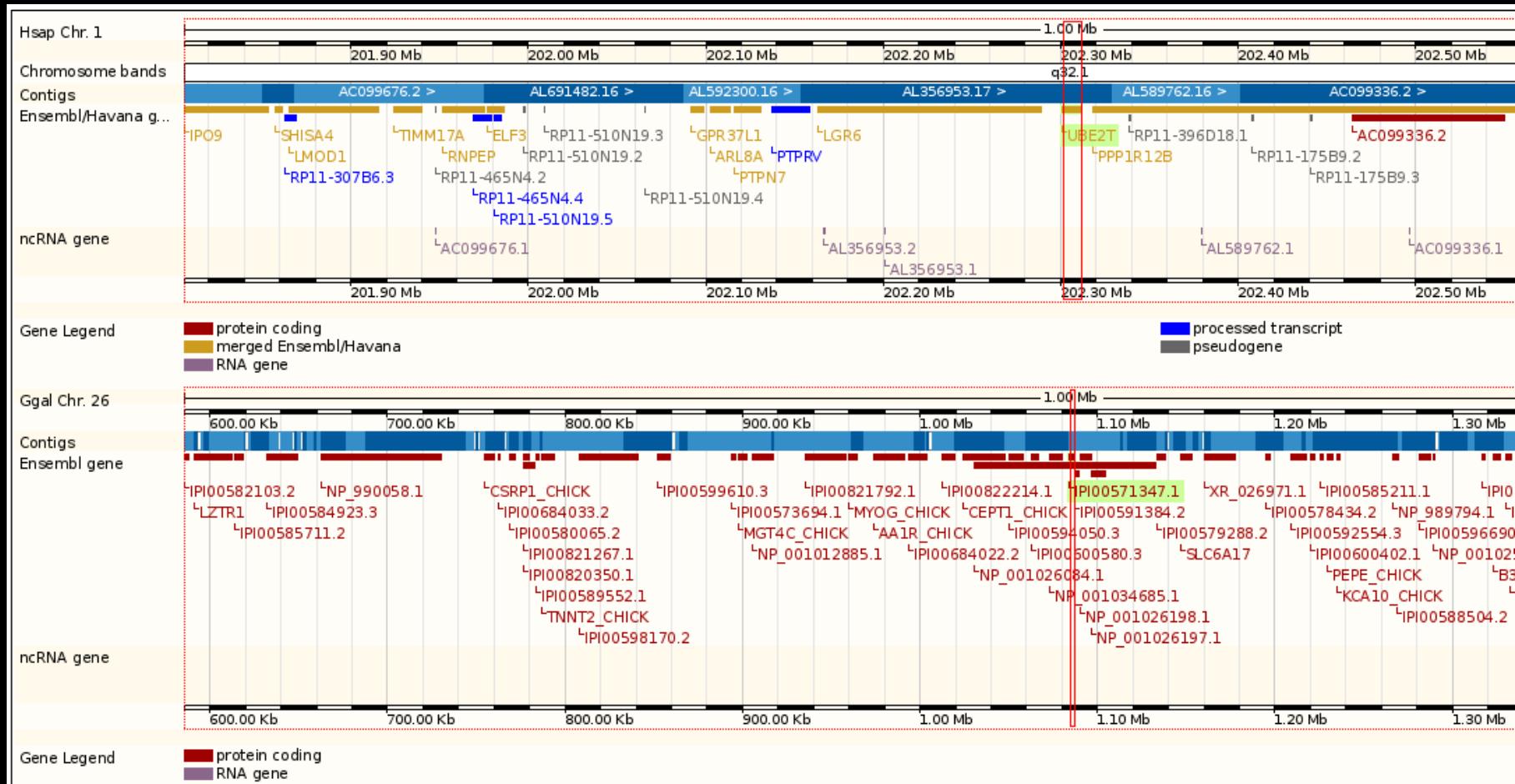
Click “Gene Tree” to get the tree for further detailed analysis

Orthology relationship table

Show All entries							Show/hide columns	Search:
Species	Type	dN/dS	Ensembl identifier	Location	Target %id	Query %id	External ref.	
Alpaca (<i>Vicugna pacos</i>)	1-to-1	n/a	ENSPAG0000000886 Multi-species view Alignment Gene Tree (image)	GeneScaffold_1422-28074-80684-1	66	65	BRCA2 breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]	
Anole Lizard (<i>Anolis carolinensis</i>)	1-to-1	n/a	ENSACAG0000004593 Multi-species view Alignment Gene Tree (image)	scaffold_312-337788-344138-1	58	5	BRCA2 breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]	
Anole Lizard (<i>Anolis carolinensis</i>)	Possible ortholog	n/a	ENSACAG0000004541 Multi-species view Alignment Gene Tree (image)	scaffold_312-315053-328668-1	59	15	Novel Ensembl prediction No description	
Armadillo (<i>Dasypus novemcinctus</i>)	1-to-1	n/a	ENSDNOG00000017393 Multi-species view Alignment Gene Tree (image)	GeneScaffold_3568-4397-64820-1	56	56	BRCA2 breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]	
Bushbaby (<i>Otolemur garnettii</i>)	1-to-1	n/a	ENSOGAG00000010588 Multi-species view Alignment Gene Tree (image)	GeneScaffold_5648-36138-48820-1	59	35	BRCA2 breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]	
Chicken (<i>Gallus gallus</i>)	1-to-1 (apparent)	n/a	ENSGALG00000017073 Multi-species view Alignment Gene Tree (image)	1:178837309-178873973-1	32	32	NP_989607.1 breast cancer 2, early onset [Source:RefSeq peptide;Acc:NP_989607]	
Chimpanzee (<i>Pan troglodytes</i>)	1-to-1	0.28906	ENSPTRG00000005766 Multi-species view Alignment Gene Tree (image)	13:32082480-32166147:1	99	99	Q8HZQ1_PANTR Breast cancer 2 Fragment [Source: UniProtKB/TrEMBL; acc: Q8HZQ1]	
Cow (<i>Bos taurus</i>)	1-to-1	0.44357	ENSBTAG0000000988 Multi-species view Alignment Gene Tree (image)	12:28415184-28466860-1	70	69	BRCA2 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

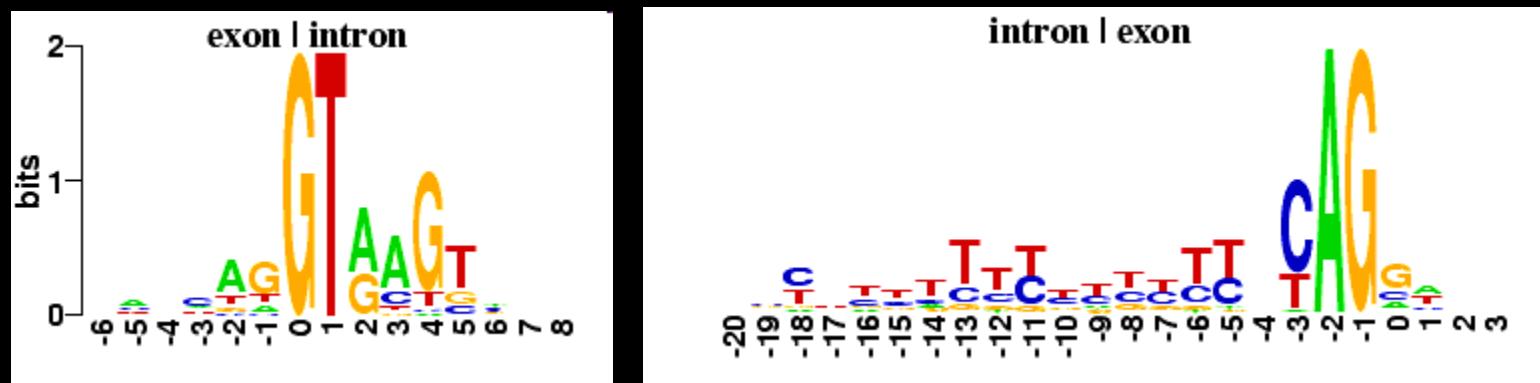
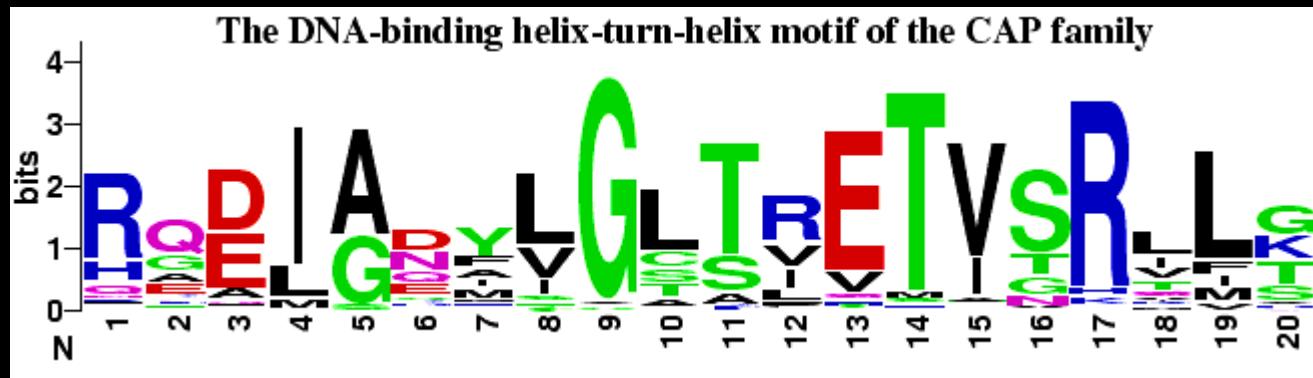
Alignment

ENSMICP00000010933/1-3908	-----	-----	-----	LGPISLNWFD	ELSSEAPPYN
ENSSTOP00000004979/1-3908	M-----	PIG SKERPTFFEI	FKTRCSGA-D	LGPISLNWFE	ELSSEASLYN
ENSOUCUP00000014514/1-3908	M-----	PIG SKERPTLFDI	FKMRCNTVD	LGPISLNWFE	ELSSEAPLYN
ENSRNOP00000001475/1-3908	M-----	TVE YKRRPTFWEI	FKARCSTA-D	LGPISLNWFE	ELSSEAPPYN
ENSTGUP00000012130/1-3908	M-----	ACK PVERPTFFEI	FAHCSSES-D	LGPISLNWFE	ELSAEAPPYE
ENSTRUP00000015030/1-3908	-----	-----	-----	-----	-----
ENSECAP00000013146/1-3908	M-----	PIG CKERPSFFDI	FKTRCNKA-D	LGPISLNWFE	ELSSEAPPYN
ENSDARP00000099674/1-3908	M-----	-----FEN FFHQIDR--E	GPLNPDWFE	ELNERAS--R	
ENSPCAP00000000440/1-3908	M-----	PVG FKERPTFFEI	FKARCSKA-D	LGPISLNWFE	ELSSKAPLYQ
ENSMODP00000033276/1-3908	M-----	SAG YQGKTTFFEV	FKTRCSSES-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
ENSOOPRP00000014082/1-3908	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
ENSLMLUP00000012516/1-3908	M-----	PIG CKERPTFFDI	FET-CSQ--	LGPISINWFE	ELSSEAPPYN
ENSMIEUP00000009812/1-3908	-----	-----	-----	-----	-----
ENSAACAP00000004460/1-3908	MAVMVKENDA	PTKKPSFFEI	YKARCCDS-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
ENSMMPUP00000009432/1-3908	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	ELSSEAPCN
ENSXETP00000037057/1-3908	-----	-----	-----	-----	-----
ENSVPAP00000000821/1-3908	M-----	PIG CKERPTFFEI	FRTRCNKA-D	LGPISLNWFE	ELSSEAPPYN
ENSPPPYP00000005997/1-3908	M-----	PVG SKERPTFFEI	FKTRCNKA-D	LGPISLHWFE	ELSSEAPPYN
ENSETEP00000003277/1-3908	L-----	VG TKERPTFFEI	FKARCSKA-D	LGPISLNWFE	ELSLEAPPYR
ENSGALP00000027524/1-3908	M-----	AYK SGKERTFFEV	FAHCSDS-D	LGPVSLDWFE	ELSSEAPPYE

Click “Alignment” to obtain alignments

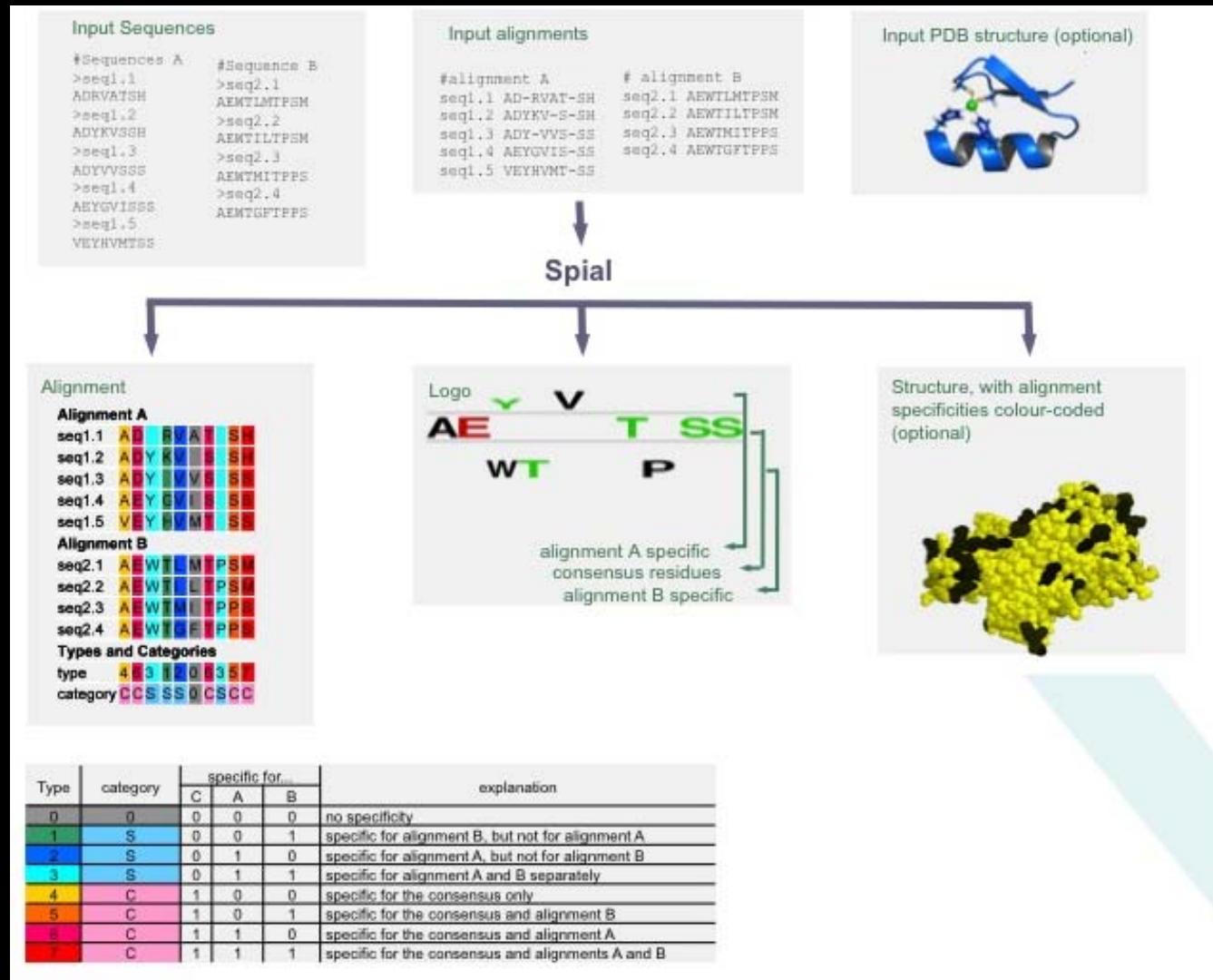
Weblogo for position specific analysis of an alignment of orthologs

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



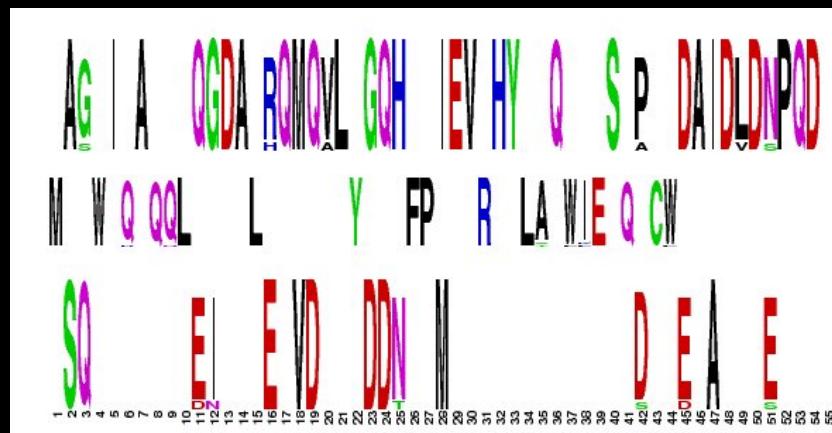
Evolutionary conservation reflects structural or a functional constrain on a position

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

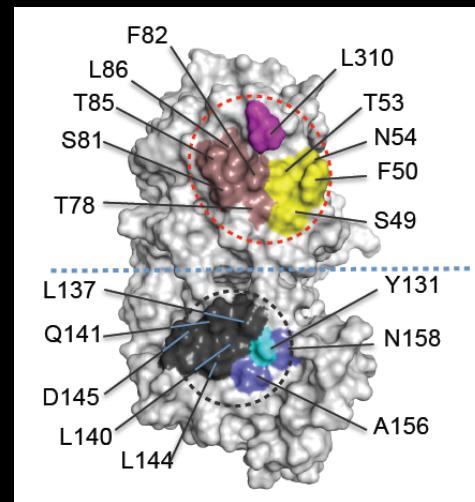


SPIAL: Specificity conferring residues in paralogous families

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

Question Set #4

How can I know more about my favourite gene?

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

Gene-based displays	
-	Gene summary
-	Splice variants (2)
-	Supporting evidence
-	Sequence
-	External references (3)
-	Regulation
-	Comparative Genomics
-	Genomic alignments (51)
-	Gene Tree (image)
-	Gene Tree (text)
-	Gene Tree (alignment)
-	Orthologues (44)

Gene: BRCA2 (ENSG00000139618)

breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]

Location [Chromosome 13: 32,889,611-32,973,347](#) forward strand.

Transcripts □ There are 2 transcripts in this gene

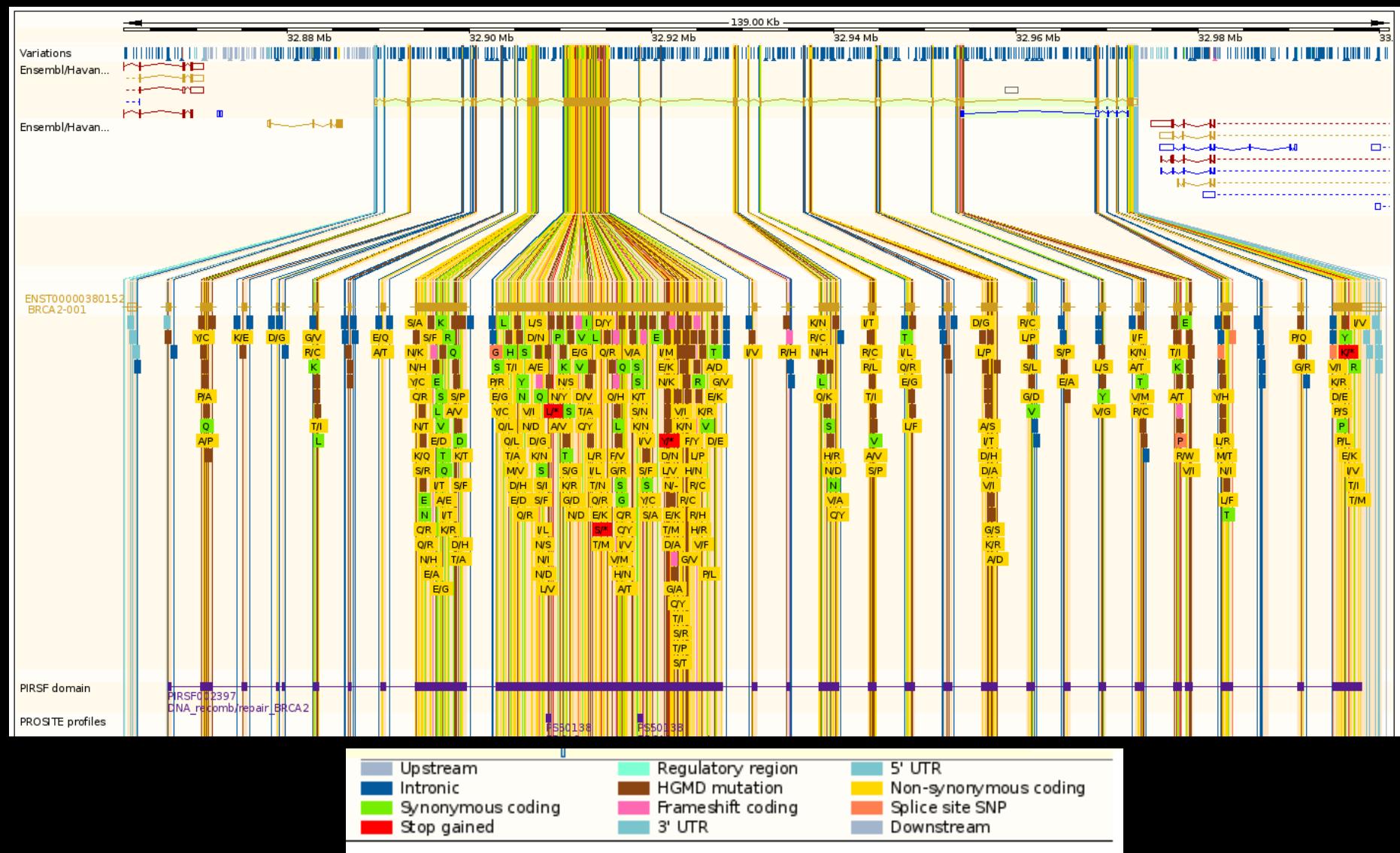
Show/hide columns						
Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
BRCA2-001	ENST00000380152	10930	ENSP00000369497	3418	Protein coding	CCDS9344
BRCA2-002	ENST00000470094	842	No protein product	-	Processed transcript	-

Gene-based displays

-	Gene summary
-	Splice variants (2)
-	Supporting evidence
-	Sequence
-	External references (3)
-	Regulation
-	Comparative Genomics
-	Genomic alignments (51)
-	Gene Tree (image)
-	Gene Tree (text)
-	Gene Tree (alignment)
-	Orthologues (44)
-	Paralogues
-	Protein families (1)
-	Genetic Variation
-	Variation Table
-	Variation Image
-	External Data
-	Personal annotation
-	ID History
-	Gene history

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>

Welcome to the Entrez cross-database search page

 PubMed: biomedical literature citations and abstracts	 Books: online books
 PubMed Central: free, full text journal articles	 OMIM: online Mendelian Inheritance in Man
 Site Search: NCBI web and FTP sites	 OMIA: online Mendelian Inheritance in Animals
 Nucleotide: Core subset of nucleotide sequence records	 dbGaP: genotype and phenotype
 EST: Expressed Sequence Tag records	 UniGene: gene-oriented clusters of transcript sequences
 GSS: Genome Survey Sequence records	 CDD: conserved protein domain database
 Protein: sequence database	 3D Domains: domains from Entrez Structure
 Genome: whole genome sequences	 UniSTS: markers and mapping data
 Structure: three-dimensional macromolecular structures	 PopSet: population study data sets
 Taxonomy: organisms in GenBank	 GEO Profiles: expression and molecular abundance profiles
 SNP: single nucleotide polymorphism	 GEO DataSets: experimental sets of GEO data
 dbVar: Genomic structural variation	 Cancer Chromosomes: cytogenetic databases
 Gene: gene-centered information	 PubChem BioAssay: bioactivity screens of chemical substances
 SRA: Sequence Read Archive	 PubChem Compound: unique small molecule chemical structures
 BioSystems: Pathways and systems of interacting molecules	 PubChem Substance: deposited chemical substance records
 HomoloGene: eukaryotic homology groups	 Protein Clusters: a collection of related protein sequences

How to use NCBI website effectively

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

The screenshot shows the NCBI "How To" section of their website. At the top, there's a blue header bar with the NCBI logo, a "Resources" dropdown, and a "How To" dropdown. Below the header is the NCBI logo and the text "National Center for Biotechnology Information". A search bar is present with a dropdown menu set to "All Databases" and a "Search" button.

The left sidebar has a "How To" heading and a list of links:

- NCBI Home
- All How To
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Small Molecules
- Taxonomy
- Training & Tutorials
- Variation

The main content area has tabs for "Resources" (highlighted in green) and "How To". Under "DATA & SOFTWARE", there are three items:

- Download a large, custom set of records from NCBI
- Download NCBI Software
- Download the complete genome for an organism

Under "DNA & RNA", there are nine items:

- Download a large, custom set of records from NCBI
- Retrieve all sequences for an organism or taxon
- Design PCR primers and check them for specificity
- Find published information on a gene or sequence
- Save a text search and/or receive regular search results by e-mail
- Find transcript sequences for a gene
- Find a curated version of a sequence record (NCBI Reference Sequence)
- Link from an object on a map to another resource
- View/download features around an object or between two objects on a chromosome
- Obtain a genomic DNA clone for a gene

Under "DOMAINS & STRUCTURES", there are three items:

- Save a text search and/or receive regular search results by e-mail
- Find the function of a gene or gene product
- View a mutation site in a 3D structure

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>

The screenshot shows the BioGPS homepage. At the top, there's a navigation bar with the GNF logo, 'Sign Up or Login' buttons, and links for 'BioGPS account' (username), 'OpenID account (what's that?)' (Google, YAHOO!, more), and a search bar. Below the header, the BioGPS logo ('BioGPS The Gene Portal Hub') is displayed next to a descriptive text: 'A free extensible and customizable gene annotation portal, a complete resource for learning about gene and protein function.' To the right, there's a thumbnail image of a detailed gene report interface. The main content area is divided into sections: 'Simple to use' (with four numbered steps: search, view, browse, build) and 'Search genes by Symbol or Accession' (with a search input field containing 'myod1', a 'Search by Symbol or Accession' button, and a link for 'Advanced Search'). On the right side, there's a sidebar titled 'Example Searches' with links to 'Gene Symbol(s)', 'Wildcard queries', 'Gene Ontology', 'Affymetrix IDs', and 'Interpro'. A 'Details »' link is located at the bottom left of the main content area.

GNF

Sign Up or Login

BioGPS account

username

OpenID account (what's that?)

Google™ YAHOO! more

BioGPS
The Gene Portal Hub

A free *extensible* and *customizable* gene annotation portal, a complete resource for learning about gene and protein function.

Simple to use

- 1 Search for your gene of interest
- 2 View the gene annotation report
- 3 Browse the gene report layouts
- 4 Build your own gene report

Details »

Search genes by Symbol or Accession

myod1

Press Ctrl-Enter or click **Search by Symbol or Accession**

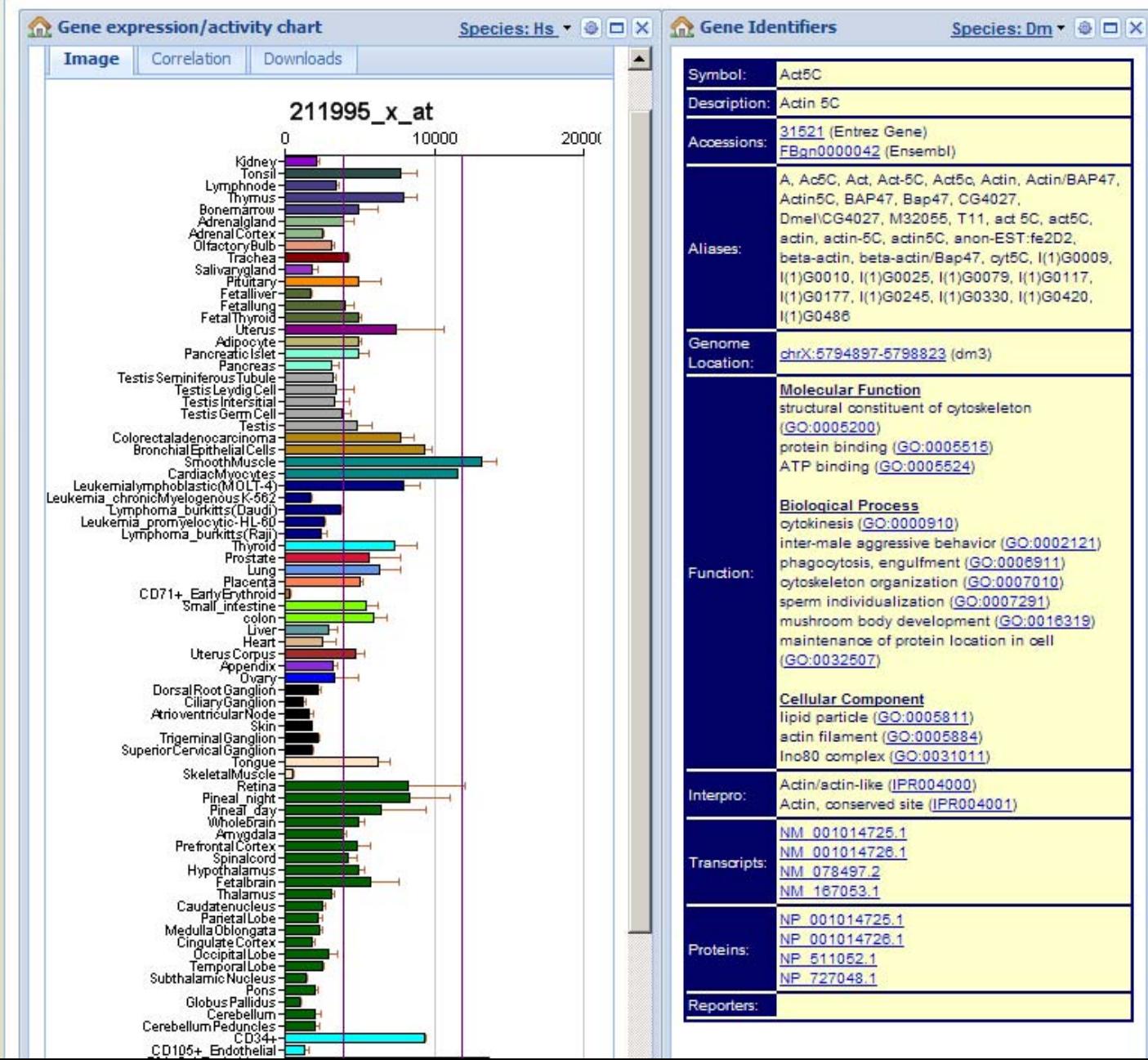
Example Searches
(click to try these samples)

- Gene Symbol(s)
- Wildcard queries
- Gene Ontology
- Affymetrix IDs
- Interpro

Advanced Search »
(keyword, interval, etc.)

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

Act5C (Actin 5C)



The screenshot shows the homepage of the ArrayExpress website. At the top left is the logo 'ARRAYEXPRESS' with a green hexagonal icon. At the top right is a stylized 'AE' logo. Below the header, a text box states: 'The **ArrayExpress Archive** is a database of functional genomics experiments including gene expression where you can query and download data collected to MIAME and MINSEQE standards. **Gene Expression Atlas** contains a subset of curated and re-annotated Archive data which can be queried for individual gene expression under different biological conditions across experiments.'

Experiments Archive
12282 experiments, 339730 assays

Experiment, citation, sample and factor annotations

Browse experiments Advanced query interface

Submitter/reviewer login ArrayExpress Query Help

Gene Expression Atlas
Information is unavailable at the moment

Genes Conditions

Any species (loading options)

Gene Expression Atlas Home

News

- 22 Apr 2010 - **Global 'Expression Space'**
EBI-Helsinki Team Integrates Array Data from Thousands of Samples to Map Global 'Expression Space' ...more
- 09 Apr 2010 - **A global map of human gene expression**
By integrating gene expression data from a large variety of human tissue samples, a global map of human gene expression is produced. For more details, please see the Nature Biotechnology [PDF - 676KB] or EMBL press release [PDF - 148KB].

Links

- [ArrayExpress User Survey](#)
- [Help | Training | FAQ | Citing](#)
- [Submit Data \(array based and re-sequencing\)](#)
- [Programmatic Access | FTP Access](#)
- [Software Downloads and Statistics](#)
- [EFO | Bioconductor Package | Quality Metrics](#)
- [ArrayExpress Scientific Advisory Board](#)
- [Functional Genomics Group](#)

Public repository for gene expression datasets at the EBI

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

The screenshot shows the NCBI Gene Expression Omnibus (GEO) homepage. At the top left is the NCBI logo, and at the top right is the GEO logo with the text "Gene Expression Omnibus". The top navigation bar includes links for "HOME", "SEARCH", "SITE MAP", "GEO Publications", "FAQ", "MIAME", and "Email GEO". A user status message "Not logged in | Login" is also present.

The main content area is divided into several sections:

- GEO navigation:** This section contains two main buttons: "QUERY" and "BROWSE".
 - The "QUERY" button leads to four categories: "DataSets", "Gene profiles", "GEO accession", and "GEO BLAST", each with a "GO" button.
 - The "BROWSE" button leads to "DataSets" and "GEO accessions". "DataSets" branches into "Platforms", "Samples", and "Series". "GEO accessions" branches into "Samples" and "Series".
- Site contents:** A sidebar on the right side of the page lists various site statistics and links.
 - Public data:** Platforms: 7,471, Samples: 445,380, Series: 17,386.
 - Documentation:** Overview, FAQ, Find, Submission guide, Linking & citing, Journal citations, Programmatic access, DataSet clusters, GEO announce list, Data disclaimer, GEO staff.
 - Query & Browse:** Repository browser, Submitters, SAGEmap, FTP site, GEO Profiles, GEO DataSets.
 - Submit:** Submit, New account.
- Submitter login:** A form for users to log in, featuring fields for "User id" and "Password", a "LOGIN" button, and links for "» New account" and "» Recover password".

Public repository for gene expression datasets at the NCBI

HUMAN PROTEIN ATLAS

The human protein atlas shows expression and localization of proteins in a large variety of normal human tissues, cancer cells and cell lines with the aid of immunohistochemistry (IHC) images and immunofluorescence (IF) confocal microscopy images.

Enter search:
[Advanced search](#)

Select a chromosome:

1	5	9	13	17	21
2	6	10	14	18	22
3	7	11	15	19	X
4	8	12	16	20	Y

OTHER

Or a protein class:
[Enzymes](#) | [GPCRs excl olfactory receptors](#) | [Kinases](#) | [Peptidases](#) |
[Transcription factors](#) | [Transporters](#) | [Membrane proteins](#) | [More...](#)

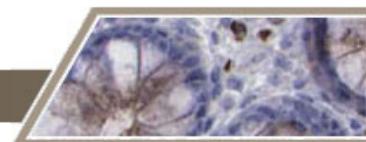
Filter search to show genes with tissue profiles

Version: 6.0 Atlas updated: 2010-03-26 ([release history](#))
Atlas content: 11,274 antibodies and 9,103,793 images.

Knut och Alice Wallenbergs Stiftelse

The HPR project is funded by the Knut & Alice Wallenberg foundation. The atlas is part of the HUPO Human Antibody Initiative ([HAI](#)).

Send questions, comments or suggestions to: contact@hpr.se. | [FAQ / Help](#)



Human protein expression levels (Immunohistochemistry)

Immunohistochemistry in normal cells

alph. sort order ▲

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

Search

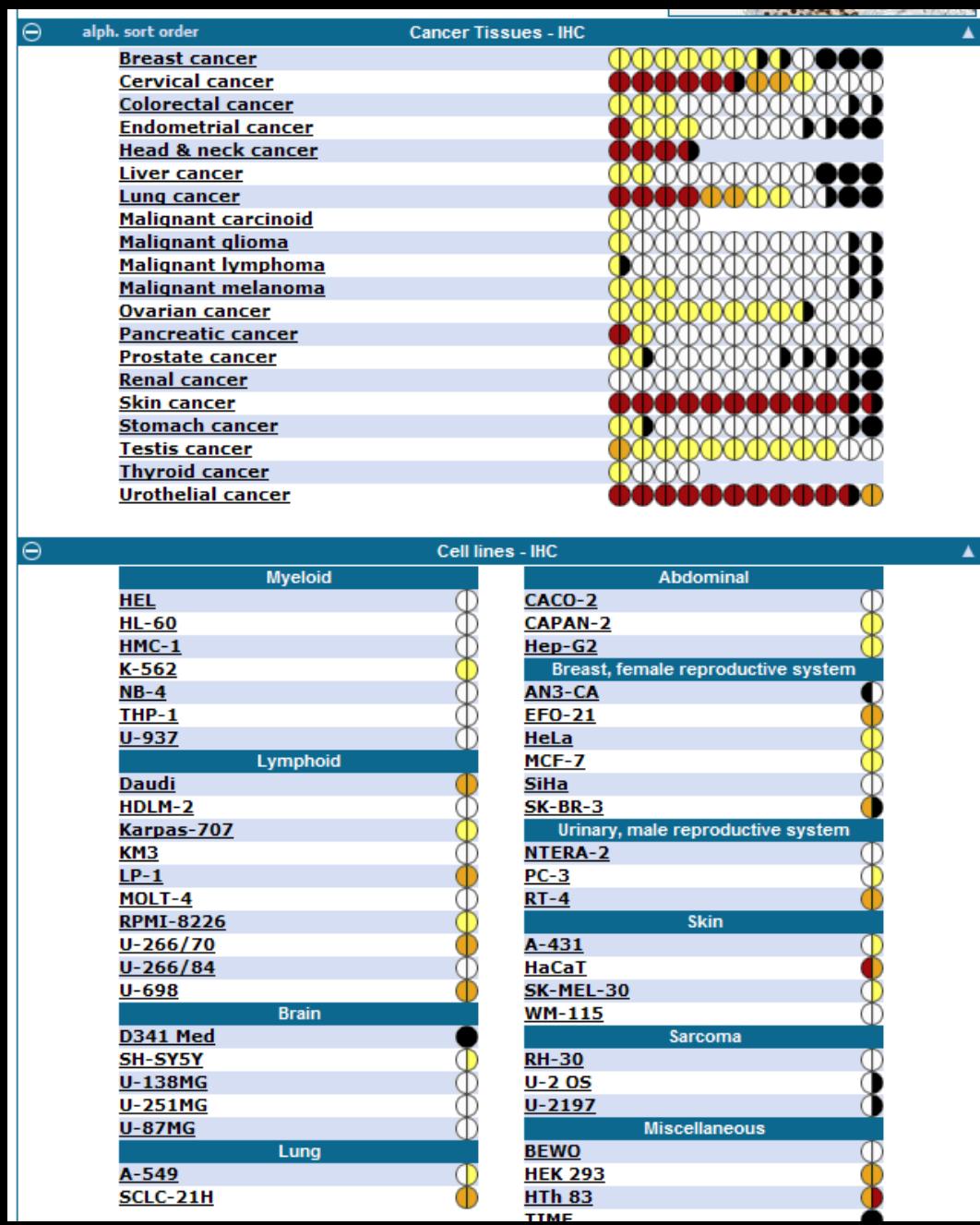
search

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)?

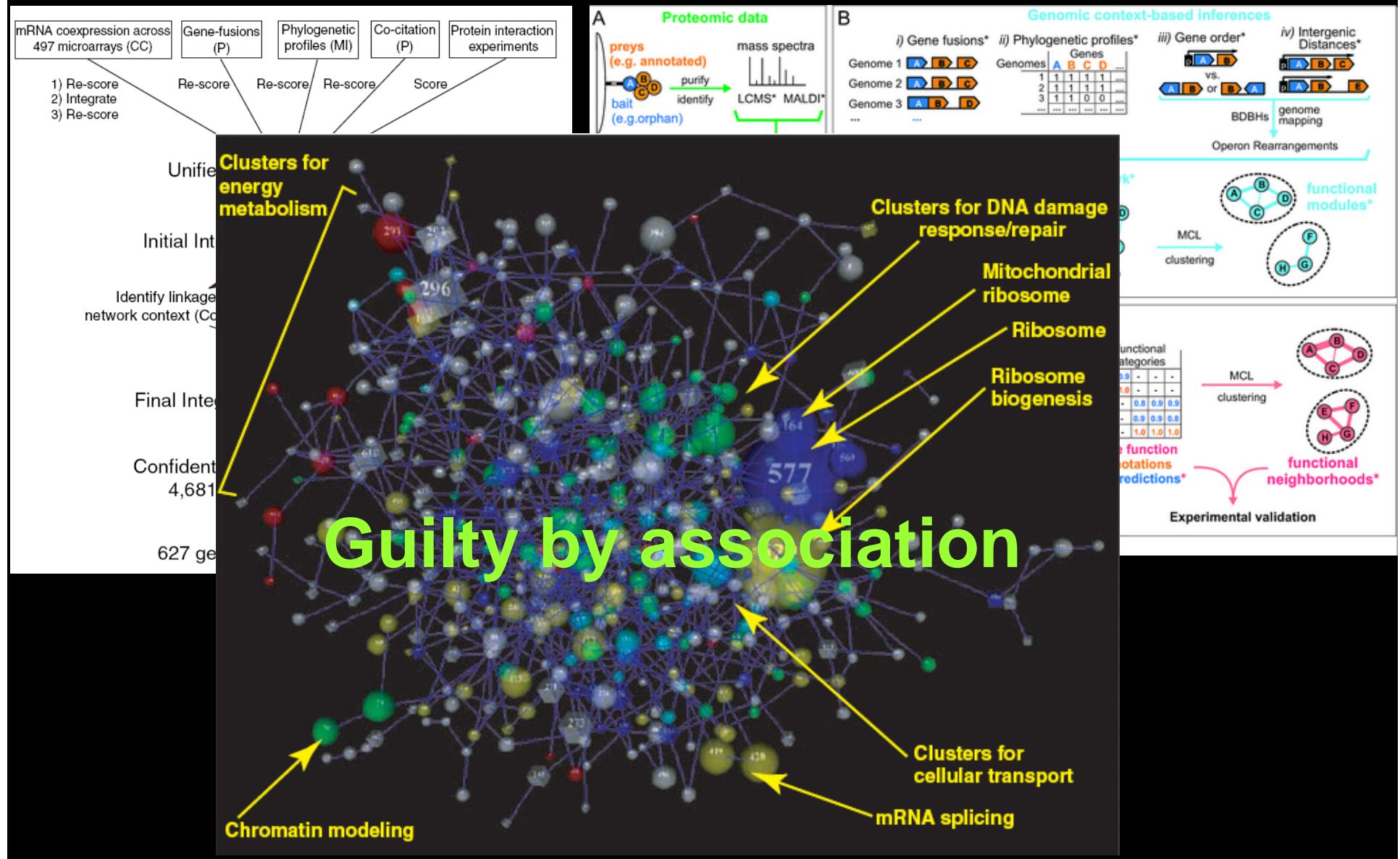
What is a functional interaction network?



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

Home · Download · Help/Info  **STRING 8.3**

STRING - Known and Predicted Protein-Protein Interactions

search by name search by protein sequence multiple names multiple sequences

protein name: (examples: #1 #2 #3)

(STRING understands a variety of protein names and accessions; you can also try a [random entry](#))

organism: auto-detect ▾

interactors wanted: COGs Proteins Reset GO !

please enter your protein of interest...

What it does ...

STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources:

Genomic Context High-throughput Experiments (Conserved) Coexpression Previous Knowledge 

STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently covers 2,590,259 proteins from 630 organisms.

More Info Funding / Support Acknowledgements Use Scenarios

STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) is being developed at [CPR](#), [EMBL](#), [SIB](#), [KU](#), [TUD](#) and [UZH](#).
STRING references: [Jensen et al. 2009](#) / [2007](#) / [2005](#) / [2003](#) / [Snel et al. 2000](#).
Miscellaneous: [Access Statistics](#), [Robot Access Guide](#), [STRING/STITCH Blog](#), [Supported Browsers](#).

What's New? This is version 8.3 of STRING - June 2010: the latest interaction data, updated textmining, and bugfixes ...
Sister Projects: check out [STITCH](#) and [eggNOG](#) - two sister projects built on STRING data!
Previous Releases: Trying to reproduce an earlier finding? Confused? Refer to our [old releases](#).

<http://string-db.org/>

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

 **FUNCOUP** *networks of functional coupling*

[What is FunCoup?](#) [Getting started](#) [FAQ](#) [Release notes](#) [Input data](#) [Download](#) [Citation](#)

Query **network**

for

[Which gene/protein IDs to use?](#)
[Example queries](#)

find genes/gene groups:
Gfd1

[More options >>>](#) Questions? Ask [Andrey Alexeyenko](#)

<http://sonorus.princeton.edu/hefalmpl>

HEFaIMp Download | Help | About

Welcome to HEFaIMp, a functional map of the human genome! Functional maps offer a way to interactively explore gene, pathway, process, and disease associations predicted from integrating hundreds of publicly available genomic datasets. You can focus on each entity - genes, processes, or diseases - and examine functional associations predicted from this data globally or in the context of a specific biological process. Please note that HEFaIMp does a great deal of calculation in real time, so if you request a complicated analysis, it might take up to a minute or two to complete.

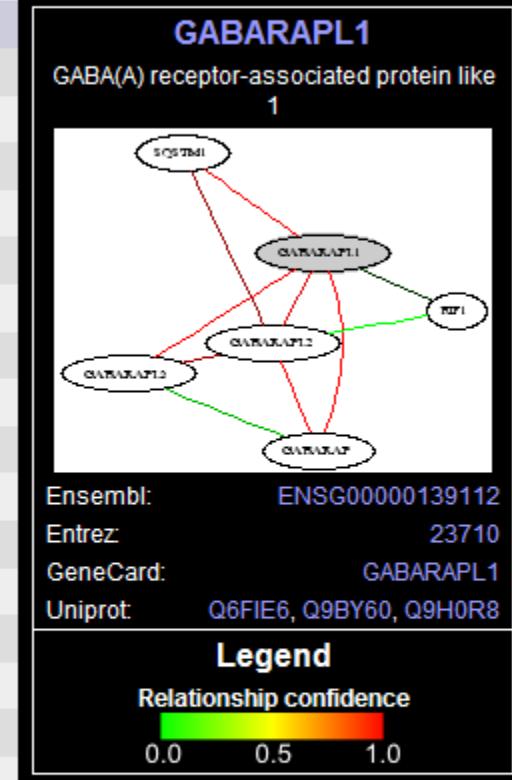
I would like to investigate a ?
and see how it relates to ?
in the context of ?
What gene? ?

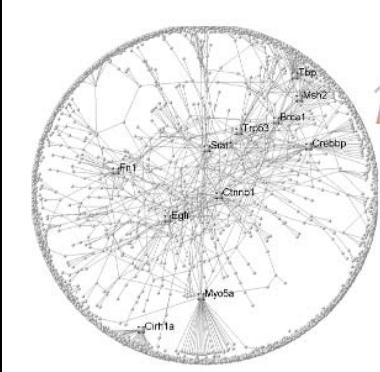
Looking for inspiration? Try some of the sample queries listed below, and you can click on any question mark popup (?) or browse through the [Help](#) linked from the top of every page. You can always find out what specific predictions or data generated a score by clicking on it, right down to the values for a single gene pair in a particular experiment. Feel free to [contact us](#) with any questions, comments, or suggestions, and thanks for your interest!

- If you're interested in finding new genes that might be associated with Alzheimer's, you can investigate a **disease** as it relates to **all genes** in the context of **protein processing** (or **learning and/or memory**) and request **Alzheimer disease**.
- Similarly, you can get a list of genes predicted to function in autophagy by focusing on a **biological process** as it relates to **all genes** in the context of **autophagy**, and investigate the specific process of **autophagy**.
- To predict what function(s) the gene *ALOX5AP* might have, investigate a **gene** to see how it relates to **biological processes** in the context of **all biological processes**, and ask for the specific gene **alox5ap**.

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](#)



mouseNET

MouseNET is a functional network for laboratory mouse based on integration of diverse genetic and genomic data. It allows you to accurately predict novel functional assignments and network components.



Legend	
PHOBS	Characterized Gene
BNII	Input Gene
PWPT	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.	

[Maximize Graph](#)
 [Download Ranked list of Genes In Graph](#)
 [View All Interactions](#)
 [View Disease-related genes](#)

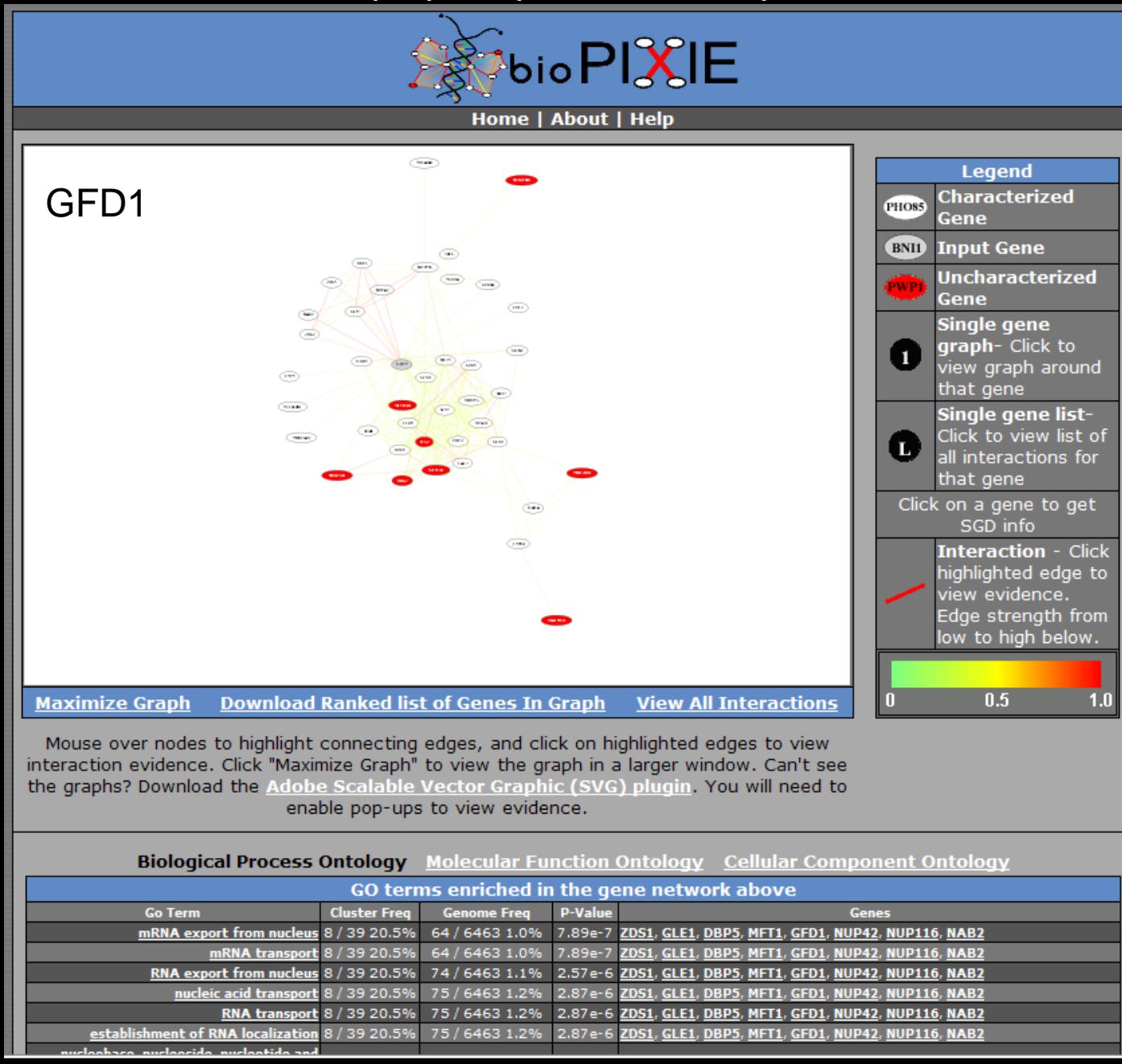
Mouse over nodes to highlight connecting edges, and click on highlighted edges to view interaction evidence. Click "Maximize Graph" to view the graph in a larger window. Can't see the graphs? Download the [Adobe Scalable Vector Graphic \(SVG\) plugin](#). You will need to enable pop-ups to view evidence.

[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>

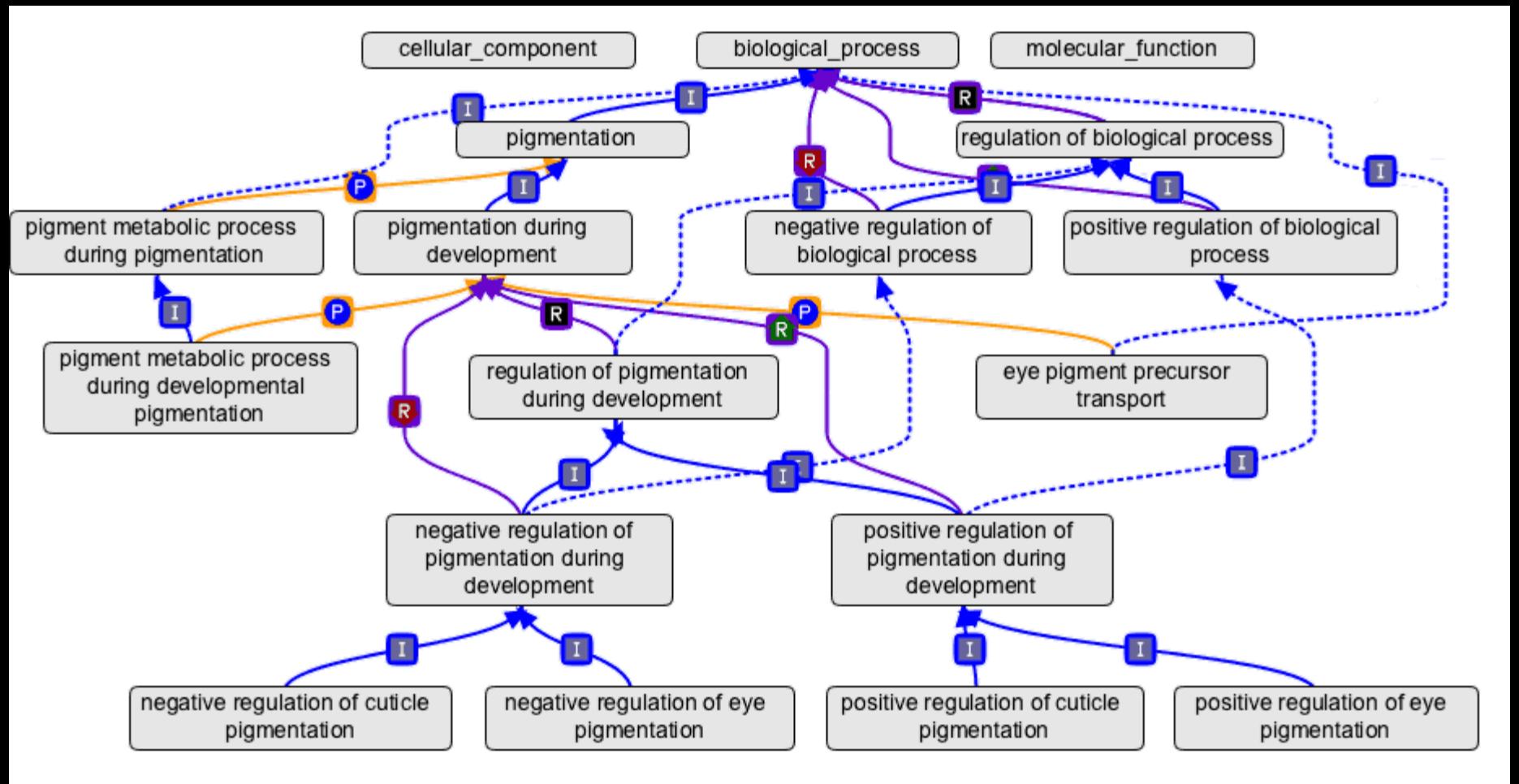


Question Set #7

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?

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 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

DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About Us

Shortcut to DAVID Tools

Functional Annotation
Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and [more](#)

Gene Functional Classification
Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. [More](#)

Gene ID Conversion
Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. [More](#)

Gene Name Batch Viewer
Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. [More](#)

Recommending: A [paper](#) published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7
[See release announcement for details](#)

2003 - 2010

Search

⊕ What's Important in DAVID?

- [New requirement to cite DAVID](#)
- [IDs of Affy Exon and Gene arrays supported](#)
- [Novel Classification Algorithms](#)
- [Pre-built Affymetrix and Illumina backgrounds](#)
- [User's customized gene background](#)
- [Enhanced calculating speed](#)

⊕ Statistics of DAVID

DAVID Citations per year (cumulative)

Based on *Google Scholar*
Updated in Jan. 2010

Year	Citations
2003	8
2004	48
2005	147
2006	296
2007	545
2008	930
2009	1672

Identify enriched biological themes, particularly GO terms
 Discover enriched functional-related gene groups
 Cluster redundant annotation terms
 Visualize genes on BioCarta & KEGG pathway maps
 Display related many-genes-to-many-terms on 2-D view.
 Search for other functionally related genes not in the list
 List interacting proteins
 Explore gene names in batch
 Link gene-disease associations
 Highlight protein functional domains and motifs
 Redirect to related literatures
 Convert gene identifiers from one type to another.

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

DAVID BIOINFORMATICS DATABASE

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](#)

Box A (empty)

Select Species

List Manager Help

Box B (empty)

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.

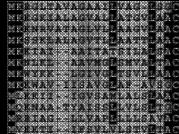
An example:

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

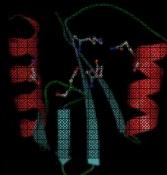
Approach

- Are they evolutionarily conserved?



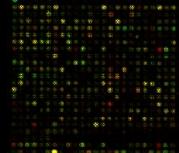
Presence/absence in
Humans, worm, fly

- What are their molecular functions?



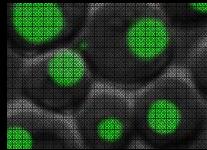
Sequence &
structure

- What are their expression patterns?



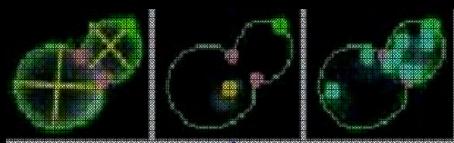
Expression pattern &
co-expression analysis

- What are their localization patterns?



Membrane
Nucleus

- What are the morphological changes?



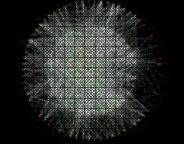
No. Nucleus
Vesicle size

- What are their interaction partners?



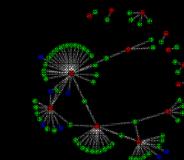
Protein complexes
Evolutionary conservation of partners

- Which TFs regulate their expression?



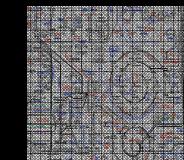
Transcription factors involved
Evolutionary conservation of regulators

- Which signaling proteins are involved?



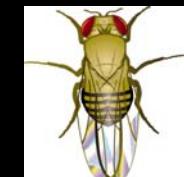
Signaling proteins involved
Evolutionary conservation

- 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