

MRC Laboratory of Molecular Biology

Investigating questions in biology using computational approaches

Balaji Santhanam
MRC Laboratory of Molecular Biology, Cambridge

At what level is this talk pitched at?

Data

Development of methods
Algorithms, programs, etc

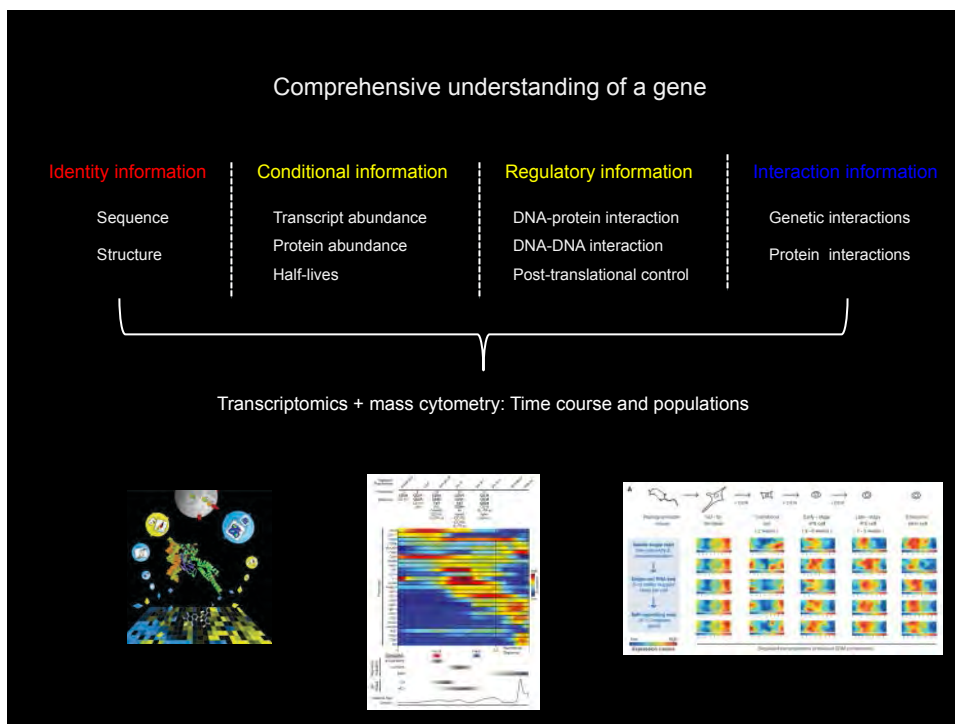
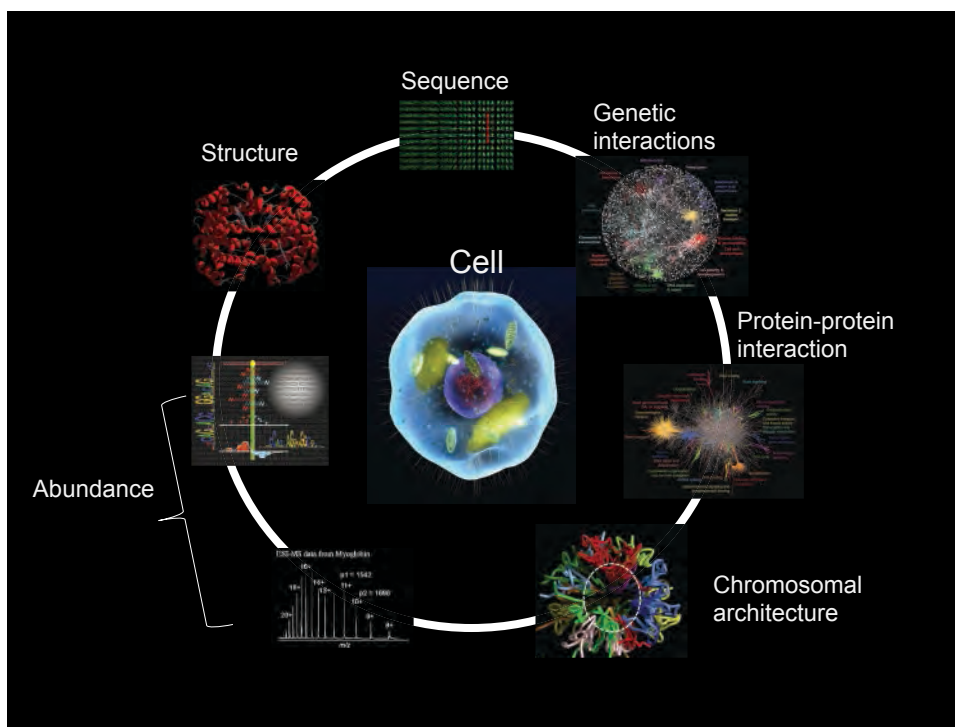
Uncovering general principles
Discovery using computational approaches

Prioritising experiments
Interpreting experimental results

“Computationally” inclined

“Biologically” inclined

	Focused	Genome-scale
Qualitative	Sequences	Protein interactions
Quantitative	qPCR	Transcriptome data

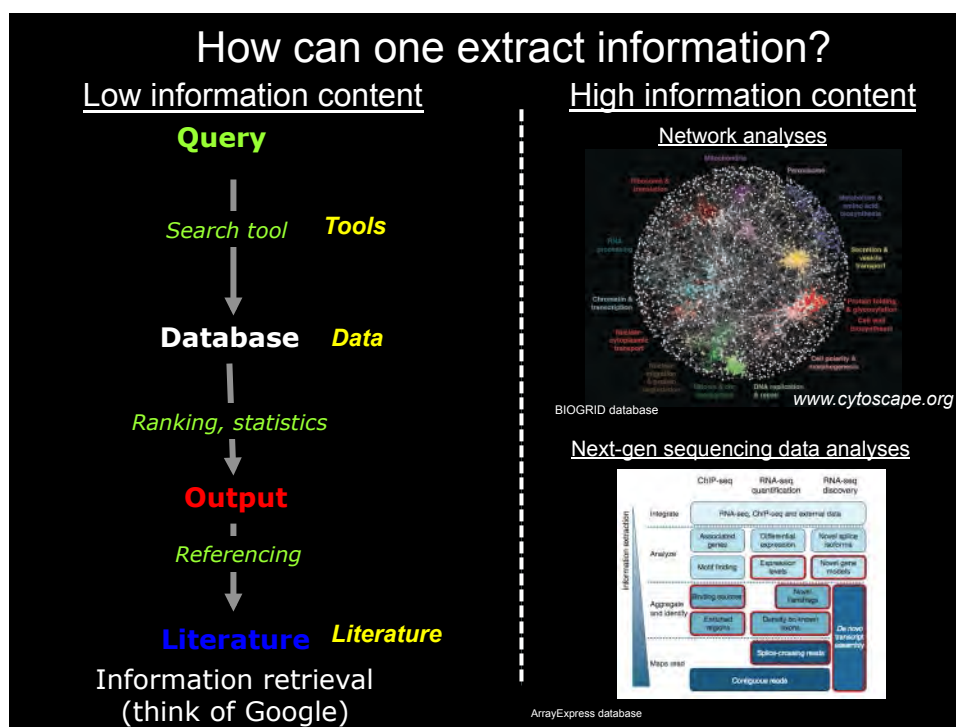


<p>Gene</p>	<p>Conducted high-throughput Experiments</p>
<p>No clue or want to more</p>	<p>Aim to make sense of the data</p>
	<p>And/or</p>
	<p>Identify most relevant set of genes</p>



Outline

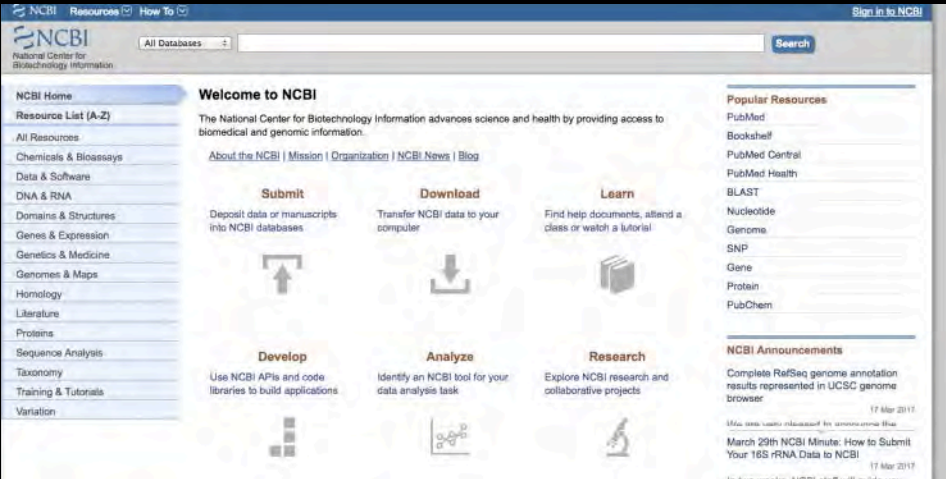
- Introduction to resources and tools (20-25 minutes)
- Some case studies (15 mins)
- High-throughput data



Question #1: Data

What is the list of databases that is currently available?

NCBI, NIH databases



The screenshot displays the NCBI homepage with the following elements:

- Navigation:** NCBI logo, Resource, How To, and Sign In to NCBI links.
- Search:** A search bar with a dropdown menu for "All Databases" and a "Search" button.
- Left Sidebar:** A vertical menu listing categories such as "Resource List (A-Z)", "All Resources", "Chemicals & Bioassays", "Data & Software", "DNA & RNA", "Domains & Structures", "Genes & Expression", "Genetics & Medicine", "Genomes & Maps", "Homology", "Literature", "Proteins", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation".
- Main Content:**
 - Welcome to NCBI:** A central message stating the center's mission to advance science and health through access to biomedical and genomic information.
 - Service Tiles:** Six main service tiles: "Submit" (Deposit data or manuscripts into NCBI databases), "Download" (Transfer NCBI data to your computer), "Learn" (Find help documents, attend a class or watch a tutorial), "Develop" (Use NCBI APIs and code libraries to build applications), "Analyze" (Identify an NCBI tool for your data analysis task), and "Research" (Explore NCBI research and collaborative projects).
- Right Sidebar:**
 - Popular Resources:** A list of key resources including PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem.
 - NCBI Announcements:** Recent news items, such as "Complete RefSeq genome annotation results represented in UCSC genome browser" (dated 17 Mar 2011) and "March 23rd NCBI Minute: How to Submit Your 16S rRNA Data to NCBI" (dated 17 Mar 2011).


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

<https://www.ensembl.org/>

<https://www.ensembl.org/>

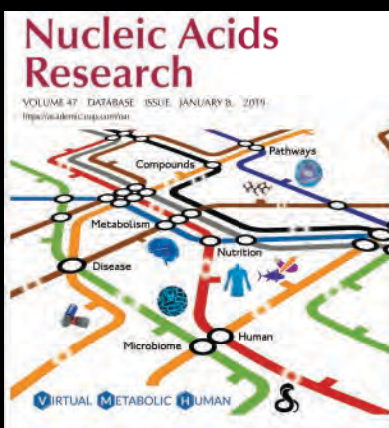
www.uniprot.org

www.uniprot.org



Over 1700 databases covering various aspects of molecular and cell biology

<https://academic.oup.com/nar/issue/47/D1>



Nucleic Acids Research
VOLUME 47 DATABASE ISSUE JANUARY 3, 2019
<https://academic.oup.com/nar>

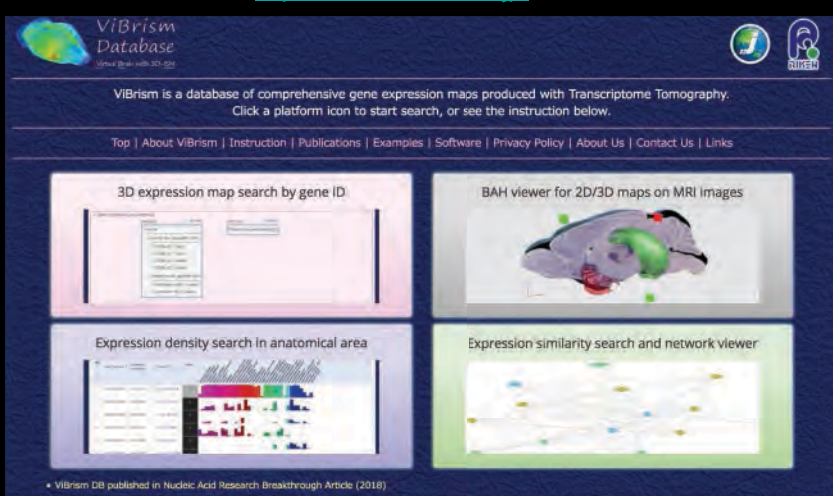
- I. Nucleic acid sequence, structure, and regulation
- II. Protein sequence and structure, motifs and domains
- III. Metabolic and signalling pathways, enzymes
- IV. Viruses, bacteria, protozoa, and fungi
- V. Human genome, model organisms, comparative genomics
- VI. Genomic variation, diseases and drugs
- VII. Plant databases
- VIII. Metagenomics

Prominent recent databases

- Cell Model Passports: Human Cancer Cell Models
- Cancer SEA: Cancer Single-Cell Atlas
- Editome Disease Knowledgebase: Curated collection of RNA editing events
- ViBrismDB: Tomographic transcriptome

ViBrism Database: Expression meets Tomography

<https://vibrism.neuroinf.jp/>



ViBrism Database
(First Open-Access 2018)

ViBrism is a database of comprehensive gene expression maps produced with Transcriptome Tomography.
Click a platform icon to start search, or see the instruction below.

[Top](#) | [About ViBrism](#) | [Instruction](#) | [Publications](#) | [Examples](#) | [Software](#) | [Privacy Policy](#) | [About Us](#) | [Contact Us](#) | [Links](#)

3D expression map search by gene ID

BAH viewer for 2D/3D maps on MRI images

Expression density search in anatomical area

Expression similarity search and network viewer

• ViBrism DB published in Nucleic Acid Research Breakthrough Article (2018)

https://www.drugbank.ca/

DRUGBANK

Browse Search Download About Help Blog Contact Us

WHAT ARE YOU LOOKING FOR?

Tylenol

Drug Targets Pathways Indications

DRUGBANK

The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.

The latest release of DrugBank (version 5.0.11, released 2017-12-20) contains 11,002 drug entries including 2,503 approved small molecule drugs, 943 approved biotech (protein/peptide) drugs, 109 nutraceuticals and over 5,110 experimental drugs. Additionally, 4,910 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target information.

https://gpcrdb.org/

GPCRdb Receptors Signal Proteins Ligands Drugs Structure Constructs Tutorial

Build diagrams

Get helix box and/or strand diagrams with custom coloring for your publications.

GPCRdb contains data, diagrams and web tools for G protein-coupled receptors (GPCRs). Users can browse all GPCR structures and the largest collections of receptor mutants. Diagrams can be produced and downloaded to illustrate receptor residues (snake-plot and helix box diagrams) and relationships (phylogenetic trees). Reference (structure) structure-based sequence alignments take into account helix bulges and constrictions, display statistics of amino acid conservation and have been assigned generic residue numbering for equivalent residues in different receptors. For an overview read the GPCRdb poster, articles or documentation.

Tweets by @gpcrdb

David Gloriam
Recent Content
#GPCRdb structure determination resource out in Nature Methods [link](#) in @BAGDRNL. Thanks to Raymond Stevens, Xavier Daupi, Lundbeck/London and European Research

Latest release
March 11, 2019

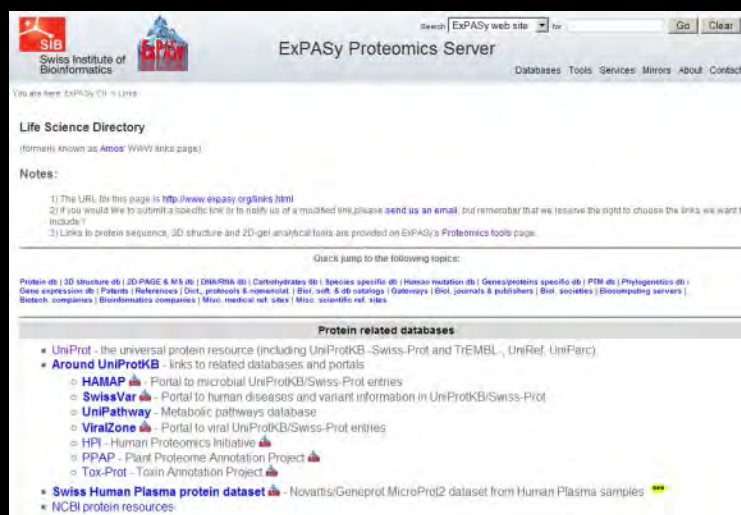
- Updated GPCR structures and statistics
- All homology and refined models updated
- Tools for structural determination [revisited](#) final and published

GPCRdb statistics

- Proteins: 15,147
- Human proteins: 420 (all non-orfatory)
- Species: 3,947
- Exp. structures: 321
- Refined structures: 233
- Ligands: 144,869
- Mutants: 31,943

Question #2: Tools

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



The screenshot shows the ExpASY Proteomics Server website. The header includes the SIB logo (Swiss Institute of Bioinformatics) and the text "ExpASY Proteomics Server". Below the header, there is a search bar and navigation links for "Databases", "Tools", "Services", "Mirrors", "About", and "Contact". The main content area is titled "Life Science Directory" and includes a "Notes" section with three numbered points. Below the notes, there is a "Protein related databases" section with a list of links to various protein resources.

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
- NCBI protein resources

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


Search [ExpASY web site] for

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 -> Proteins] [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

 **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 **HMMER** 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...](#)

MAX-PLANCK-GESSELLSCHAFT
Show results of job:

Recent jobs:
Select all Deselect all

queued
running
done
error

https://www.ebi.ac.uk/services

EMBL-EBI Services Research Training About us

Services

Overview A to Z Data submission Support

The European Bioinformatics Institute (EMBL-EBI) maintains the world's most comprehensive range of freely available and up-to-date molecular data resources.

Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically.

Tools & Data Resources

Tools

Clustal Omega
Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.
Multiple sequence alignment

Data resources

Ensembl
Genome browser, API and database, providing access to reference genome annotation

UniProt
A comprehensive resource for protein

Browse by type

DNA & RNA	Gene Expression	Proteins
Structures	Systems	Chemical Biology

https://genome.ucsc.edu/

UNIVERSITY OF CALIFORNIA SANTA CRUZ UCSC Genome Browser

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Our tools

- Genome Browser**
Interactively visualize genomic data
- BLAT**
rapidly align sequences to the genome
- Table Browser**
download data from the Genome Browser database
- Variant Annotation Integrator**
get functional effect predictions for variant calls
- Data Integrator**
combine data sources from the Genome Browser database
- Gene Sorter**
find genes that are similar by expression and other metrics
- Genome Browser in a Box (GBIB)**
run the Genome Browser on your laptop or server
- In-Silico PCR**
rapidly align PCR primer pairs to the genome
- LiftOver**
convert genome coordinates between assemblies
- VisiGene**
Interactively view in situ images of mouse and frog

More tools...

Our story

On June 22, 2000, UCSC and the other members of the International Human Genome Project consortium completed the first working draft of the human genome assembly, forever

What's new

Feb. 20, 2018 - New video: Visibility control in the Browser
Feb. 16, 2018 - New search support for chromosome aliases

DAVID Bioinformatics Resources 6.8
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

*** Welcome to DAVID 6.8 with updated Knowledgebase ([more info](#)).***
*** If you are looking for DAVID 6.7, please visit our [development site](#).***

Recommending: A paper published in *Nature Protocols* describes step-by-step procedure to use DAVID!

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 context 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](#)

Welcome to DAVID 6.8

2003 - 2017

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- 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

What's Important in DAVID?

- 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 Bioinformatic Resources Citations

4418

0 04 2005 06 2007 08 2009 10 2011 12 2013 14 2015 16

<https://david.ncifcrf.gov/>

Broad Institute data and tools

BROAD INSTITUTE

HOME | PEOPLE | SCIENCE | SERVICES | CONTACT | NEWS AND MEDIA

DATA, SOFTWARE AND TOOLS

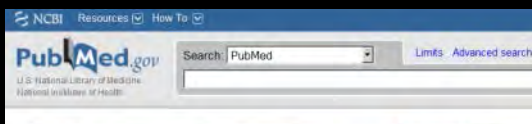
SEARCH

PROJECT	TYPE
ABSOLUTE	Copy number variation, Cancer Genome Analysis
Actinobacter	Human Microbiome Project
Actinobacter agglomerans	
Actinobacteroides	Human Microbiome Project
Actinobacteroides sp. D21	BACTERIA
Actinobacter	
Actinobacter (HMP)	Human Microbiome Project
Actinobacter group	BACTERIA
Actinobacterium	Human Microbiome Project
Actinoptea	Human Microbiome Project
Actinoptea (various CG)	BACTERIA

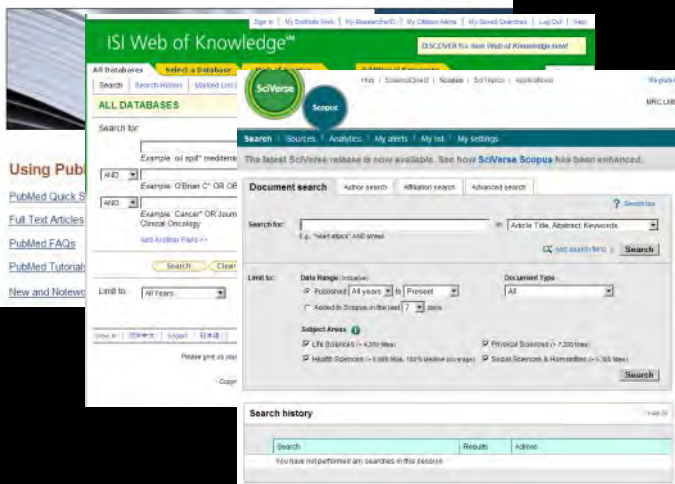
Question #3: Literature

What is the list of databases, web-servers and programs that are currently available to explore the literature?

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




<http://apps.isiknowledge.com>



<http://www.scopus.com/home.url>

http://www.ihop-net.org/UniPub/iHOP/





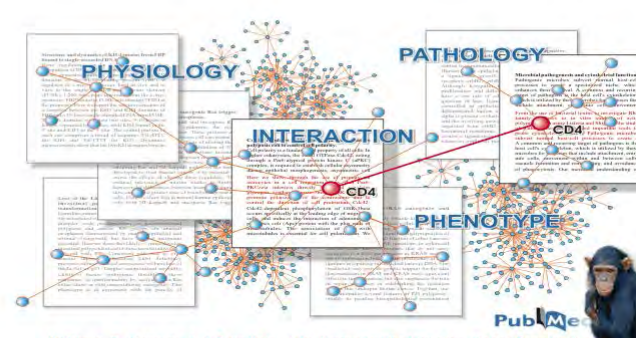
**Information Hyperlinked
Over proteins**

Search Gene

Gene Model
Developer's Zone ***
How to cite iHOP

Contact
Links
Help




Hoffmann, R., Valencia, A. A Gene Network for Navigating the Literature. *Nature Genetics* 36
more than 2,800 organisms, 110,000 genes, 23.4 million sentences.
...always up to date - every day.

Search for a gene *synonym* or *accession number*... [\(Click here for an example: SNF1\)](#)

all fields in All organisms

[SEARCH]



**Information Hyperlinked
Over proteins**

Search Gene

Show sentences
Find in this Page

Filter and options
Gene Model

Developer's Zone ***
Help

Copyright & Implementation
by Richard Hoffmann

Symbol / Name	Synonyms	Organism
TP53 tumor protein p53	Antigen NY-CO-13, Cellular tumor antigen p53, FLJ92943, LFS1, p53, P53, Phosphoprotein p53, TRP53, Tumor suppressor p53	Homo sapiens

[WikiGenes](#) [UniProt](#) [IntAct](#) [PDB Structure](#) [OMIM](#) [NCBI Gene](#) [NCBI RefSeq](#) [NCBI UniGene](#) [NCBI Accession](#)

more than 2,800 organisms, 110,000 genes, 23.4 million sentences.
...always up to date - every day.

[Homologues of TP53](#) ...
[Definitions for TP53](#) ...
[Most recent information for TP53](#) ...
[Enhanced PubMed/Google query](#) ...

WARNING: Please keep in mind that gene selection is done automatically and can exhibit a certain error. [Please note about synonym ambiguity and the iHOP identifier rules](#)

[Find in this Page](#)

Genes in this view interact with TP53 in the literature - Interaction Information is available whenever you see this symbol - [Read more](#).

Click on a gene below to see sentences on a specific interaction or click here to see all sentences.

Symbol	Name	Organism	N° of sentences with TP53
RAD51	RAD51 homolog (RecA homolog, E. coli) (S. cerevisiae)	Homo sapiens	68
TBP	TATA box binding protein	Homo sapiens	38
HMGB1	high-mobility group box 1	Homo sapiens	21
Mdm2	Mdm2 p53 binding protein homolog (mouse)	Homo sapiens	265
CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	Homo sapiens	2437
MKI67	antigen identified by monoclonal antibody Ki-67	Homo sapiens	1421
CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	Homo sapiens	1389

Explosion of information about living systems

Sequence

> 45,000,000 sequences from
> 160,000 organisms (Genbank, NCBI, UniProt)



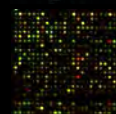
Structure

150,861 structures from
> 1500 organisms (PDB, MSD)



Expression

>100,000 different conditions
> 200 organisms (SMD, GEO, ArrayExpress)



Interaction

>800,000 interactions
60 organisms (Bind, DIP, BIOGRID, publications)



Literature

Over 50 million abstracts and papers
Numerous organisms (PubMed, ISI, Scopus)



Major challenge – How to exploit this information?

Sequence



Query sequence

BLAST, PSI-BLAST, etc

Sequence database (NCBI,
ENSEMBL)

E-value, score, etc

**Sequence
Alignment**

Structure



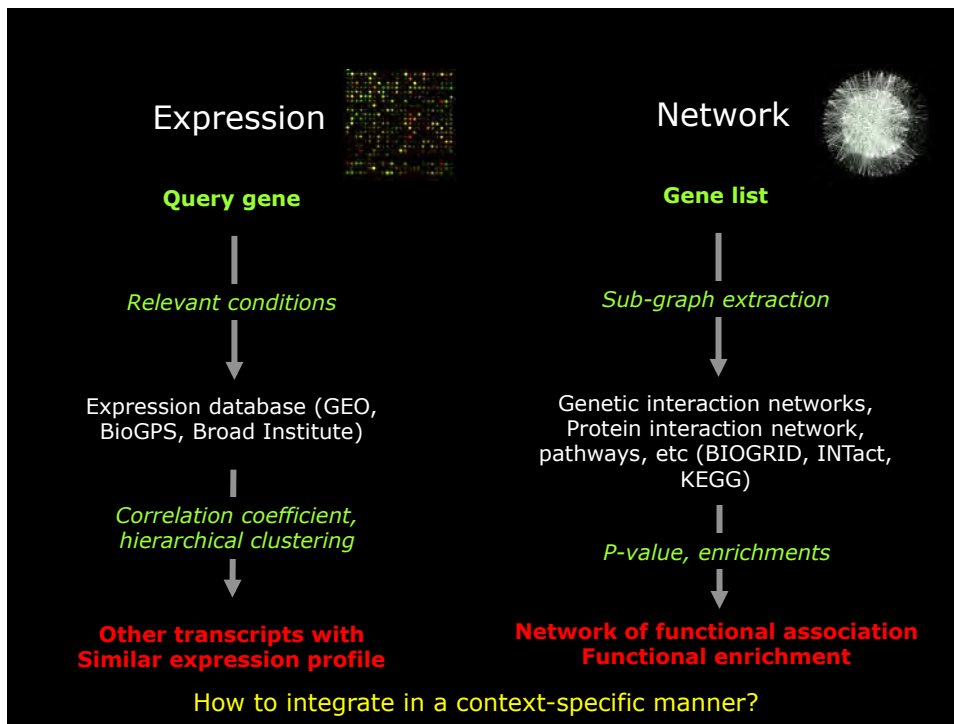
Query structure

DALI, TopSearch, VAST, etc

Structure database (PDB)

p-value, score, etc

**Structure
Alignments**



Outline

- Introduction to resources and tools (20-25 minutes)
- A case studies (15 mins)
- High-throughput data

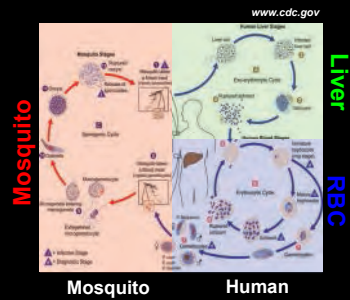
Previous comparative genomic analysis of eukaryotes suggested lack of detectable transcription factors in Plasmodium

Large number of genes



5300 genes with over 700 metabolic enzymes
 Extensive complement of chromosomal regulatory proteins
 Extensive complement signaling proteins (GTPases, kinases)

Complex life cycle



The Problem!

How does this pathogen regulate gene expression?

Possible explanations for the paradoxical observation

Alternative regulatory mechanisms

Chromatin-level regulation
 Post-translational modification
 RNA based regulation

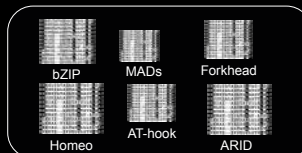
Undetected transcription factors

Distantly related or unrelated to known DNA binding domains



Proteome of Plasmodium

+



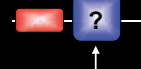
Profiles & HMMs of known DBDs



AT-Hook

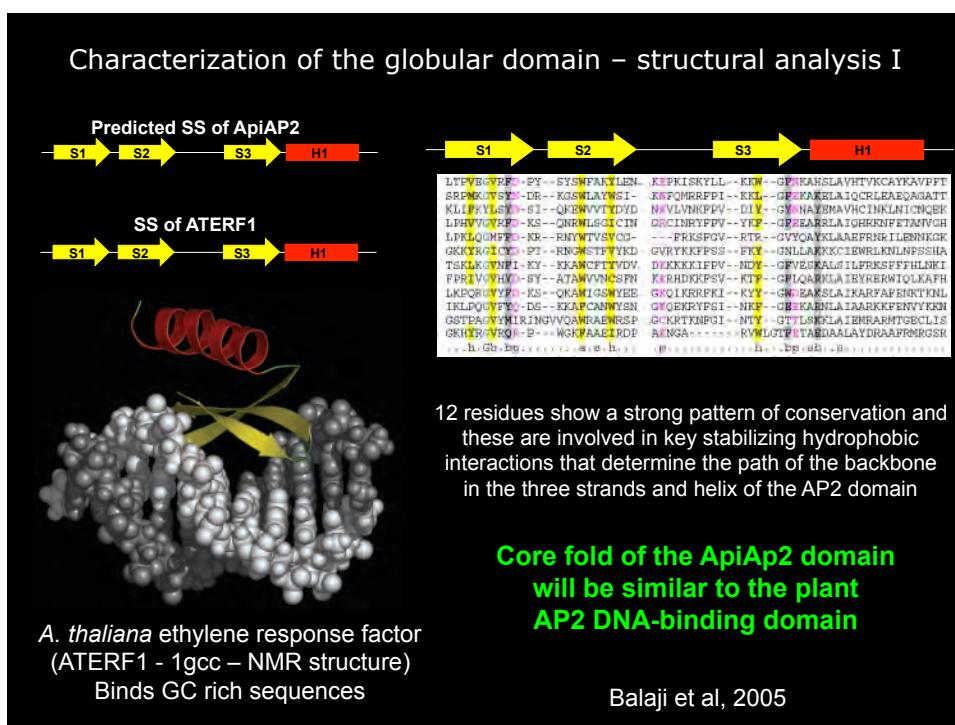
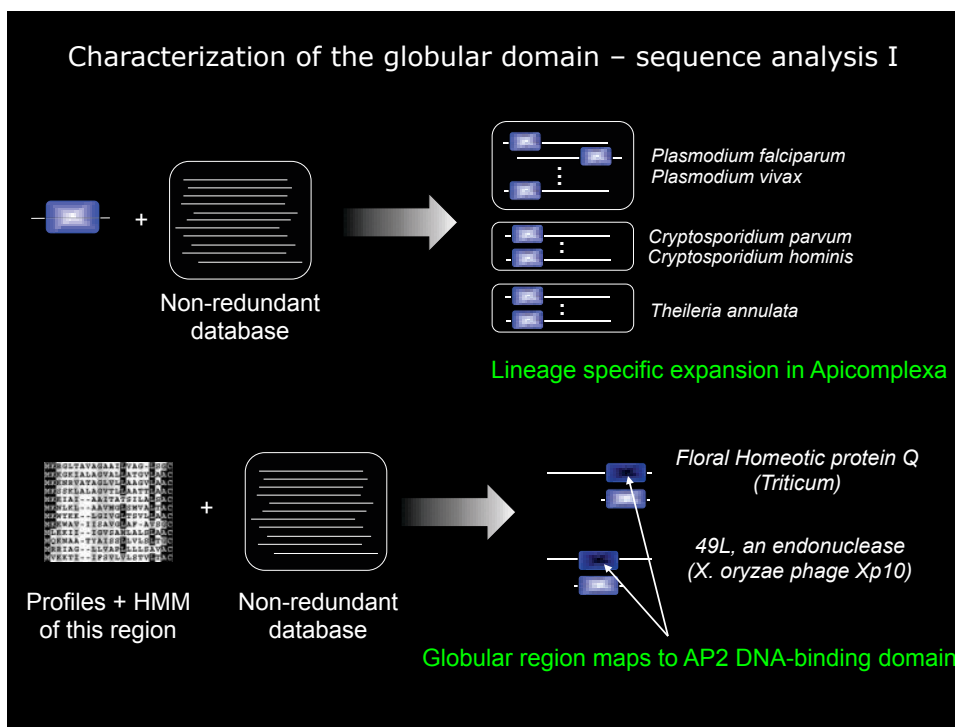


PF14_0633

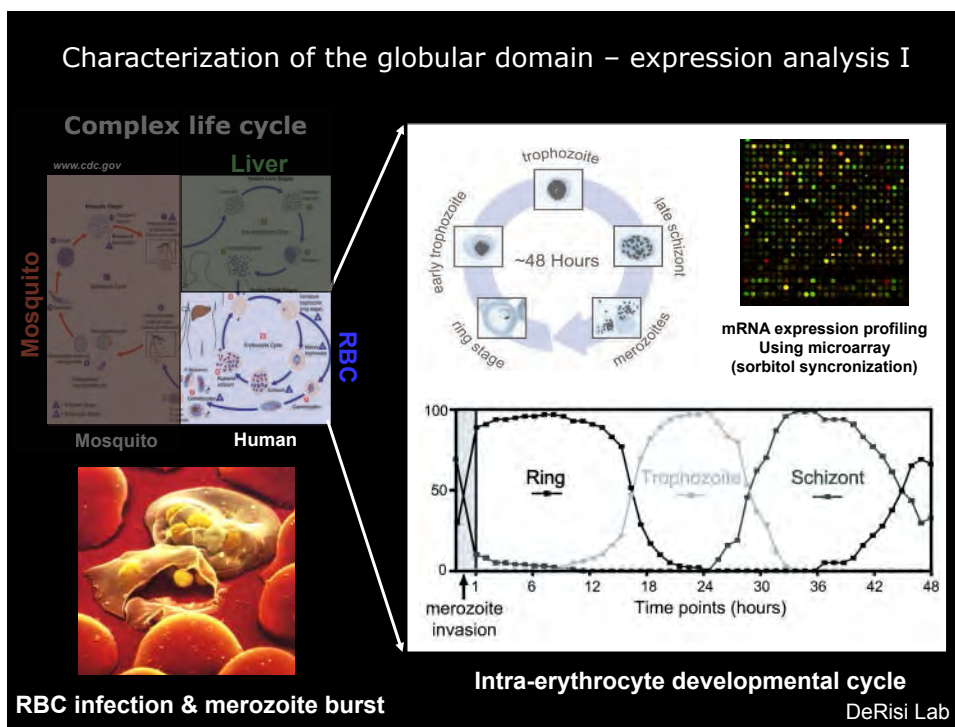


Uncharacterized Globular domain ~60 aa

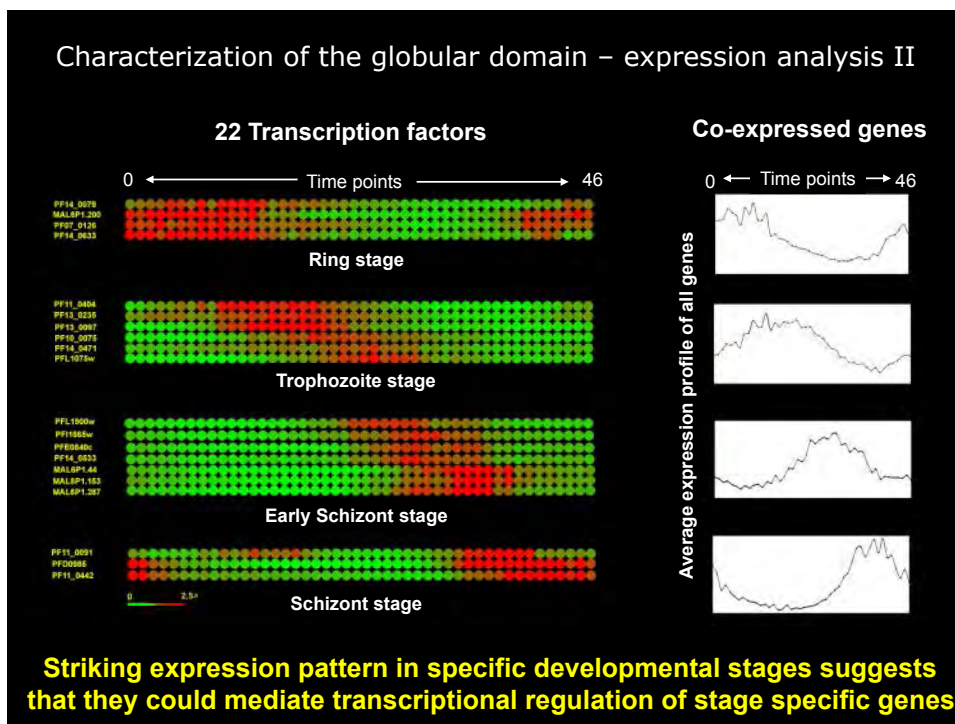
The suspect!



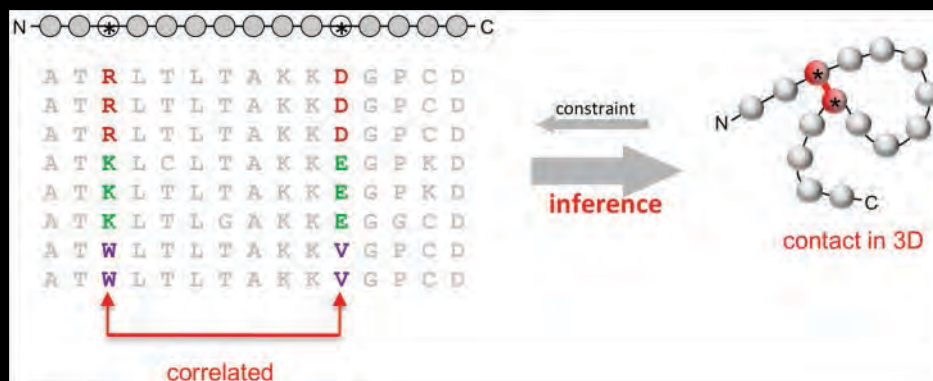
Characterization of the globular domain – expression analysis I



Characterization of the globular domain – expression analysis II



Coevolution to identify protein functions



Debora S. Marks*, Lucy J. Colwell*, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, Chris Sander. PLoS One. 2011

Thomas A. Hopf, Lucy J. Colwell, Robert Sheridan, Burkhard Rost, Chris Sander, Debora S. Marks Cell, 2012

Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science, 1999

Evcouplings.org

Home | Evcouplings | Evcouplings | Publications | Data Sets | Code | Tutorial | FAQ | About | Contact

Try the new submission server at evcouplings.org

EVolutionary Couplings

New! Evcouplings+EVfold

- Run your protein of interest
- Generate an alignment (if you don't have one)
- Calculate EC scores for all pairs of residues
- Map high ranking ECs onto a contact map
- Compare ECs to contacts in 3D structure (if you want)
- Calculate cumulative EC strength for individual residues (for functional interpretation)
- Point strong ECs onto a 3D structure
- Fold the protein when unknown structure (module enough sequences)
- Relax the models
- Move alternative membrane topology input

dx.doi.org/10.1093/bioinformatics/btt100

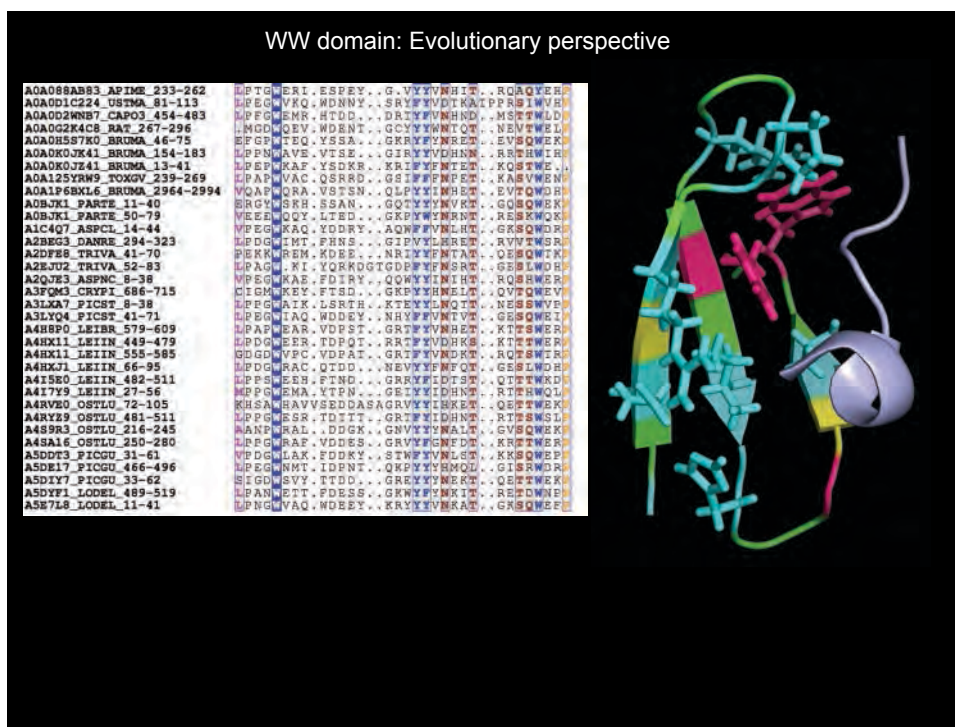
New! EVcomplex

- Analyze two known interacting domains
- Find and join partner sequences from the two MSAs
- Distinguish and compare inter-ECs into inter-ECs
- Produce a sequence map for docking use
- Produce visualizations and comparisons to known 3D PDB
- Evaluate in-situ test with coupling data www.evcouplings.org


Legacy Evcouplings

- Run your protein of interest
- Generate an alignment (if you don't have one)
- Calculate EC scores for all pairs of residues
- Map high ranking ECs onto a contact map
- Compare ECs to contacts in 3D structure (if you want)
- Calculate cumulative EC strength for individual residues (for functional interpretation)
- Point strong ECs onto a 3D structure

dx.doi.org/10.1093/bioinformatics/btt100



Intrinsically unstructured or disordered proteins/regions



<http://elm.eu.org/>

ELM The Eukaryotic Linear Motif resource for Functional Sites in Proteins

ELM Home ELM Prediction ELM DB ELM Candidates ELM Information ELM downloads Help

Welcome to the Eukaryotic Linear Motif (ELM) resource

This computational biology resource mainly focuses on annotation and detection of eukaryotic linear motifs (ELMs) by providing both a repository of annotated motif data and an exploratory tool for motif prediction. ELMs, or short linear motifs (SLMs), are compact protein interaction sites composed of short stretches of adjacent amino acids. They are enriched in intrinsically disordered regions of the proteome and provide a wide range of functionality to proteins (Davey, 2011; Van Roey, 2014). They play crucial roles in cell regulation and are also of clinical importance, as aberrant SLM function has been associated with several diseases and SLM mimics are often used by pathogens to manipulate their hosts' cellular machinery (Davey, 2011; Uyar, 2014).

ELM Prediction

The ELM prediction tool scans user-submitted protein sequences for matches to the regular expressions defined in ELM. Distinction is made between matches that correspond to experimentally validated motif instances already curated in the ELM database and matches that correspond to putative motifs based on the sequence. Since SLMs are short and degenerate, overprediction is likely and many putative SLMs will be false positives. However, predictive power is improved by using additional filters based on contextual information, including taxonomy, cellular compartment, evolutionary conservation and structural features.

Protein sequence

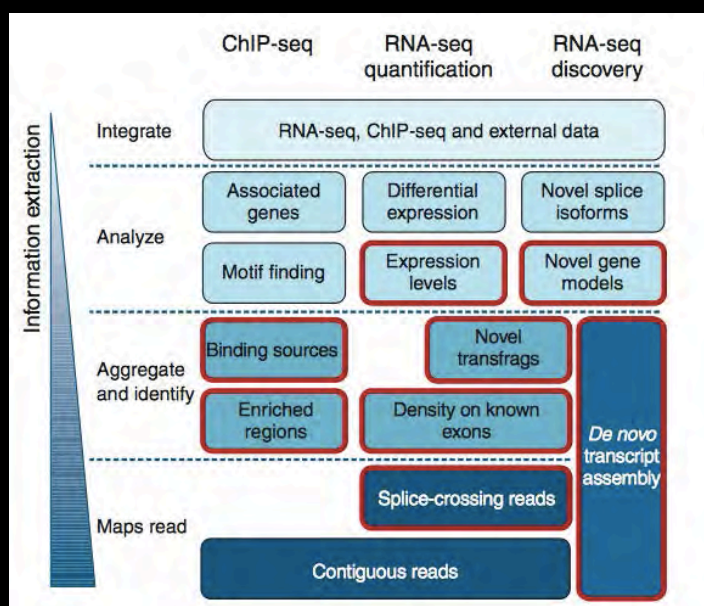
Enter UniProt identifier or accession number. (auto-completion)
e.g. EPN1_HUMAN, P04637_TAU_HUMAN, [RANDOM]

POB-Structure : 2HE2 showing a peptide from ELM class:LiG_PDX_Class_1

- ELM database update
Added 9 new instances for DOC_CYCLIN_Rxl_1
- ELM database update
Updated the motif DOC_CYCLIN_Rxl_1 and 19 motif instances
- ELM database update
Added a new motif: LiG_PROFILIN_1 and news instances for MOD_CAABox and MOD_CDK_SPxK_1
- ELM database update
22 New Fungal instances added for MOD_Pik_1

- Introduction to resources and tools (20-25 minutes)
- A case studies (15 mins)
- High-throughput data

Next-gen sequencing data analysis



Web server intergrated platform:

- Galaxy server: <https://galaxyproject.org>

The screenshot shows the Galaxy Project website interface. At the top, there is a navigation bar with links for 'Use', 'Learn', 'Teach', 'Support', 'Community', and 'Deploy & Develop', along with a search bar and a link to 'Edit on GitHub'. The main content area is titled 'Latest Tutorials' and features a promotional graphic for a metagenomics tutorial. The graphic includes a mouse on the left, an arrow labeled '150 days' pointing to a larger, fatter mouse on the right, and the text 'Eat, get fat, and be merry'. Below the graphic, there are sections for 'News', 'Events', and a social media feed for '@galaxyproject'.

Software packages for differential gene expression

- ◆ Cufflinks: <http://cole-trapnell-lab.github.io/cufflinks/> uses TopHat and estimates differential expression based on a reference genome
- ◆ RSEM (RNA-Seq by Expectation-Maximization) <http://deweylab.github.io/RSEM/> integrated to EBSeq

Reference genome mapping and exon-exon junction identification

- TopHat: <http://ccb.jhu.edu/software/tophat/index.shtml> uses Bowtie to map and identify splice junctions between exons

Global characteristics of the mapping tools

<i>Tool</i>	<i>Format</i>	<i>Algorithm</i>	<i>Threads</i>	<i>Gaps</i>	<i>Mismatches</i>
BWA	SAM	BWT	yes	yes	yes
Novoalign	SAM	hash the ref.	yes	yes	yes
Bowtie	SAM	BWT	yes	no	yes
SOAP2	perso	BWT	yes	no	at most 2
BFAST	SAM	hash the ref.	yes	yes	yes
SSAHA2	SAM	hash the ref.	no	no	yes
MPscan	perso	suffix tree	no	no	no
GASSST	SAM	hash the ref.	yes	yes	yes
PerM	SAM	hash the ref.	no	no	yes

Web server intergrated platform:

Galaxy server: <https://galaxyproject.org>Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the [free public server](#) or [your own instance](#), you can perform, reproduce, and share complete analyses.

Use Galaxy



Use [project's free server](#) or other public servers

Get Galaxy



Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy



Screencasts, [Galaxy 101](#), ...


Get Involved



[Mailing lists](#), [Tool Shed](#), [wiki](#)

[Search all resources](#)

Protein interactions databases




BioGRID 3.4
 Welcome to the Biological General Repository for Interaction Datasets
 BioGRID is an interaction repository with data compiled through comprehensive curation efforts. Our current index is version 3.4.148 and associates 58,254 publications for 1,418,874 protein and genetic interactions, 27,748 chemical associations and 39,569 post-translational modifications from major model organism species. All data are freely provided via our search tools and available for download in standardized formats.

AREAS OF INTEREST TO HELP YOU GET STARTED

- Build and Download Interaction Datasets
- Link To Us or Submit Interactions
- Online Tools and Resources
- View Our Interaction Statistics

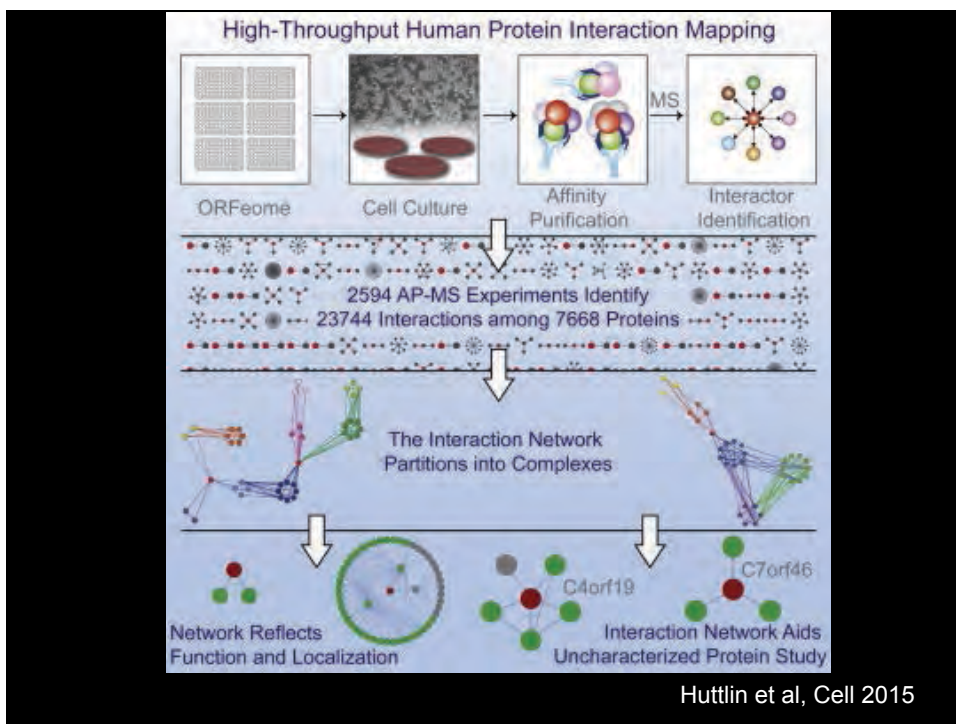
<https://thebiogrid.org/>



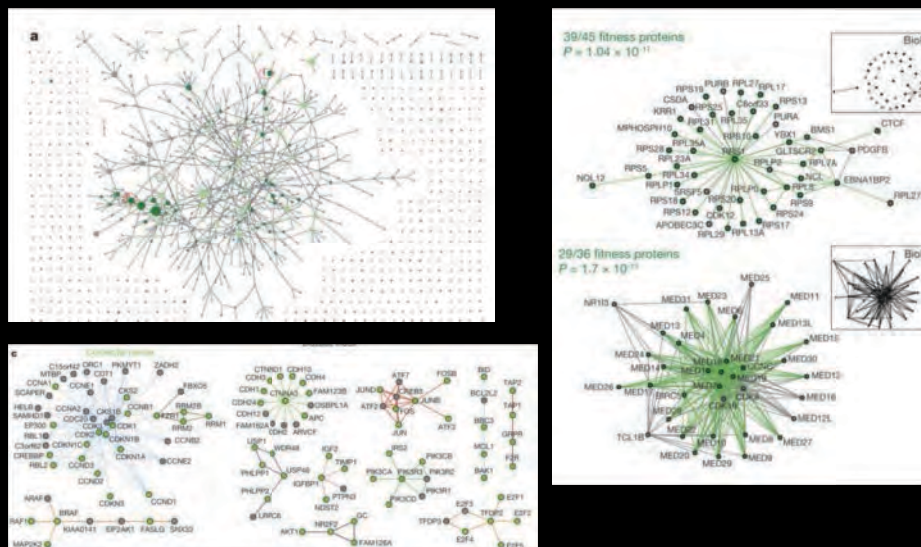
IntAct
 IntAct Molecular Interaction Database

Dataset of the month: March
 The CCA targeting complex is highly regulated and provides two distinct binding sites for client, non-substrate proteins.

<http://www.ebi.ac.uk/intact/>



Protein-protein interaction data set – BioPlex network



Huttlin et al, Nature 2017

www.cytoscape.org

The screenshot shows the homepage of the Cytoscape website. The page includes the following elements:

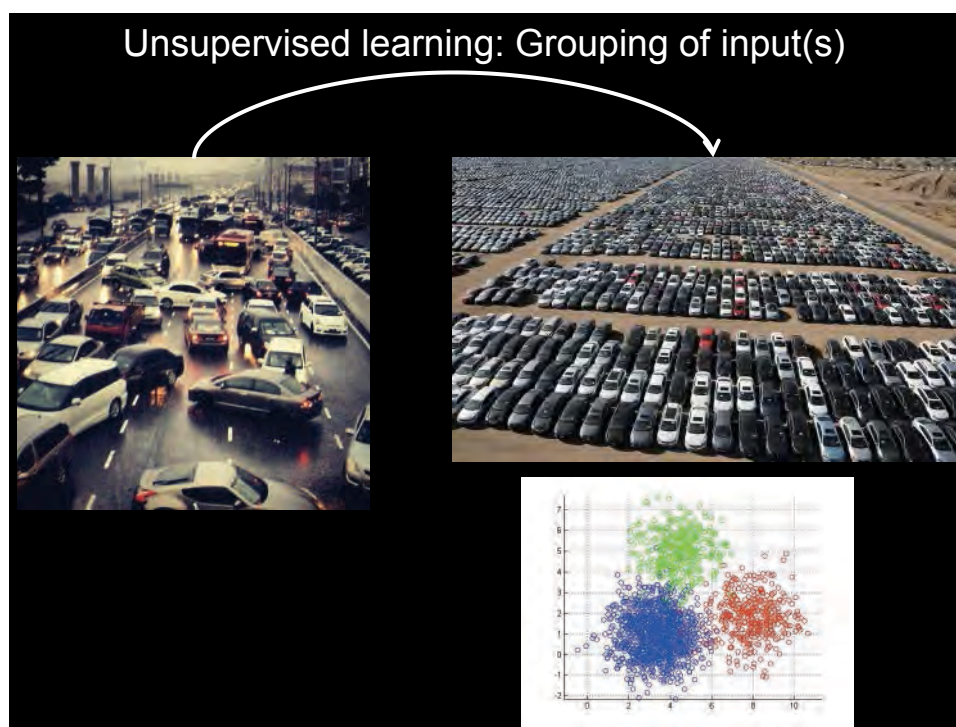
- Header:** The website URL www.cytoscape.org and navigation links for Home, Download, Apps, Documentation, Jobs, Community, and Report a Bug.
- Logo:** The Cytoscape logo, which is a stylized orange network graph.
- Tagline:** "Network Data Integration, Analysis, and Visualization in a Box".
- Buttons:** Two prominent buttons: "Introduction" and "Download 3.3.0".
- Background:** A dark background with a network graph visualization.

Machine Learning

- Supervised learning
- Unsupervised learning

Supervised learning: Classification of input(s)





Machine Learning 101: IRIS flower classification



<https://machinelearningmastery.com/machine-learning-in-python-step-by-step/>

The `caret` Package in R

General approach to investigate biological questions using a computational approach

WHY

1. Formulate the big question and have valid reasons

WHAT

1. Come up with several specific questions
2. Prioritise questions and prepare a checklist

HOW

1. Identify the database
2. Identify the tools
3. Be aware of the basic statistics
4. Retrieve and integrate the information

FRAME MORE WHY

1. Formulate hypothesis and READ A LOT!
2. Design experiments
3. Publish work & be happy ever after ☺