Regulatory Genomics III: Pathway Analysis

Shamith Samarajiwa
email: ss861@mrc-cu.cam.ac.uk

Integrative Systems Biomedicine Group,
MRC Cancer Unit,
University of Cambridge.

Computational Biology MPhil: Functional Genomics Lectures
18 Nov. 2016

UNIVERSITY OF CAMBRIDGE

MRC Cancer Unit

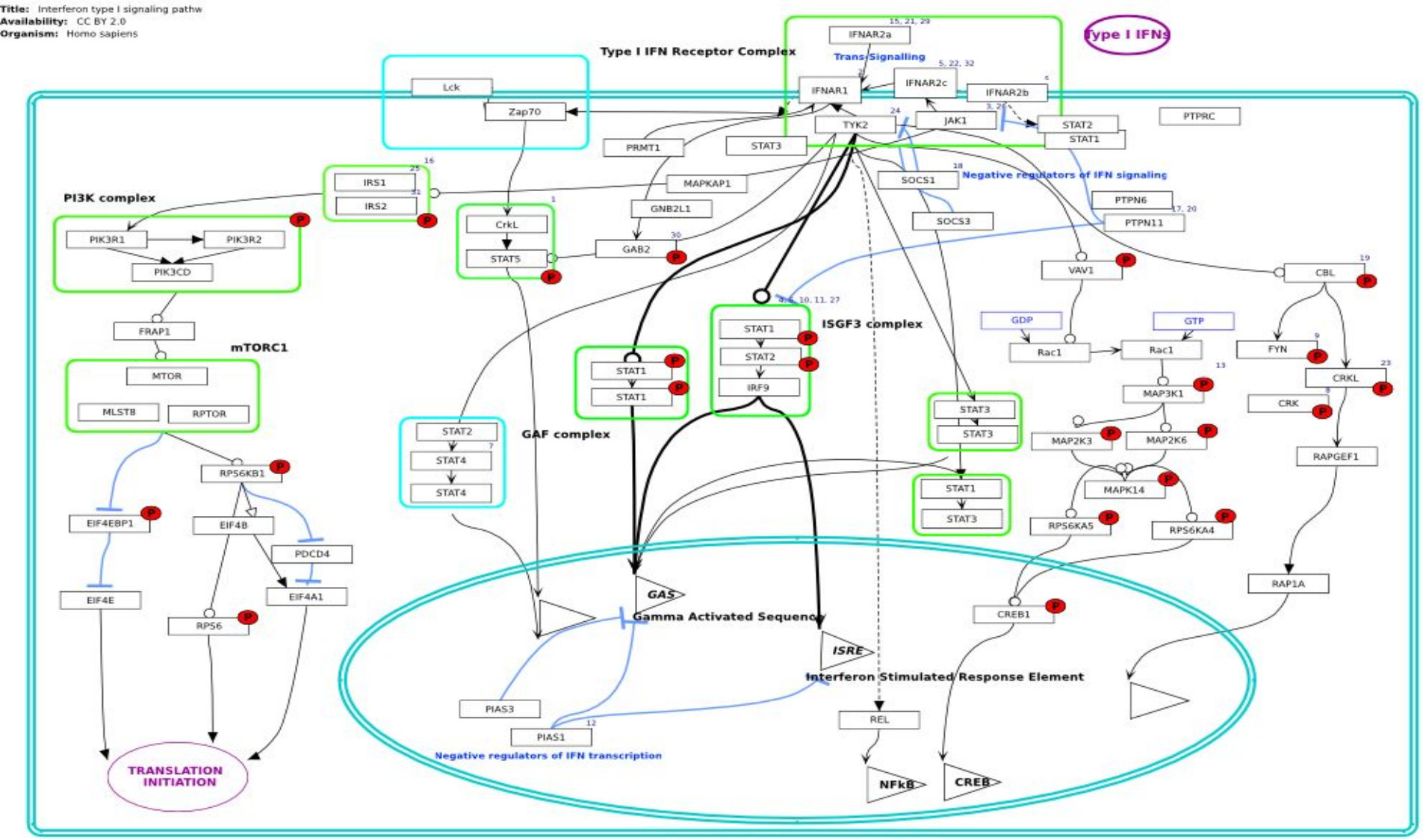
Overview

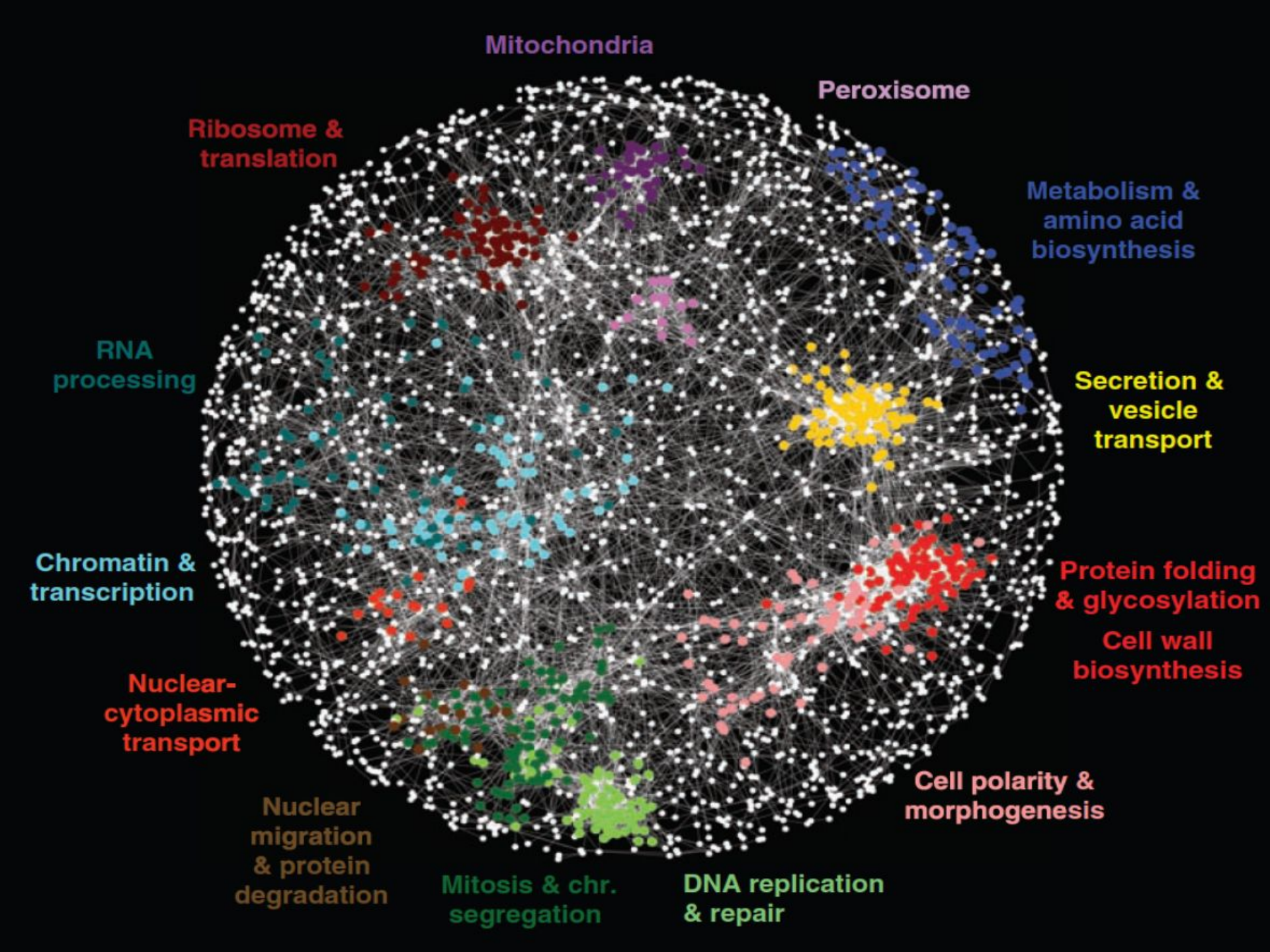
- What is a biological pathway?
- Statistical methods for pathway enrichment analysis
- Pitfalls of enrichment analysis
- Gene name disambiguation
- Gene Ontology
- Pathway
- GSEA
- Interactomes

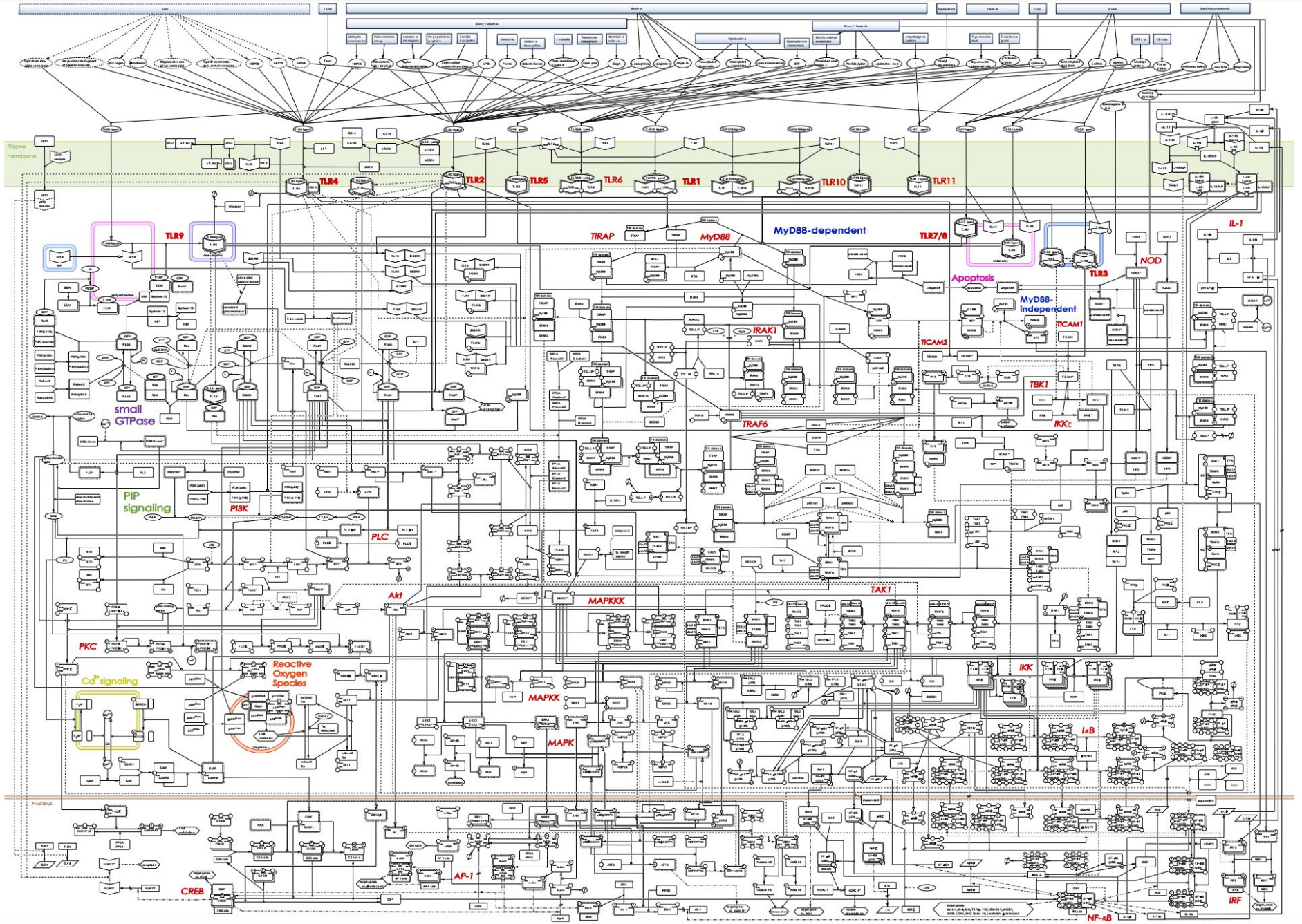
What are biological pathways?

- Biological Pathways are a series of consecutive interactions between biochemical molecules acting in concert to maintain and control of cellular information flow, energy and biochemical compounds in the cell leading to a change in molecular products or a cellular states.
- What's known as "pathways" are usually parts of more complex networks.
- Information flow in pathways have directionality.
- Feedforward and feedback loops, negative and positive regulation.
- There are different types of pathways:
  - Signalling
  - Genetic / Transcriptional / Regulatory
  - Metabolic (anabolic, catabolic, transport and energy)

What is a pathway?




Mitochondria
Peroxisome
Ribosome & translation
Metabolism & amino acid biosynthesis
RNA processing
Secretion & vesicle transport
Chromatin & transcription
Protein folding & glycosylation
Cell wall biosynthesis
Nuclear-cytoplasmic transport
Cell polarity & morphogenesis
Nuclear migration & protein degradation
Mitosis & chr. segregation
DNA replication & repair



Oda and Kitano, Mol. Sys. Biol., 2006

Enrichment tests

Enrichment of gene lists:

- Fisher's exact test / Hypergeometric test
- Chi-squared
- Binomial

Enrichment of ranked lists:

- Kolmogorov-Smirnov (implemented in GSEA)
- Minimum hypergeometric test
- Wilcoxon Rank Sum

Types of enrichment analysis

- Gene Lists
  - Are any gene sets significantly enriched or depleted in my list?
  - Differential enrichment between two lists
  - Fisher's exact test

- Ranked list of genes
  - Are any gene sets ranked significantly high or low in my ranked list of genes?
  - minHG

- Where do the gene list come from?
  - Omics data: Transcriptomics, Proteomics, ChIP-seq etc.
- How do we assess significance?
- Compare to a background set
- Correct for multiple testing

Enrichment of ranked lists

- Arbitrary thresholds
- Possible problems with gene lists
  - No “natural” value for filter thresholds (Fold Change >= 2 fold adj.p-value <= 0.01 etc.)
  - Results change with different threshold settings
  - Results change with background set
  - Loss of statistical power due to thresholding (no resolution of significant signal with different strengths, weak signals neglected)

Comparison to a background set

- All possible genes that could appear in your gene list
  - If using high throughput sequencing technology use all genes
  - ~21000 (20, 441 in Ensembl 86) protein coding genes
  - ~22 000 (22 219 in Ensembl 86) non coding genes
  - If using microarrays - the “universe of genes” limited by the number of probes on platform.

- Annotation sources
  - RefSeq, Gencode, Ensembl
  - Some annotation terms may be subsets of other terms (GO ontology)

Sampling Bias

Multiple sources of bias confound functional enrichment analysis of global -omics data

James A. Timmons, Krzysztof J. Szkop and Iain J. Gallagher

Genome Biology 2015 16:186 | DOI: 10.1186/s13059-015-0761-7 | © Timmons et al. 2015

Published: 7 September 2015

- Technology bias
- Detection bias: In a transcriptomic experiment not all genes can be detected with equal reliability, to the extent that some genes are never detected as being 'regulated'.
- Biological bias: The transcriptome of a given cell type or tissue is highly specialized, to the point that it can be used to determine the identity of an unknown RNA profile efficiently.

Some useful advice

Nature Reviews Genetics **9**, 509-515 (July 2008) | doi:10.1038/nrg2363

Use and misuse of the gene ontology annotations

Seung Yon Rhee[1], Valerie Wood[2], Kara Dolinski[3] & Sorin Draghici[4] About the authors

An introduction to effective use of enrichment analysis software

Hannah Tipney and Lawrence Hunter*

Center for Computational Pharmacology, University of Colorado Denver, Aurora, CO 80045, USA
**Correspondence to*: Tel: +1 303 724 3369; E-mail: hannah.tipney@ucdenver.edu

JOURNAL OF THE ROYAL SOCIETY Interface

Separate enrichment analysis of pathways for up- and downregulated genes

Guini Hong[1], Wenjing Zhang[1], Hongdong Li[1], Xiaopei Shen[1] and Zheng Guo[1,2]

Gene pairs with various types of functional links defined in pathways tend to have positively correlated expression levels.

Gene name disambiguation

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg[†], Joseph Riss[†], David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barrett and John N Weinstein

[†] Contributed equally

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren and Assam El-Osta

Genome Biology 2016 17:177 | DOI: 10.1186/s13059-016-1044-7 | © The Author(s). 2016

Published: 23 August 2016

Getting the latest gene nomenclature

HGNC
HUGO Gene Nomenclature Committee

Search everything | Search symbols, keywords or IDs

Use * to search with a root symbol (eg ZNF*)

Home | Downloads | Gene Families | Tools | Useful links | About | Newsletters | Contact Us | Help | VGNC | Request Symbol

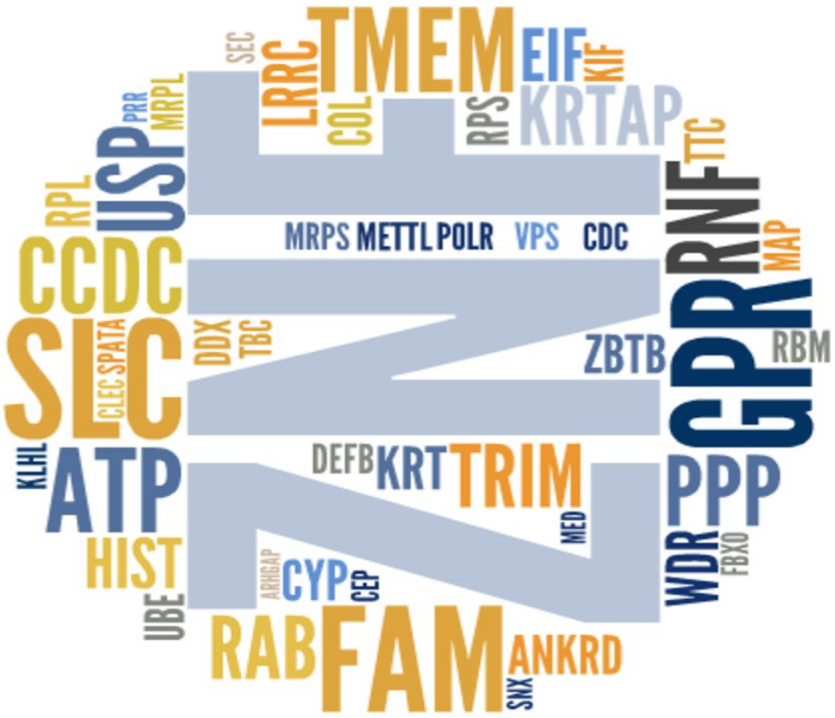
HGNC is responsible for approving unique symbols and names for human loci, including protein coding genes, ncRNA genes and pseudogenes, to allow unambiguous scientific communication.

genenames.org is a curated online repository of HGNC-approved gene nomenclature, gene families and associated resources including links to genomic, proteomic and phenotypic information.

Search our catalogue of more than 40,000 symbol reports using our improved search engine (see Search help), search lists of symbols using our Multi-symbol checker and identify possible orthologs using our HCOP tool.

Download our ready-made data files from our Statistics and Downloads page, create your own datasets using either our Custom Downloads tool or BioMart service, or write a script/program utilising our REST service.

Submit your gene symbol and name proposals to us to be accredited with HGNC approved nomenclature for use in publications, databases and presentations.



Extracting meaning from biomedical data

- Gene Ontology enrichment
- Pathway enrichment
- Gene Set Enrichment Analysis
- Interactome analysis
- Networks
- Utilizing natural language processing

Gene Ontology

The Gene Ontology project provides an ontology (a machine readable controlled vocabulary) of defined terms representing properties of gene products. The ontology covers three domains:

- CC- **cellular component**, the parts of a cell or its extracellular environment;
- MF- **molecular function**, the elemental activities of a gene product at the molecular level, such as binding or catalysis;
- BP- **biological process**, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

Each domain has 9 levels (generic to specific) organised as a directed di-acyclic graph.

A collection of GO resources

- NIH DAVID
- AMIGO (GO consortium)
- WebGestalt
- BINGO (Cytoscape)
- GOrilla and Revigo
- topGO (Bioconductor)
- STEM (short time series expression miner)

DAVID Bioinformatics Resources 6.8

National Institute of Allergy and Infectious Diseases (NIAID), NIH

*** 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 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

Welcome to DAVID 6.8

2003 - 2016

Search

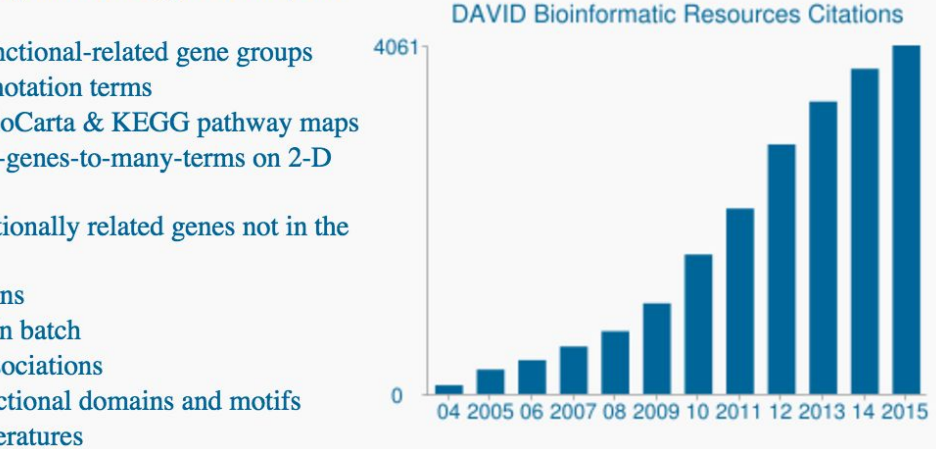
The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand 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
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.
- And more

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



- > 21,000 Citations
- Average Daily Usage: ~2,600 gene lists/sublists from ~800 unique researchers.

Amigo

AmiGO 2 | Home | Search | Browse | Tools & Resources | Help | Feedback | About | AmiGO 1.8

AmiGO 2

More information on quick search

Quick search

Search

Search Templates

Use predefined **templates** to explore Gene Ontology data.

Go »

Advanced Search

Interactively **search** the Gene Ontology data for annotations, gene products, and terms using a powerful search syntax and filters.

Search

Browse the Ontology

Use the drill-down **browser** to view the ontology structure with annotation counts.

Go »

GOOSE

Use GOOSE to query the legacy GO database with **SQL**.

Go »

Term Enrichment Service

Your genes here...

biological process

Homo sapiens

Statistics

View the most recent **statistics** about the Gene Ontology data in AmiGO.

Go »

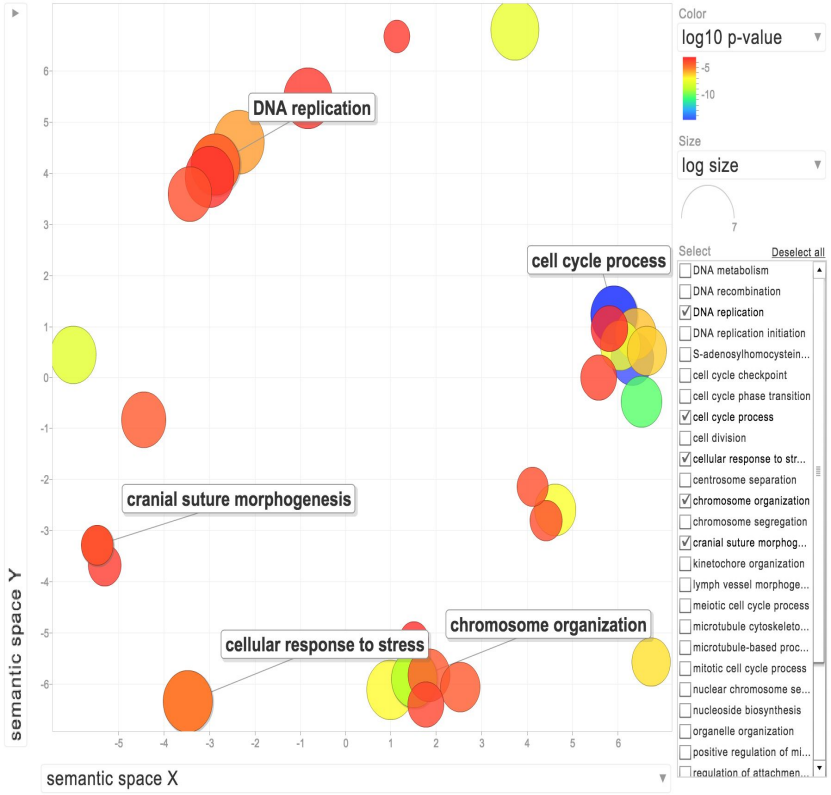
And Much More...

Many **more tools** are available from the software list, such as alternate searching modes, Visualize, non-JavaScript pages.

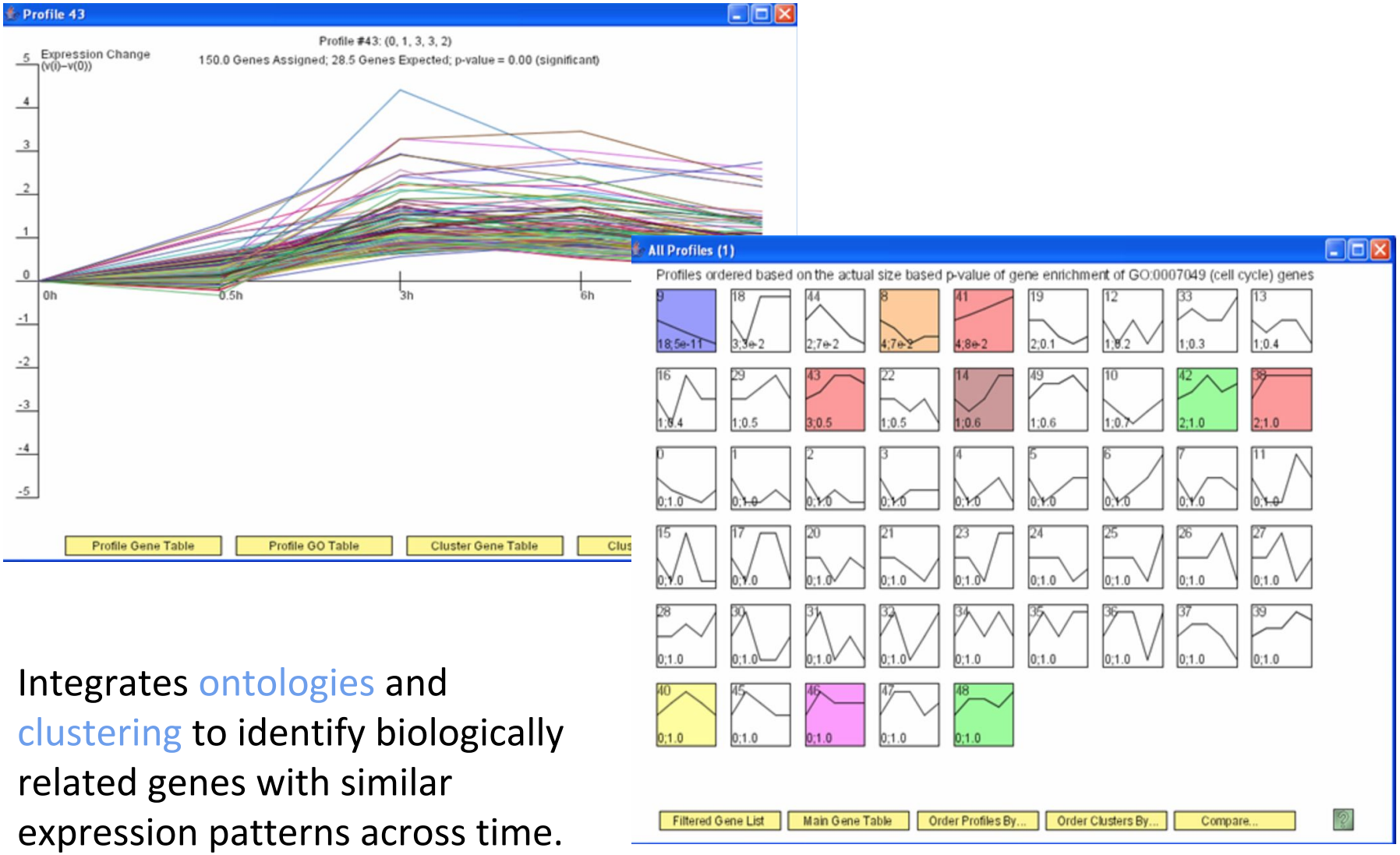
Go »

Build your own annotated gene sets

GOrilla and Revigo



STEM: Short Time Series Expression Miner



Integrates ontologies and clustering to identify biologically related genes with similar expression patterns across time.

A collection of pathway analysis resources

- KEGG
- Reactome
- Panther
- BioCyc
- Wikipathways
- Pathway Commons
- Pathguide
- Ingenuity
- Metacore
- Pathway studio

OPEN ACCESS Freely available online

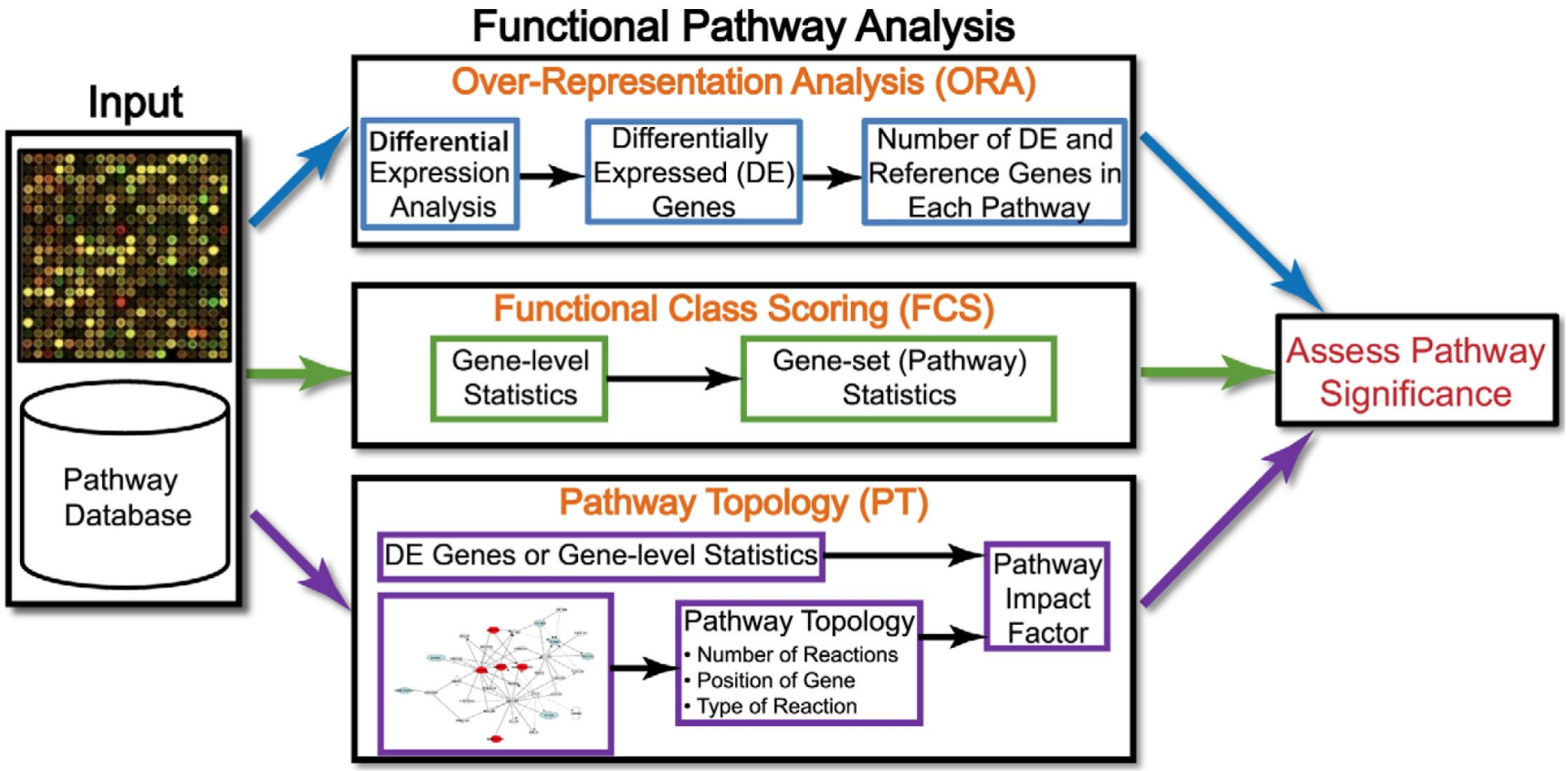
PLoS COMPUTATIONAL BIOLOGY

Review

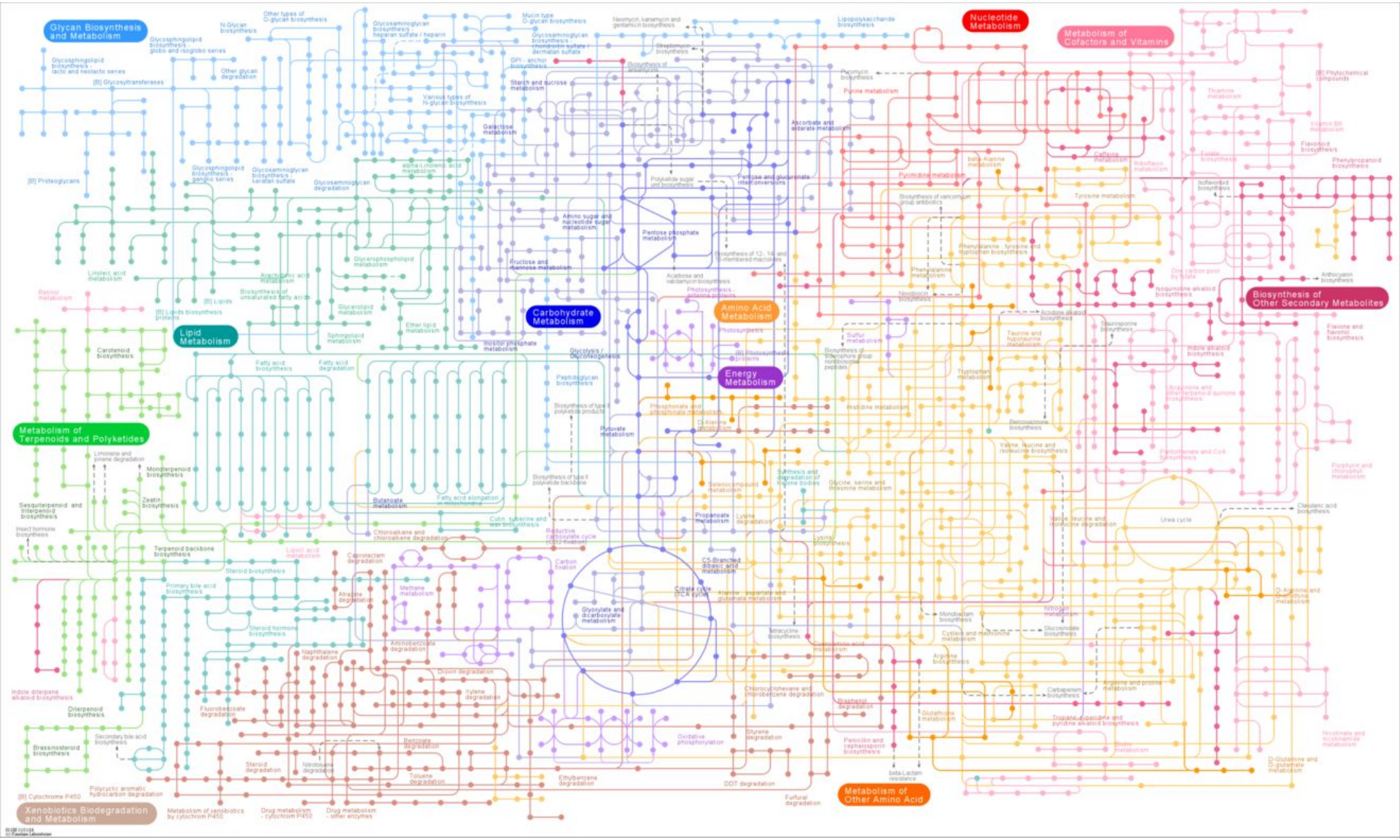
Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges

Purvesh Khatri[1,2]*, Marina Sirota[1,2], Atul J. Butte[1,2]*

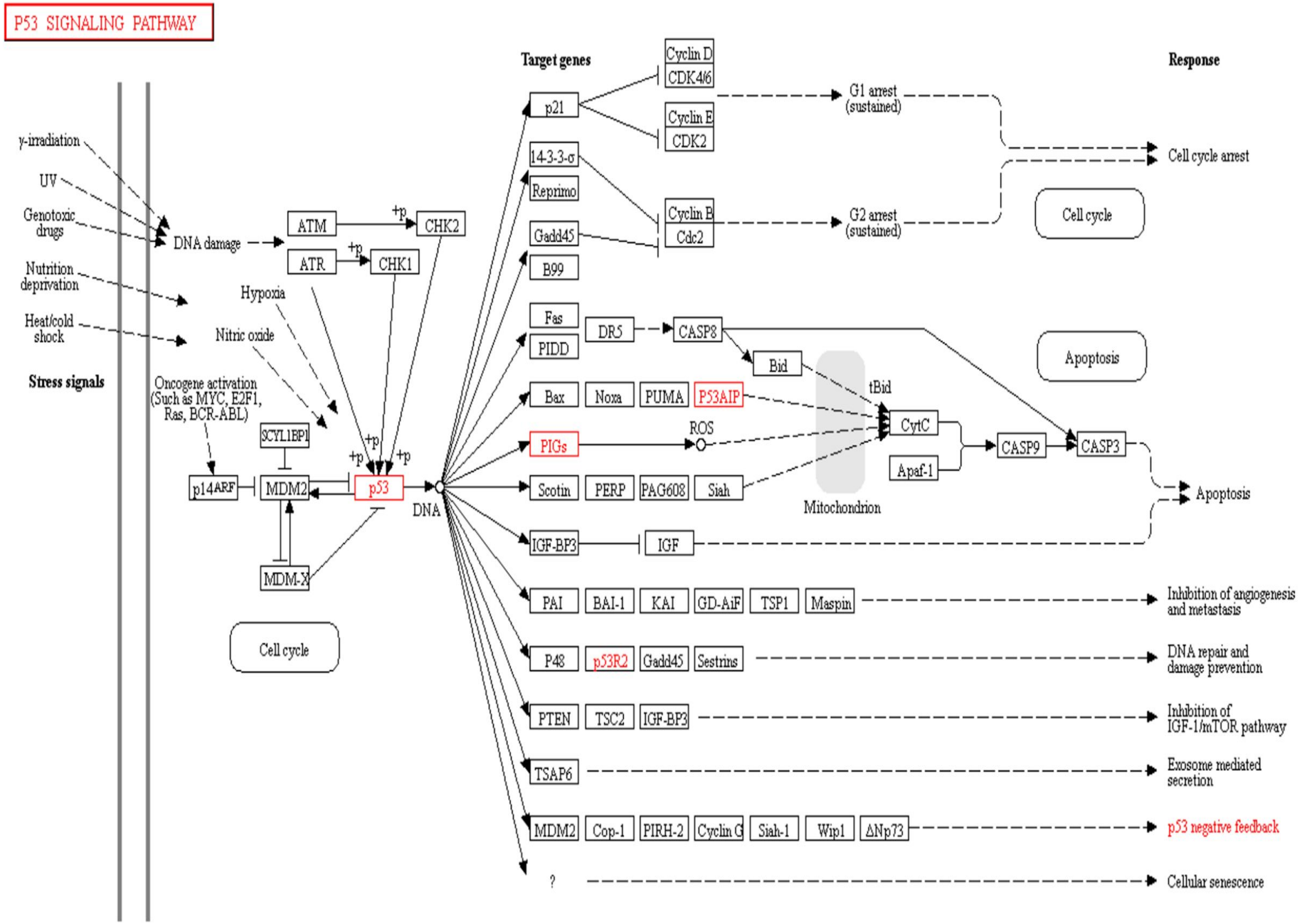
1 Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, United States of America, 2 Lucile Packard Children's Hospital, Palo Alto, California, United States of America



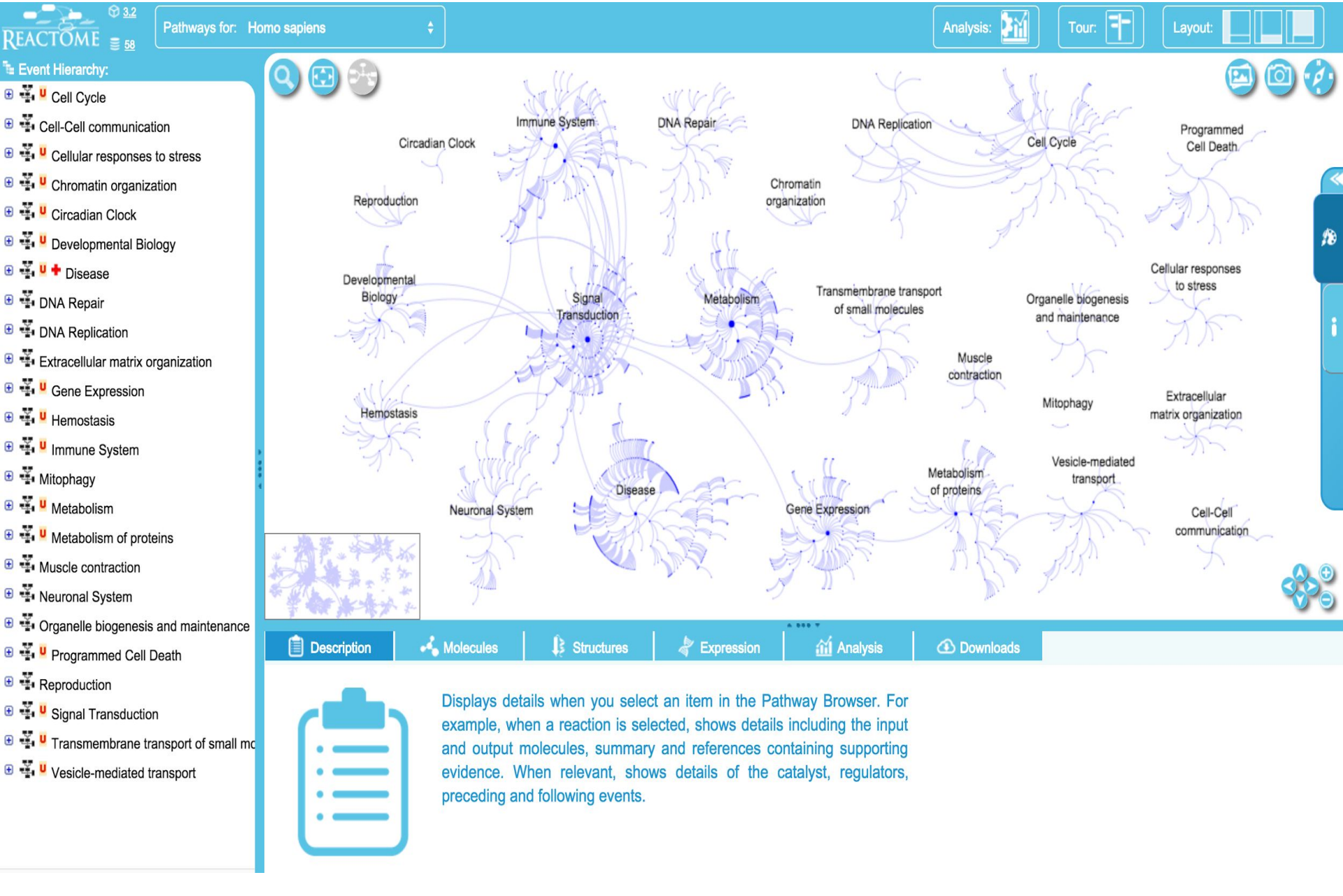
KEGG: Kyoto Encyclopedia of Genes and genomes



Global Metabolic Map



Reactome


REACTOME
3.2
58
Pathways for: Homo sapiens
Analysis:
Tour:
Layout:
Event Hierarchy:
Cell Cycle
Cell-Cell communication
Cellular responses to stress
Chromatin organization
Circadian Clock
Developmental Biology
Disease
DNA Repair
DNA Replication
Extracellular matrix organization
Gene Expression
Hemostasis
Immune System
Mitophagy
Metabolism
Metabolism of proteins
Muscle contraction
Neuronal System
Organelle biogenesis and maintenance
Programmed Cell Death
Reproduction
Signal Transduction
Transmembrane transport of small mc
Vesicle-mediated transport
Immune System
DNA Repair
DNA Replication
Cell Cycle
Programmed Cell Death
Circadian Clock
Chromatin organization
Reproduction
Developmental Biology
Signal Transduction
Metabolism
Transmembrane transport of small molecules
Organelle biogenesis and maintenance
Cellular responses to stress
Muscle contraction
Mitophagy
Extracellular matrix organization
Hemostasis
Disease
Metabolism of proteins
Vesicle-mediated transport
Neuronal System
Gene Expression
Cell-Cell communication
Description
Molecules
Structures
Expression
Analysis
Downloads
Displays details when you select an item in the Pathway Browser. For example, when a reaction is selected, shows details including the input and output molecules, summary and references containing supporting evidence. When relevant, shows details of the catalyst, regulators, preceding and following events.

Pathguide » the pathway resource list

Navigation
- Protein-Protein Interactions
- Metabolic Pathways
- Signaling Pathways
- Pathway Diagrams
- Transcription Factors / Gene Regulatory Networks
- Protein-Compound Interactions
- Genetic Interaction Networks
- Protein Sequence Focused
- Other

Search
Organisms: All
Availability: All
Standards: All
Reset | Search

Analysis
- Statistics
- Database Interactions

Contact
- Comments, Questions, Suggestions are Always Welcome!

Complete Listing of All Pathguide Resources

Pathguide contains information about **547** biological pathway related resources and molecular interaction related resources. Click on a link to go to the resource home page or 'Details' for a description page. Databases that are free and those supporting BioPAX, CellML, PSI-MI or SBML standards are respectively indicated.

If you know of a pathway resource that is not listed here, or have other questions or comments, please send us an e-mail.

News

Major new update of Pathguide August 2013
We now have information about ~550 resources!.

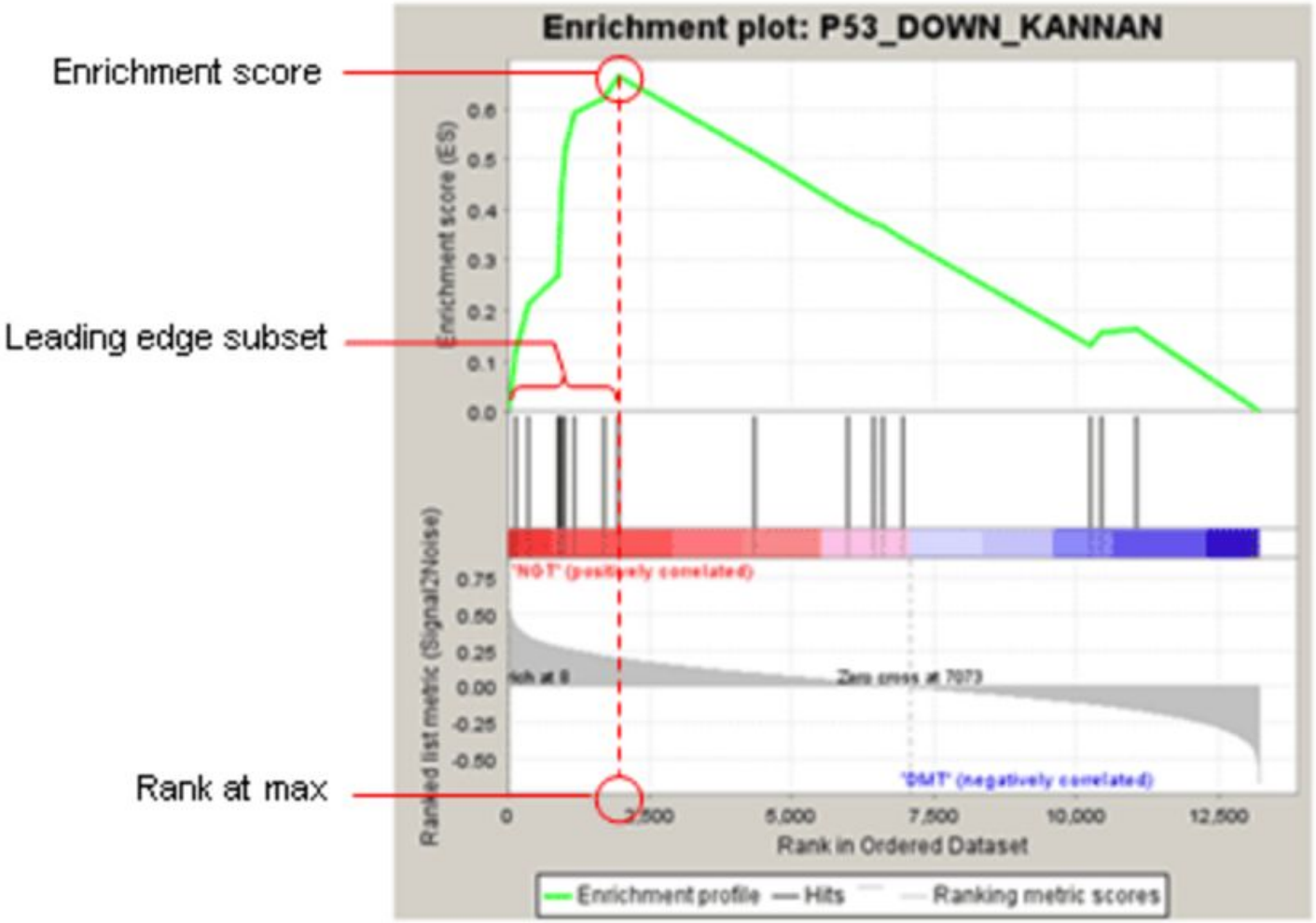
Visual navigation added May 2010
Click the 'Database interactions' link on the left menu to access.

Protein-Protein Interactions

Database Name (Order: alphabetically | by web popularity)	Full Record	Availability	Standards
2P2Idb - The Protein-Protein Interaction Inhibition Database	Details	Free	
3D-Interologs - 3D-Interologs	Details	Free	
3DID - 3D interacting domains	Details	Free	
ADAN - Prediction of protein-protein interaction of modular domains	Details	X	
AHD2.0 - Arabidopsis Hormone Database 2.0	Details	Free	
AllFuse - Functional Associations of Proteins in Complete Genomes	Details	X	
aMAZE - Protein Function and Biochemical Pathways Project	Details	X	
ANAP - Arabidopsis Network Analysis Pipeline	Details	Free	
AnimalTFDB - Animal Transcription Factor Database	Details	Free	
AntiJen - AntiJen a Kinetic, Thermodynamic and Cellular Database	Details	Free	
APID - Agile Protein Interaction DataAnalyzer	Details	Free	
AS-ALPS - Alternative Splicing - induced ALteration of Protein Structure	Details	Free	
ASD - Allosteric Database	Details	Free	
ASEdb - Alanine Scanning Energetics Database	Details	Free	
ASPD - Artificial Selected Proteins/Peptides Database	Details	Free	
ATDB - Animal Toxin Database	Details	Free	
AtPID - Arabidopsis thaliana Protein Interactome Database	Details	Free	
AtPIN - Arabidopsis thaliana Protein Interactome Network	Details	Free	
Bacteriome.org - Bacterial Protein Interaction Database for Escherichia Coli	Details	Free	
BIANA - Biologic Interaction and Network Analysis	Details	Free	
BID - Binding Interface Database	Details	Free	

Gene Set Enrichment Analysis (GSEA)

- Enrichment Score (ES), which reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes.



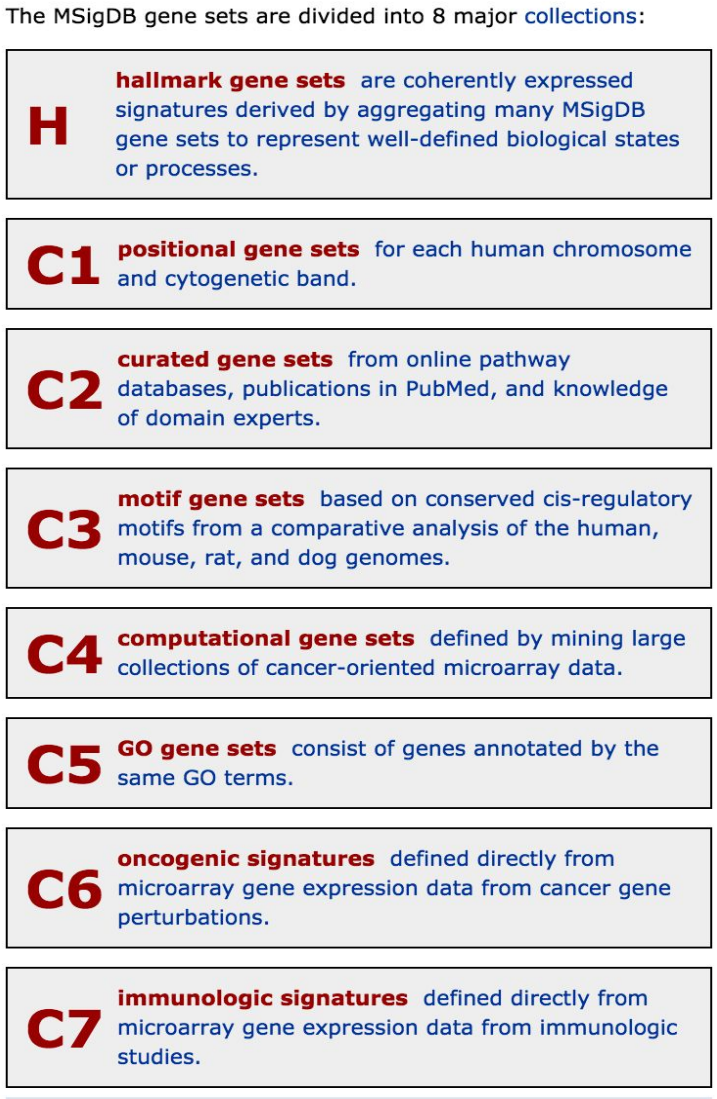
Interpreting GSEA

- GSEA calculates the ES by walking down the ranked list of genes, increasing a running-sum statistic when a gene is in the gene set and decreasing it when it is not.
- The magnitude of the increment depends on the correlation of the gene with the phenotype. The ES is the maximum deviation from zero encountered in walking the list. A positive ES indicates gene set enrichment at the top of the ranked list; a negative ES indicates gene set enrichment at the bottom of the ranked list.
- Gene sets with a distinct peak at the beginning or end of the ranked list are generally the most interesting.
- The **leading edge subset** of a gene set is the subset of members that contribute most to the ES.
- Use fdr of 25%

MSigDB

- The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software.
- GMT file format
- Used together with BROAD institute GSEA java package or web app.

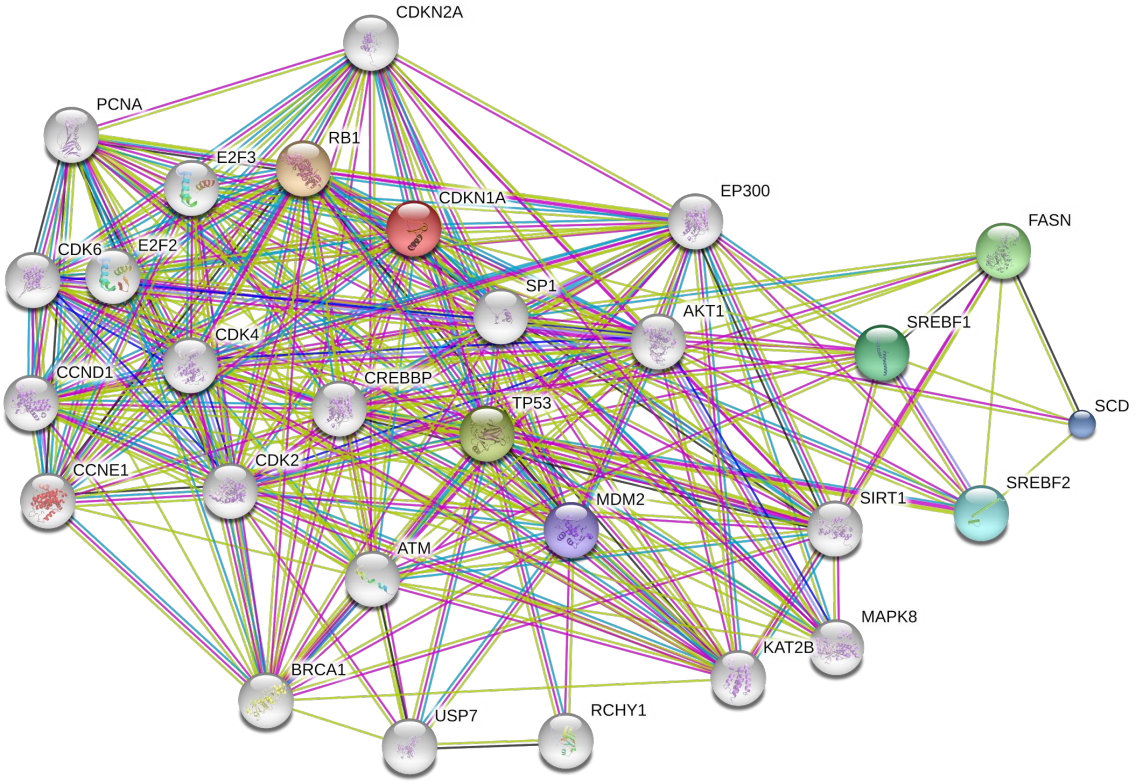
- Can also be used with GSEABase and EGSEA Bioconductor analysis tools.



Exploring Connectivity

- Macromolecules (DNA, metabolites, other entities) and Gene products (Proteins, Non-coding RNA) interact in specific ways to regulate phenotypes.
- Evolutionarily defines and functional properties such as Protein-Protein interactions, subcellular localization, Co-Expression, Protein domains, Post-translational modification etc can be used for determining connectivity.
- Tools to explore connectivity can help us understand complex biological processes.
    - STRING
    - GeneMania
    - Cytoscape

STRING

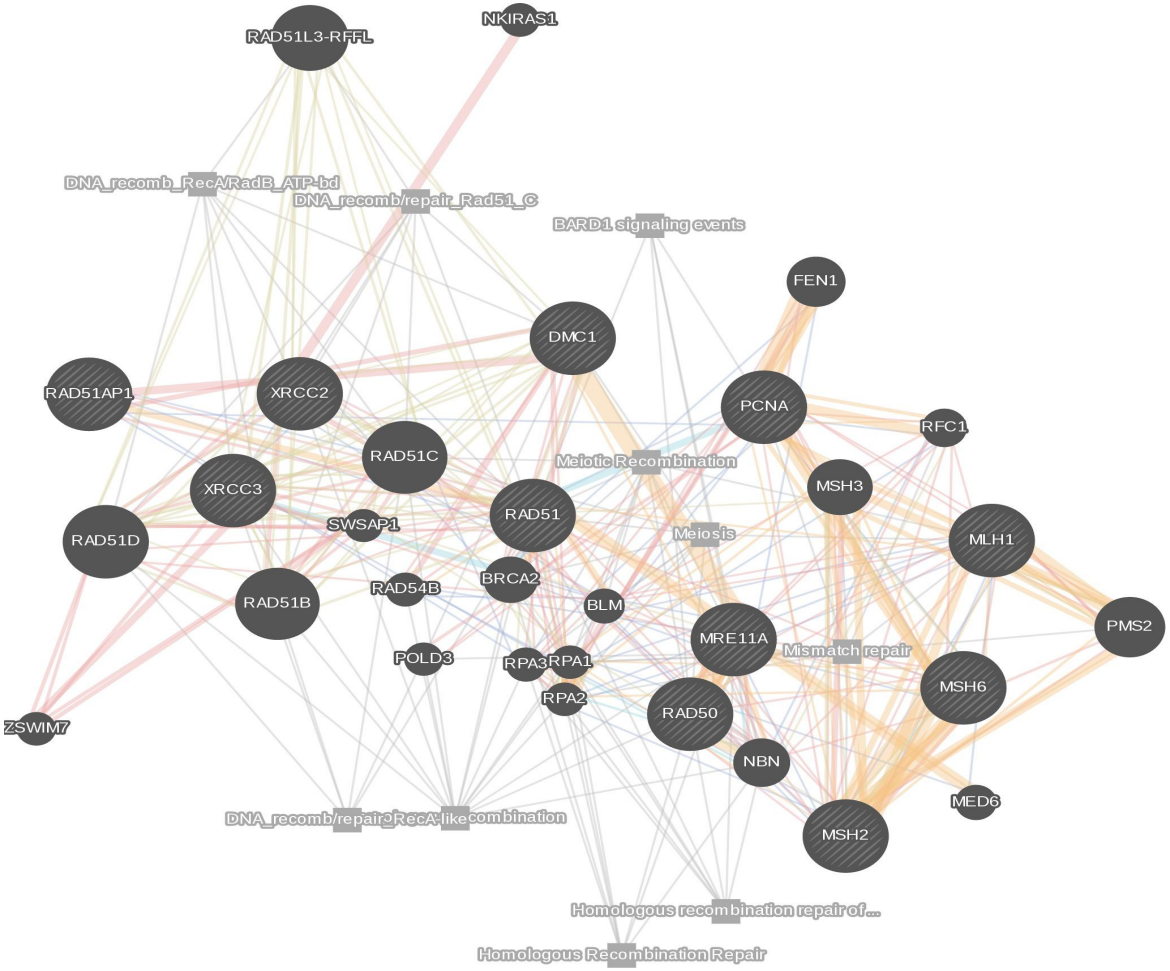


- Web app and STRINGdb -bioconductor package

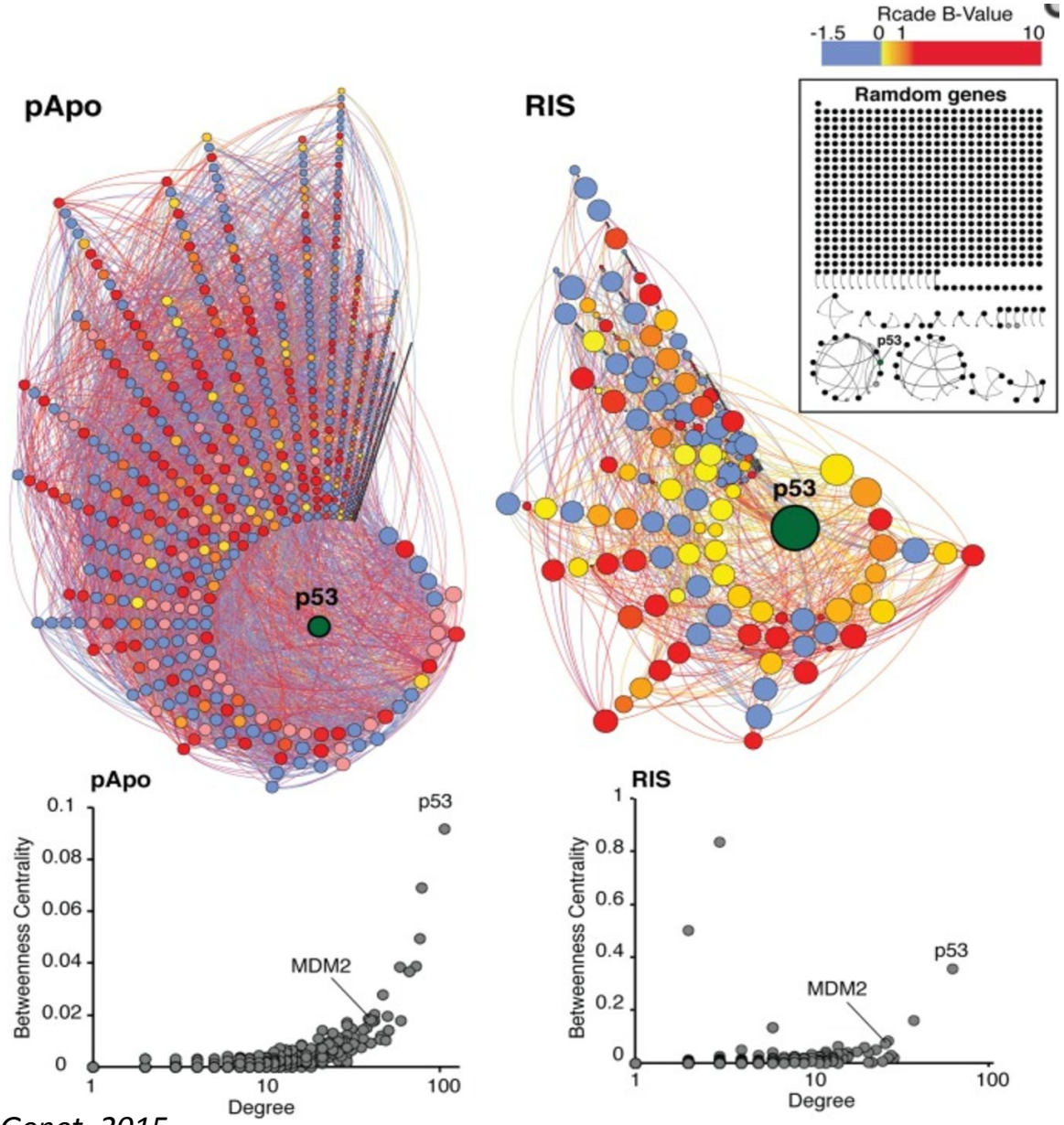
- Critical assessment and integration of protein–protein interactions, including direct (physical) as well as indirect (functional) associations. The new version 10.0 of STRING covers more than 2000 organisms, which has necessitated novel, scalable algorithms for transferring interaction information between organisms.

GeneMania

- Web and Cytoscape apps.
- Utilizes connectivity of 1500 external genomic and proteomic datasets.
- Semi-supervised learning algorithm.



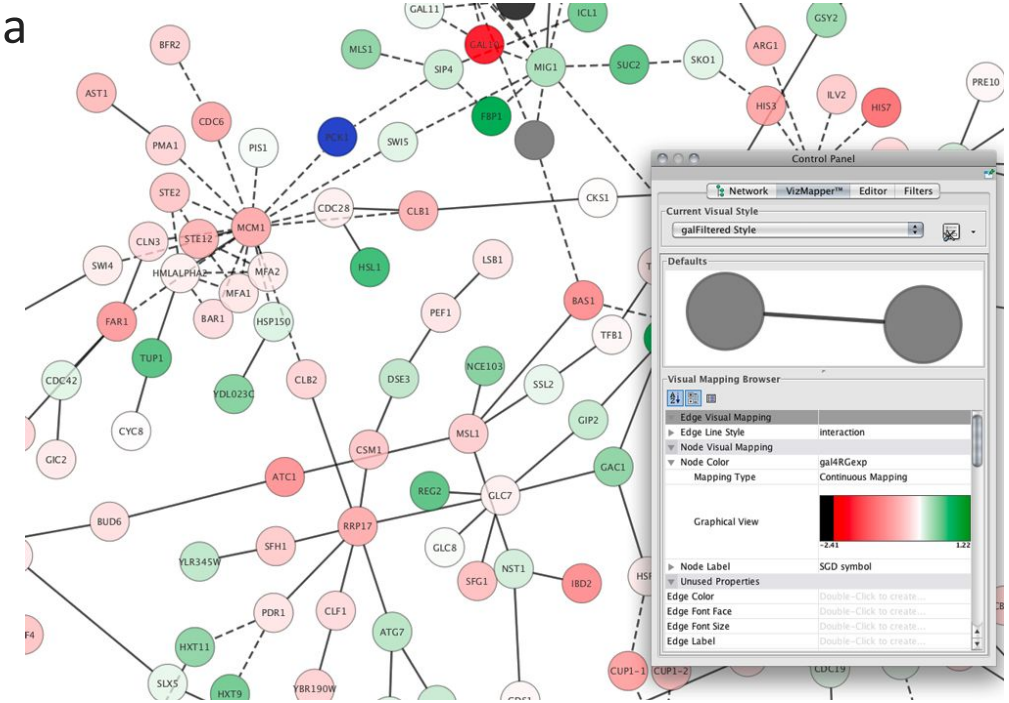
Network topology of p53 targets



Samarajiwa et al., PLoS Genet. 2015

Cytoscape

- *Cytoscape* is an open source software platform for *visualizing* molecular interaction networks and biological pathways and *integrating* these networks with annotations, gene expression profiles and other state data. Although Cytoscape was originally designed for biological research, now it is a general platform for complex network analysis and visualization.
- Large number of analysis apps and algorithms.
- Network topology analysis methods.
- Many layout a



References

1. Szklarczyk et al., STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015 Jan;43(Database issue):D447-52. 1
2. Shannon P et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks Gen. Res. 2003 Nov; 13(11):2498-504
3. Montojo et al., GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics. 2010 Nov 15;26(22):2927-8.