

# Canadian Bioinformatics Workshops

[www.bioinformatics.ca](http://www.bioinformatics.ca)

This page is available in the following languages:

Afrikaans বাংলা Azərbaycanca Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto Español Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE) Euskara Suomi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Malayu Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska 中文 華語 (台灣) isiZulu



## Attribution-Share Alike 2.5 Canada

### You are free:



**to Share** — to copy, distribute and transmit the work



**to Remix** — to adapt the work



### Under the following conditions:



**Attribution.** You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



**Share Alike.** If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full licence) available in the following languages:  
[English](#) [French](#)

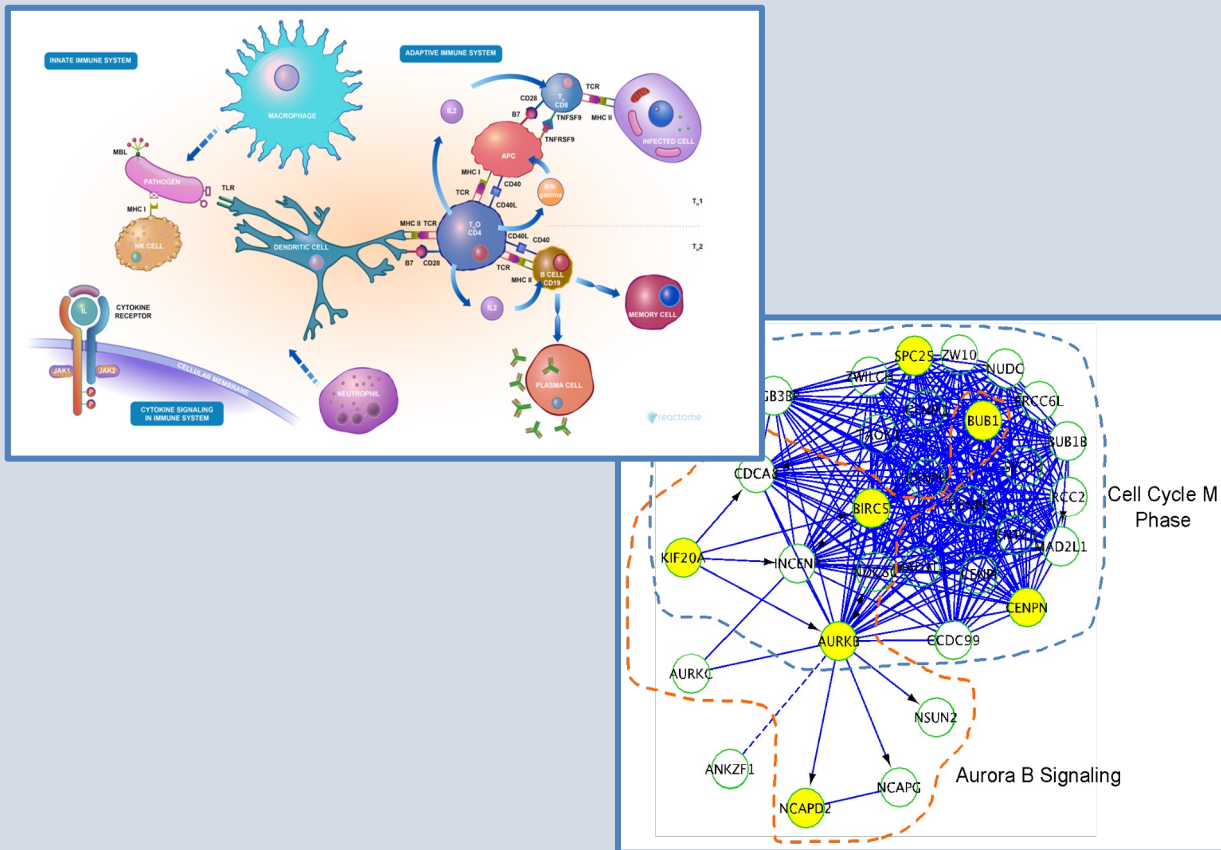
# Module 4

## More Depth on Pathway & Network Analysis

Lincoln Stein

Pathway and Network Analysis of –Omics Data

June 26, 2024



# Learning Objectives of Module

- Understand the principles of pathway and network analysis.
  - Sources of pathway and network data.
  - Analytical approaches to data analysis, visualization and integration.
  - Applications of pathway enrichment analysis.

# Why Pathway Analysis?

- Dramatic data size reduction: 1000's of genes => dozens of pathways.
- Increase statistical power by reducing multiple hypotheses.
- Find meaning in the “long tail” of rare cancer mutations.
- Tell biological stories:
  - Identifying hidden patterns in gene lists.
  - Creating mechanistic models to explain experimental observations.
  - Predicting the function of unannotated genes.
  - Establishing the framework for quantitative modeling.
  - Assisting in the development of molecular signatures.

# What is Pathway/Network Analysis?

- Any analytic technique that makes use of biological pathway or molecular network information to gain insights into a tumor or other biological system.
- A rapidly evolving field.
- Many approaches.

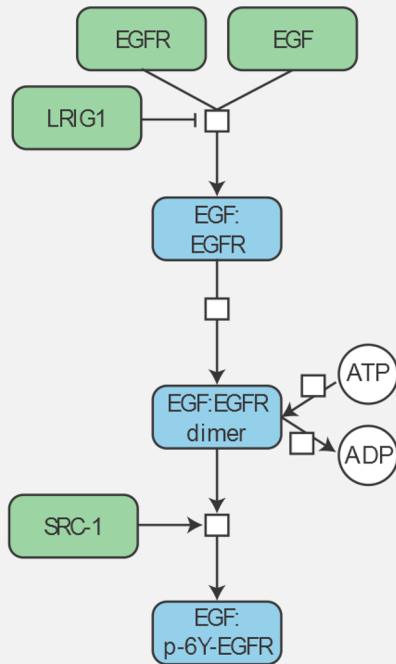
# Ingredients you will Need

1. High-throughput biological data: A list of altered genes, proteins, RNAs, etc.
2. A source of pathways or networks.

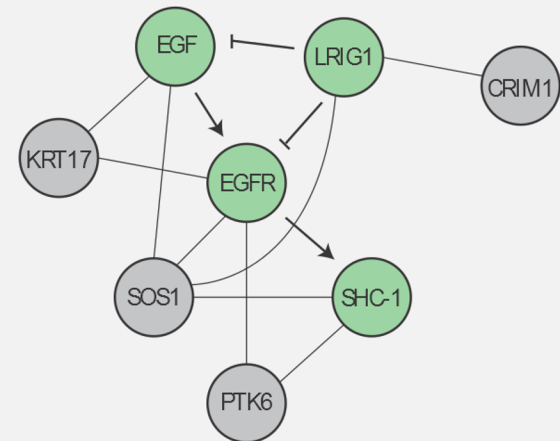


# Pathways vs Networks

EGFR-centered  
Pathway



EGFR-centered  
Network



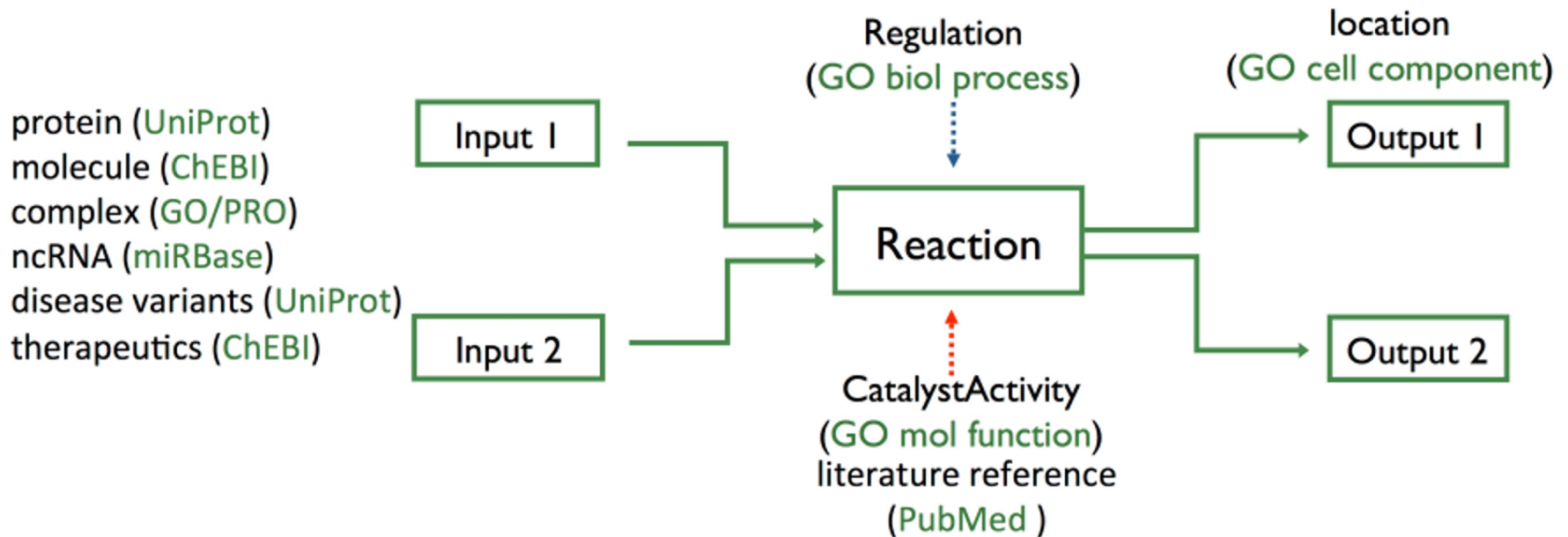


# Pathway Databases

- Advantages:
  - Usually curated.
  - Biochemical view of biological processes.
  - Cause and effect captured.
  - Human-interpretable visualizations.
- Disadvantages:
  - Sparse coverage of genome.
  - Different databases disagree on boundaries of pathways.

# Reaction-Network Databases

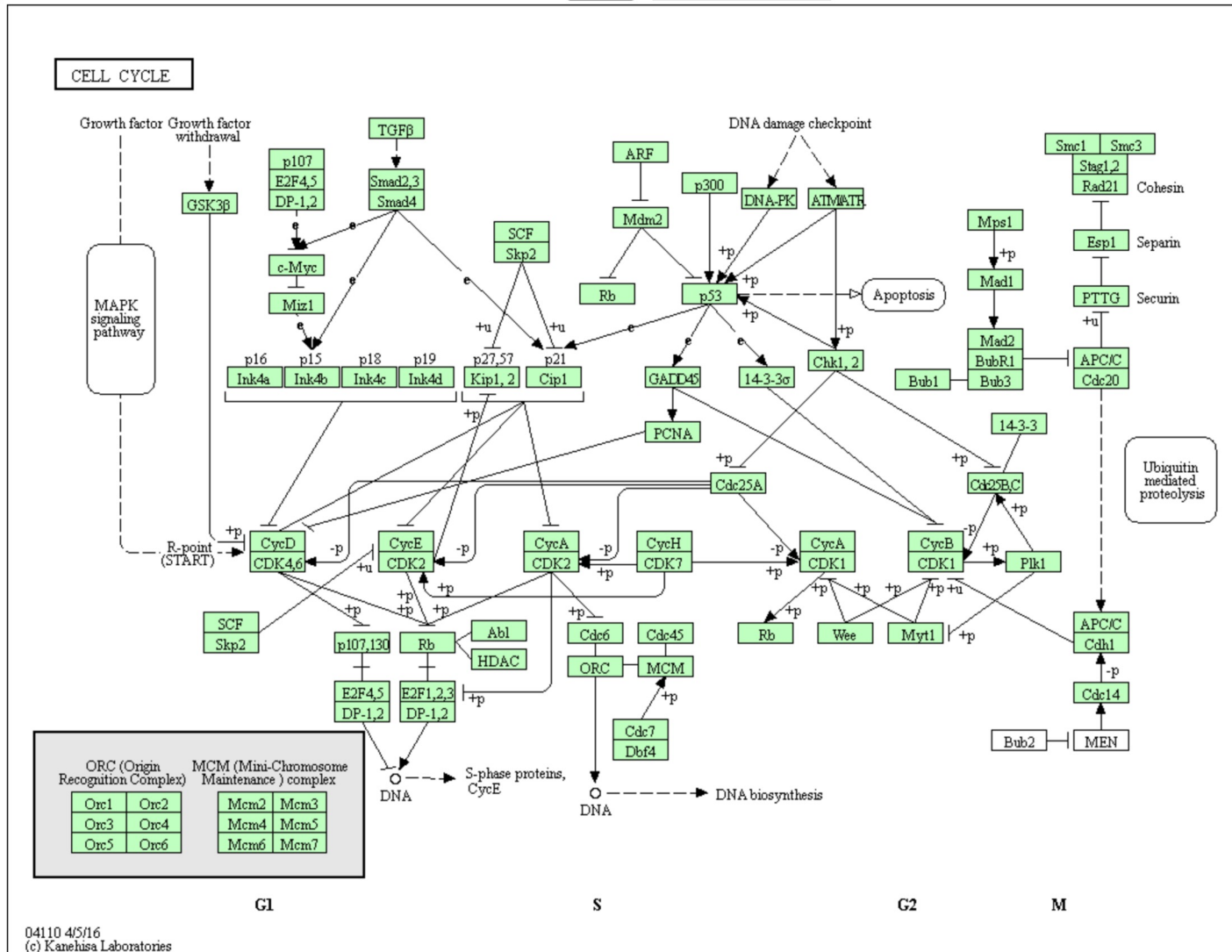
- Reactome & KEGG
  - explicitly describe biological processes as a series of biochemical reactions.
  - represents many events and states found in biology.



# KEGG

- Kyoto Encyclopedia of Genes and Genomes(KEGG):
  - A vast library of information = fully sequenced genomes, genes, proteins, pathways, and chemical compounds pertaining to over a hundred different species of both prokaryotes and eukaryotes.
  - KEGG PATHWAY is a collection of manually drawn pathway maps representing knowledge on the molecular interaction and reaction networks for Metabolism, Cellular Processes, Organismal Systems, Human Diseases and Drug Development
- Subscription required for access to underlying data for analysis use.

# KEGG Cell Cycle



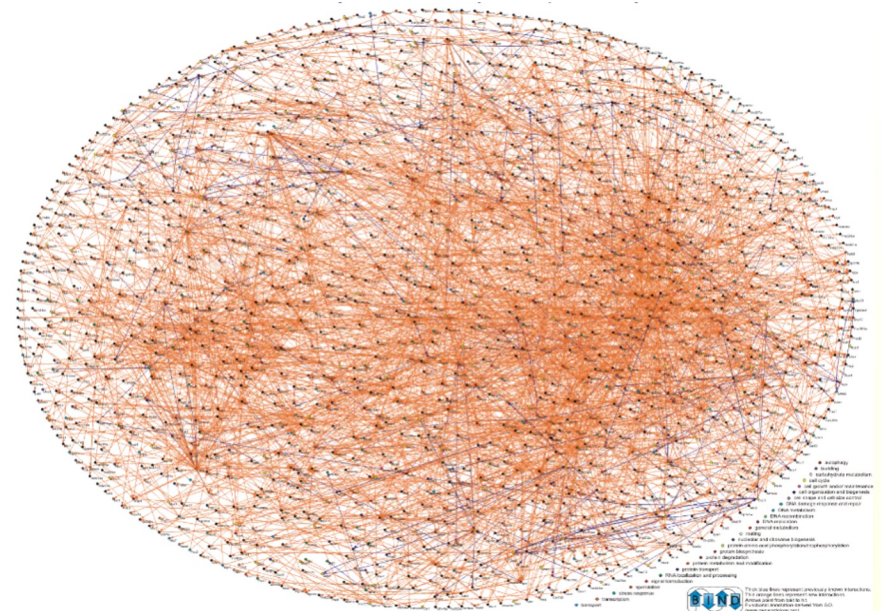
# Reactome

- Open source and open access pathway database
- Curated human pathways encompassing metabolism, signaling, and other biological processes.
- Rigorous curation standards – every pathway is traceable to primary literature.
- Cross-reference to many other bioinformatics databases.
- Provides data visualization and analysis tools
  - Google-map style reaction diagrams and textbook-style illustrations with overlays;
  - Find pathways containing your gene list;
  - Calculate gene overrepresentation in pathways;
  - Find corresponding pathways in other species.



# Networks

- Pathways capture only the “well understood” portion of biology.
- Networks cover less well understood relationships:
  - Genetic interactions
  - Physical interaction
  - Coexpression
  - GO term sharing
  - Adjacency in pathways



# Network Databases

- Can be built automatically or via curation.
- More extensive coverage of biological systems.
- Relationships and underlying evidence more tentative.
- Popular sources of curated networks:
  - BioGRID – Curated physical and genetic interactions from literature; 89K genes & 2.1M interactions from 80 species (<https://thebiogrid.org/>)
  - IntAct – Curated interactions from literature; 143K interactors & 1.5M interactions from 9000 species. (<https://www.ebi.ac.uk/intact/home>)
  - GeneMANIA - Compendium of 2.8K gene association networks representing 167K genes and 660M interactions from 9 species



# IntAct - Search for TP53

Filters  Interactor Type  Interaction Detection Method Interaction Host Organism Mutation Expansion Positive MI Score

**Layout**

Force directed

Circular

Bubbles

**Edges**

Expand

Affected By Mutation

**Group By**

Species

**Nodes**

Color ~ Species

- Homo sapiens
- Mus musculus
- Other mammals
- Other bacteria

Shape ~ Type

- protein
- molecule set

**Edges**

Color ~ MI Score

0.1   0.3   0.5   0.7   0.9   1.0

Width ~ #Evidence

1   25

Showing 1 to 25 of 394 entries

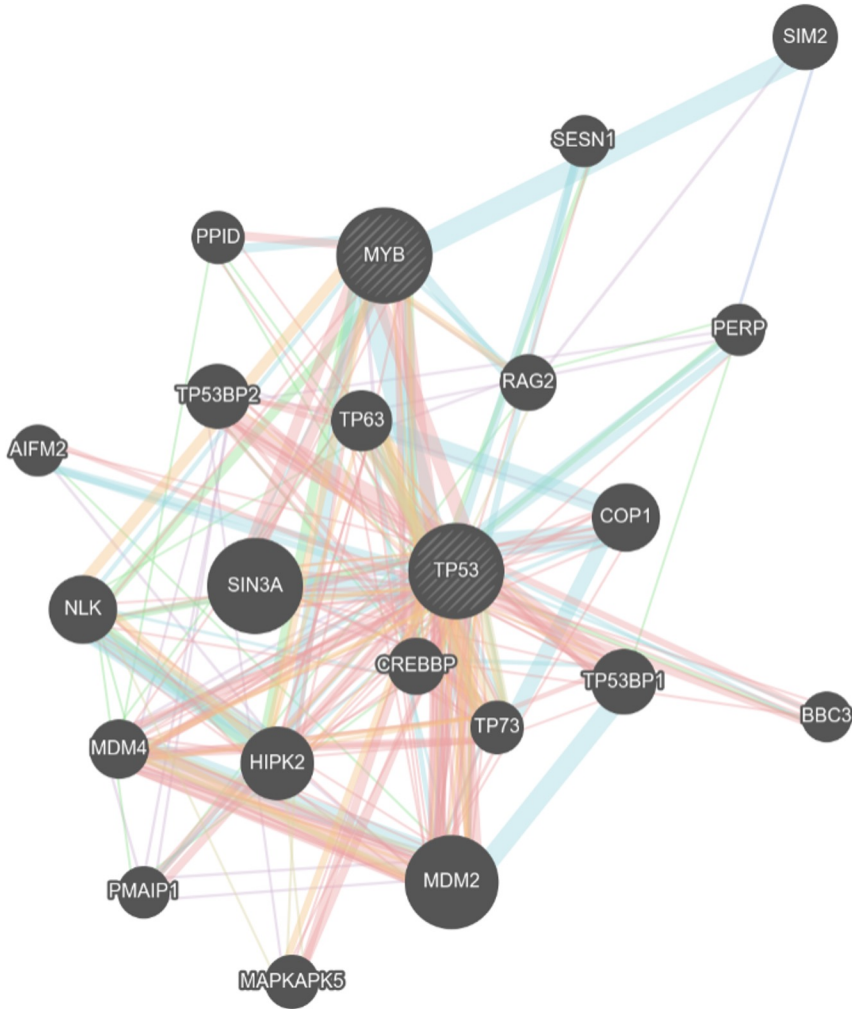
...

Show  entries

| Select                   | Molecule A | Molecule B | Identifier A                   | Identifier B                   | Type A  | Type B  | Species A                    | Species B                    | Host Organism  | Positive Interaction | Detection Method | Publication IDs          | Interac Typ |
|--------------------------|------------|------------|--------------------------------|--------------------------------|---------|---------|------------------------------|------------------------------|--|----------------------|------------------|--------------------------|-------------|
| <input type="checkbox"/> | MDM2       | TP53       | <a href="#">UniProt Q00987</a> | <a href="#">UniProt P04637</a> | protein | protein | <a href="#">Homo sapiens</a> | <a href="#">Homo sapiens</a> | <a href="#">Homo sapiens epitheloid cervix carcinoma cells</a> | ✓                    | pla              | <a href="#">25241761</a> | proxir      |

# GeneMANIA

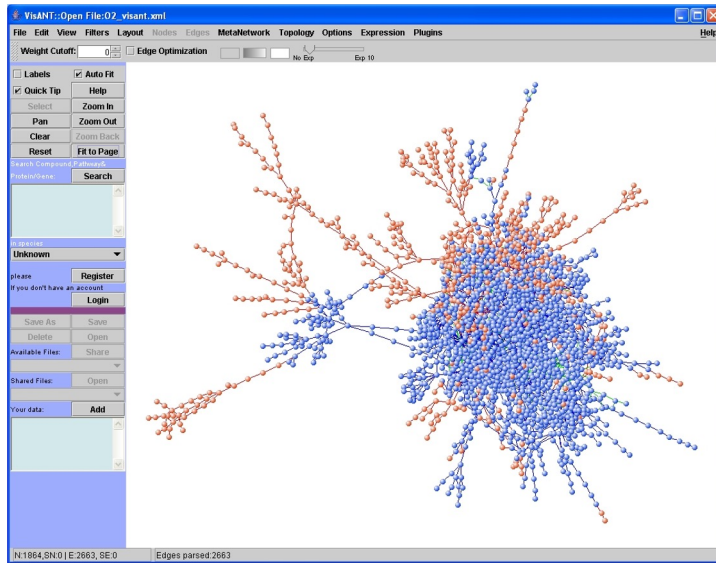
> ↑ TP53  
MYR e.g. 🔍



Networks

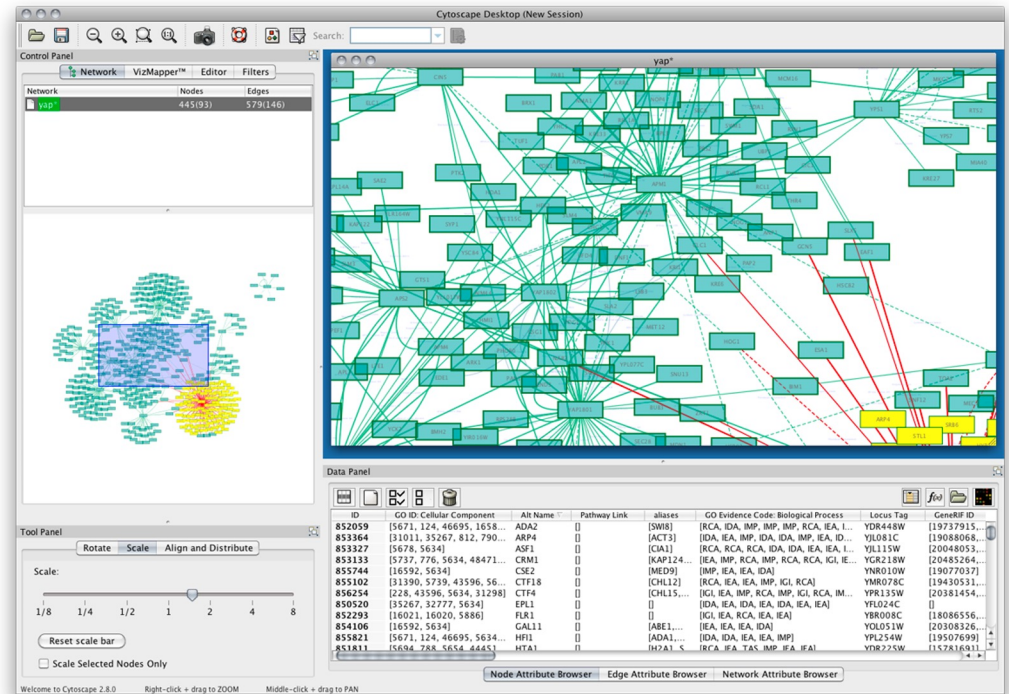
- Physical Interactions 77.64%
- Co-expression 8.01%
- Predicted 5.37%
- Co-localization 3.63%
- Genetic Interactions 2.87%
- Pathway 1.88%
- Shared protein domains 0.60%

# Visualization and Analysis Tools for Biological Networks



VisANT

Web & desktop versions  
(offline since Jan 2024)



Cytoscape

(Web & desktop versions)

# Pathway/Network Analysis

## Goal

### 1 Enrichment of fixed gene sets

Identification of pre-built pathways or networks that are enriched in a set of mutated or differentially expressed genes

### 2 De novo sub-network construction and clustering

Are new pathways altered in this cancer? Are there clinically-relevant tumour subtypes?

### 3 Pathway-based modeling

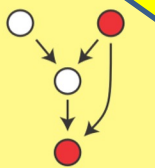
Evaluation of potential network rules that would be consistent with the identified set of mutated, differentially expressed or amplified genes

How are pathway activities altered in a particular patient? Are there targetable pathways in this patient?

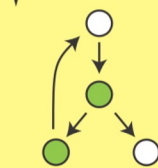
What biological processes are altered in this cancer?

## Tools

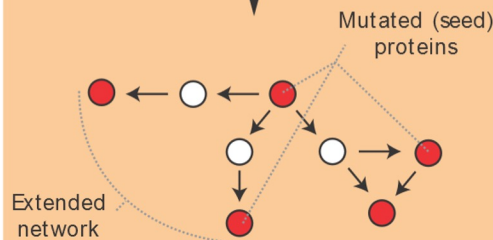
## Output



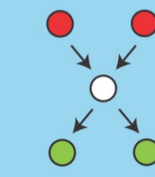
Enriched network



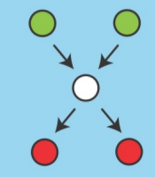
Depleted network



Extended network



Loss-Of-Function Network Signature



Gain-Of-Function Network Signature

# 1) Enrichment of Fixed Gene Sets

- Covered in Module 2.
- Most popular form of pathway/network analysis.
  - Overrepresentation analysis vs functional class scoring.
- Advantages:
  - Easy to perform.
  - Many good end-user tools.
  - Statistical model well worked out.
- Disadvantages:
  - Many possible gene sets;
  - Gene sets are heavily overlapping;
  - “Bags of genes” obscure regulatory relationships among them.

# Reactome: Pathway Enrichment Analysis

**Analysis tools**

**Your data** Options Analysis

Step 1: Select a file from your computer or paste your own data and click on the corresponding "Continue" button.

Select data file for analysis:  No file chosen

Paste your data to analyse or try example data sets:

| #Gene symbol | #Somatic | #Germline | Total |
|--------------|----------|-----------|-------|
| TP53         | 8        | 5         | 18    |
| APC          | 6        | 6         | 17    |
| TERT         | 9        | 1         | 15    |
| DICER1       | 8        | 1         | 14    |
| BRAF         | 12       | 0         | 12    |
| KDM6A        | 7        | 0         | 12    |
| CUX1         | 6        | 0         | 11    |
| KRAS         | 6        | 0         | 11    |
| CTNNB1       | 5        | 0         | 10    |
| ERBB2        | 5        | 0         | 10    |
| HRAS         | 2        | 3         | 10    |
| CDKN2A       | 2        | 2         | 9     |
| CDKN2A(p14)  | 2        | 2         | 9     |
| ETV6         | 4        | 0         | 9     |
| MDM2         | 4        | 0         | 9     |
| BRCA1        | 2        | 5         | 8     |

Some examples:

- 
- 
- 
- 
- 
- 
- 
- 
- 

**Analyse gene list**

**Analyse gene expression**

**Species Comparison**

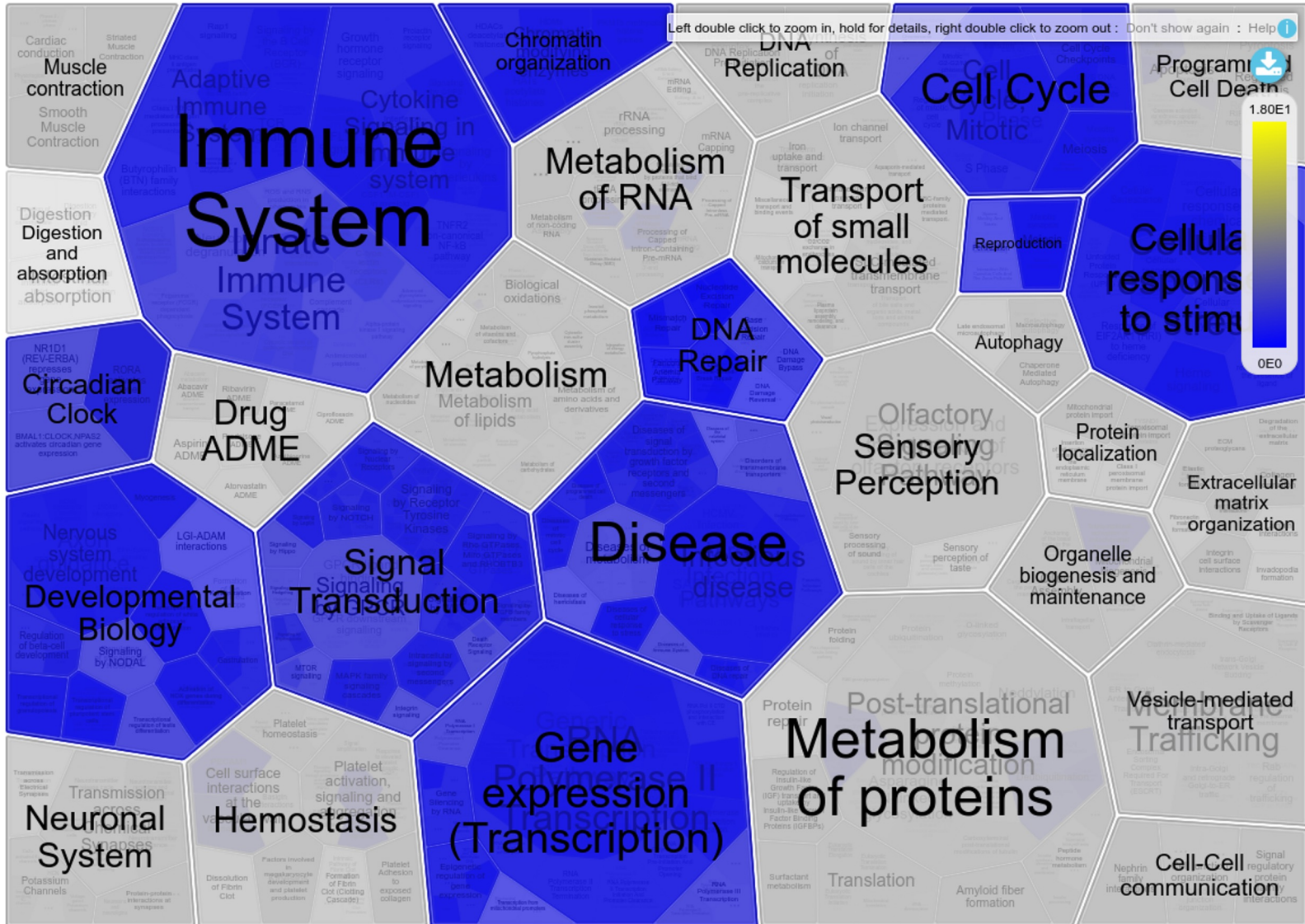
**Tissue Distribution**

**Reactome v84**

Click to learn more about our analysis tools

The analysis results are only kept for 7 days after your last usage. Afterwards you'll need to re-perform your analysis to see the results.

# Reactome: Pathway Enrichment Analysis



# Reactome: Pathway Enrichment Analysis

reactome 3.7 84 Pathways for: Homo sapiens Citation: Analysis: Tour: Layout:

Event Hierarchy:

- Metabolism (1093/644) FDR: 1E-07
- Metabolism of proteins (102/2,214) FDR: 1.2E-06
- Metabolism of RNA (33/830) FDR: 9.22E-06
- Muscle contraction (9/232) FDR: 8.13E-06
- Neuronal System (15/490) FDR: 9.84E-06
- Organelle biogenesis and maintenance (10/1,000) FDR: 1.1E-05
- Programmed Cell Death (17/237) FDR: 1.2E-05
- Protein localization (6/170) FDR: 8.47E-06
- Reproduction (15/122) FDR: 8.47E-03
- Sensory Perception (7/1,262) FDR: 1E-06
- Signal Transduction (245/3,028) FDR: 1.2E-06**
- Signaling by Receptor Tyrosine Kinases (166/1,000) FDR: 1.2E-06**
- Signaling by EGFR (13/61) FDR: 1.2E-06
- Signaling by FGFR (20/107) FDR: 1.2E-06
- Signaling by NTRKs (28/166) FDR: 1.2E-06
- Signaling by Insulin receptor (18/100) FDR: 1.2E-06
- Signaling by PDGF (15/70) FDR: 1.2E-06
- Signaling by VEGF (17/140) FDR: 1.2E-06
- Signaling by SCF-KIT (16/51) FDR: 1.2E-06
- Signaling by ERBB2 (16/68) FDR: 1.2E-06**
- ERBB3 binds neuregulins
- ERBB2 forms heterodimers
- SRC family kinases phosphorylate
- Trans-autophosphorylation of
- Trans-autophosphorylation of
- SHC1 events in ERBB2 signaling
- GRB2 events in ERBB2 signaling
- PI3K events in ERBB2 signaling
- PLCG1 events in ERBB2 signaling**
- GRB7 events in ERBB2 signaling
- ERBB2 Regulates Cell Motility
- ERBB2 Activates PTK6 Signaling

Search for a term, e.g. pten ...

1.8E1  
0E0

1/3 :: #Somatic

Description Molecules Structures Expression Analysis 1,601 Downloads

Expression analysis results for TOTAL [Data: Gene symbol]

| Pathway name   | Entities found | Entities Total | Entities ratio | Entities pValue | Entities FDR | Reactome found |
|--|----------------|----------------|----------------|-----------------|--------------|----------------|
| PLCG1 events in ERBB2 signaling                          | 3              | 6              | 0              | 3.51E-3         | 1.72E-2      |                |
| FBXW7 Mutants and NOTCH1 in Cancer                       | 3              | 6              | 0              | 3.51E-3         | 1.72E-2      |                |
| Loss of Function of FBXW7 in Cancer and NOTCH1 Signaling | 3              | 6              | 0              | 3.51E-3         | 1.72E-2      |                |
| Prolonged ERK activation events                          | 5              | 20             | 0.001          | 3.52E-3         | 1.72E-2      |                |
| SUMOylation of transcription factors                     | 5              | 20             | 0.001          | 3.52E-3         | 1.72E-2      |                |
| MAP2K and MAPK activation                                | 8              | 49             | 0.003          | 3.54E-3         | 1.72E-2      |                |

321-340 of 1,601

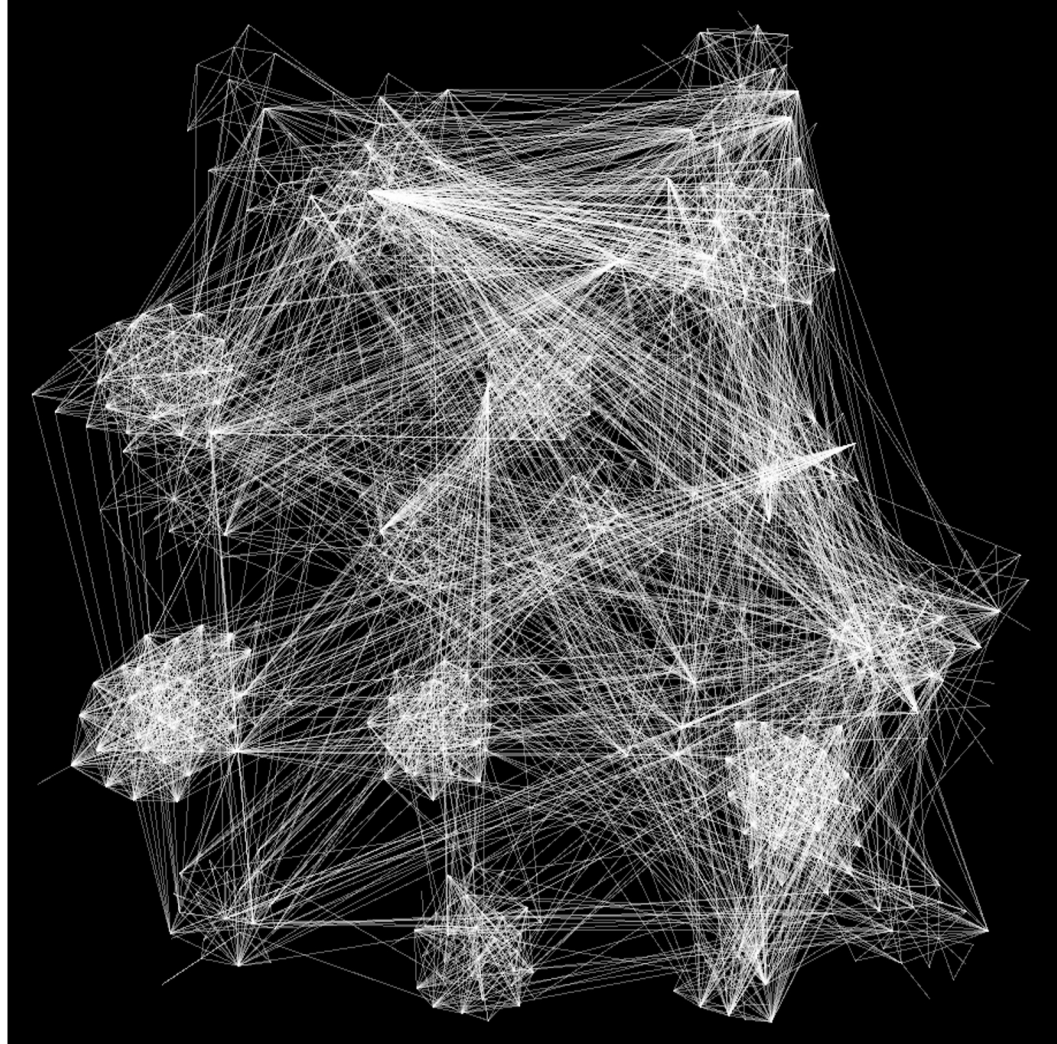


## 2) De Novo Subnetwork Construction & Clustering

- Apply list of altered {genes,proteins,RNAs} to a biological network.
- Identify “topologically unlikely” configurations.
  - E.g. a subset of the altered genes are closer to each other on the network than you would expect by chance.
- Extract clusters of these unlikely configurations.
- Annotate the clusters.

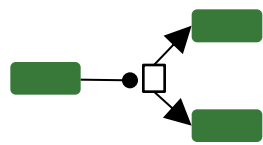
# Reactome FI Network

- 12,441 Genes
- 291,172 FIs
- ~61% coverage of genome.
- False (+) rate < 1%
- False (-) rate ~80%



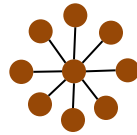
*5% of network shown here*

# Reactome FI Network



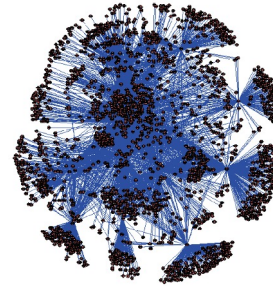
Curated  
Pathway  
Dbs

+



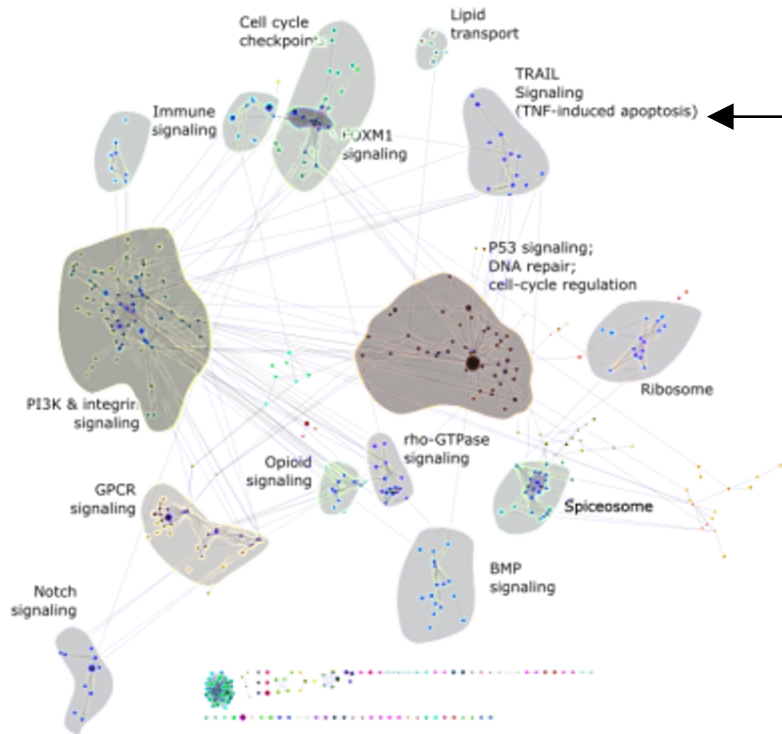
Uncurated  
Interaction  
Evidence

Machine Learning

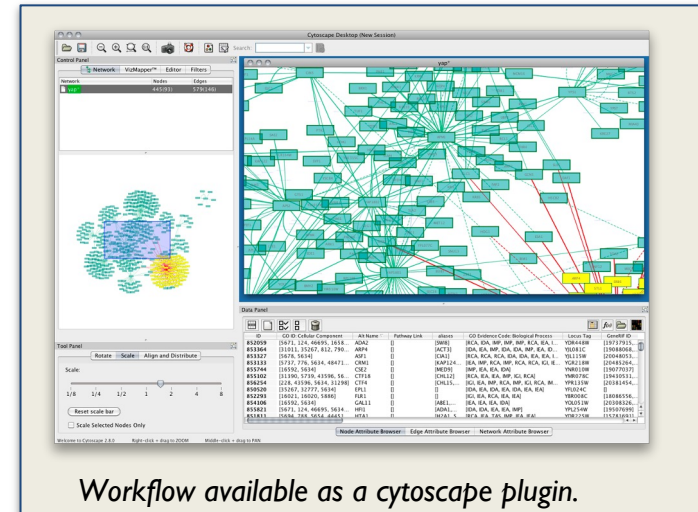


Reactome Functional Interaction Network  
(~11,000 proteins; 270,000 interactions)

Extract and Cluster "Disease Genes"



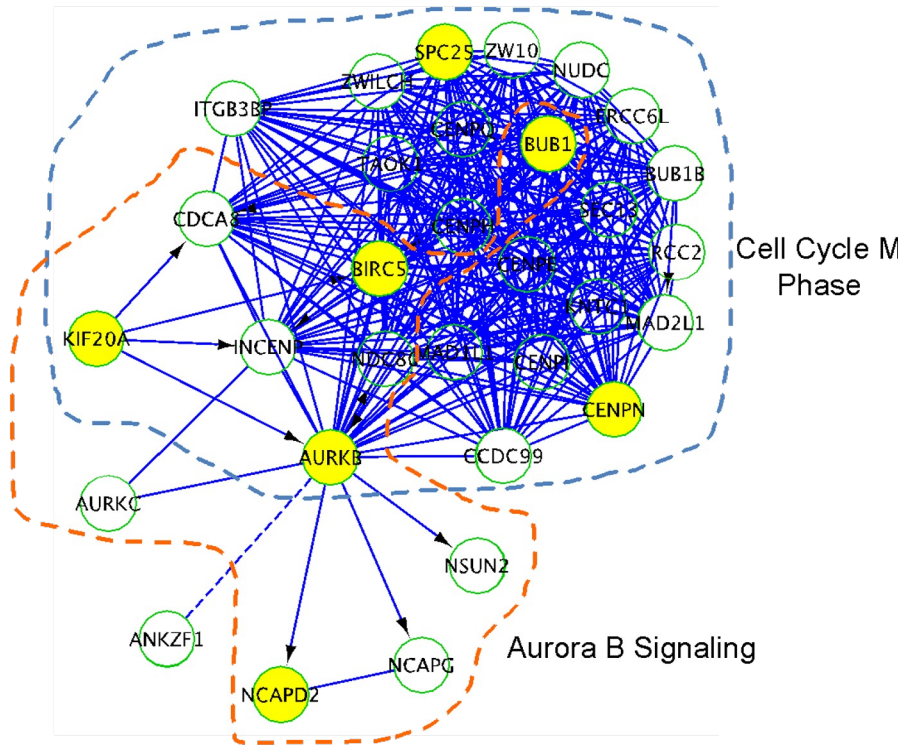
Disease "modules" (10-30)



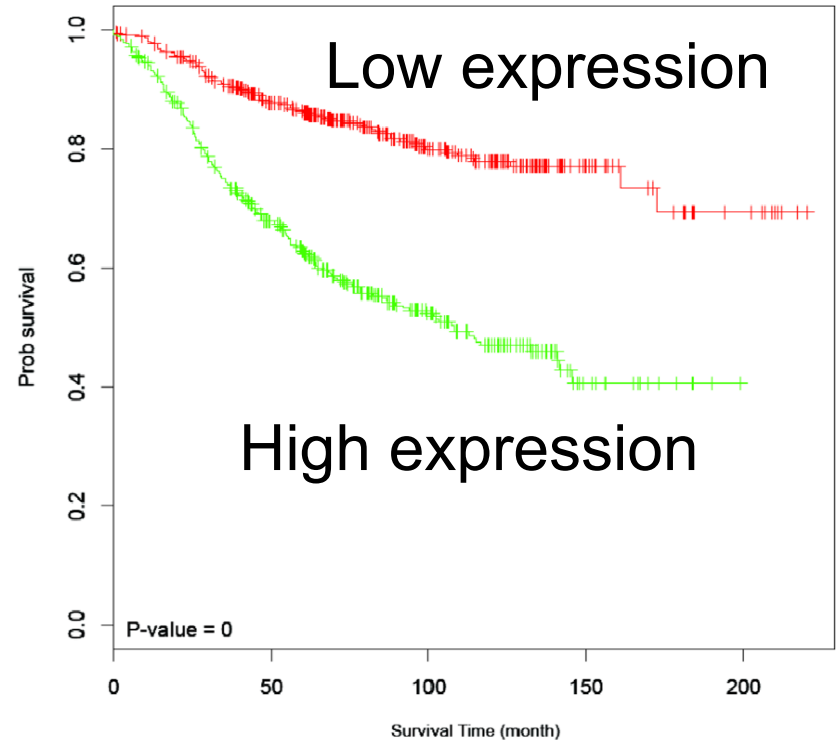
Workflow available as a cytoscape plugin.

A human functional protein interaction network and its application to cancer data analysis, [Wu et al. 2010 Genome Biology](#)

# Module-Based Prognostic Biomarker in ER+ Breast Cancer



Measure levels of expression of the genes in this network module



A human functional protein interaction network and its application to cancer data analysis, [Wu et al. 2010 Genome Biology](#)

# Popular Network Clustering Algorithms

- GeneMANIA
  - “Birds of a feather” principle.
  - Very useful for finding genes that are related to an experimentally defined set.
- HotNet
  - Finds “hot” clusters based on propagation of heat across metallic lattice.
  - Avoids ascertainment bias on unusually well-annotated genes.
- HyperModules Cytoscape App
  - Find network clusters that correlate with clinical characteristics.
- Reactome FI Network Cytoscape App
  - Offers multiple clustering and correlation algorithms (including HotNet, and survival correlation analysis)

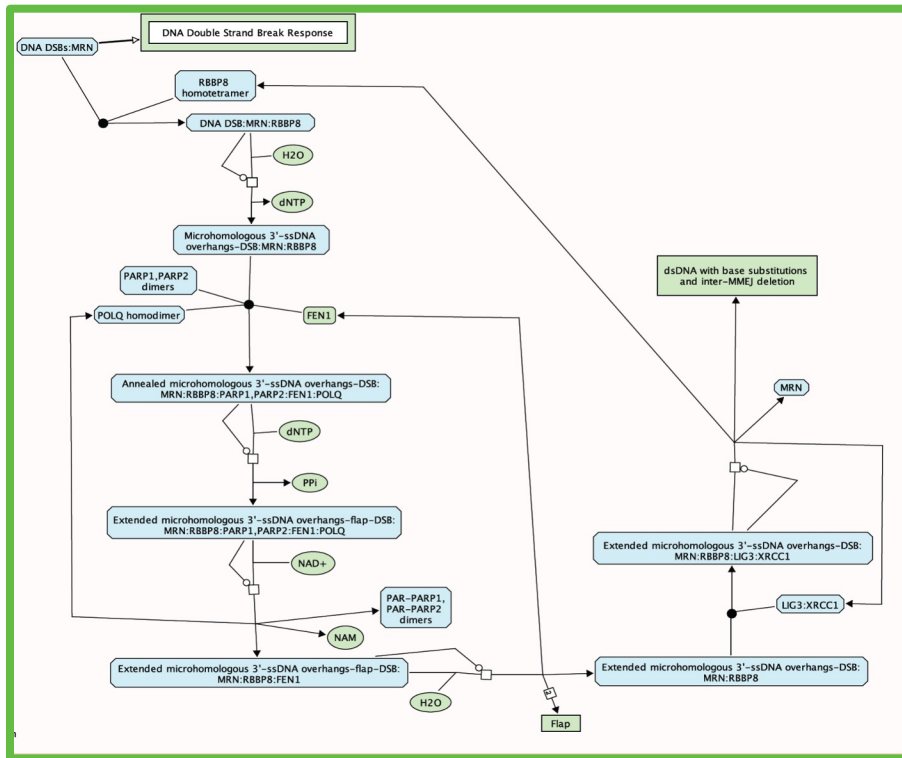
# 3) Pathway-Based Modeling

- Apply list of altered {genes,proteins,RNAs} to biological pathways.
- Preserve detailed biological relationships.
- Attempt to integrate multiple molecular alterations together to yield lists of altered pathway activities.
- Pathway modeling shades into Systems Biology

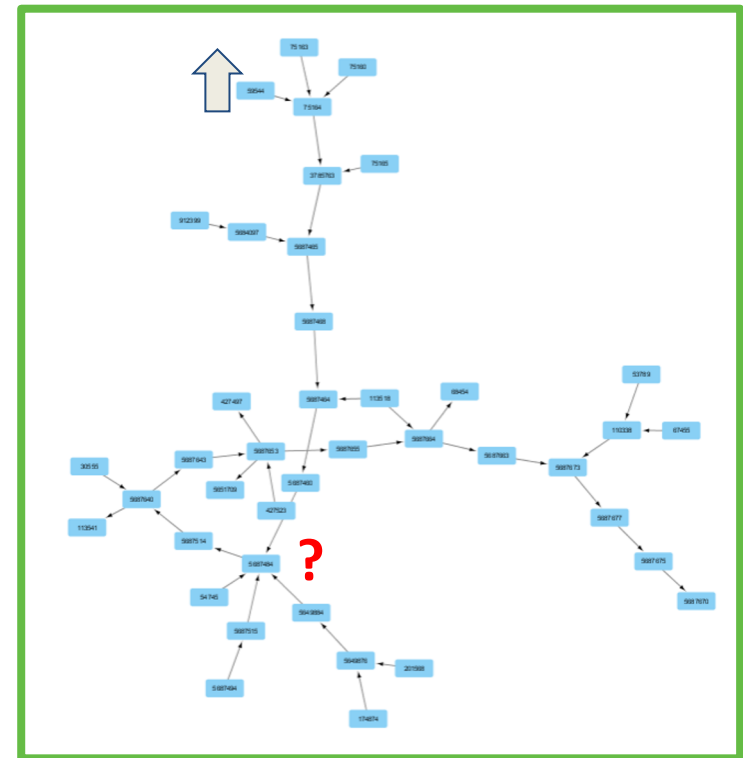
# Types of Pathway-Based Modeling

- Partial differential equations, e.g. CellNetAnalyzer
  - Mostly suited for biochemical systems (metabolomics)
- Network flow models, e.g. NetPhorest
  - Mostly suited for kinase cascades (phosphorylation info)
- Transcriptional regulatory network-based reconstruction methods, e.g. ARACNe
- Logic Graphs and probabilistic graph models (PGMs)
  - Capture the logic of a pathway without needing rate/binding constants.
- Generative AI Models

# Boolean Network Inference



Pathway View



Logic Graph View



# Generative AI Network Models

Large Language Models (GPT-3) are trained to predict masked text:

The quick brown ? jumped over the lazy dog.

Generative pathway models are trained to predict gene network perturbations:





# Can you use pathways to predict biology?



[Database \(Oxford\)](#), 2022; 2022: baac009.

PMCID: PMC9216552

Published online 2022 Mar 6. doi: [10.1093/database/baac009](https://doi.org/10.1093/database/baac009)

PMID: [35348650](https://pubmed.ncbi.nlm.nih.gov/35348650/)

## Evaluating the predictive accuracy of curated biological pathways in a public knowledgebase

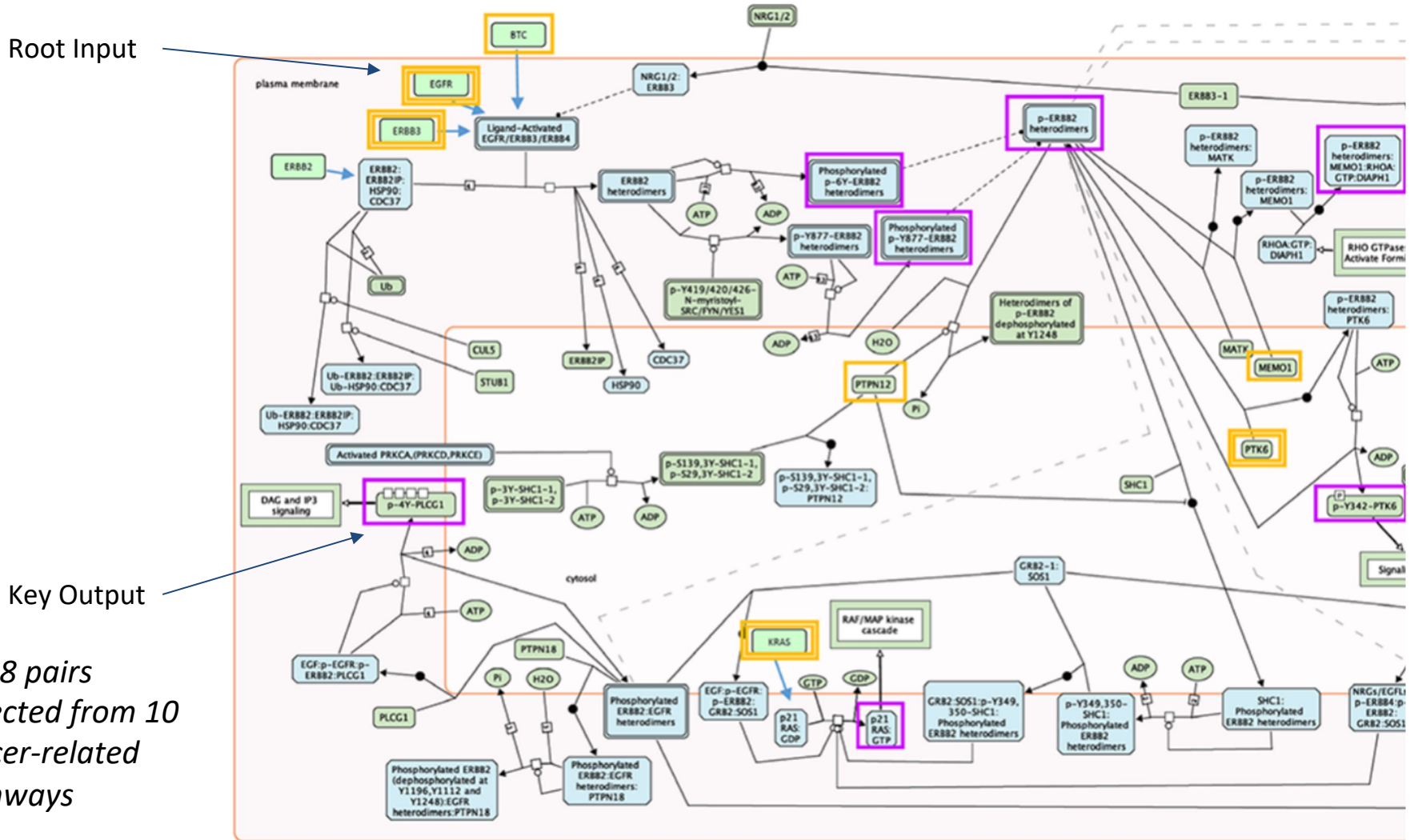
[Adam J Wright](#),<sup>✉</sup> [Marija Orlic-Milacic](#),<sup>✉</sup> [Karen Rothfels](#), [Joel Weiser](#), [Quang M Trinh](#), [Bijay Jassal](#), [Robin A Haw](#), and [Lincoln D Stein](#)<sup>✉</sup>

[▶ Author information](#) ▶ [▶ Article notes](#) ▶ [▶ Copyright and License information](#) [▶ Disclaimer](#)

Database (Oxford)

- How well can we predict the downstream effects of knocking up/down a gene using:
  - Experts gazing at pathway diagrams?
  - A graph-based inference algorithm?

# Step 1: Gather Input/Output Pairs



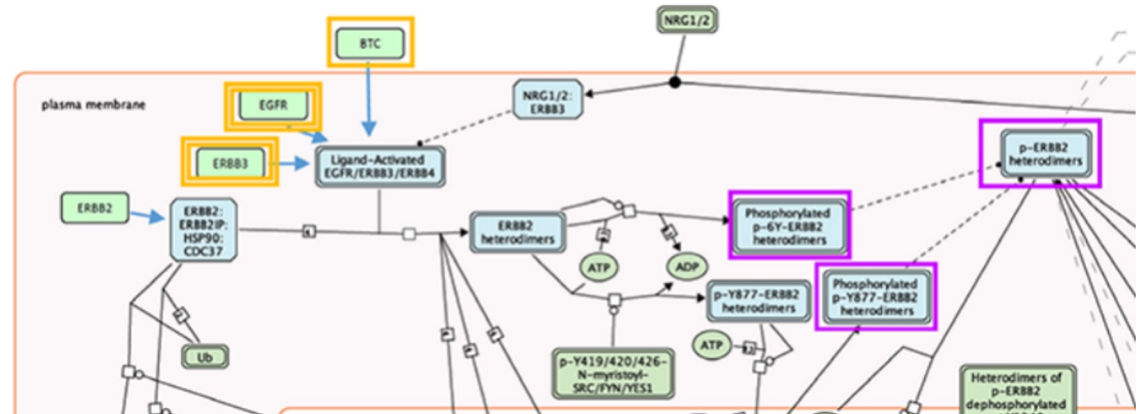
4,968 pairs collected from 10 cancer-related pathways

# Step 2: Collect Empirical Results

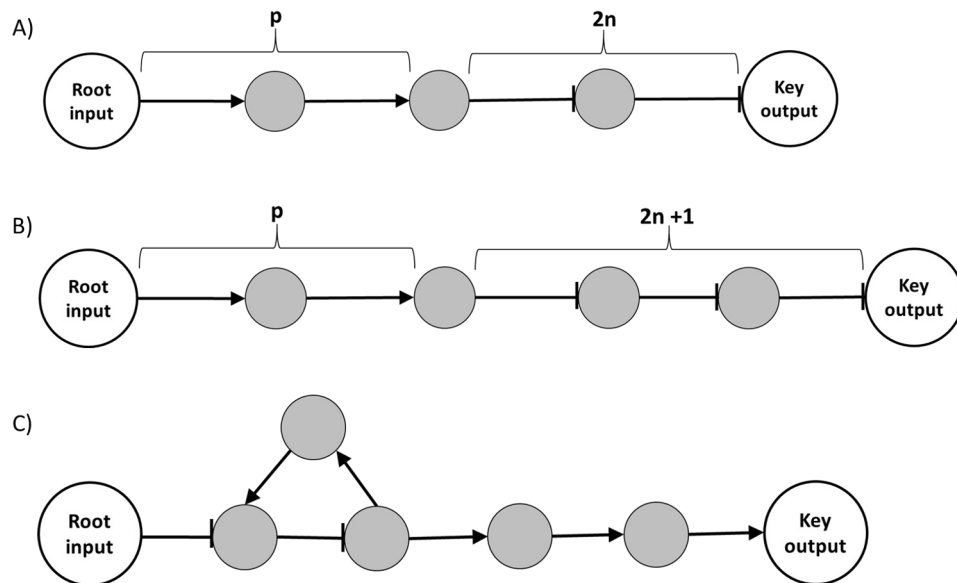
- Literature searches for functional genomics experiments in which key input was perturbed and effect on key output measured.
- 531 papers found, reporting 847 tested cases.

# Step 3: Predict Downstream Effects

1. Curators gaze at pathway diagram and apply logic rules to predict effect of perturbation.



2. Apply a boolean inference algorithm, MP-BioPath, to predict effect of perturbation.



# Step 4: Compare Predictions to Empirical

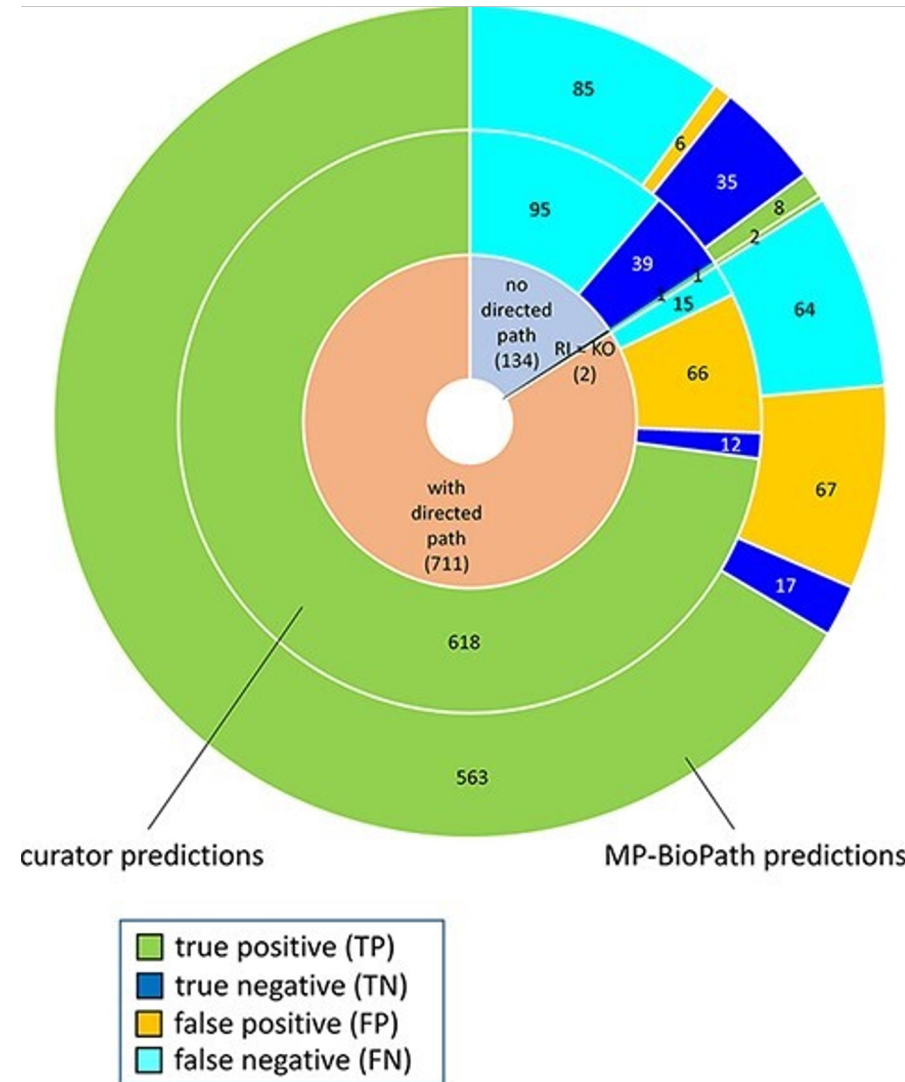
Curator Predictive Accuracy: **81%**

MP-BioPath Predictive Accuracy: **75%**

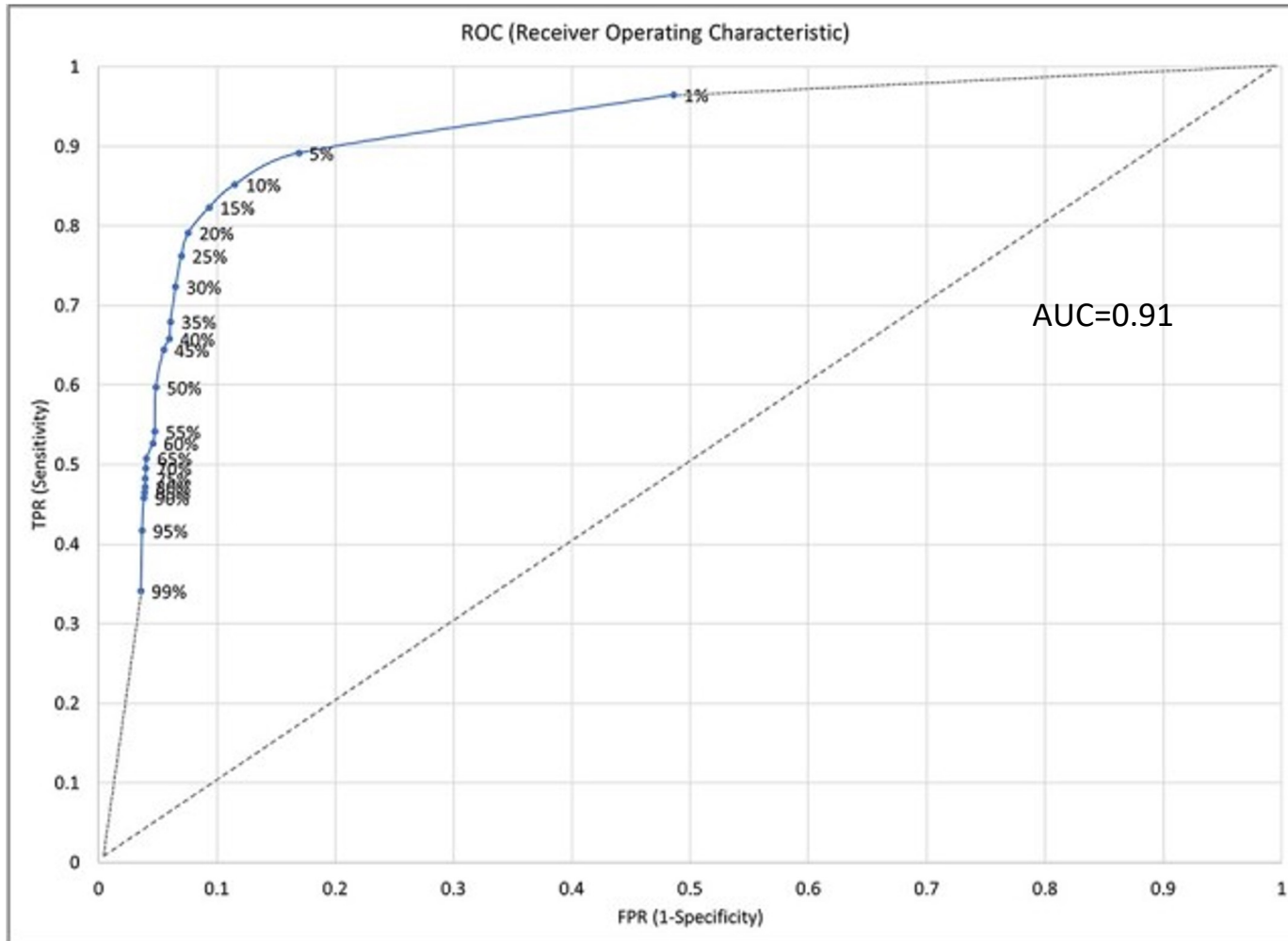
Random Guessing: **33%**

Largest source of error were false negatives due to missing elements of the pathway.

Largest source of false positives was direction of perturbation predicted incorrectly, also related to missing elements.



# Excellent Concordance between Algorithm's & Humans' Predictions





# Conclusions & Takeaways

- Pathway analysis allows discovery of biological processes hidden in large-scale data sets.
- Many databases and tools to choose from.
- Curated pathway databases now reaching levels of completeness that allow for accurate prediction of perturbations.
- Field is ripe for machine learning approaches.

# Pathway/Network Database URLs

- BioGRID
  - [http:// www.thebiogrid.org](http://www.thebiogrid.org)
- IntAct
  - <http://www.ebi.ac.uk/intact/>
- KEGG
  - <http:// www.genome.jp/kegg>
- Reactome
  - <http:// www.reactome.org>

# De novo network construction & clustering

- GeneMANIA
  - <http://www.genemania.org>
- HotNet
  - <http://compbio.cs.brown.edu/projects/hotnet/>
- HyperModules
  - <http://apps.cytoscape.org/apps/hypermodules>
- Reactome Cytoscape FIViz App
  - <http://apps.cytoscape.org/apps/reactomefis>

# Pathway Modeling

- CellNetAnalyzer
  - <http://www.ebi.ac.uk/research/saez-rodriguez/software>
- NetPhorest/NetworkKIN
  - <http://netphorest.info>, <http://networkin.info>
- ARACNe
  - <http://wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE>
- scGPT
  - <https://github.com/bowang-lab/scGPT>
- Pathway Prediction Evaluation Paper
  - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9216552/>

We are on a Coffee Break &  
Networking Session