



Canadian Bioinformatics Workshops

www.bioinformatics.ca

This page is available in the following languages:

Afrikaans Български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego ລາວ hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenščina jezik 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](#)

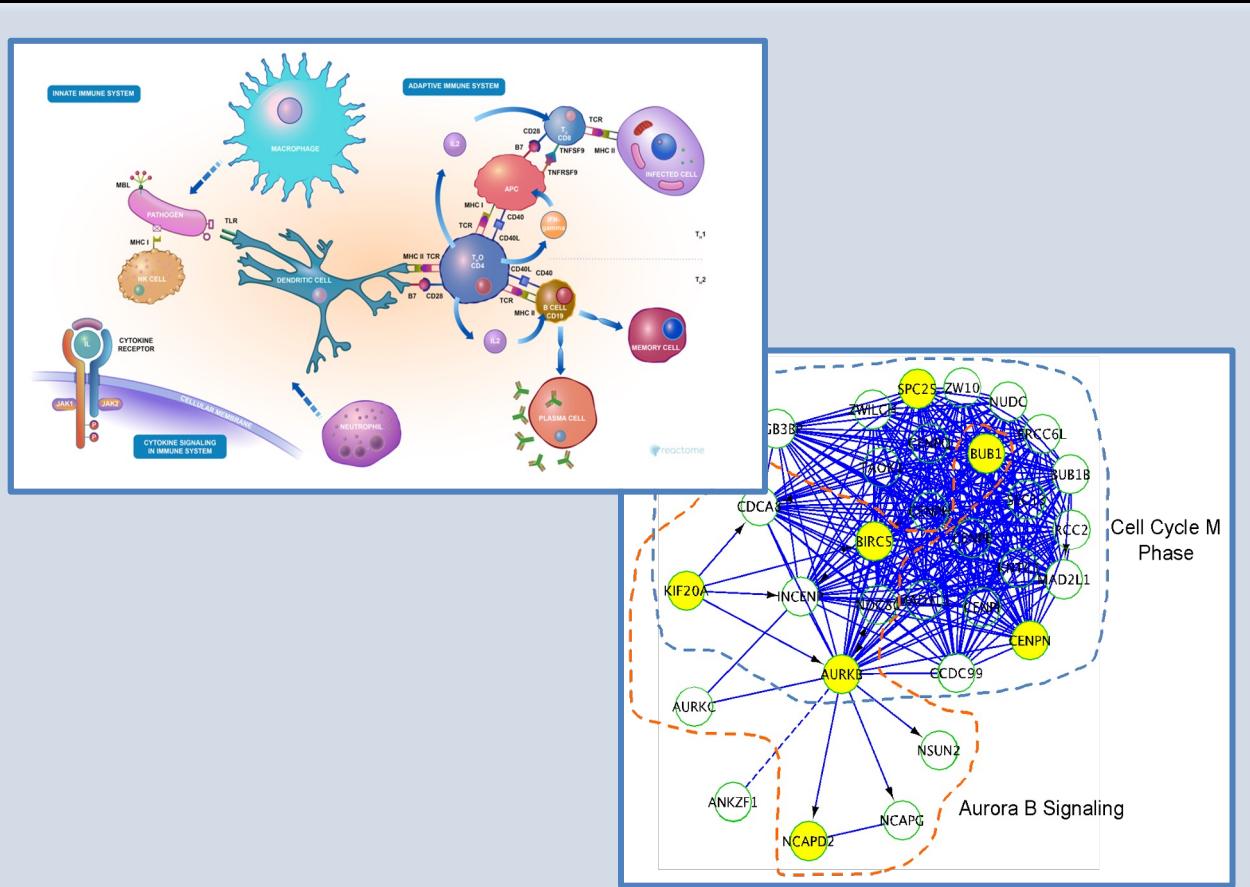
Module 4

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.

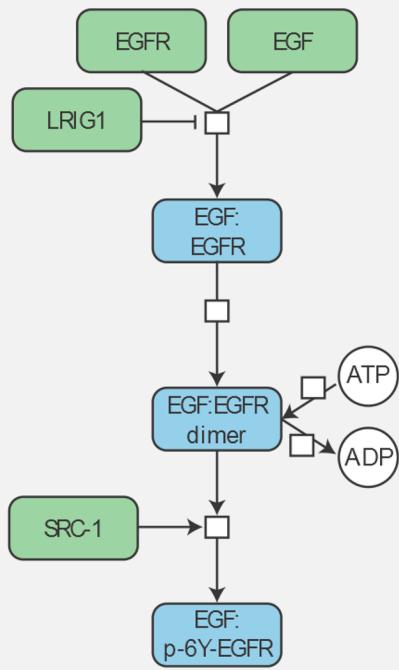
Ingredients you will Need

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

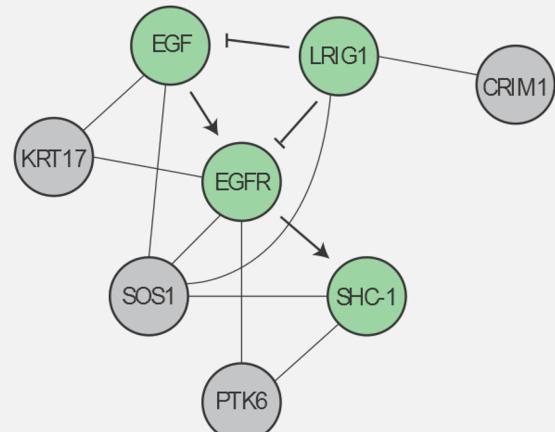


Pathways vs Networks

EGFR-centered
Pathway



EGFR-centered
Network

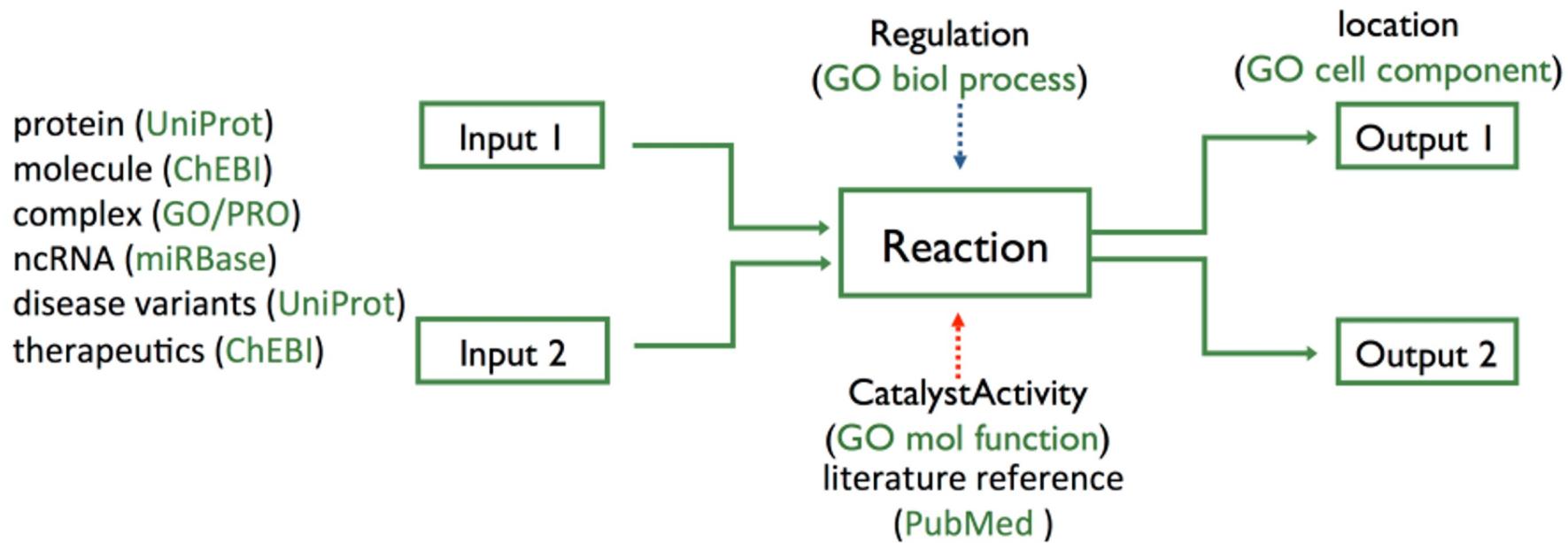


Pathway Databases

- 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

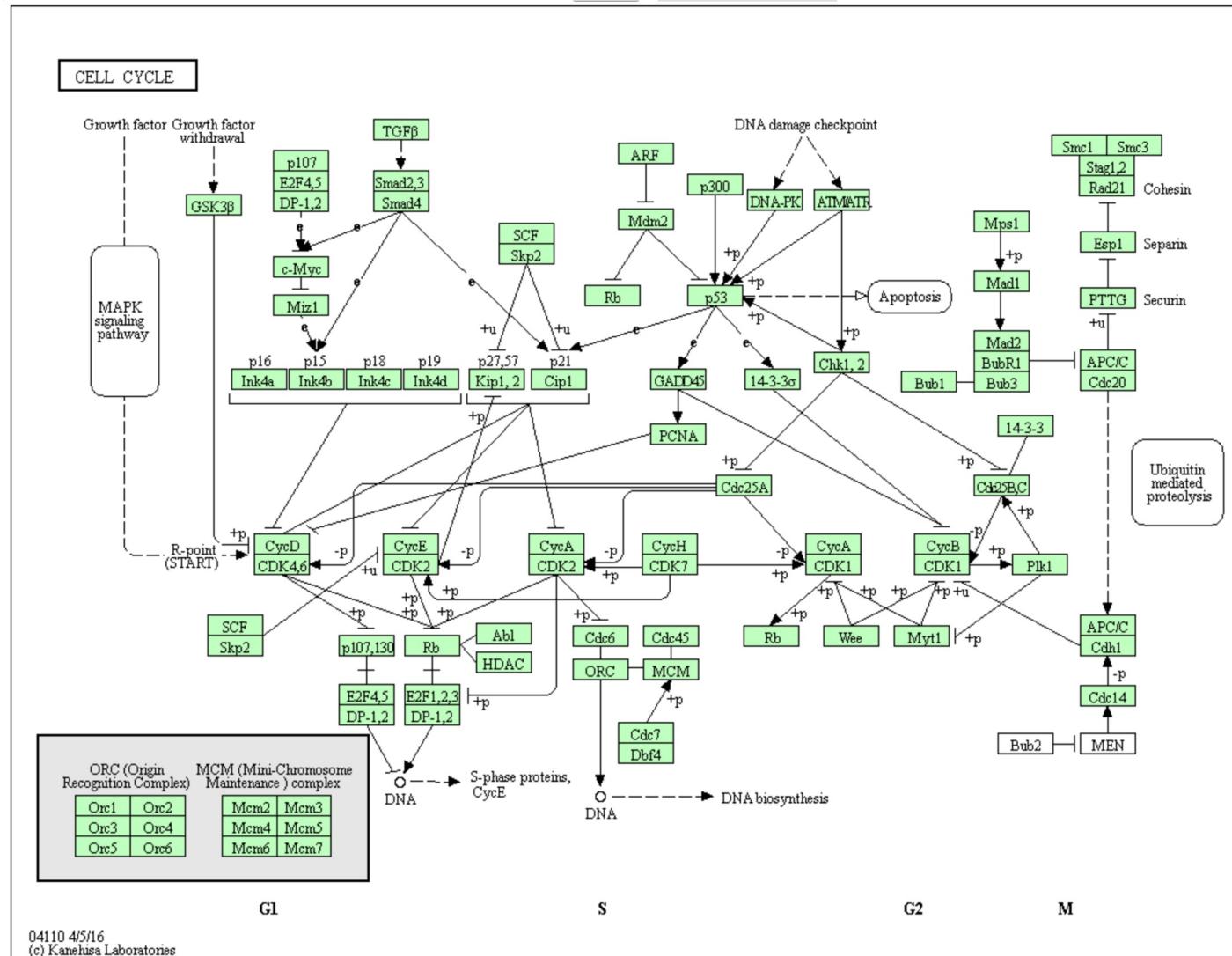
- Reactome & KEGG
 - explicitly describe biological processes as a series of biochemical reactions.
 - represents many events and states found in biology.



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.

Reactome Cell Cycle

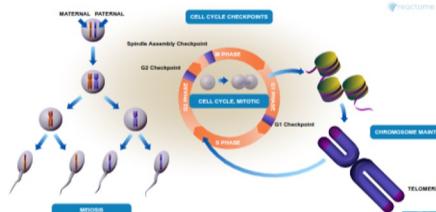
Locations in the PathwayBrowser

Cell Cycle (Homo sapiens)

- General

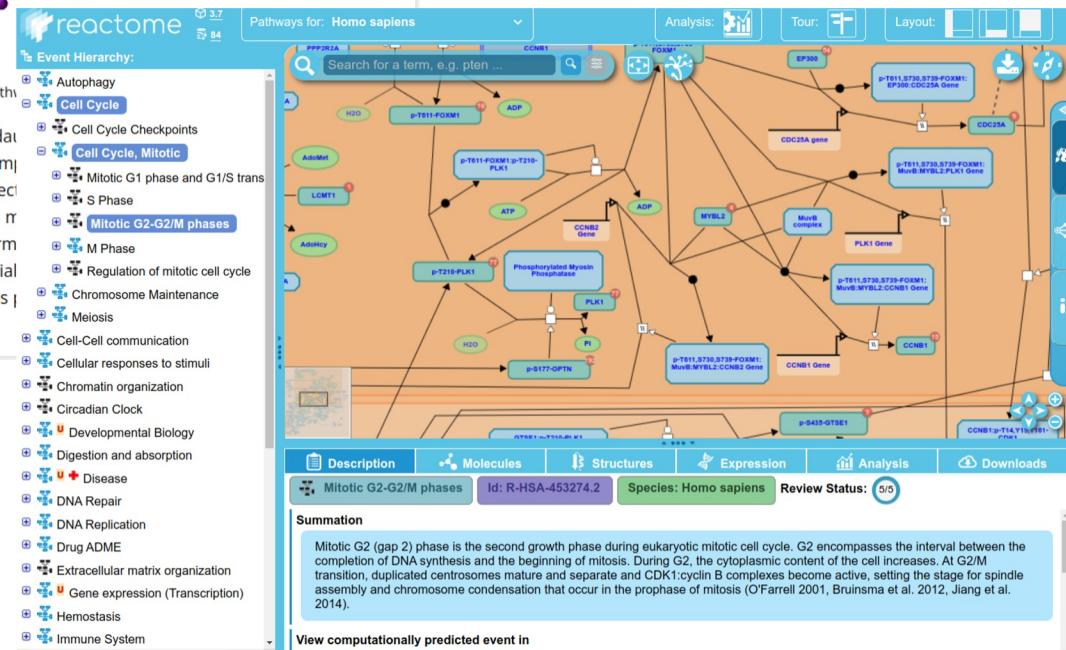
 SBML | BioPAX | PDF

 SVG | PNG



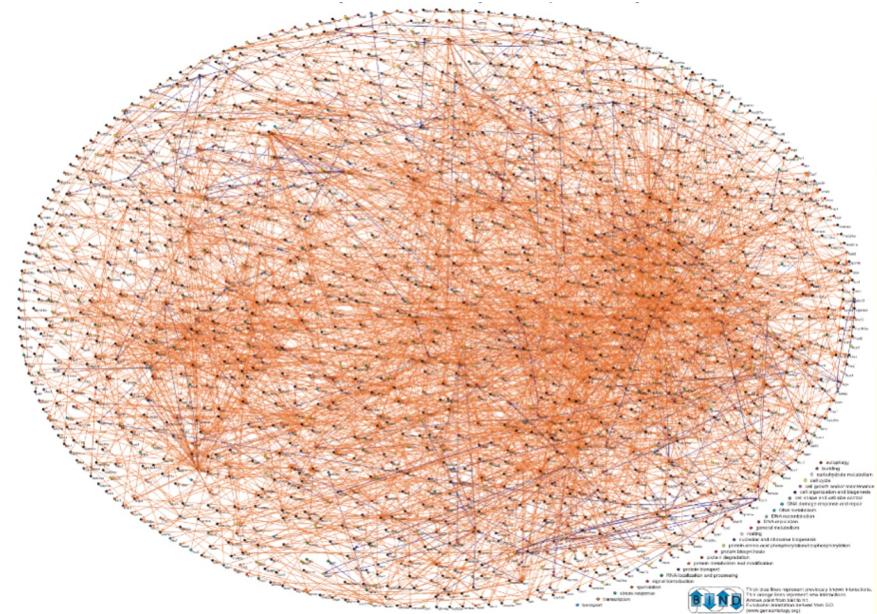
Click the Image above or [here](#) to open this pathway in the Pathway Browser.

The replication of the genome and the subsequent segregation of chromosomes into daughter cells collectively known as the cell cycle. DNA replication is carried out during a discrete temporal phase and chromosome segregation occurs during a massive reorganization to cellular architecture. The major cell cycle events: G1 between mitosis and S-phase, and G2 between S-phase and mitosis. Cells can exit the cell cycle for a period and enter a quiescent state known as G0, or terminally differentiate again, but undergo morphological development to carry out the wide variety of specialized functions. Protein serine/threonine kinases known as the cyclin-dependent kinases (CDKs) controls progression through the cell cycle. The CDKs are composed of a regulatory cyclin subunit and a catalytic subunit. The catalytic subunit suggests, the activity of the catalytic subunit is dependent... [Read more](#)



Networks

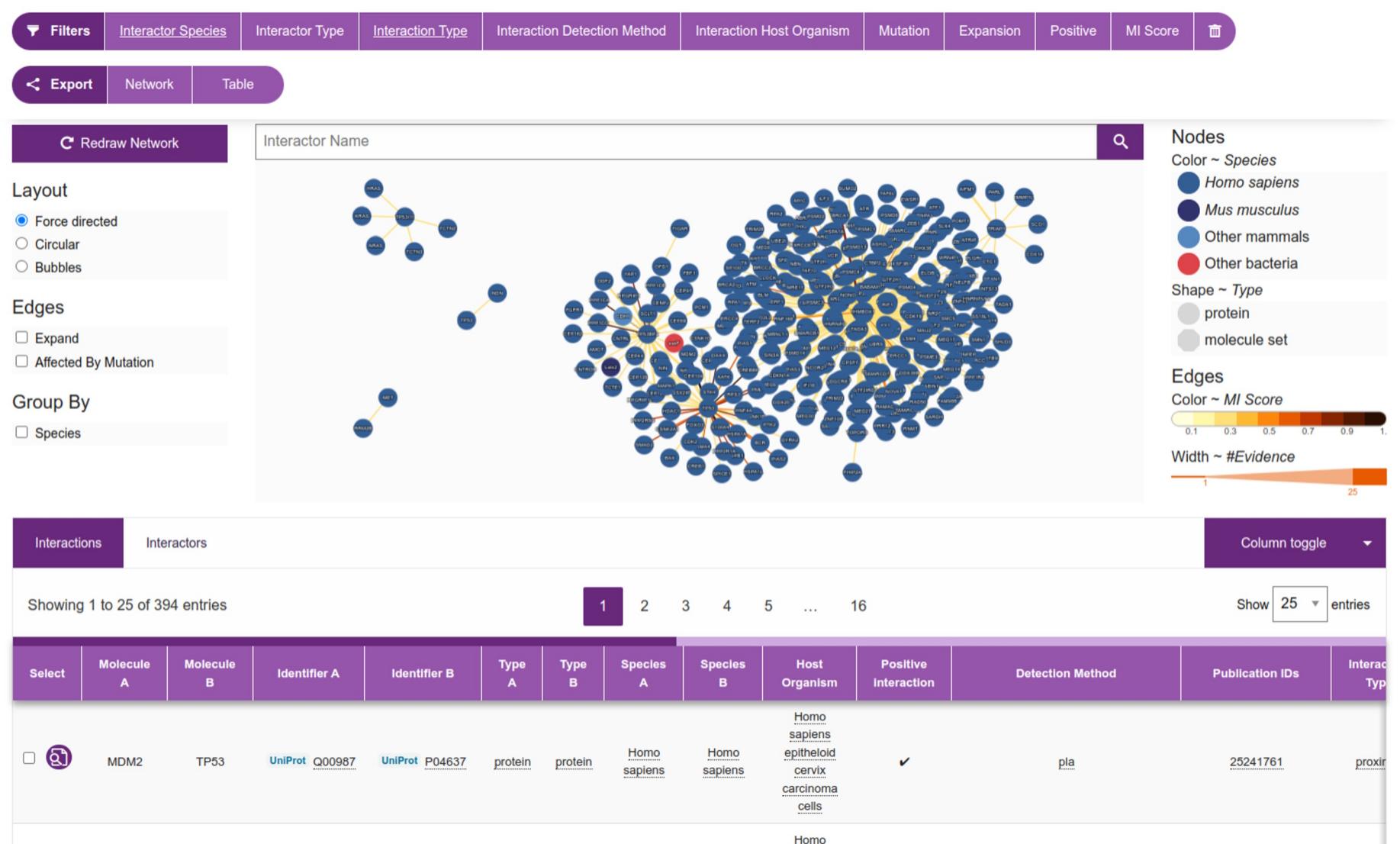
- Pathways capture only the “well understood” portion of biology.
- Networks cover less well understood relationships:
 - Genetic interactions
 - Physical interaction
 - Coexpression
 - GO term sharing
 - Adjacency in pathways



Network Databases

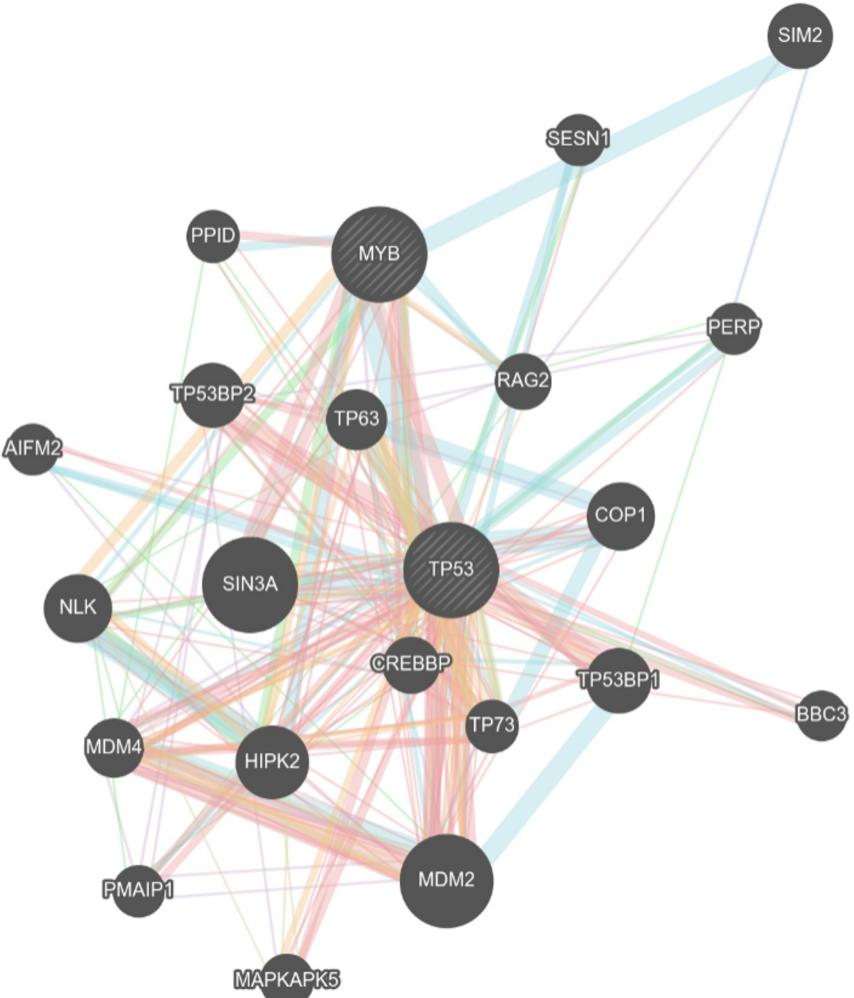
- Can be built automatically or via curation.
- More extensive coverage of biological systems.
- Relationships and underlying evidence more tentative.
- Popular sources of curated networks:
 - BioGRID – Curated physical and genetic interactions from literature; 89K genes & 2.1M interactions from 80 species (<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



GeneMANIA

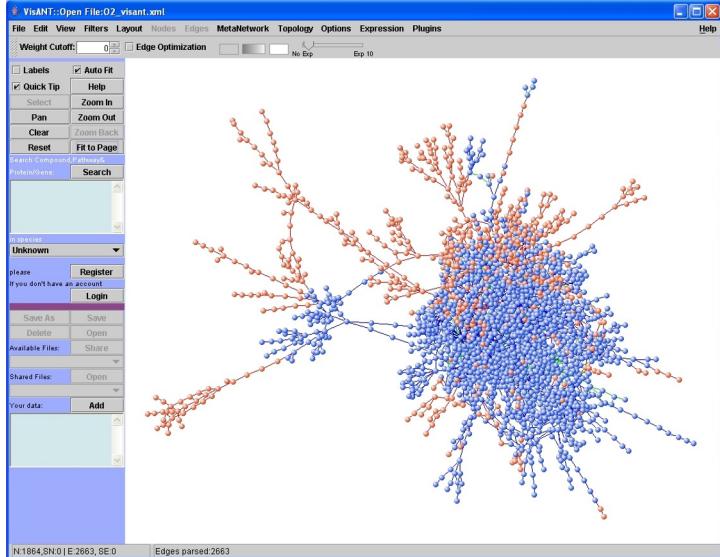
> | TP53
MYB e.g. :



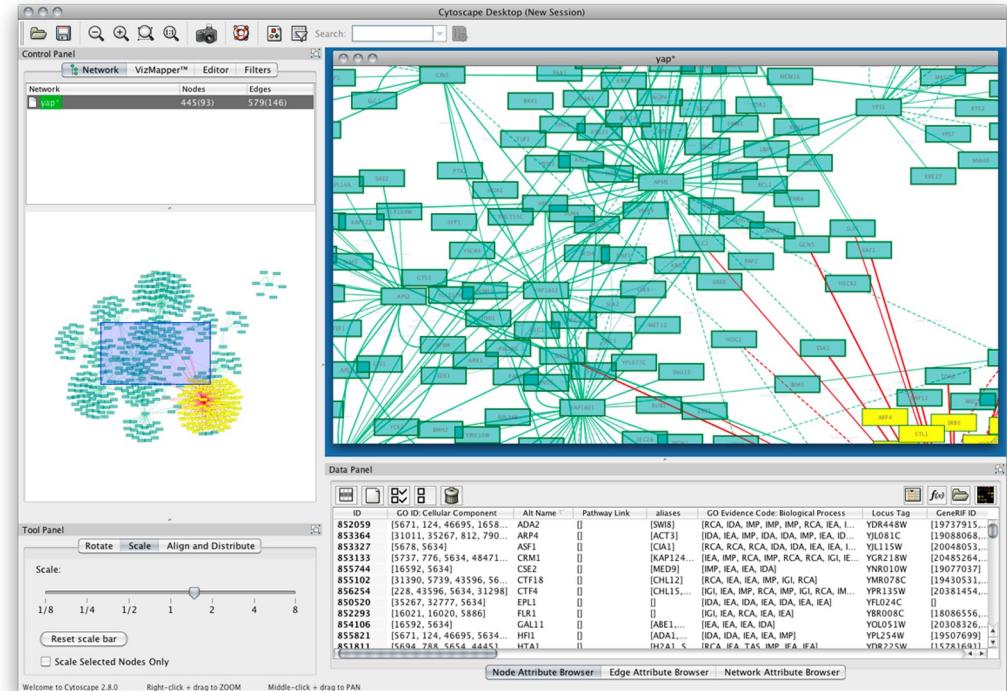
Networks	
<input checked="" type="checkbox"/>	Physical Interactions 77.64%
<input checked="" type="checkbox"/>	Co-expression 8.01%
<input checked="" type="checkbox"/>	Predicted 5.37%
<input checked="" type="checkbox"/>	Co-localization 3.63%
<input checked="" type="checkbox"/>	Genetic Interactions 2.87%
<input checked="" type="checkbox"/>	Pathway 1.88%
<input checked="" type="checkbox"/>	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

Tools

What biological processes are altered in this cancer?

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

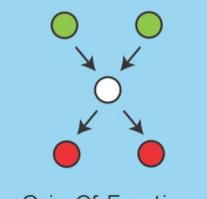
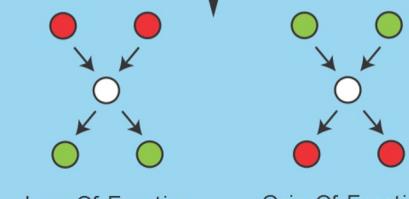
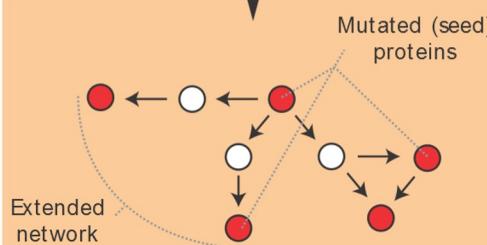
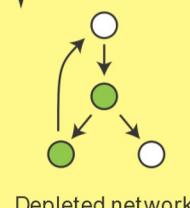
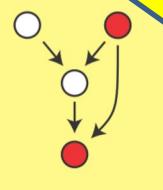
2 De novo sub-network construction and clustering

3 Pathway-based modeling

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

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

Output



1) Enrichment of Fixed Gene Sets

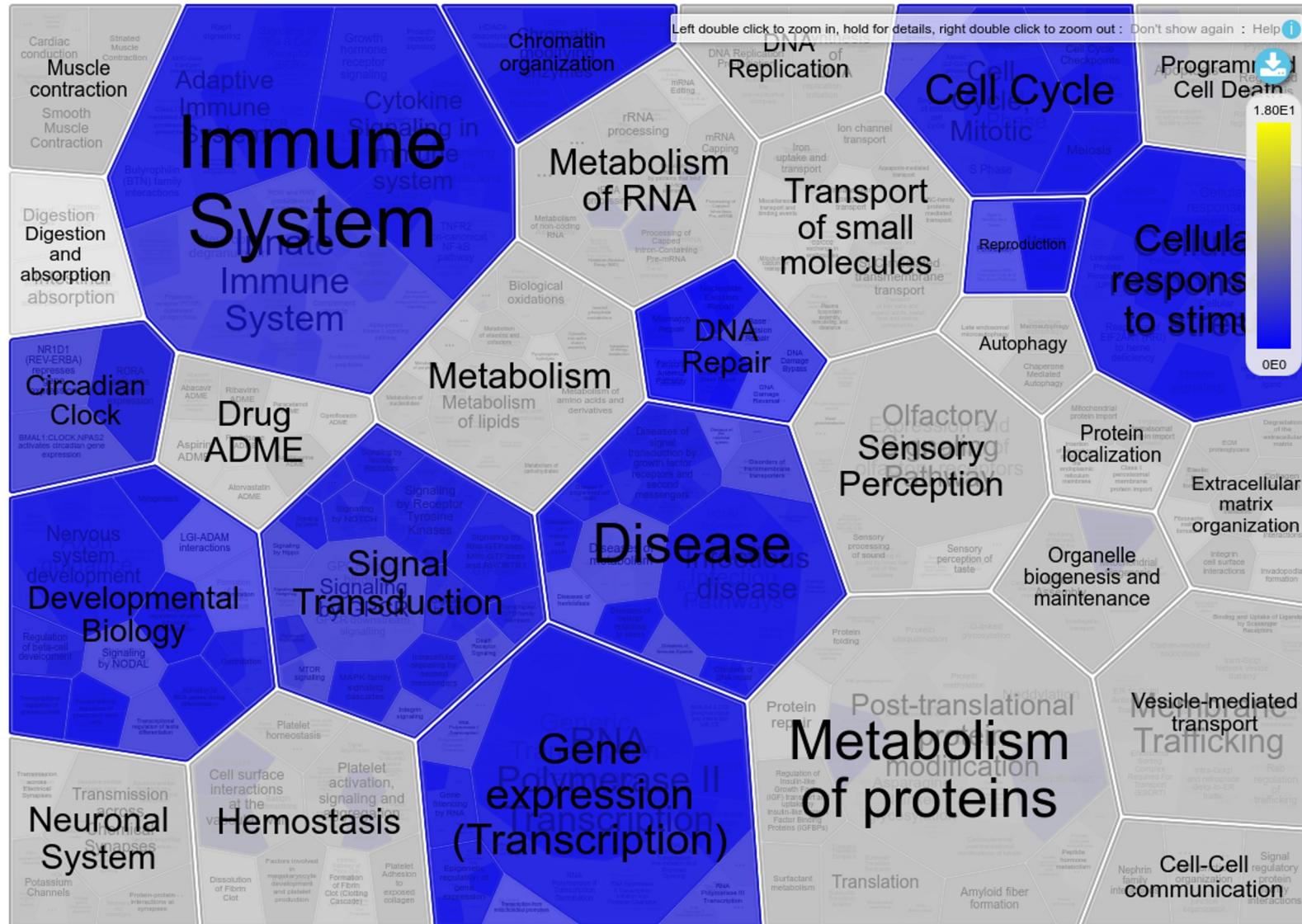
- Covered in Module 2.
- Most popular form of pathway/network analysis.
 - Overrepresentation analysis vs functional class scoring.
- Advantages:
 - Easy to perform.
 - Many good end-user tools.
 - Statistical model well worked out.
- Disadvantages:
 - Many possible gene sets;
 - Gene sets are heavily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

Reactome: Pathway Enrichment Analysis

The screenshot shows the Reactome Pathway Enrichment Analysis tool. The interface is divided into several sections:

- Analysis tools sidebar:** On the left, there are four main buttons: "Analyse gene list" (gear and chart icon), "Analyse gene expression" (DNA helix and magnifying glass icon), "Species Comparison" (two animal icons), and "Tissue Distribution" (organ icons).
- Header:** At the top, it says "Analysis tools" and has three tabs: "Your data" (selected), "Options", and "Analysis".
- Step 1:** A blue banner at the top says "Step 1: Select a file from your computer or paste your own data and click on the corresponding "Continue" button." It includes a "Choose file" button, a "No file chosen" message, and a "Continue" button.
- Data entry:** A large text area labeled "Paste your data to analyse or try example data sets:" contains a table of gene counts across somatic and germline samples. The table includes rows for TP53, APC, TERT, DICER1, BRAF, KDM6A, CUX1, KRAS, CTNNB1, ERBB2, HRAS, CDKN2A, CDKN2A(p14), ETV6, MDM2, and PRCA2.
- Some examples:** To the right of the data entry area, there is a list of pre-defined data sources:
 - UniProt accession list
 - Gene name list
 - Gene NCBI / Entrez list
 - Small molecules (ChEBI)
 - Small molecules (KEGG)
 - Microarray data
 - Metabolomics data
 - Cancer Gene Census (COSMIC)
 - Tissue Specific Expression (HPA)
- Buttons:** At the bottom of the data entry area are "Clear" and "Continue" buttons.
- Bottom text:** At the very bottom left, it says "Reactome v84" and "Click to learn more about our analysis tools". Below that, it states: "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



View: flattened | hierarchical

Reactome: Pathway Enrichment Analysis

reactome 3.7
84

Pathways for: Homo sapiens

Citation: Analysis: Tour: Layout:

Event Hierarchy: Metabolism (69/3,644) FDR: 1E0

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

Download W N S

trastuzumab pertuzumab

plasma membrane

ERBB2:TKis:ERBIN: HSP90:CDC37

ERBB2:trastuzumab: ERBIN:HSP90: CDC37

Ligand-Activated EGFR/ERBB3/ERBB4

ERBB2:ERBIN:HSP90:CDC37

NRG1/2: ERBB3

1.8E1

0E0 ?

ERBB2 heterodimers

ATP

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	pValue	FDR	Reacti 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	

Results 110 Not found Downloads

321-340 of 1,601

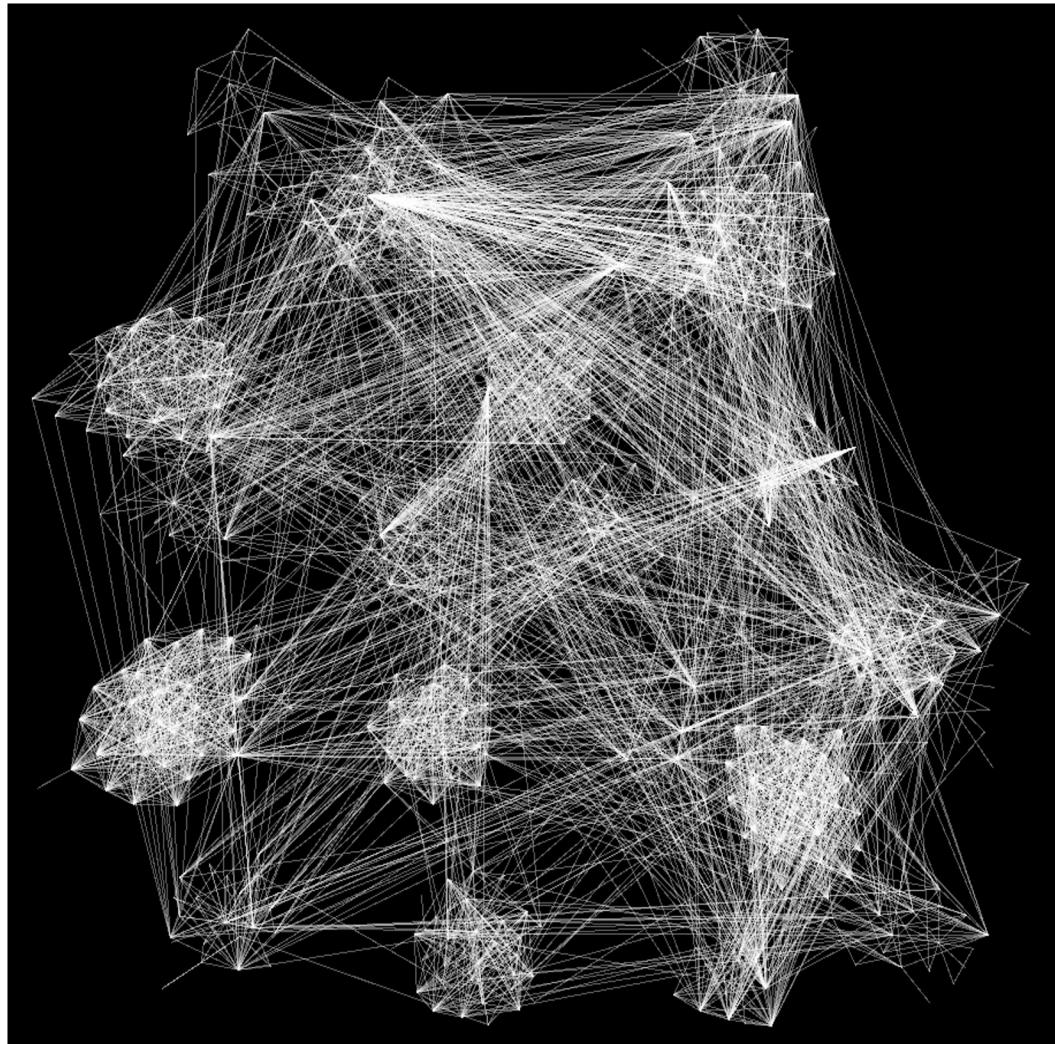
Module bioinformatics.ca

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.

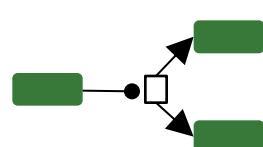
Reactome FI Network

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

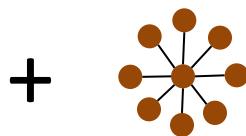


5% of network shown here

Reactome FI Network

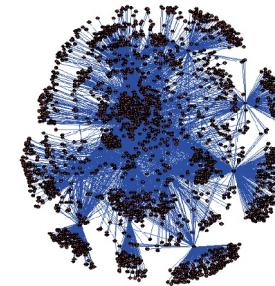


Curated
Pathway
Dbs

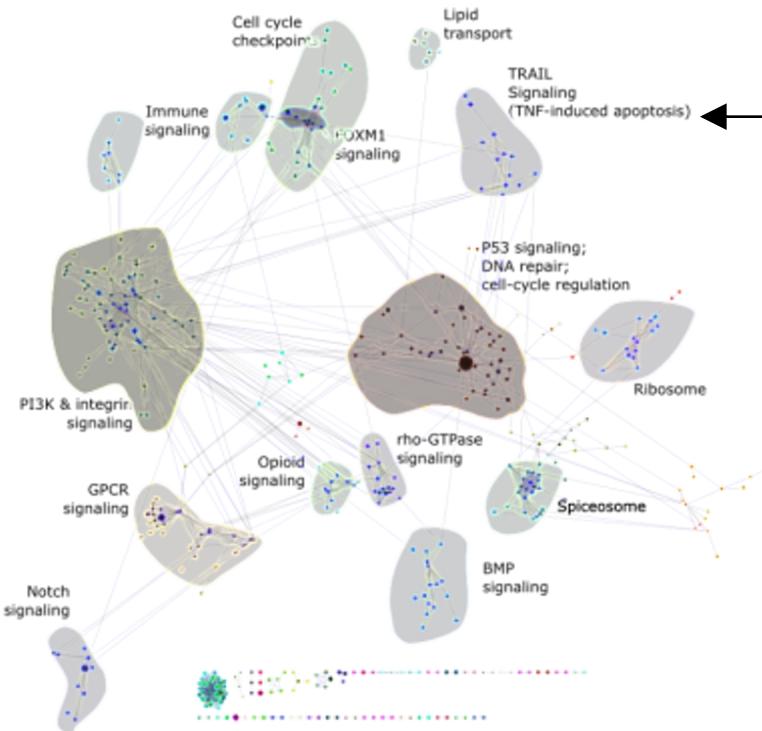


Uncurred
Interaction
Evidence

Machine Learning

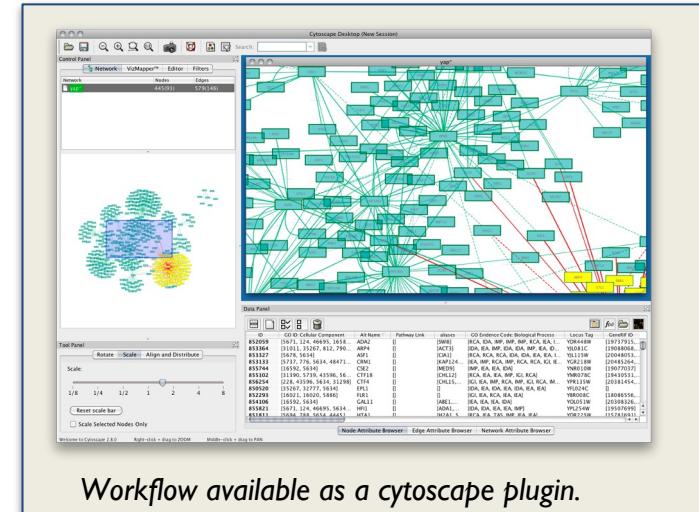


Reactome Functional Interaction Network
(~11,000 proteins; 270,000 interactions)



Disease “modules” (10-30)

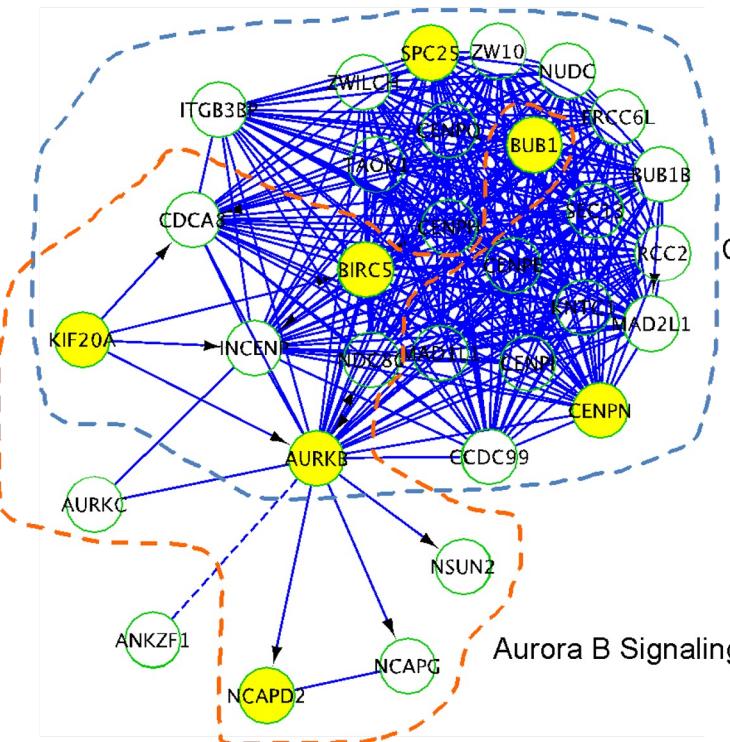
Extract and Cluster “Disease Genes”



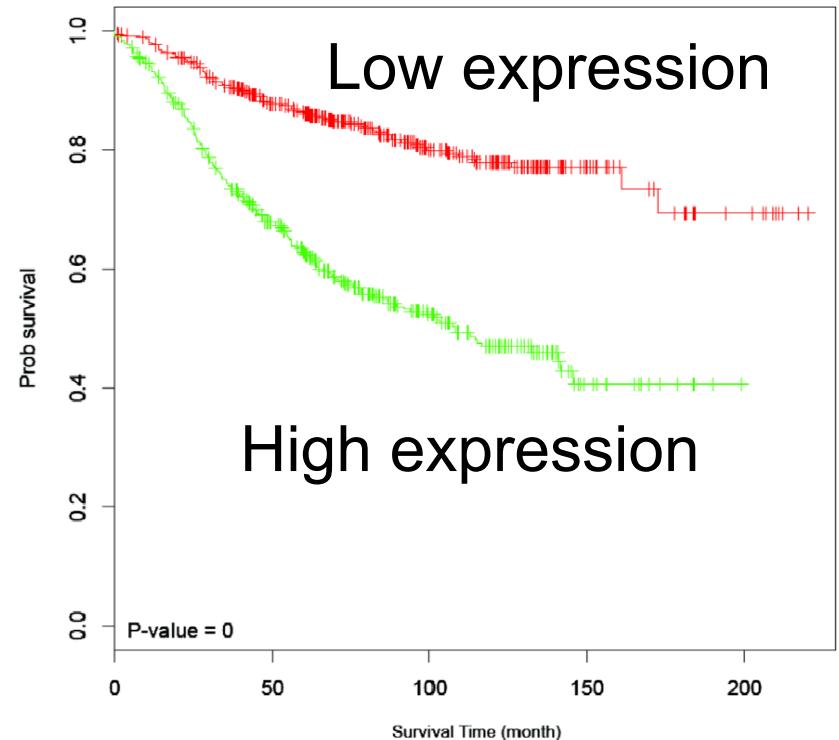
Workflow available as a cytoscape plugin.

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 metabolic 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)

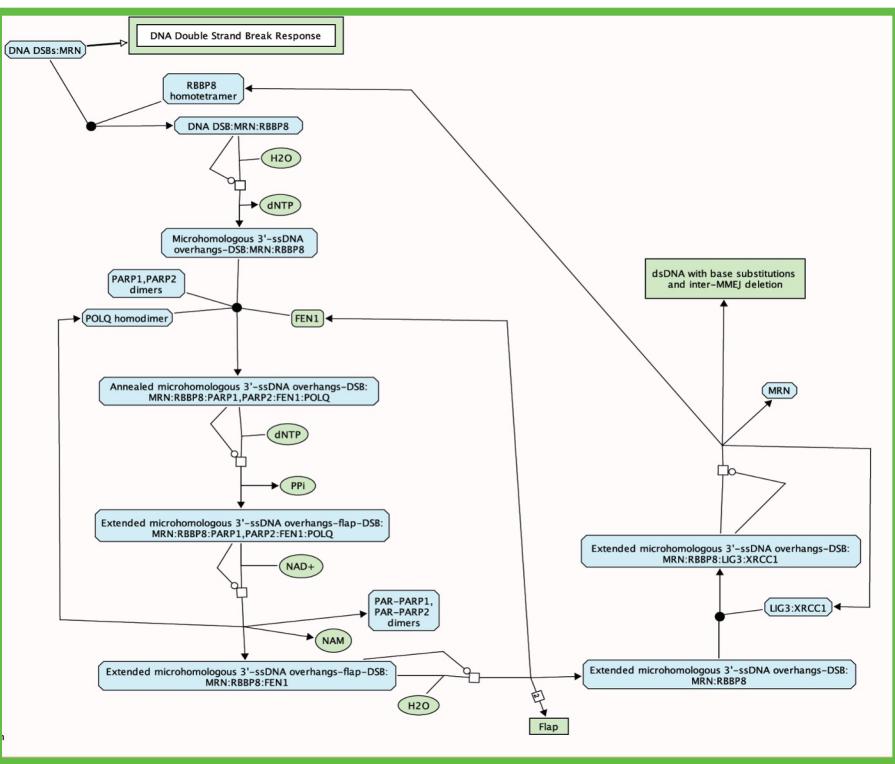
3) Pathway-Based Modeling

- 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

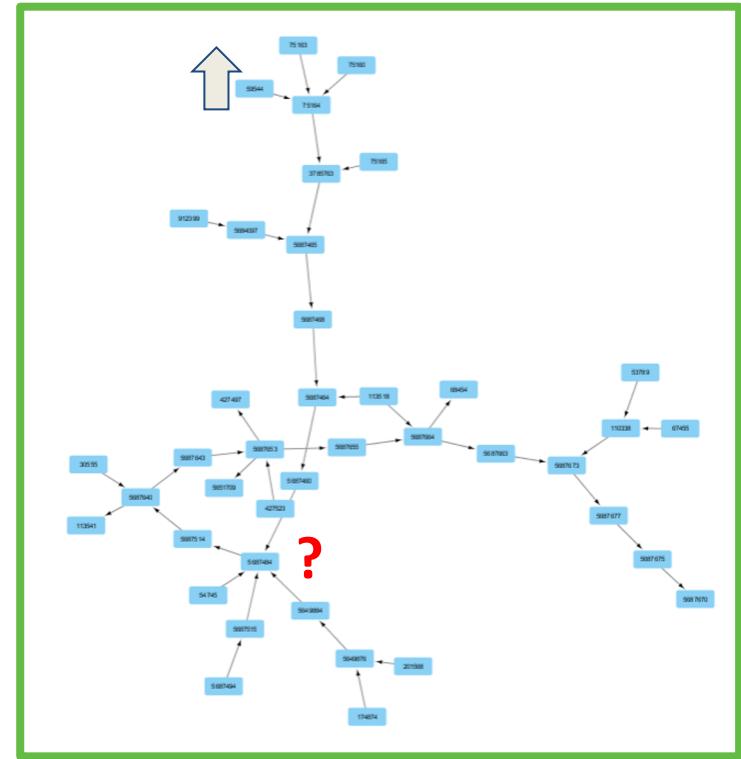
Types of Pathway-Based Modeling

- 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

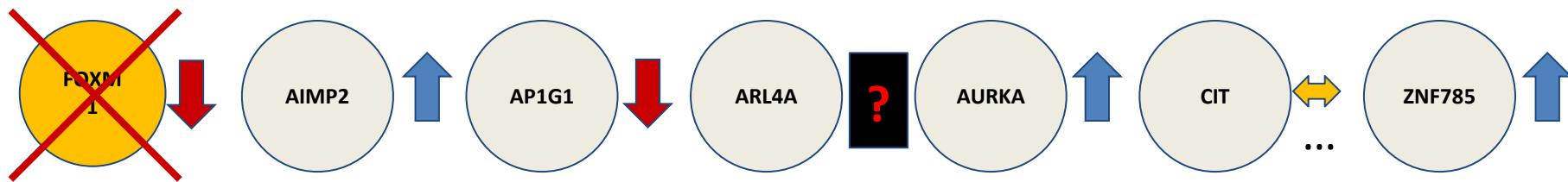


Generative AI Network Models

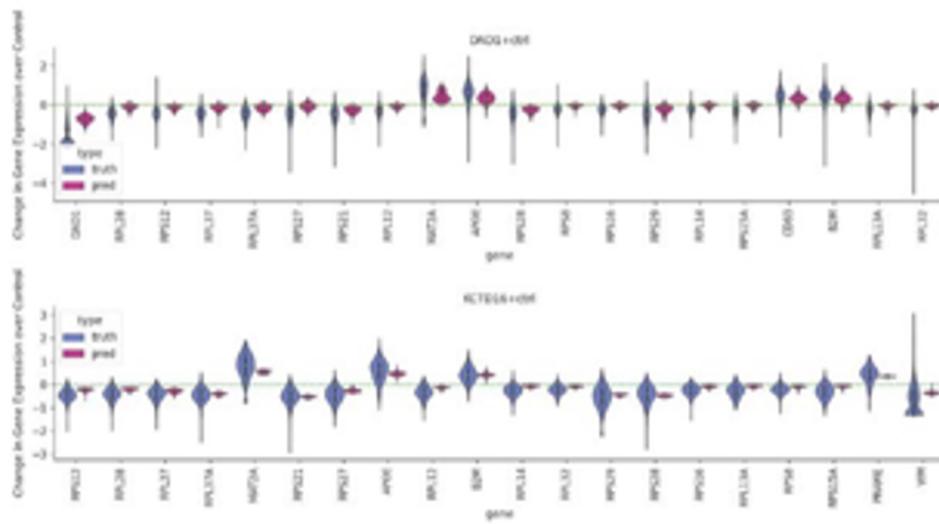
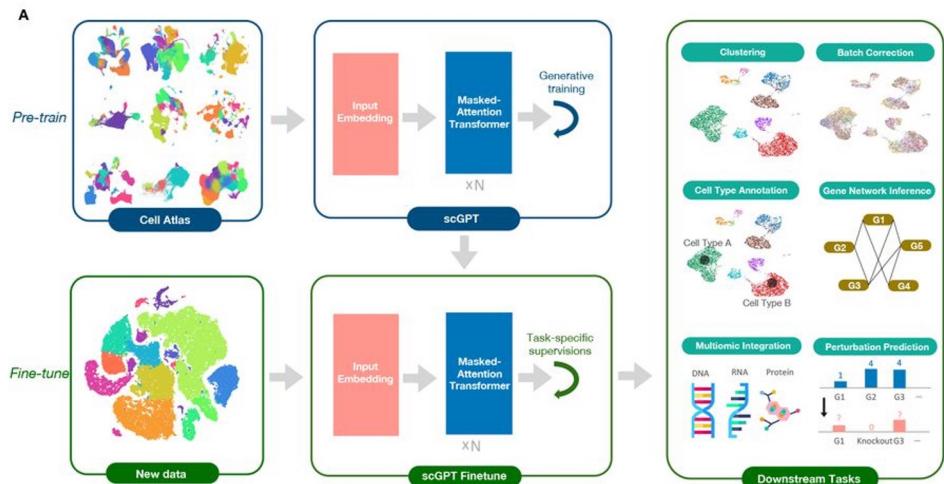
Large Language Models (GPT-3) are trained to predict masked text:

The quick brown ? jumped over the lazy dog.

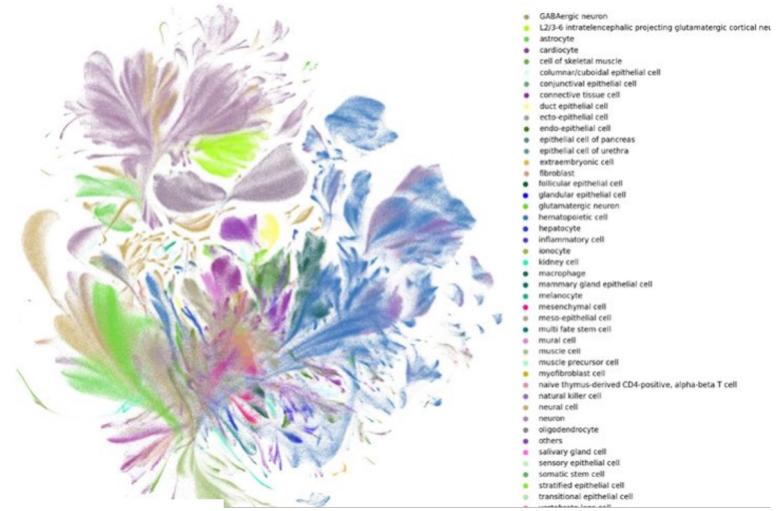
Generative pathway models are trained to predict gene network perturbations:



scGPT - Trained on scRNASeq of 33M Cells



E UMAP of sampled normal human cells using scGPT emb



Cui et al 2023
Nature Methods 10.1038/s41592-024-02201-0

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](#)

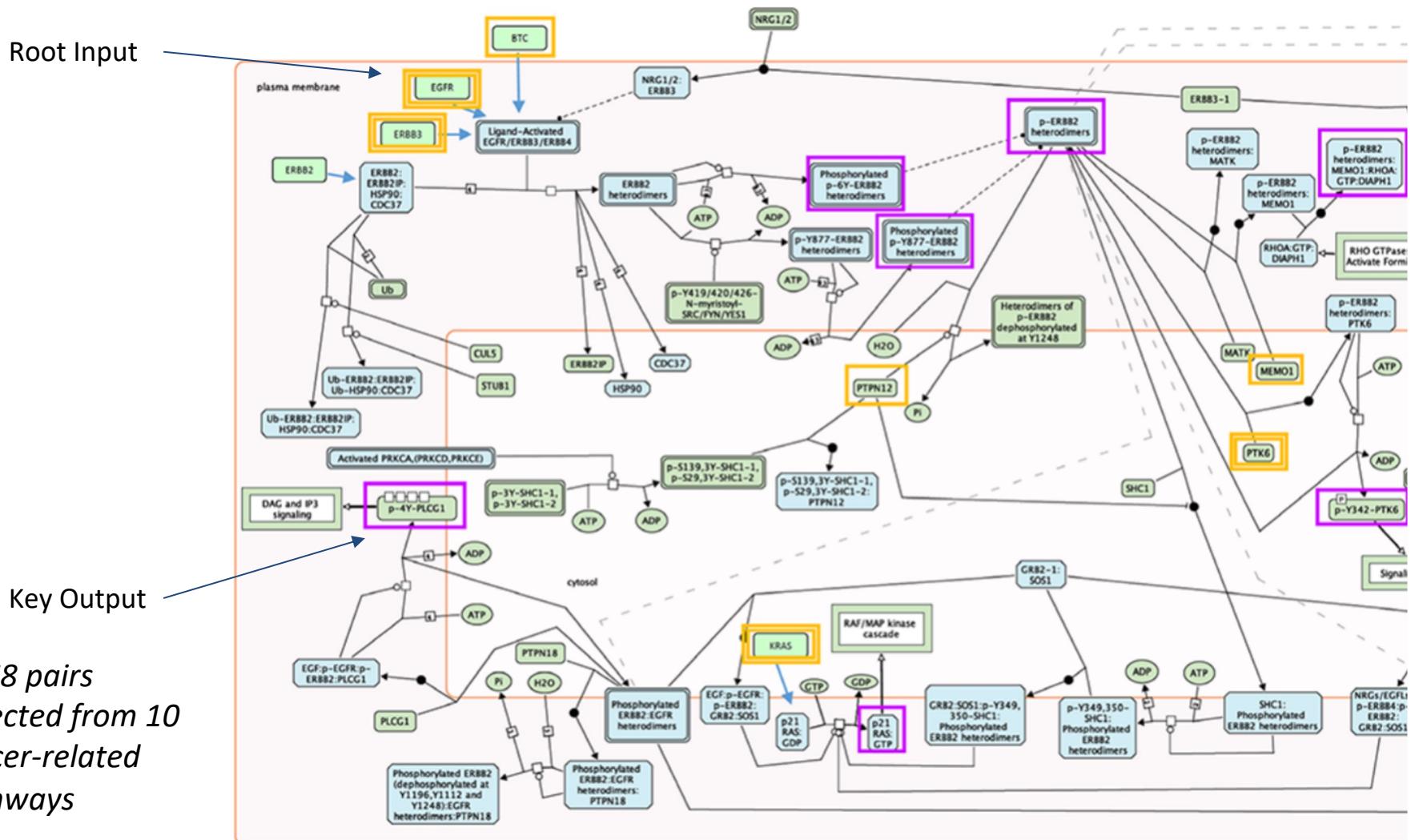
Evaluating the predictive accuracy of curated biological pathways in a public knowledgebase

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

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

- 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

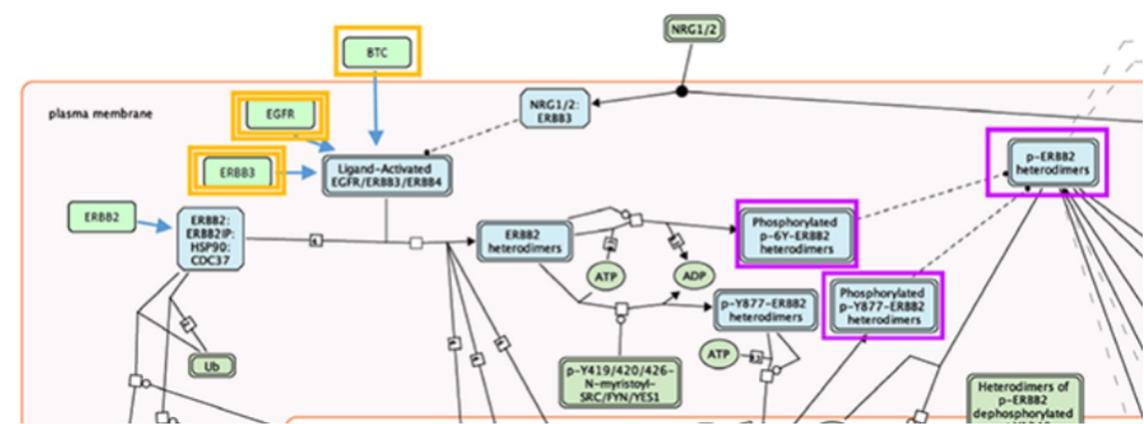


Step 2: Collect Empirical Results

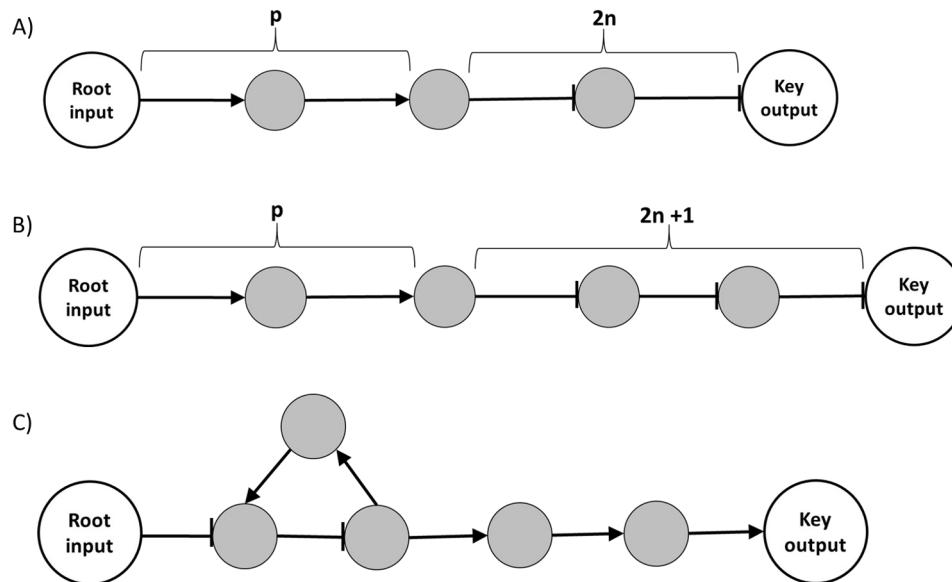
- Literature searches for functional genomics experiments in which key input was perturbed and effect on key output measured.
- 531 papers found, reporting 847 tested cases.

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

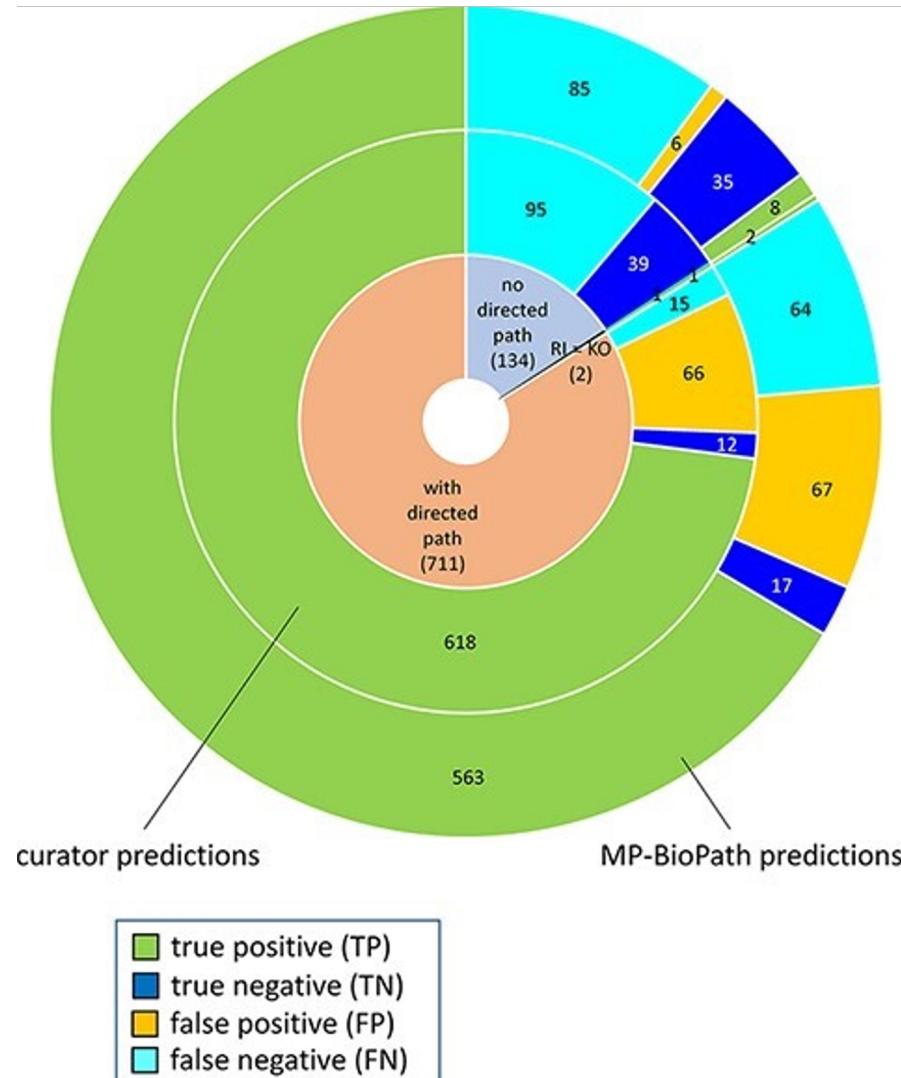
Curator Predictive Accuracy: **81%**

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

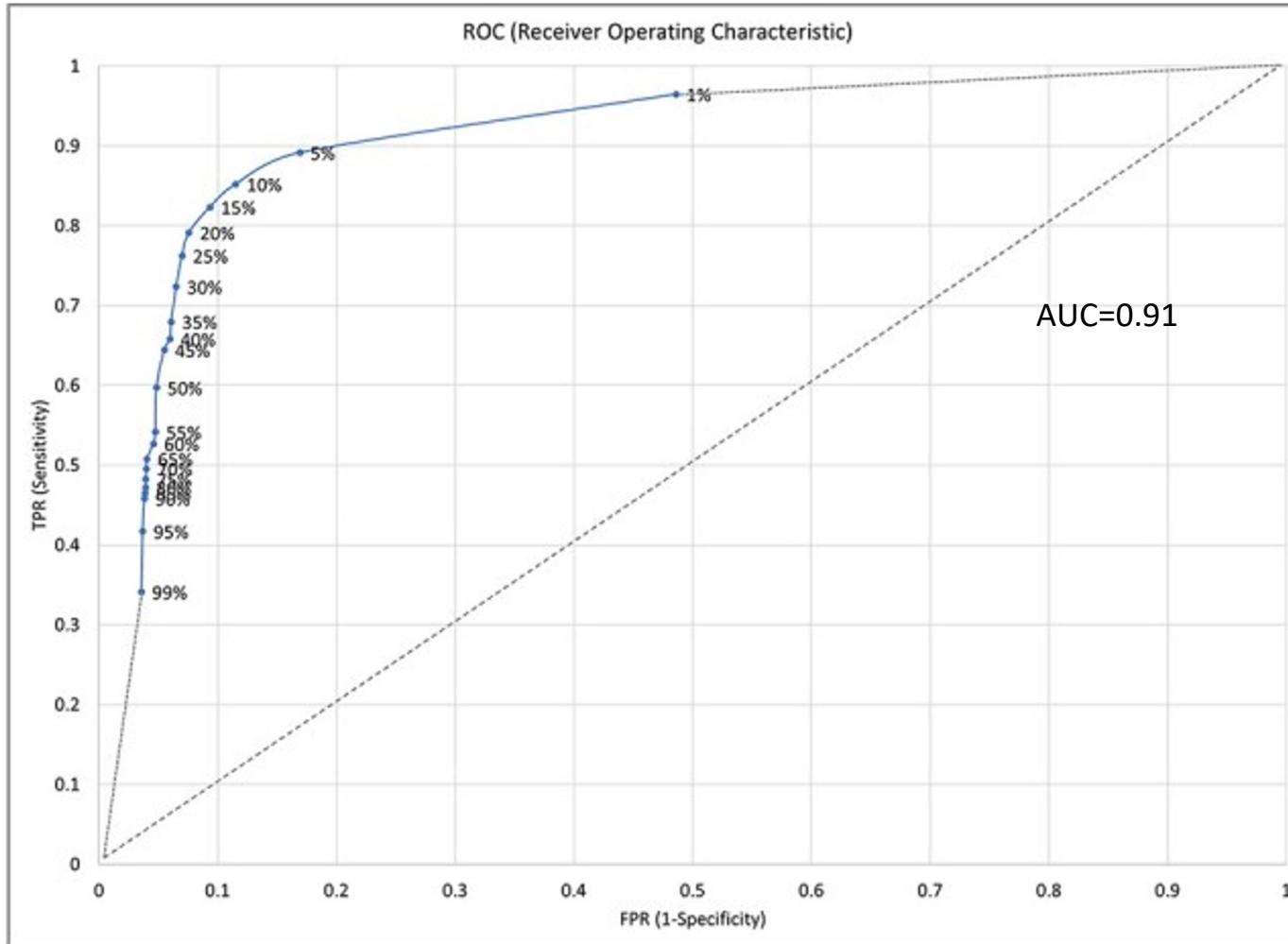
Random Guessing: **33%**

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>
- 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 FlViz App
 - <http://apps.cytoscape.org/apps/reactomefis>

Pathway Modeling

- CellNetAnalyzer
 - <http://www.ebi.ac.uk/research/saez-rodriguez/software>
- NetPhorest/NetworKIN
 - <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