knoweng BIG DATA TO KNOWLEDGE CENTER OF EXCELLENCE

NIH National Institutes of Health

MAYO CLINIC

Knowledge-guided Algorithms in Systems Biology

Charles Blatti
Research Scientist
National Center for Supercomputing Applications
University of Illinois Urbana-Champaign

Computational Genomics Course

Some Slides By **Amin Emad**
Assistant Professor at McGill University
http://www.ece.mcgill.ca/~aemad2/

Plan for this Lecture

Topic: Methods for analyzing omics datasets while integrating prior knowledge

- Systems Biology and Knowledge Networks
- Sample Clustering
- Gene Prioritization
- Gene Set Characterization

Emphasis: tools that take advantage of prior knowledge networks (KnowEnG)

Goal: understand basic concepts and aware of approaches and resources

Systems Biology

- Systems biology is the computational and mathematical modeling of complex biological systems.

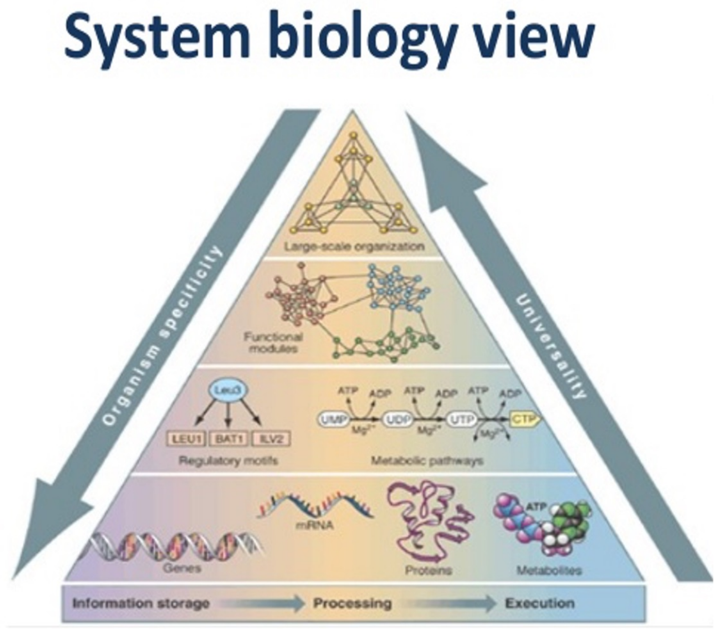


Figure from Oltvai , Z.N. and Barabasi Life's complexity pyramid.

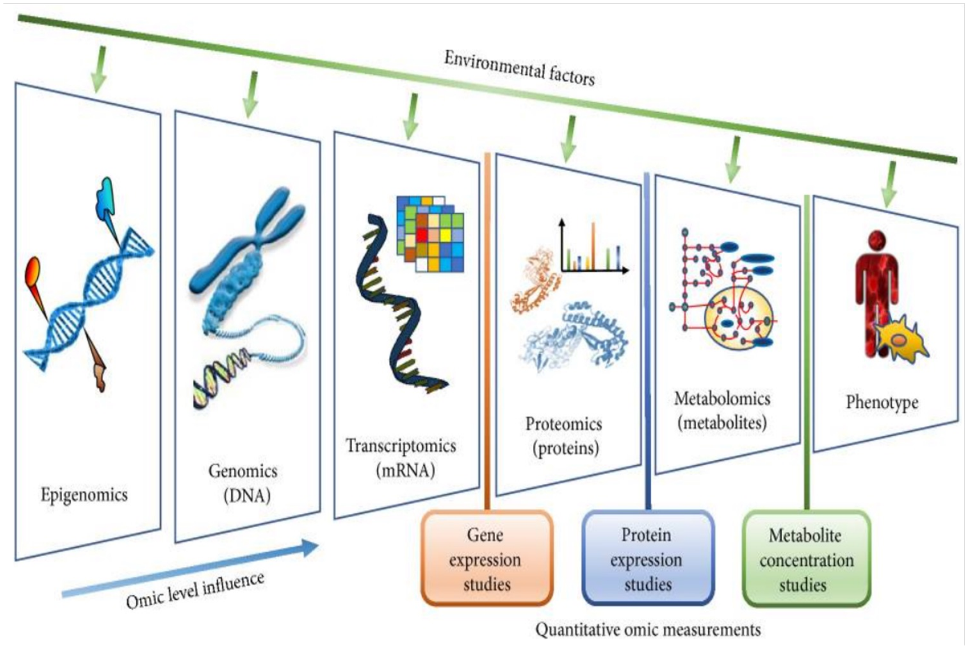


Figure from Angione, C. Human Systems Biology and Metabolic Modelling: A Review-From Disease Metabolism to Precision Medicine. Biomed Res Int 2019.

- Studies the interactions between the components of biological systems such as genes, proteins, metabolites, etc. (i.e. biological networks), and how these interactions give rise to the function and behavior of that system (phenotype)

Statistical and Machine Learning Methods

Applied to heterogeneous ‘omics and phenotype data and prior knowledge

Unsupervised Learning

Supervised Learning

- No training example exists and the goal is to learn structure in the data
- Training examples are provided with desired inputs and outputs to help learning the desired rule

Clustering
(subtyping)

Dimensionality Reduction
(data visualization)

Classification
(resistance group)

Regression
(survival time)

Supervised Feature Selection
(biomarkers)

Some Example Applications

Clustering
(subtyping)

Supervised Feature Selection
(biomarkers)

Classification
(resistance group)

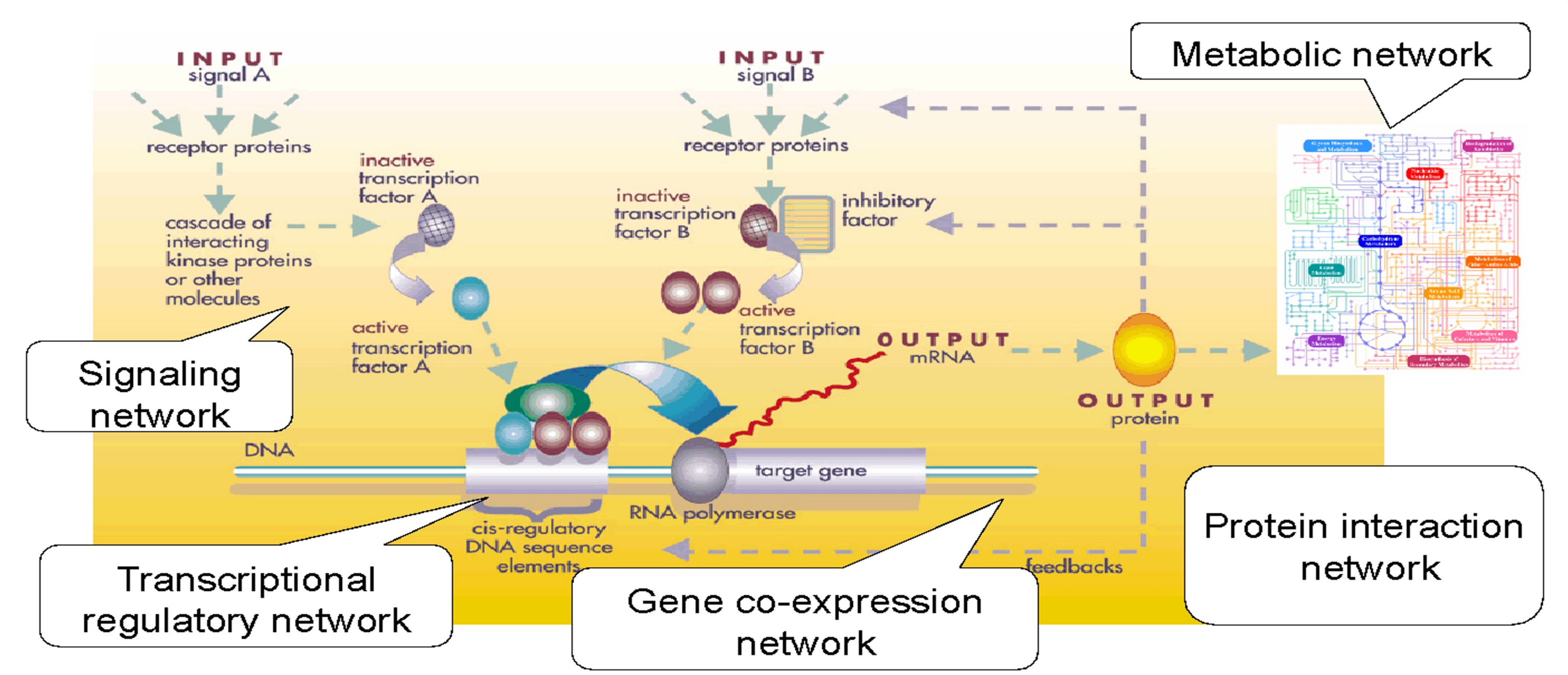
Regression
(survival time)

- Identifying the subtypes of a disease
- Identifying genes associated with a disease
- Predicting whether a patient is sensitive or resistant to a drug
- Predicting the survival probability of a cancer patient
- etc.

Prior Knowledge as Biological Networks

- Existing **prior knowledge** in literature captures known interactions within and across different levels of the biological systems

- **Knowledge Network** - a graphical representation of the interactions of the components of a biological systems


INPUT
signal A
receptor proteins
cascade of interacting kinase proteins or other molecules
inactive transcription factor A
active transcription factor A
INPUT
signal B
receptor proteins
inactive transcription factor B
inhibitory factor
active transcription factor B
OUTPUT
mRNA
OUTPUT
protein
DNA
target gene
RNA polymerase
cis-regulatory DNA sequence elements
feedbacks

Metabolic network

Signaling network

Transcriptional regulatory network

Gene co-expression network

Protein interaction network

Figure from https://commons.wikimedia.org/wiki/File:Gene_Regulatory_Network.jpg

Directed Biological Networks

Gene regulatory networks

- Nodes represent genes, proteins, etc.
- Edges show regulatory relationships between the nodes
- The network shows which entities (e.g. transcription factors) regulate the expression of each gene
- Edges can have meaningful weights

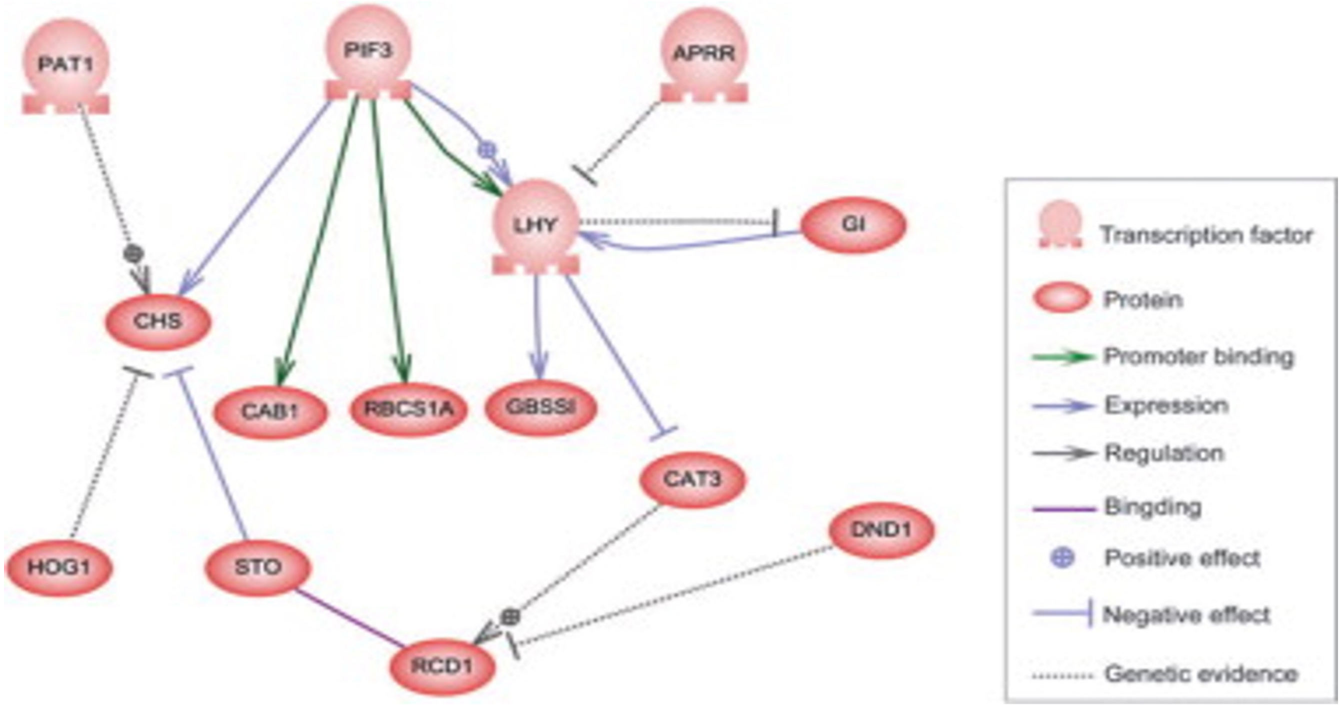
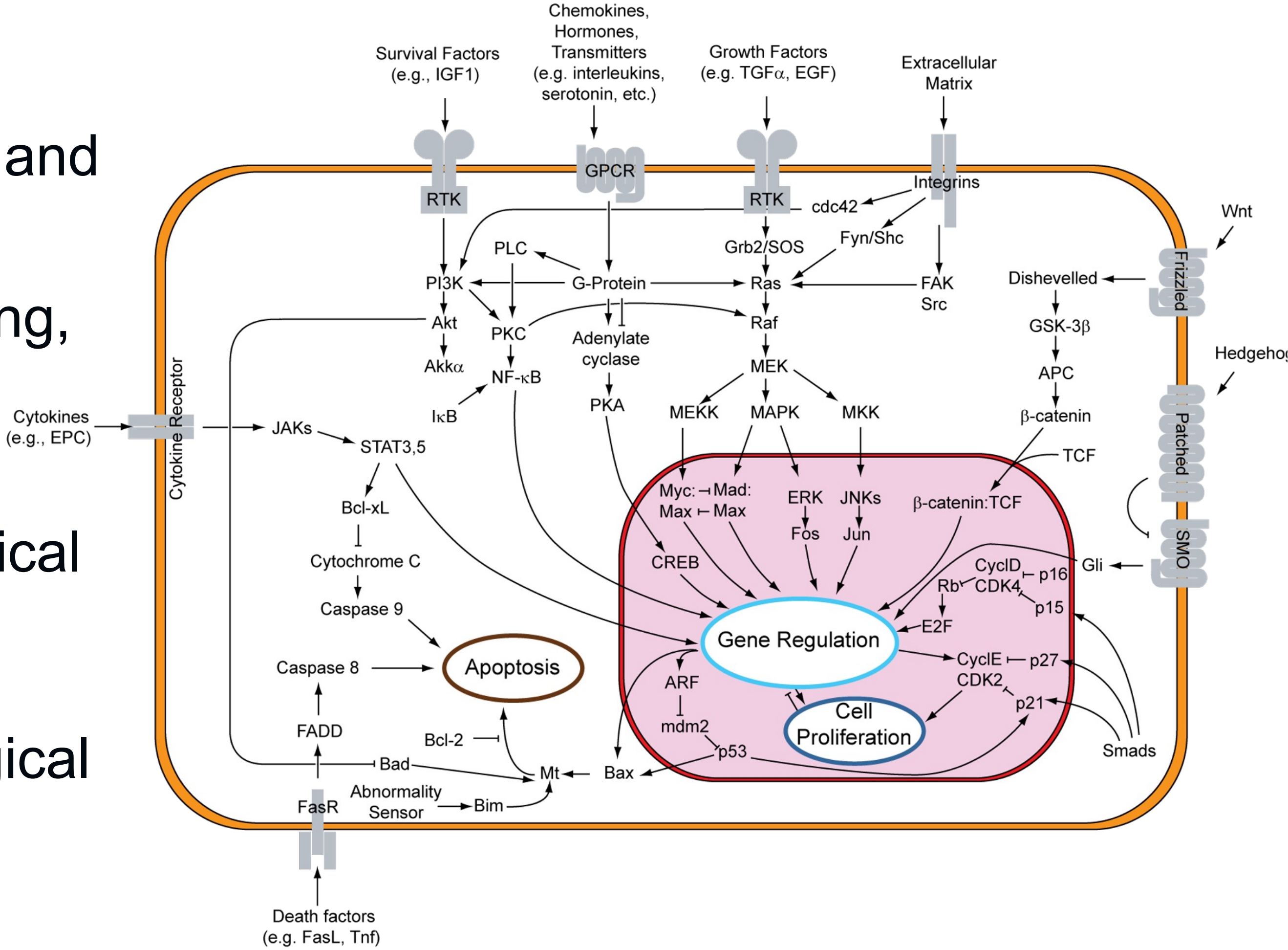


Figure from Song, et al. "Comparative transcriptional profiling and preliminary study on heterosis mechanism of super-hybrid rice." *Molecular plant* 3.6 (2010).

Directed Biological Networks

Signaling Networks

- Represents communications within and between cells
- Responsible for receiving, transmitting and processing information
- The network is a graphical representation of the interactions of the components of a biological systems



Signal Transduction Pathway

https://commons.wikimedia.org/wiki/File:Signal_transduction_pathways.svg

Experimental Networks

Protein-protein interaction networks

- Nodes represent proteins
- Edges show interactions between proteins
- Interactions usually refer to different levels of physical contact and proximity of protein molecules

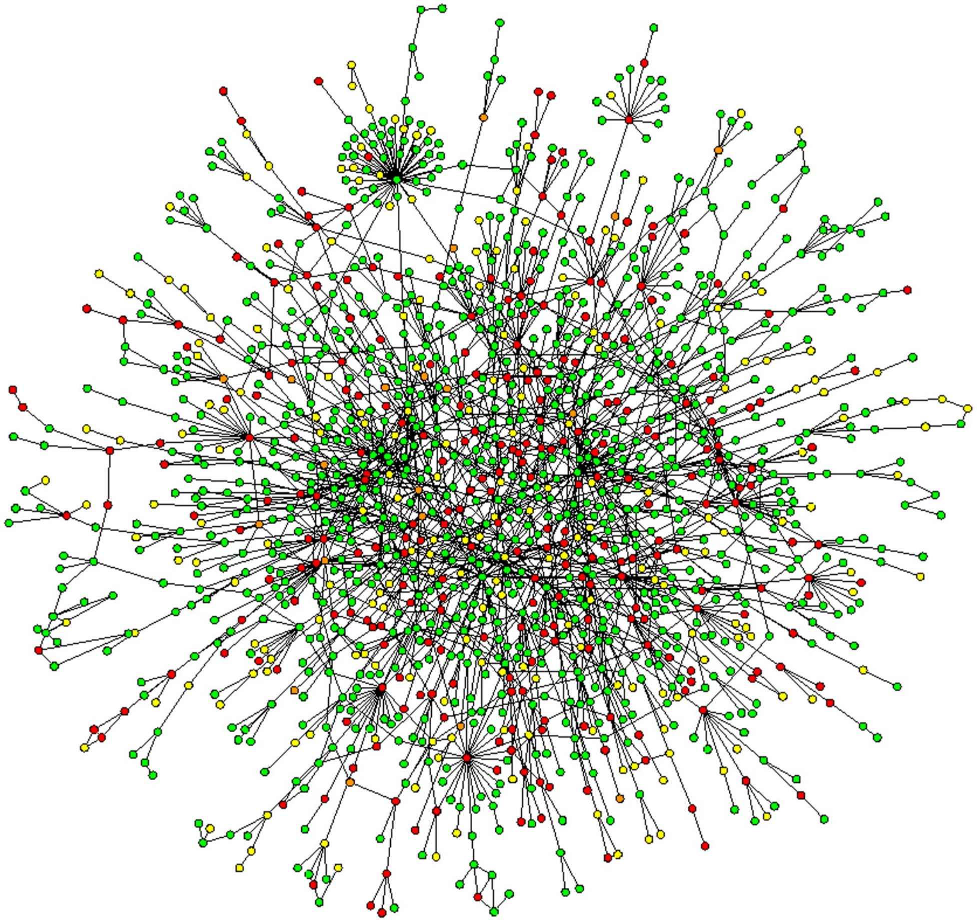
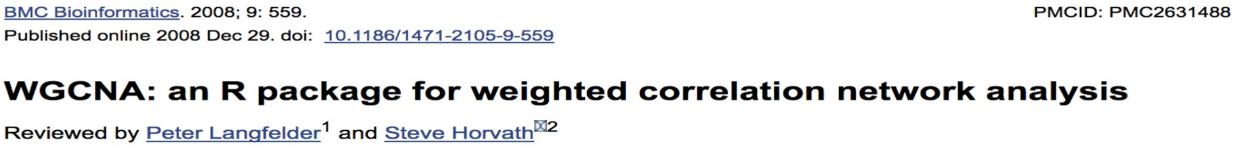


Figure from Jeong, Hawoong, et al. "Lethality and centrality in protein networks." *Nature* 411.6833 (2001).

Experimental Networks

Gene co-expression networks

- Nodes represent genes
- An edge exists between two genes that are highly co-expressed across different samples



BMC Bioinformatics. 2008; 9: 559. PMCID: PMC2631488
Published online 2008 Dec 29. doi: 10.1186/1471-2105-9-559

WGCNA: an R package for weighted correlation network analysis

Reviewed by Peter Langfelder[1] and Steve Horvath☒[2]

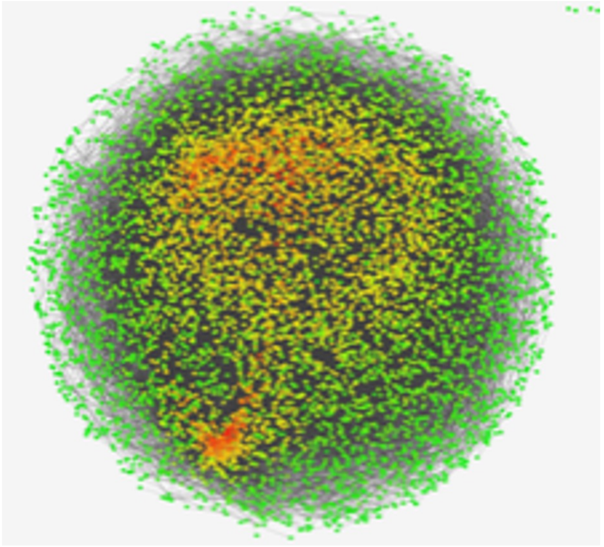


Figure from https://commons.wikimedia.org/wiki/File:Gene_co-expression_network_with_7221_genes_for_18_gastric_cancer_patients.png

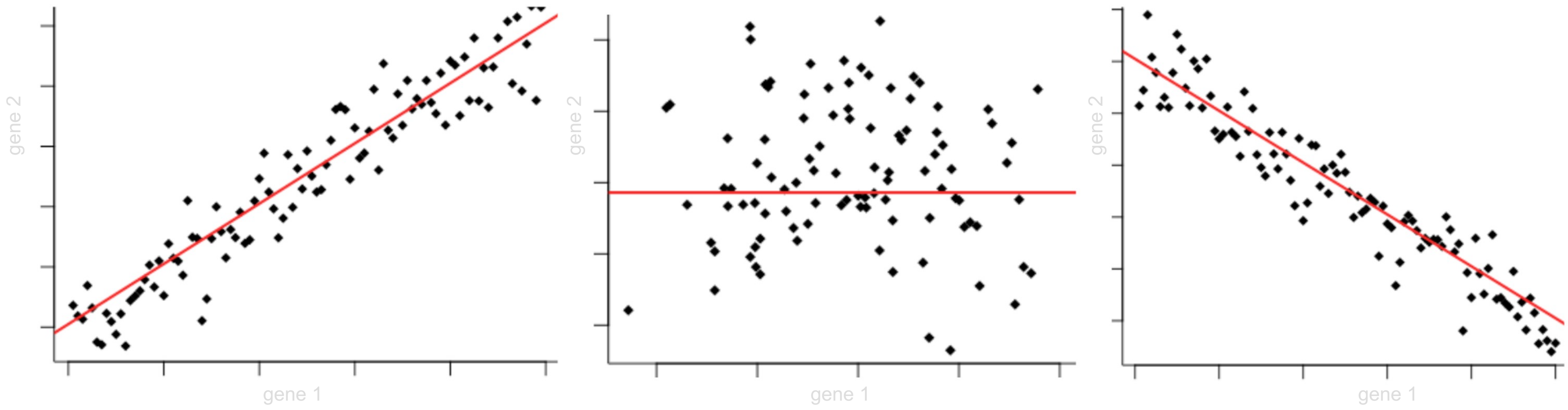
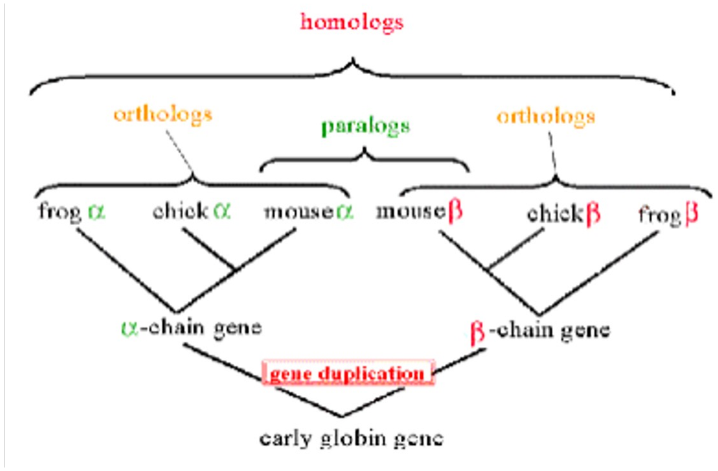


Figure from https://www.freecodecamp.org/news/how-machines-make-predictions-finding-correlations-in-complex-data-dfd9f0d87889/

Computational Networks

Evolutionary Conservation networks

- Nodes represent gene DNA or protein amino acid sequences
- Edges represent the similarity between the pair of sequences, the more similarly the more recently the nodes share an evolutionary history



https://www.ncbi.nlm.nih.gov/books/NBK1762/pdf/Bookshelf_NBK1762.pdf

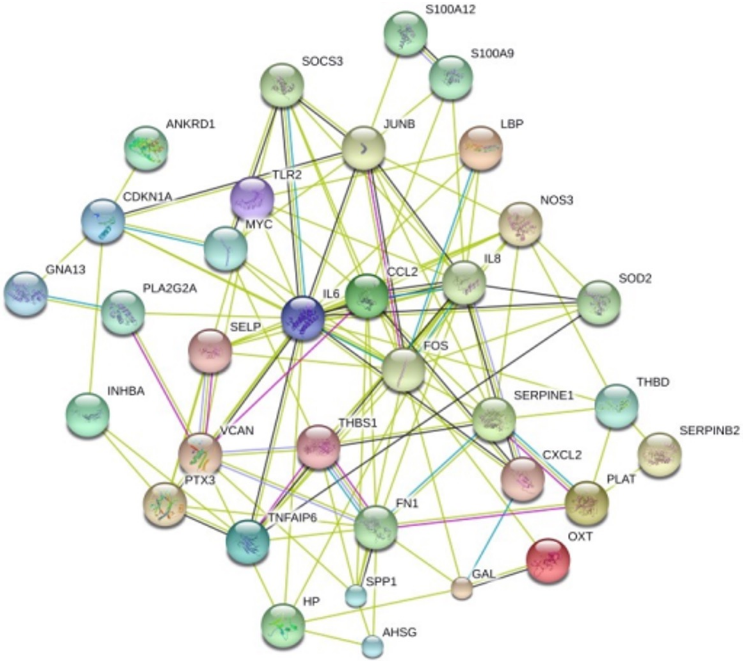


Figure from Yahaya, et al. "Gene expression changes associated with the airway wall response to injury." *PloS one* 8.4 (2013).

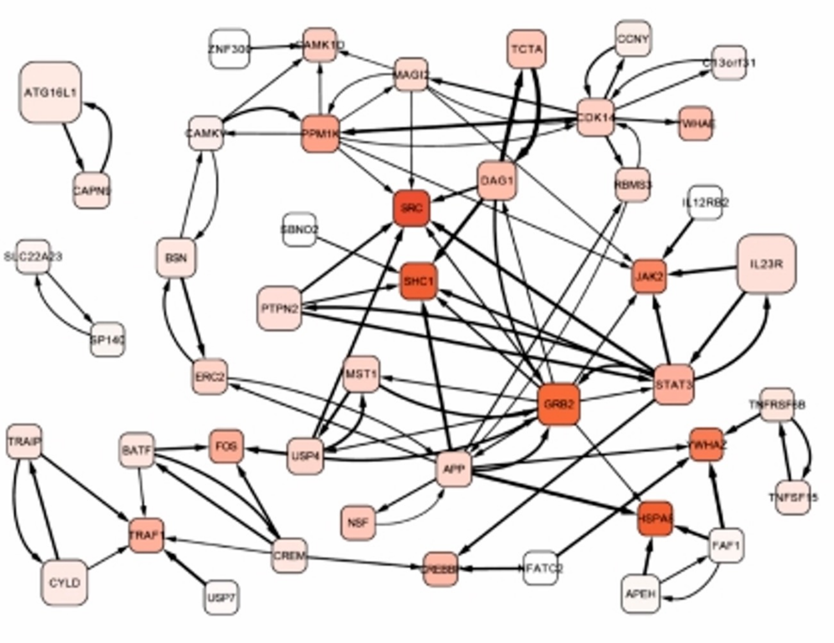
Text Mining networks

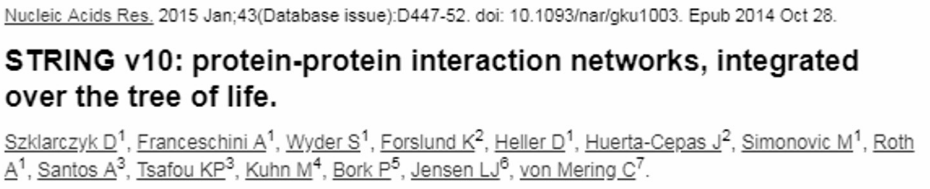
- Nodes represent gene entities
- Edges represent the frequency names, aliases, and synonyms for a pair of genes co-occur in literature abstracts

Computational Networks

Integrated networks

- Nodes represent gene or proteins
- Edges represent the weighted combination of normalized edge weights from many different types of network edges based on some predetermined criteria

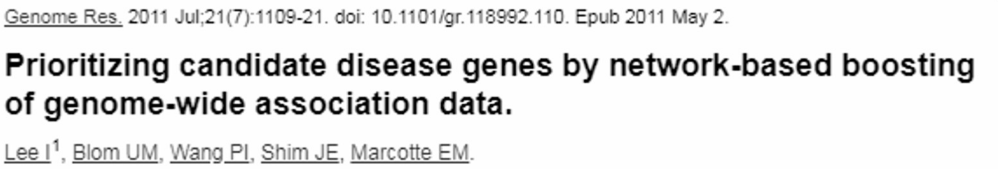




Nucleic Acids Res. 2015 Jan;43(Database issue):D447-52. doi: 10.1093/nar/gku1003. Epub 2014 Oct 28.

STRING v10: protein-protein interaction networks, integrated over the tree of life.

Szklarczyk D[1], Franceschini A[1], Wyder S[1], Forslund K[2], Heller D[1], Huerta-Cepas J[2], Simonovic M[1], Roth A[1], Santos A[3], Tsafou KP[3], Kuhn M[4], Bork P[5], Jensen LJ[6], von Mering C[7].



Genome Res. 2011 Jul;21(7):1109-21. doi: 10.1101/gr.118992.110. Epub 2011 May 2.

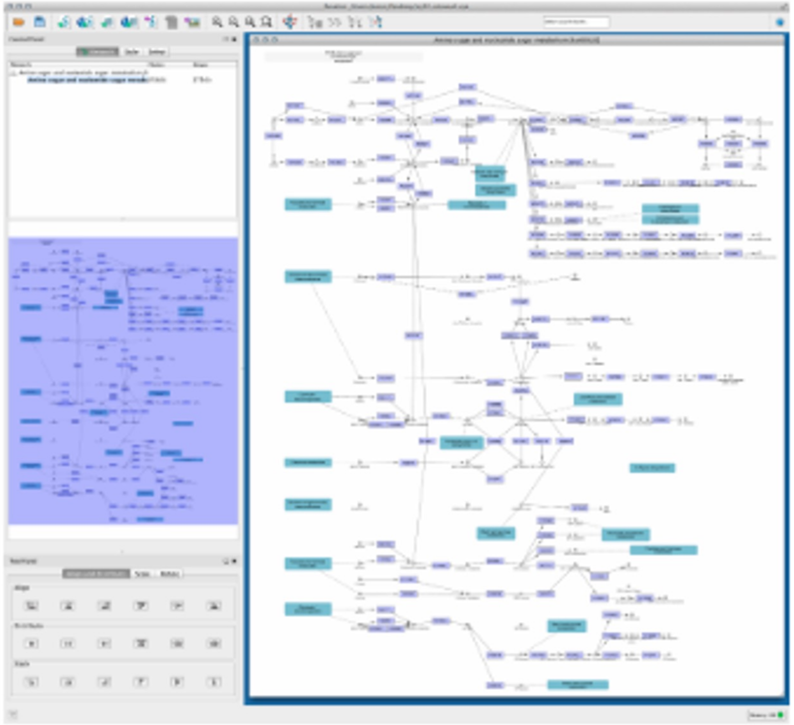
Prioritizing candidate disease genes by network-based boosting of genome-wide association data.

Lee I[1], Blom UM, Wang PI, Shim JE, Marcotte EM.

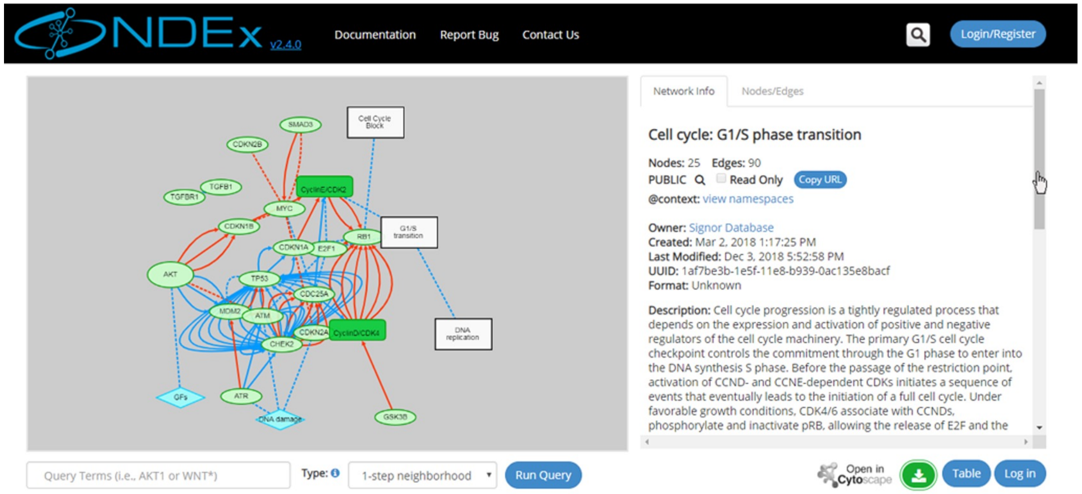
Visualizing / Sharing Biological Networks

Cytoscape
Network Data Integration, Analysis, and Visualization in a Box

NDEx



https://cytoscape.org/release_notes_3_2_1.html



Cell cycle: G1/S phase transition

Nodes: 25 **Edges:** 90
PUBLIC □ Read Only Copy URL
@context: view namespaces

Owner: Signor Database
Created: Mar 2, 2018 1:17:25 PM
Last Modified: Dec 3, 2018 5:52:58 PM
UUID: 1af7be3b-1e5f-11e8-b939-0ac135e8bacf
Format: Unknown

Description: Cell cycle progression is a tightly regulated process that depends on the expression and activation of positive and negative regulators of the cell cycle machinery. The primary G1/S cell cycle checkpoint controls the commitment through the G1 phase to enter into the DNA synthesis S phase. Before the passage of the restriction point, activation of CCND- and CCNE-dependent CDKs initiates a sequence of events that eventually leads to the initiation of a full cell cycle. Under favorable growth conditions, CDK4/6 associate with CCNDs, phosphorylate and inactivate pRB, allowing the release of E2F and the

Query Terms (i.e., AKT1 or WNT*) **Type:** 1-step neighborhood

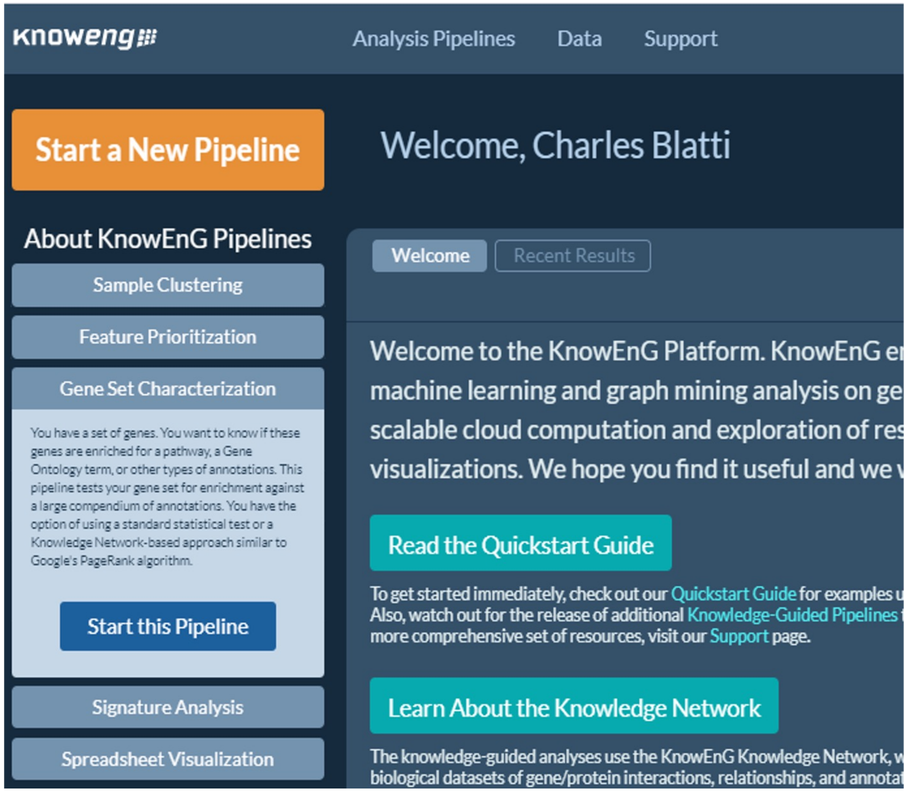
https://home.ndexbio.org/quick-start/

knoweng BIG DATA TO KNOWLEDGE CENTER OF EXCELLENCE

NIH National Institutes of Health

MAYO CLINIC

KnowEnG: Platform for Network-guided Analysis



knoweng

Welcome, Charles Blatti

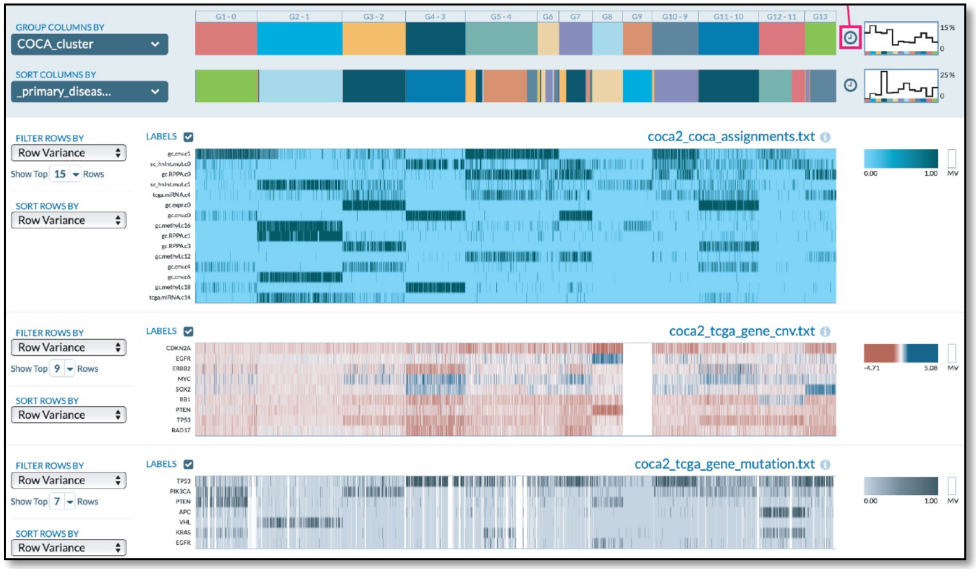
About KnowEnG Pipelines

You have a set of genes. You want to know if these genes are enriched for a pathway, a Gene Ontology term, or other types of annotations. This pipeline tests your gene set for enrichment against a large compendium of annotations. You have the option of using a standard statistical test or a Knowledge Network-based approach similar to Google's PageRank algorithm.

Welcome to the KnowEnG Platform. KnowEnG e
machine learning and graph mining analysis on ge
scalable cloud computation and exploration of res
visualizations. We hope you find it useful and we v

To get started immediately, check out our Quickstart Guide for examples u
Also, watch out for the release of additional Knowledge-Guided Pipelines t
more comprehensive set of resources, visit our Support page.

The knowledge-guided analyses use the KnowEnG Knowledge Network, w
biological datasets of gene/protein interactions, relationships, and annotat



GROUP COLUMNS BY: COCA_cluster

SORT COLUMNS BY: _primary_diseas...

coca2_coca_assignments.txt — FILTER ROWS BY: Row Variance; Show Top 15 Rows; SORT ROWS BY: Row Variance; LABELS ☑

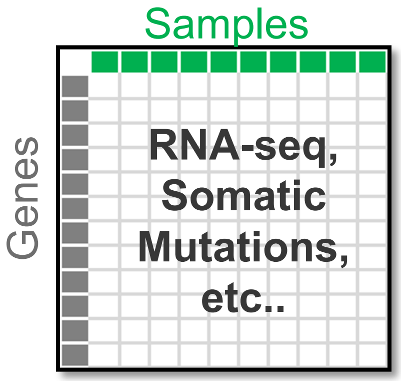
coca2_tcga_gene_cnv.txt — FILTER ROWS BY: Row Variance; Show Top 9 Rows; SORT ROWS BY: Row Variance; LABELS ☑

coca2_tcga_gene_mutation.txt — FILTER ROWS BY: Row Variance; Show Top 7 Rows; SORT ROWS BY: Row Variance; LABELS ☑

https://knoweng.org/analyze/

KnowEnG: *Knowledge Engine for Genomics*

- ‘omics Data Analysis Pipelines

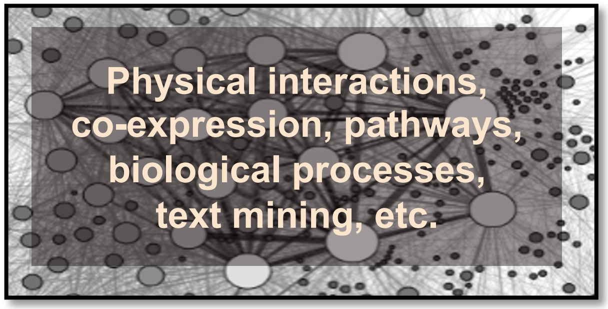


Samples

Genes

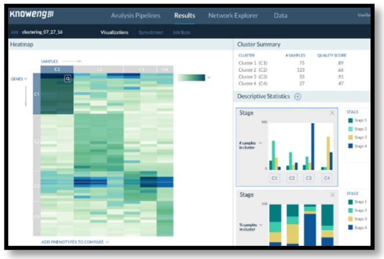
RNA-seq, Somatic Mutations, etc..

- Using Prior Knowledge



Physical interactions, co-expression, pathways, biological processes, text mining, etc.

- In a Scalable Cloud Platform



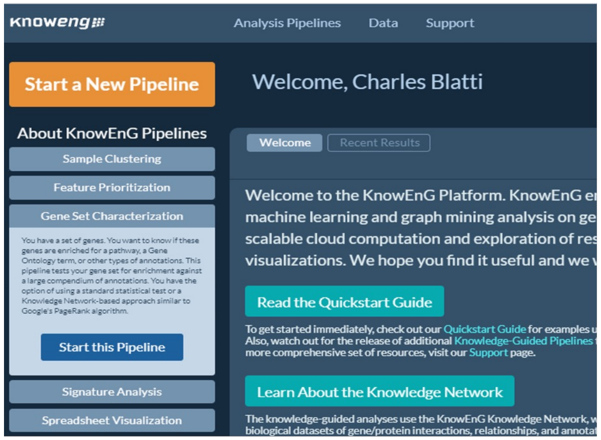
Cluster Summary

CLUSTER	# SAMPLES	QUALITY SCORE
Cluster 1 (C1)	75	.89
Cluster 2 (C2)	123	.66
Cluster 3 (C3)	55	.91
Cluster 4 (C4)	27	.47

KnowEnG Pipelines and User Interface

knoweng BIG DATA TO KNOWLEDGE CENTER OF EXCELLENCE

- Sample Clustering
  - What are the separate transcriptomic subtypes of patients and how do they relate to outcome?
- Feature(Gene) Prioritization
  - What genes are differentially expressed with respect to viral shedding
- Gene Set Characterization
  - What pathways do these differentially expressed genes relate to?
- Signature Analysis
  - Given a new patient, what subtype does their profile most resemble?
- Spreadsheet Visualization
  - Given multiple omics and clinical datasets on patient samples, what features relate to selected phenotypes?



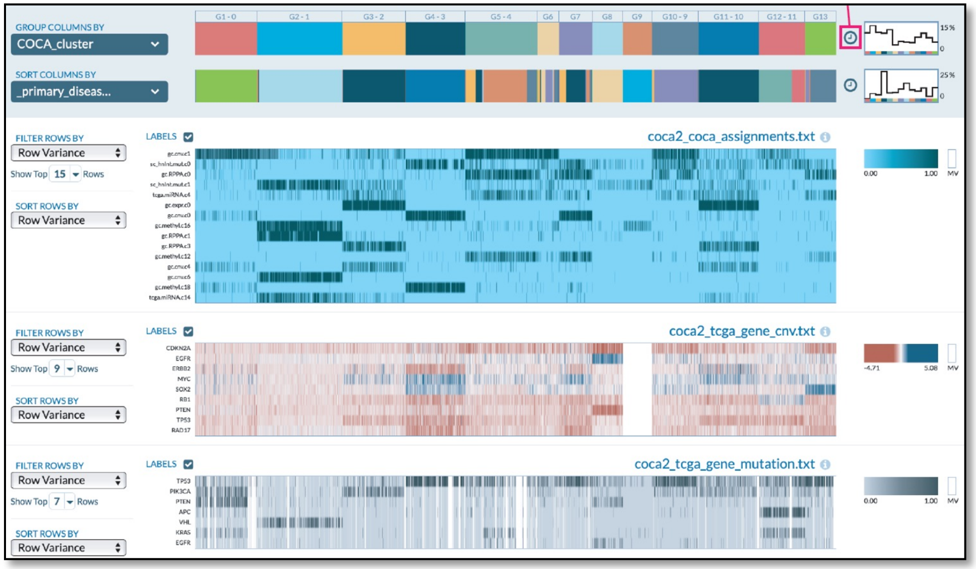
Welcome, Charles Blatti

You have a set of genes. You want to know if these genes are enriched for a pathway, a Gene Ontology term, or other types of annotations. This pipeline tests your gene set for enrichment against a large compendium of annotations. You have the option of using a standard statistical test or a Knowledge Network-based approach similar to Google's PageRank algorithm.

Welcome to the KnowEnG Platform. KnowEnG e
machine learning and graph mining analysis on ge
scalable cloud computation and exploration of re
visualizations. We hope you find it useful and we

To get started immediately, check out our Quickstart Guide for examples u
Also, watch out for the release of additional Knowledge-Guided Pipelines
more comprehensive set of resources, visit our Support page.

The knowledge-guided analyses use the KnowEnG Knowledge Network, w
biological datasets of gene/protein interactions, relationships, and annota



GROUP COLUMNS BY: COCA_cluster

SORT COLUMNS BY: _primary_diseas...

coca2_coca_assignments.txt

FILTER ROWS BY: Row Variance

Show Top 15 Rows

SORT ROWS BY: Row Variance

coca2_tcga_gene_cnv.txt

FILTER ROWS BY: Row Variance

Show Top 9 Rows

SORT ROWS BY: Row Variance

coca2_tcga_gene_mutation.txt

FILTER ROWS BY: Row Variance

Show Top 7 Rows

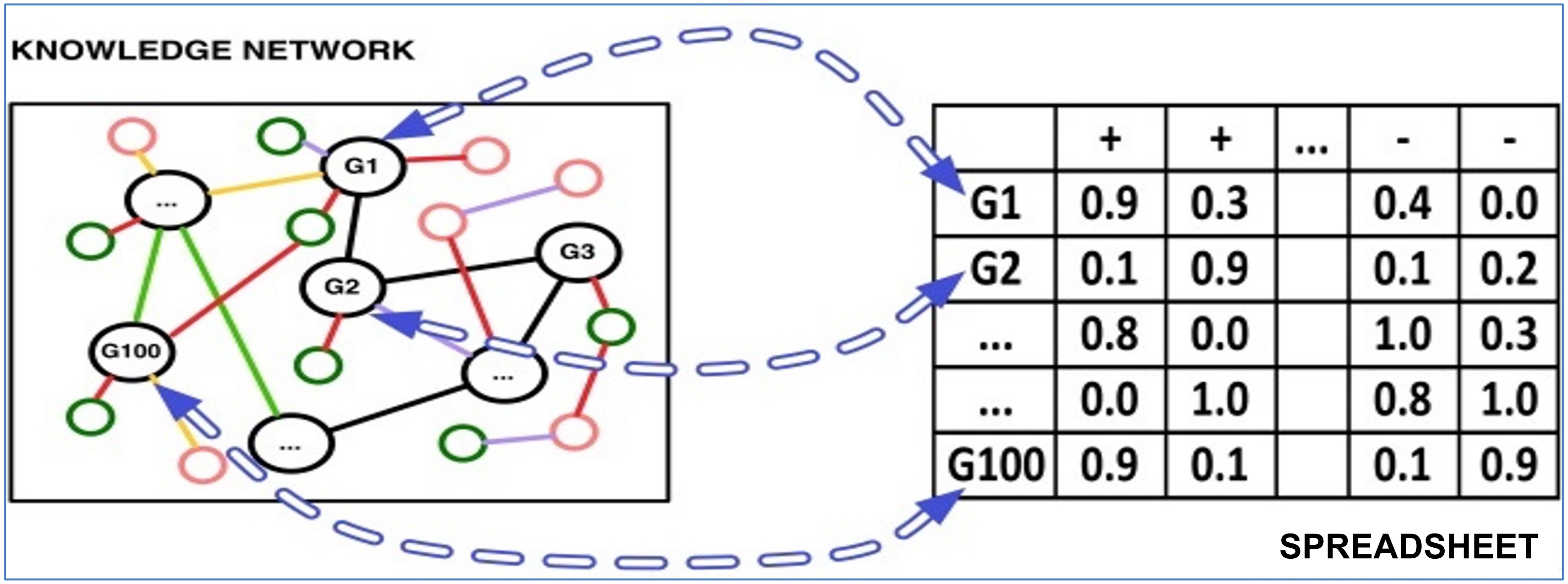
SORT ROWS BY: Row Variance

Analysis Pipelines Using Prior Knowledge

- **Knowledge Network (KN)**: heterogeneous graph whose nodes and edges encodes major public data sets as a network represented by genes/proteins, their properties, and relationships

- **Omics data:** a **spreadsheet** (rows = genes or proteins) to be analyzed

Graph Mining

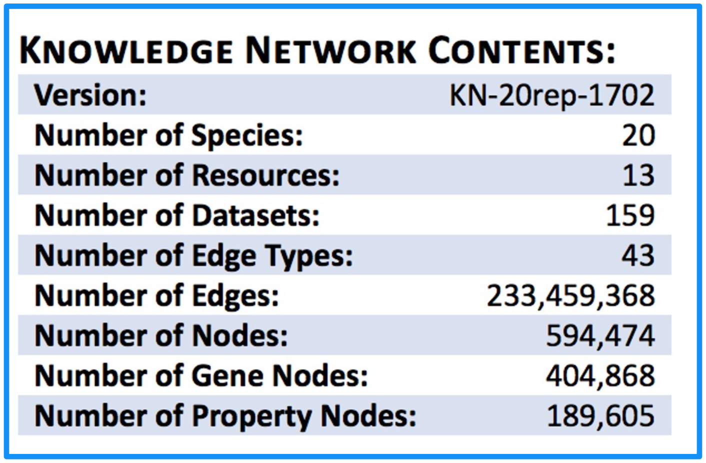


	+	+	...	-	-
G1	0.9	0.3		0.4	0.0
G2	0.1	0.9		0.1	0.2
...	0.8	0.0		1.0	0.3
...	0.0	1.0		0.8	1.0
G100	0.9	0.1		0.1	0.9

Machine learning

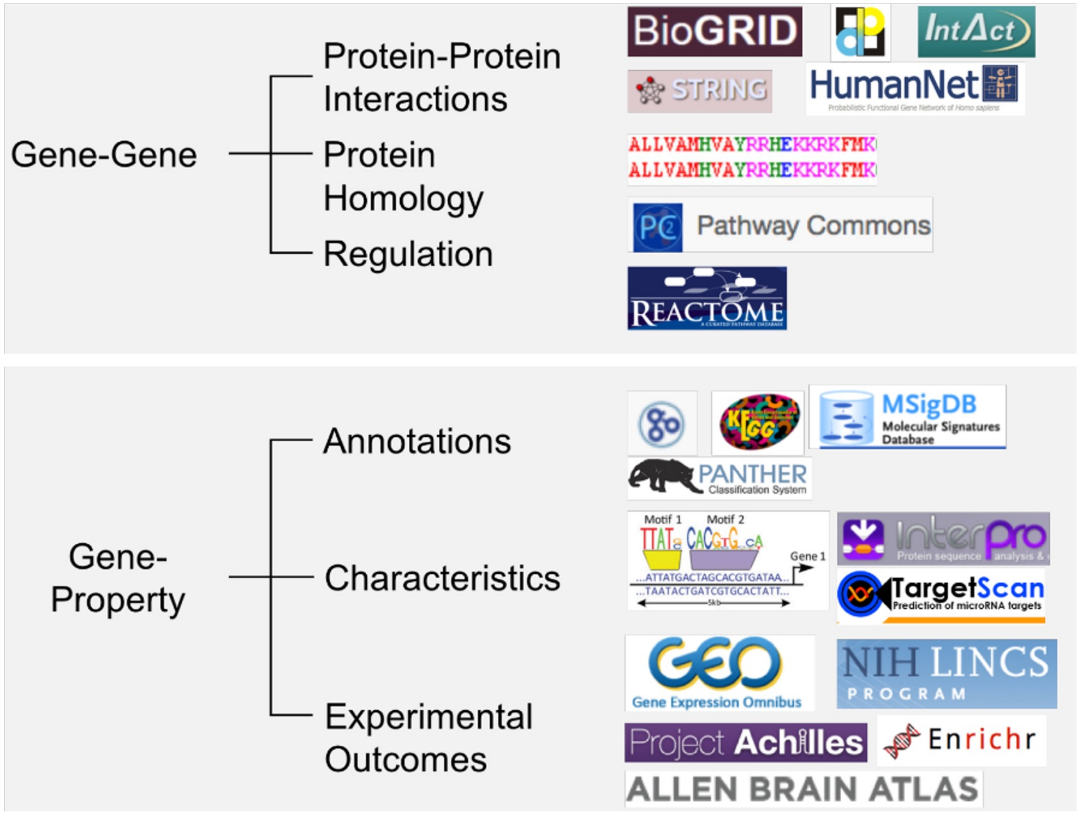
Knowledge network + user spreadsheet

KnowEnG Prior Knowledge Networks



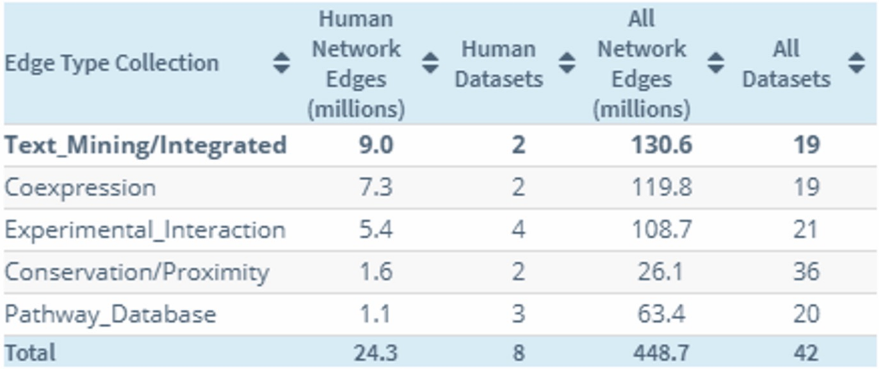
Knowledge Network Contents:

Version:	KN-20rep-1702
Number of Species:	20
Number of Resources:	13
Number of Datasets:	159
Number of Edge Types:	43
Number of Edges:	233,459,368
Number of Nodes:	594,474
Number of Gene Nodes:	404,868
Number of Property Nodes:	189,605

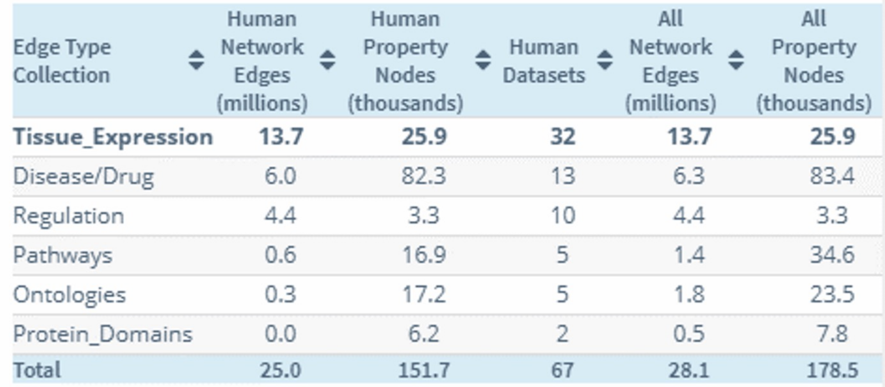


Gene-Gene
- Protein-Protein Interactions
- Protein Homology
- Regulation

Gene-Property
- Annotations
- Characteristics
- Experimental Outcomes



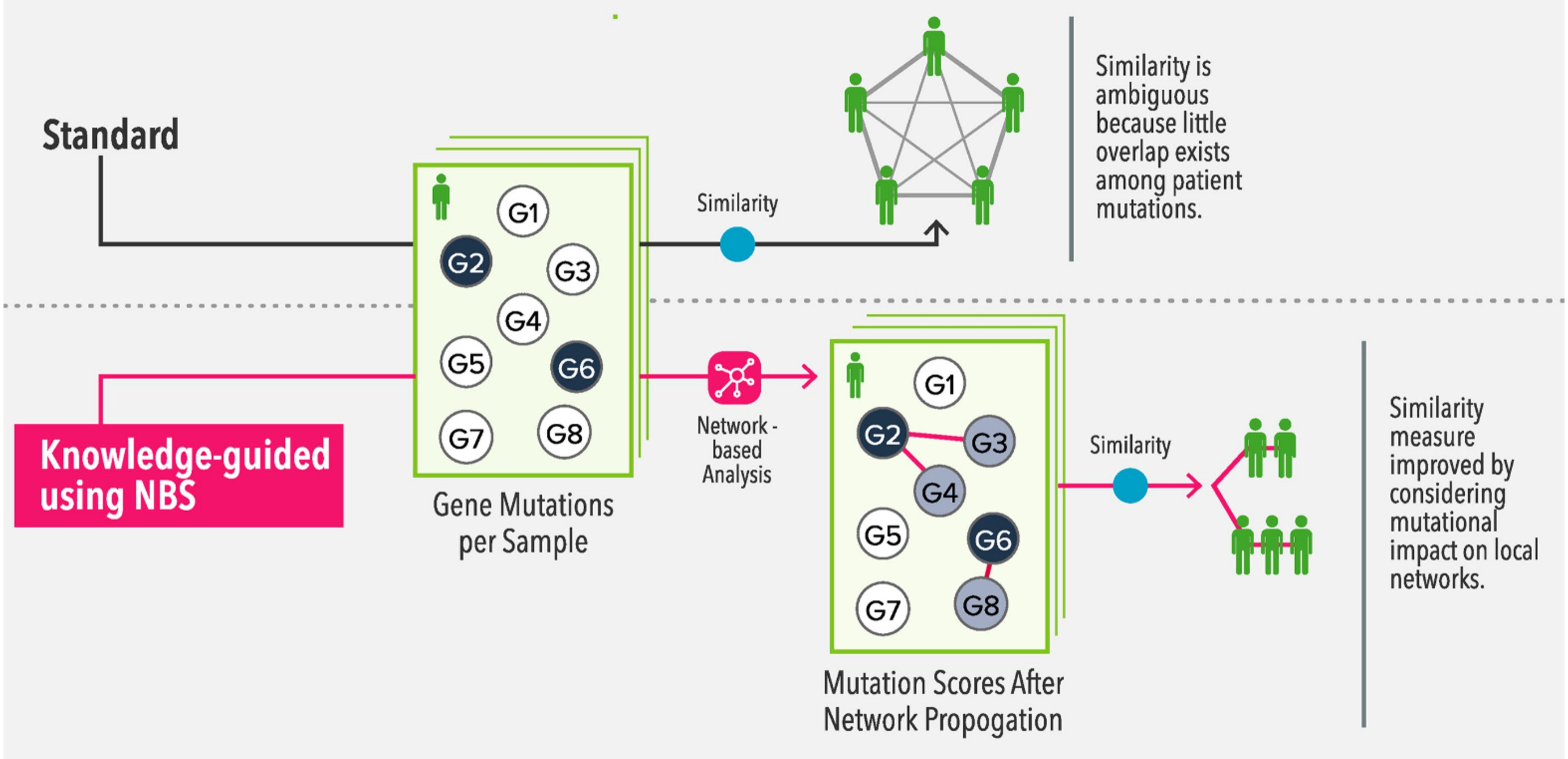
Edge Type Collection	Human Network Edges (millions)	Human Datasets	All Network Edges (millions)	All Datasets
Text_Mining/Integrated	**9.0**	**2**	**130.6**	**19**
Coexpression	7.3	2	119.8	19
Experimental_Interaction	5.4	4	108.7	21
Conservation/Proximity	1.6	2	26.1	36
Pathway_Database	1.1	3	63.4	20
Total	24.3	8	448.7	42



Edge Type Collection	Human Network Edges (millions)	Human Property Nodes (thousands)	Human Datasets	All Network Edges (millions)	All Property Nodes (thousands)
Tissue_Expression	**13.7**	**25.9**	**32**	**13.7**	**25.9**
Disease/Drug	6.0	82.3	13	6.3	83.4
Regulation	4.4	3.3	10	4.4	3.3
Pathways	0.6	16.9	5	1.4	34.6
Ontologies	0.3	17.2	5	1.8	23.5
Protein_Domains	0.0	6.2	2	0.5	7.8
Total	25.0	151.7	67	28.1	178.5

https://github.com/KnowEnG/KN_Fetcher/blob/master/Contents.md

Network-guided Sample Clustering


Standard
G1
G2
G3
G4
G5
G6
G7
G8
Gene Mutations per Sample
Similarity
Similarity is ambiguous because little overlap exists among patient mutations.
Knowledge-guided using NBS
Network - based Analysis
Mutation Scores After Network Propogation
Similarity
Similarity measure improved by considering mutational impact on local networks.

Network-Guided Sample Clustering

Goal:

- Stratification (clustering) of tumor samples based on somatic mutation profiles

Main Issue:

- The mutation data is very sparse and most conventional clustering techniques fail to identify reasonable patterns

- Although two tumors may not share the same somatic mutations, they may affect the same pathways and interaction networks

Knowledge-Guided Sample Clustering

- Problem: Data sparsity in gene-level somatic mutation data

- Toy Example
  - Due to the sparsity of the data, all samples are at equal distance of each other

	S1	S2	S3	S4	S5	S6
G01			🟩			
G02		🟩				
G03			🟩			
G04	🟩					
G05		🟩				
G06	🟩					
G07				🟩		
G08				🟩		
G09					🟩	
G10						🟩
G11						🟩
G12					🟩	
G13		🟩				
G14				🟩		
G15						🟩
G16					🟩	
G17	🟩					
G18			🟩			

Knowledge-Guided Sample Clustering

- **Problem:** Data sparsity in gene-level somatic mutation data

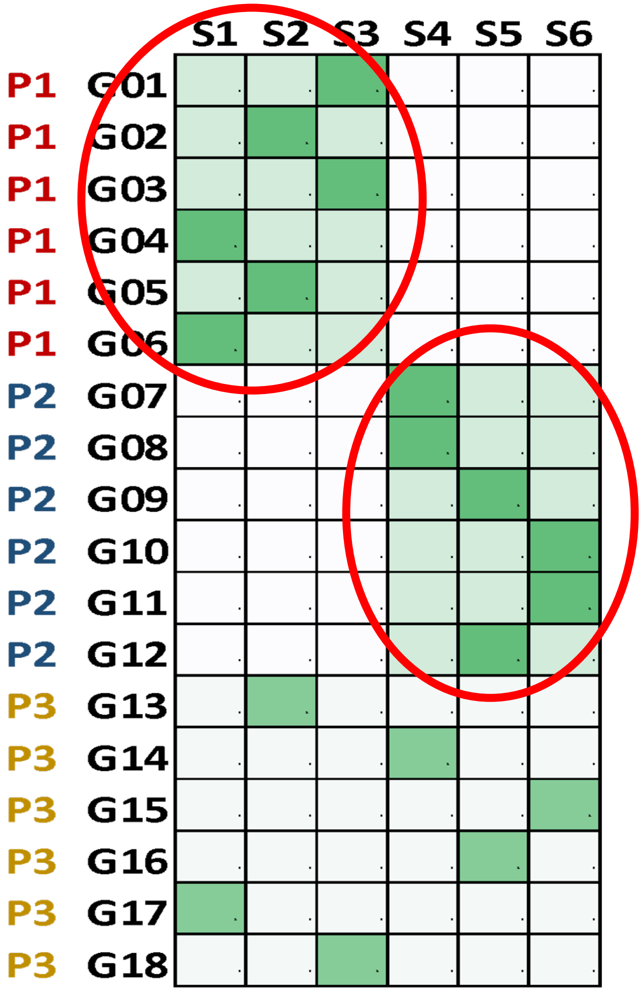
- **Toy Example**
  - Due to the sparsity of the data, all samples are at equal distance of each other
  - Pathway information clarifies the similarity among some samples

		S1	S2	S3	S4	S5	S6
P1	G01			🟩			
P1	G02		🟩				
P1	G03			🟩			
P1	G04	🟩					
P1	G05		🟩				
P1	G06	🟩					
P2	G07				🟩		
P2	G08				🟩		
P2	G09					🟩	
P2	G10						🟩
P2	G11						🟩
P2	G12					🟩	
P3	G13		🟩				
P3	G14				🟩		
P3	G15						🟩
P3	G16					🟩	
P3	G17	🟩					
P3	G18			🟩			

Knowledge-Guided Sample Clustering

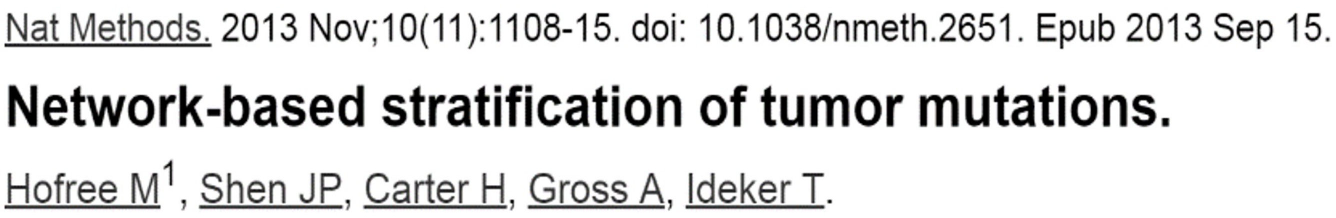
- Problem: Data sparsity in gene-level somatic mutation data

- Toy Example
  - Due to the sparsity of the data, all samples are at equal distance of each other
  - Pathway information clarifies the similarity among some samples
  - Conventional clustering methods can then identify clusters based on network-smoothed features



		S1	S2	S3	S4	S5	S6
P1	G01	▒	▒	🟩			
P1	G02	▒	🟩	▒			
P1	G03	▒	▒	🟩			
P1	G04	🟩	▒	▒			
P1	G05	▒	🟩	▒			
P1	G06	🟩	▒	▒			
P2	G07				🟩	▒	▒
P2	G08				🟩	▒	▒
P2	G09				▒	🟩	▒
P2	G10				▒	▒	🟩
P2	G11				▒	▒	🟩
P2	G12				▒	🟩	▒
P3	G13		🟩				
P3	G14				🟩		
P3	G15						🟩
P3	G16					🟩	
P3	G17	🟩					
P3	G18			🟩			

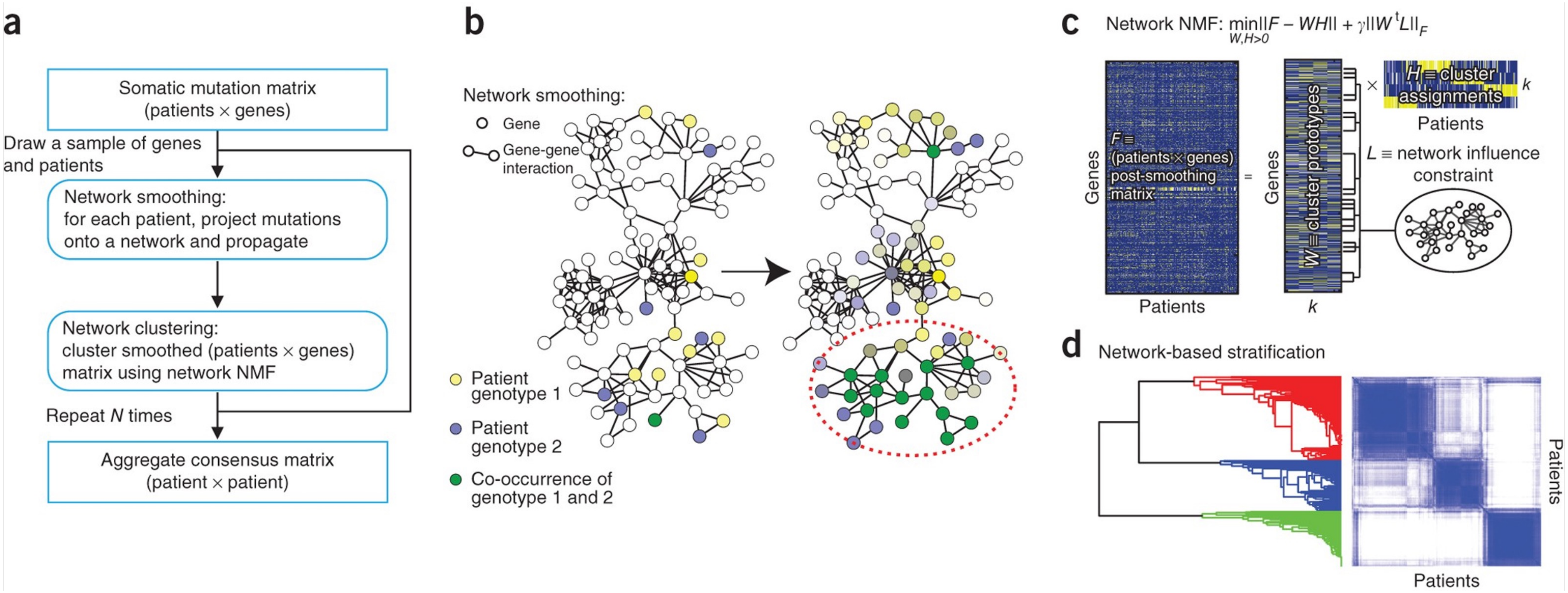
Network-based Stratification (NBS)



Nat Methods. 2013 Nov;10(11):1108-15. doi: 10.1038/nmeth.2651. Epub 2013 Sep 15.

Network-based stratification of tumor mutations.

Hofree M[1], Shen JP, Carter H, Gross A, Ideker T.



a

Somatic mutation matrix (patients × genes)

Draw a sample of genes and patients

Network smoothing: for each patient, project mutations onto a network and propagate

Network clustering: cluster smoothed (patients × genes) matrix using network NMF

Repeat *N* times

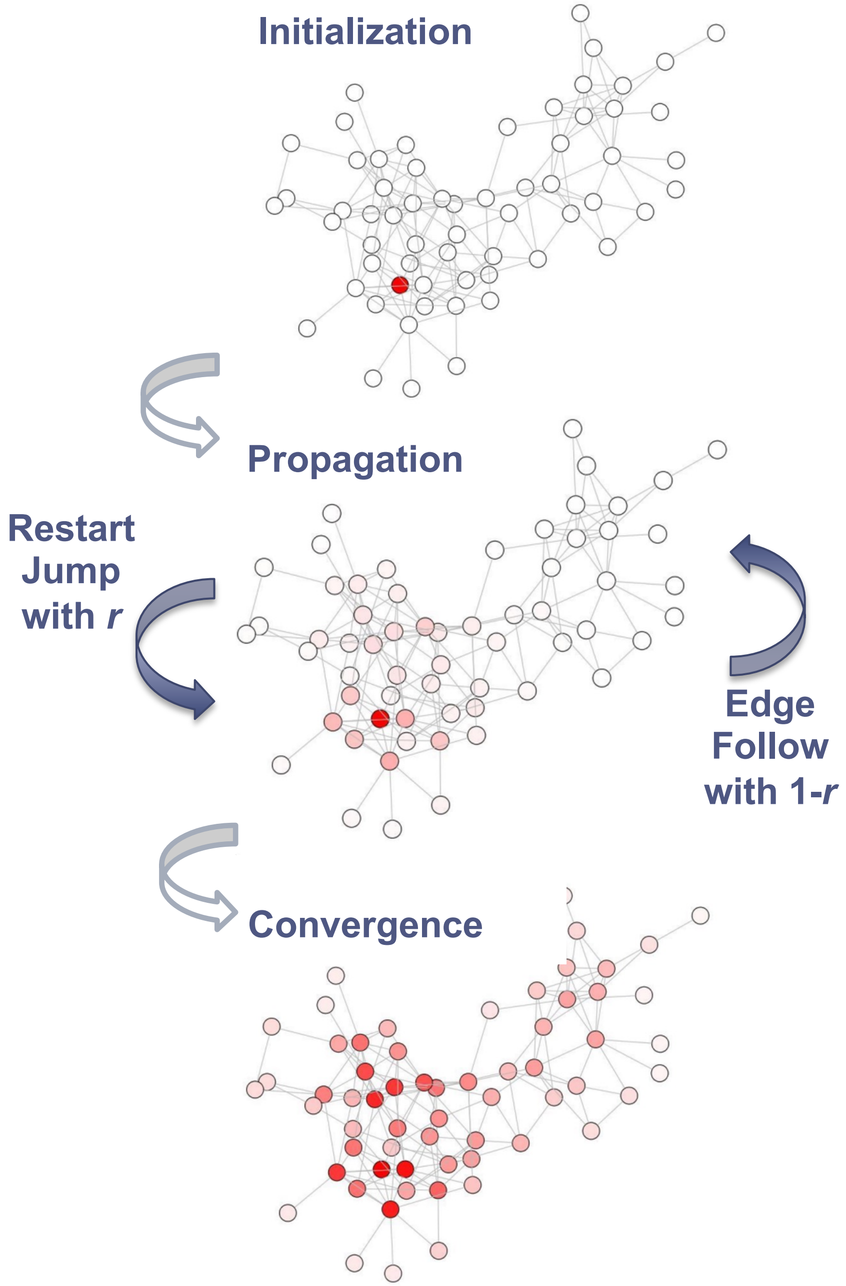
Aggregate consensus matrix (patient × patient)

b
Network smoothing:
Gene
Gene-gene interaction
Patient genotype 1
Patient genotype 2
Co-occurrence of genotype 1 and 2
c
Network NMF: $\min_{W,H>0} \|F - WH\| + \gamma \|W^t L\|_F$
Genes
F ≡ (patients × genes) post-smoothing matrix
=
Genes
W ≡ cluster prototypes
×
H ≡ cluster assignments
k
Patients
L ≡ network influence constraint
Patients
k
d
Network-based stratification
Patients
Patients

- Network Smoothing – Random Walk with Restart
- Patient Sampling for Robust Clustering

Random Walk With Restart Algorithm

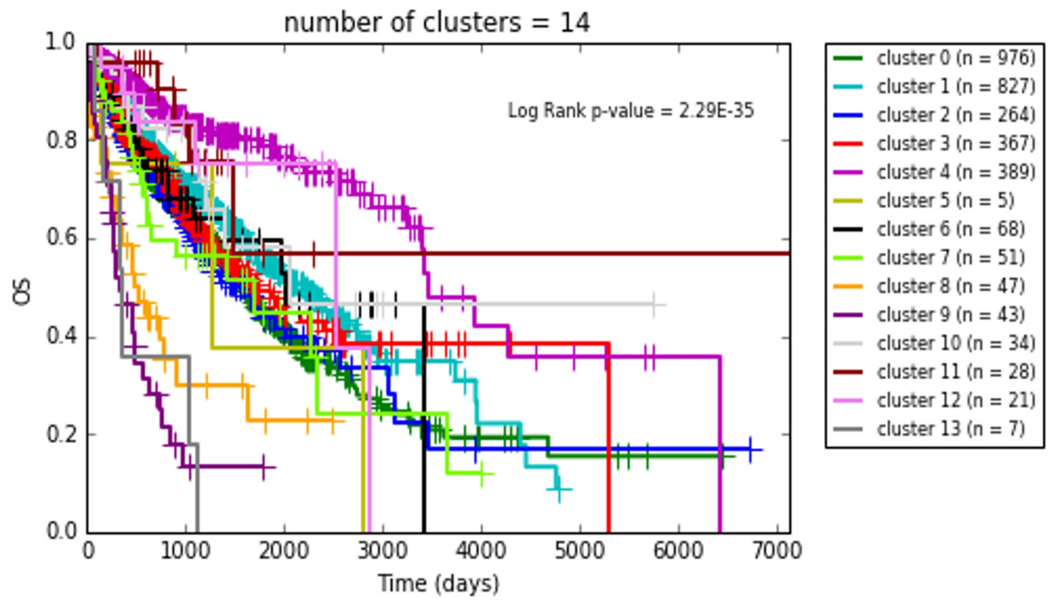
- Fast, scalable guilt-by-association method
  - Same ideas as personalized PageRank
- Intuition
  - Walker at a node either
    - With probability *1-r*, follows an outgoing edge
    - With restart probability *r*, returns to node in restart set
  - Converges to long run "stationary" distribution of the walker over the nodes
- Final node ranking based on distribution incorporates
  - Connectedness of node in network
  - Proximity of node to restart set



https://cseweb.ucsd.edu/~atsiatas/pr_diffusion_slides.pdf

NBS Sample Clustering with KnowEnG

- 3276 tumor samples from TCGA from 12 cancer projects with sparse non-synonymous somatic mutation
- Perform standard and network-guided **Sample Clustering** in platform
- Knowledge-guided clusters significantly relate to survival outcome

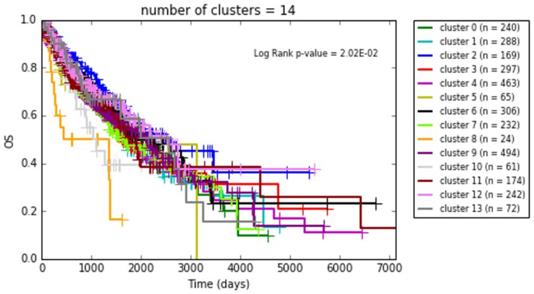


Knowledge-guided analysis of "omics" data using the KnowEnG cloud platform

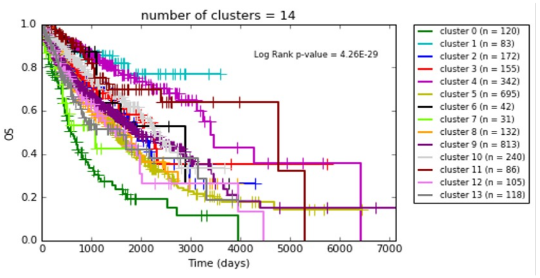
Charles Blatti III, Amin Emad, Matthew J. Berry, Lisa Gatzke, Milt Epstein, Daniel Lanier, Pramod Rizal, Jing Ge, Xiaoxia Liao, Omar Sobh, Mike Lambert, Corey S. Post, Jinfeng Xiao, [...], Saurabh Sinha [view all]

Published: January 23, 2020 • https://doi.org/10.1371/journal.pbio.3000583

- Much better than standard methods that do not incorporate prior knowledge

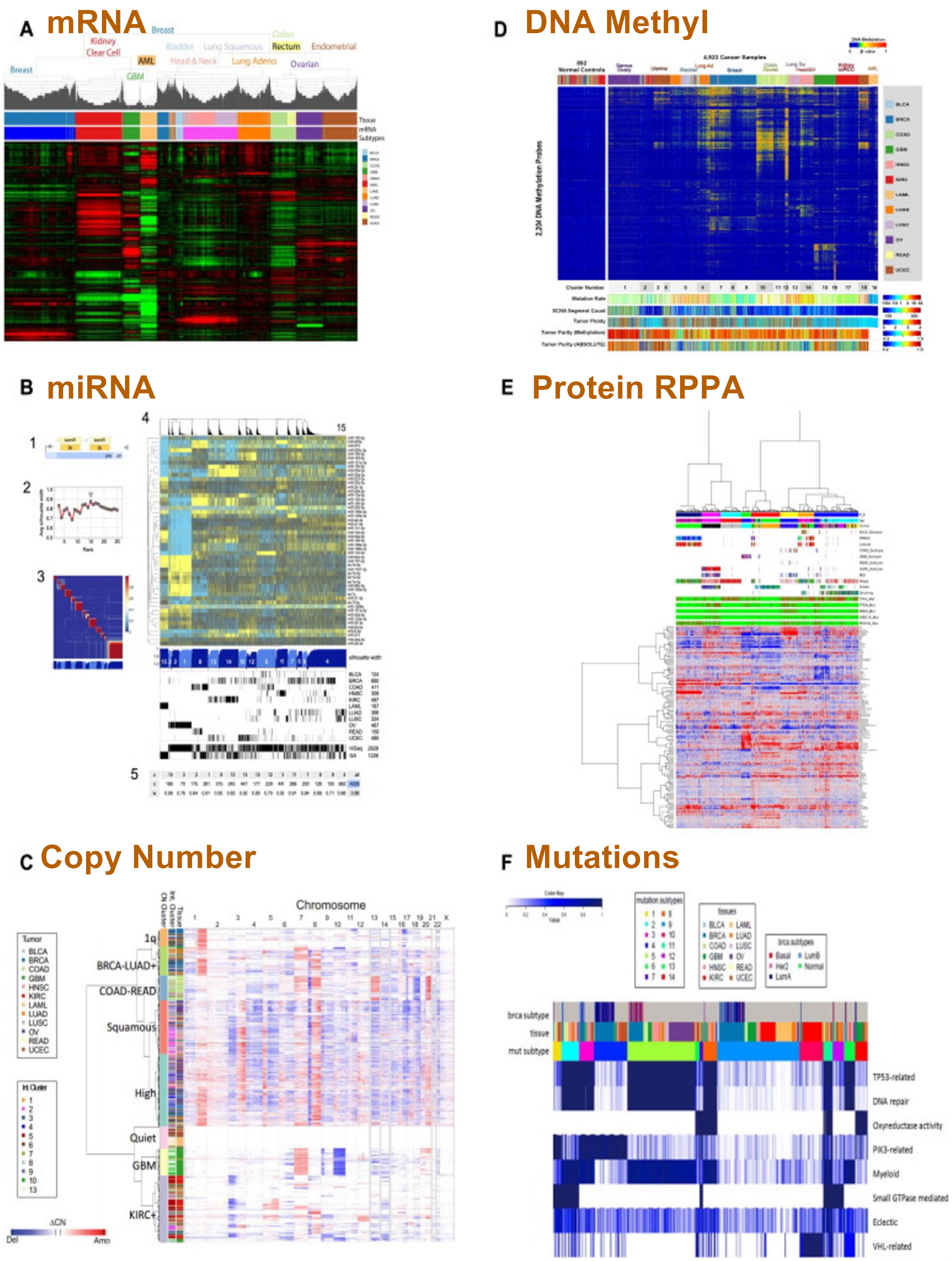


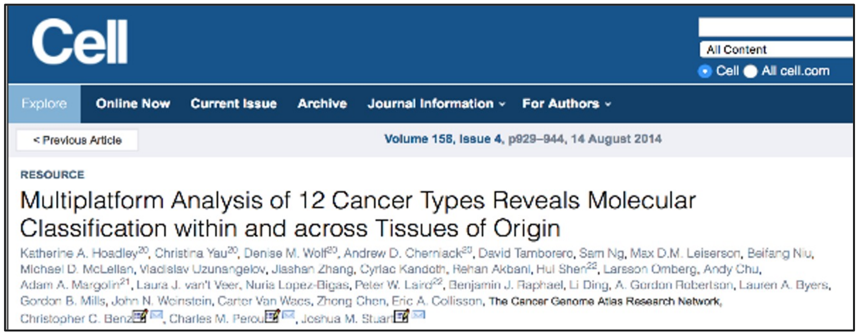
- In line with specialized method developed in TCGA paper that would be very difficult to reproduce



Integrating Experimental Assays for Stratification

- Data from each experimental assay is subjected to sample clustering to find cancer subtypes per assay
- Mutation data required specialized knowledge guided methods (panel F)





Volume 158, Issue 4, p929–944, 14 August 2014

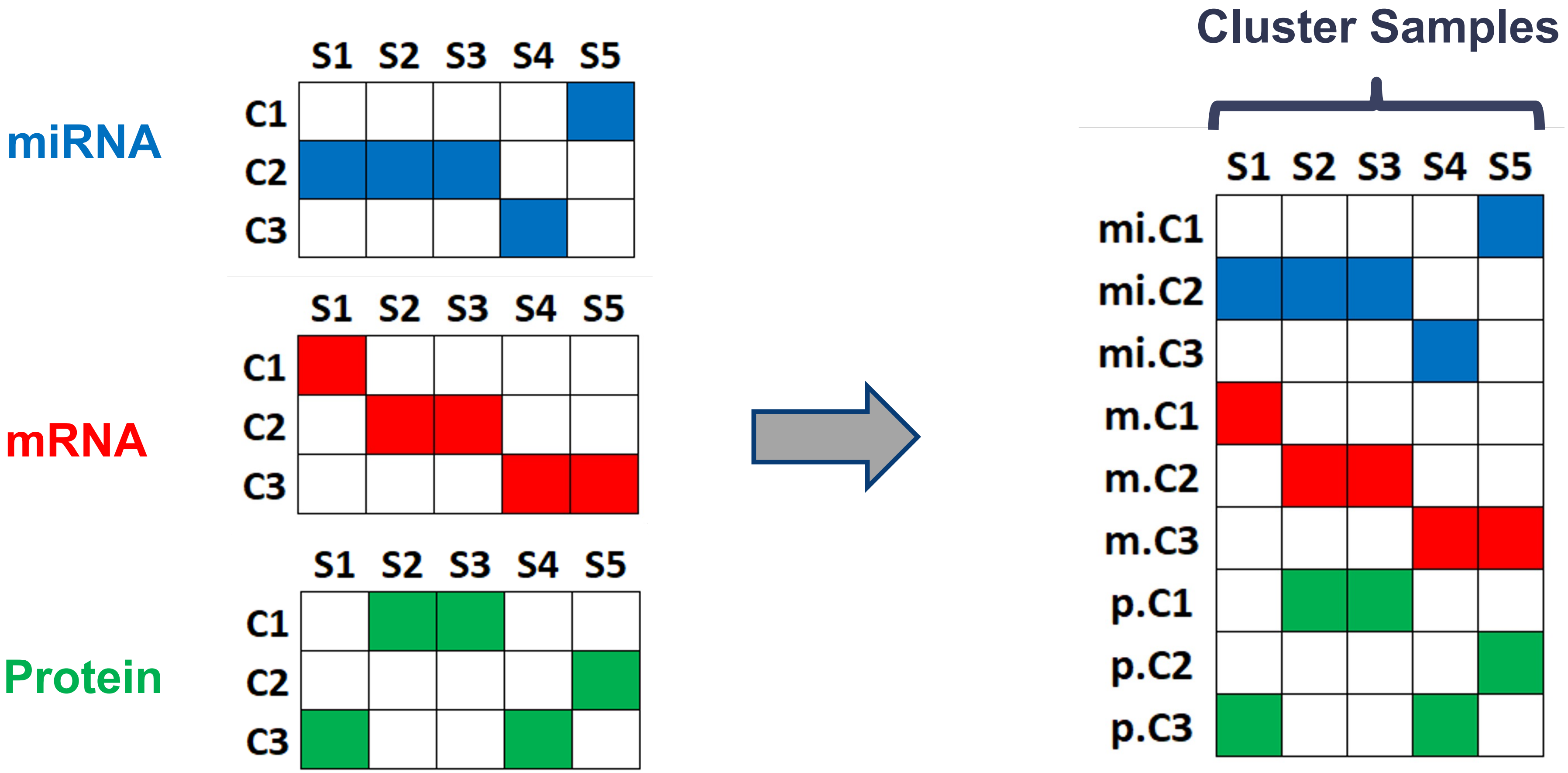
RESOURCE

Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin

Katherine A. Hoadley[20], Christina Yau[20], Denise M. Wolf[20], Andrew D. Cherniack[20], David Tamborero, Sam Ng, Max D.M. Leiserson, Beifang Niu, Michael D. McLellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen[22], Larsson Omberg, Andy Chu, Adam A. Margolin[21], Laura J. van't Veer, Nuria Lopez-Bigas, Peter W. Laird[22], Benjamin J. Raphael, Li Ding, A. Gordon Robertson, Lauren A. Byers, Gordon B. Mills, John N. Weinstein, Carter Van Waes, Zhong Chen, Eric A. Collisson, The Cancer Genome Atlas Research Network, Christopher C. Benz, Charles M. Perou, Joshua M. Stuart

Cluster-Of-Cluster-Assignments (COCA)

- Merge cluster assignments x samples matrices
- Cluster the samples in the multi-omics matrix



miRNA

	S1	S2	S3	S4	S5
C1					🟦
C2	🟦	🟦	🟦		
C3				🟦	

mRNA

	S1	S2	S3	S4	S5
C1	🟥				
C2		🟥	🟥		
C3				🟥	🟥

Protein

	S1	S2	S3	S4	S5
C1		🟩	🟩		
C2					🟩
C3	🟩			🟩	

Cluster Samples

	S1	S2	S3	S4	S5
mi.C1					🟦
mi.C2	🟦	🟦	🟦		
mi.C3				🟦	
m.C1	🟥				
m.C2		🟥	🟥		
m.C3				🟥	🟥
p.C1		🟩	🟩		
p.C2					🟩
p.C3	🟩			🟩	

13 Cancer Subtypes from 6 Assays

- Strong relationship between subtypes & disease
- Interesting relations between clusters of different data types

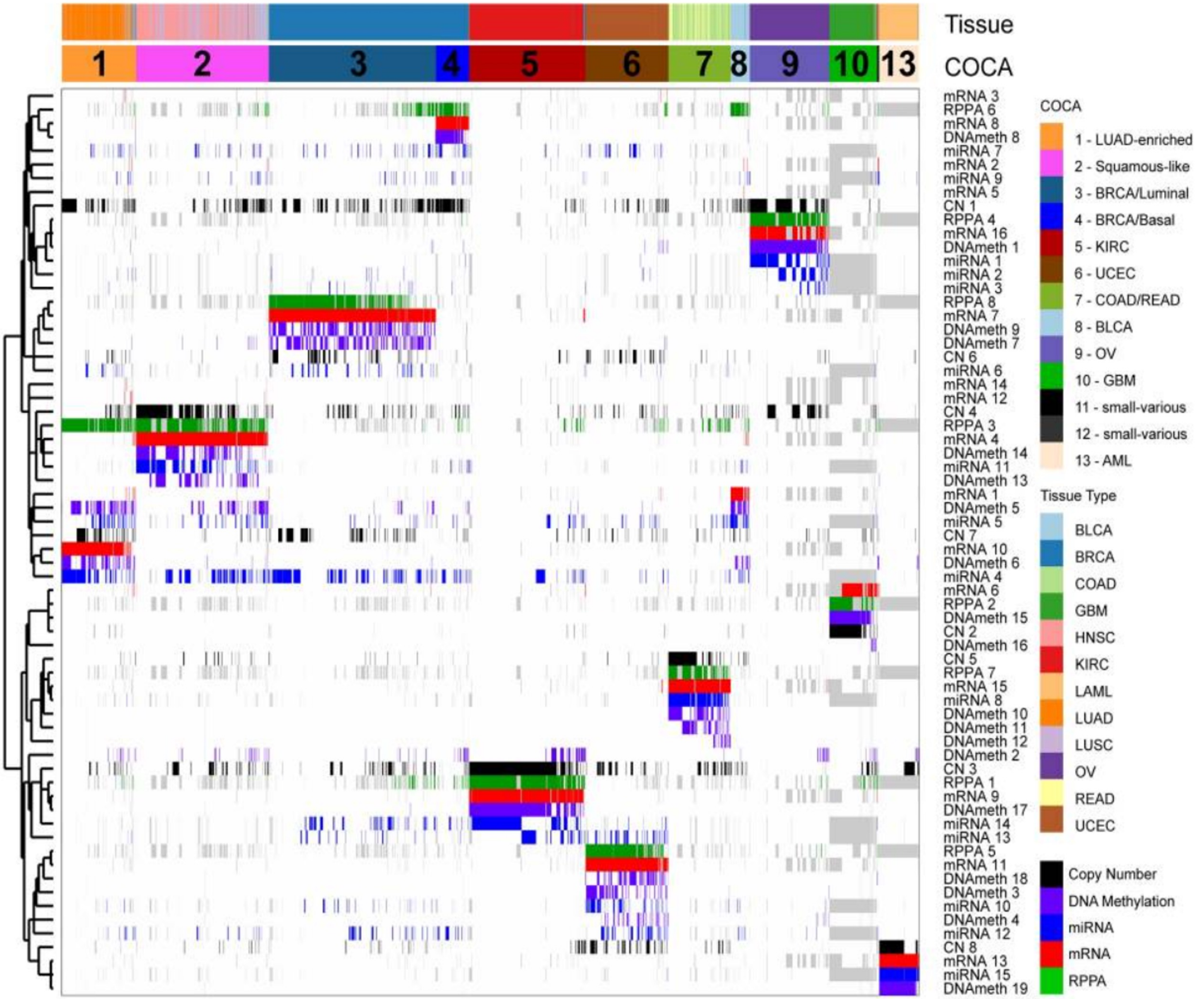
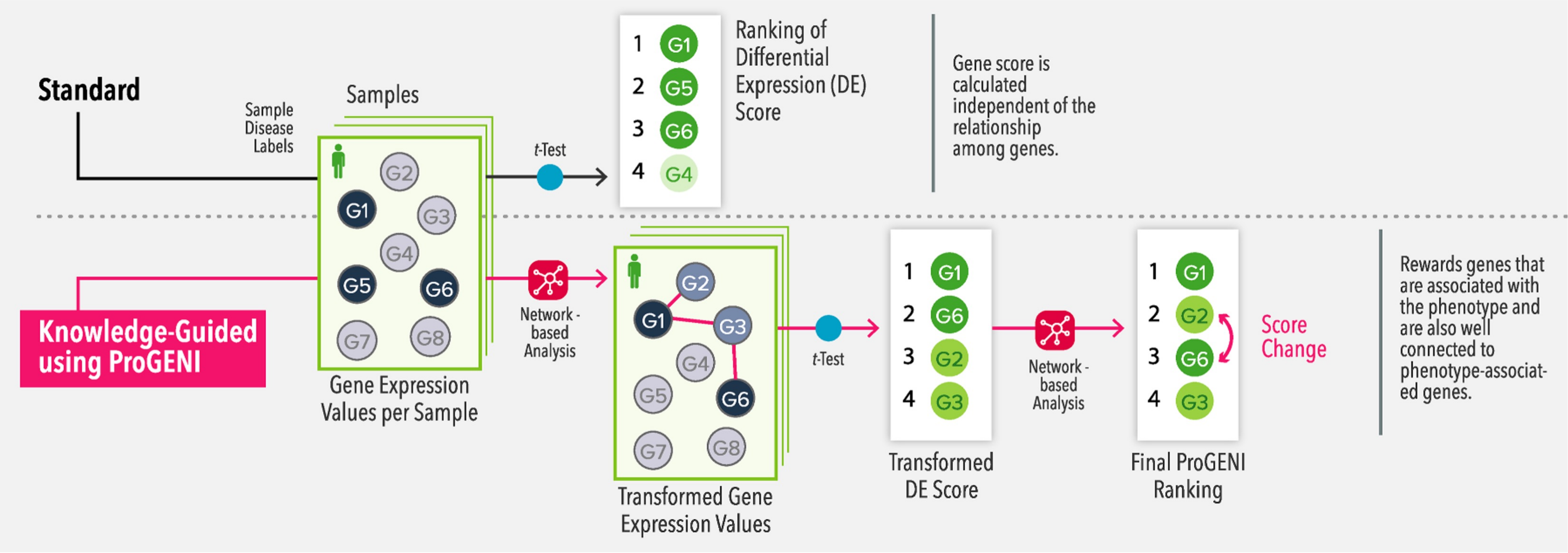


Figure from Hoadley, et al. "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin." *Cell* 158.4 (2014).

Network-Guided Gene Prioritization



Gene score is calculated independent of the relationship among genes.

Rewards genes that are associated with the phenotype and are also well connected to phenotype-associated genes.

Characterizing Cancer Subtypes

- Find top related mutations and copy number alterations
- Compare each subtype vs `all others`
- KnowEnG calls this `**Gene Prioritization**`

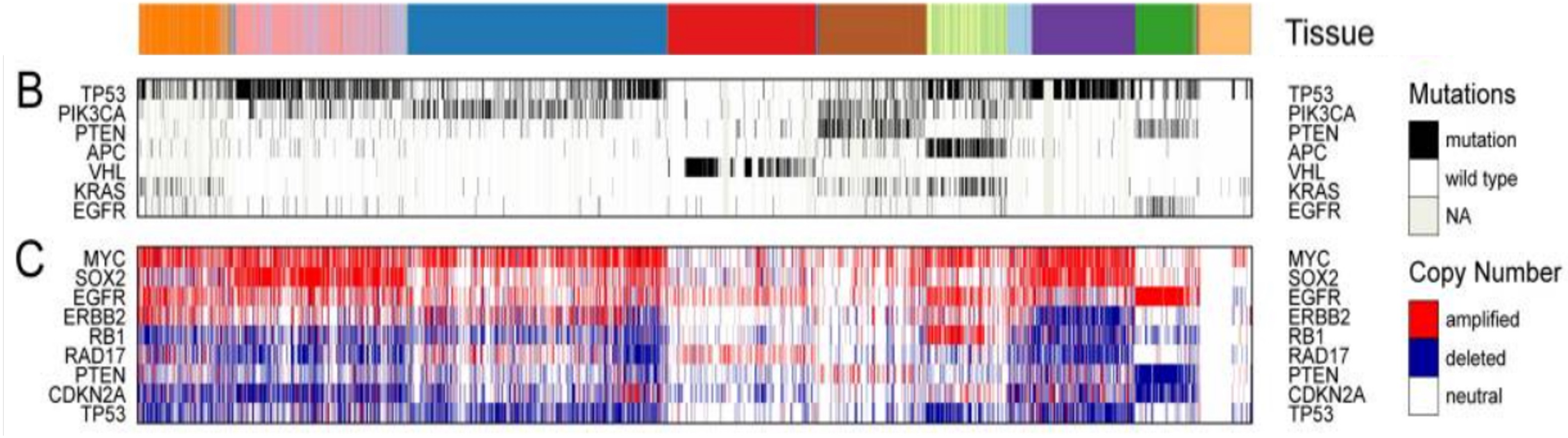


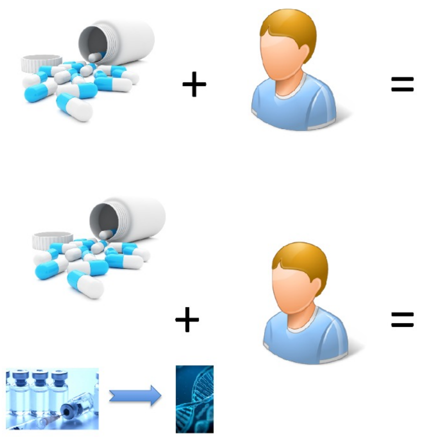
Figure from Hoadley, et al. "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin." *Cell* 158.4 (2014).

Towards Network-Guided Gene Prioritization

Drug Sensitivity Example

- Goal:
  - Identifying genes whose basal mRNA expression determines the drug sensitivity in different samples (supervised feature selection)

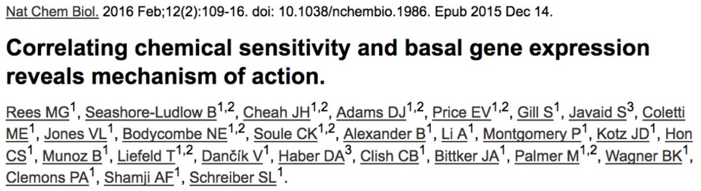
- Motivations:
  - Overcoming drug resistance
  - Revealing drug mechanism of action
  - Identifying novel drug targets
  - Predicting drug sensitivity of individuals



Standard Gene Prioritization

Examples of current methods:

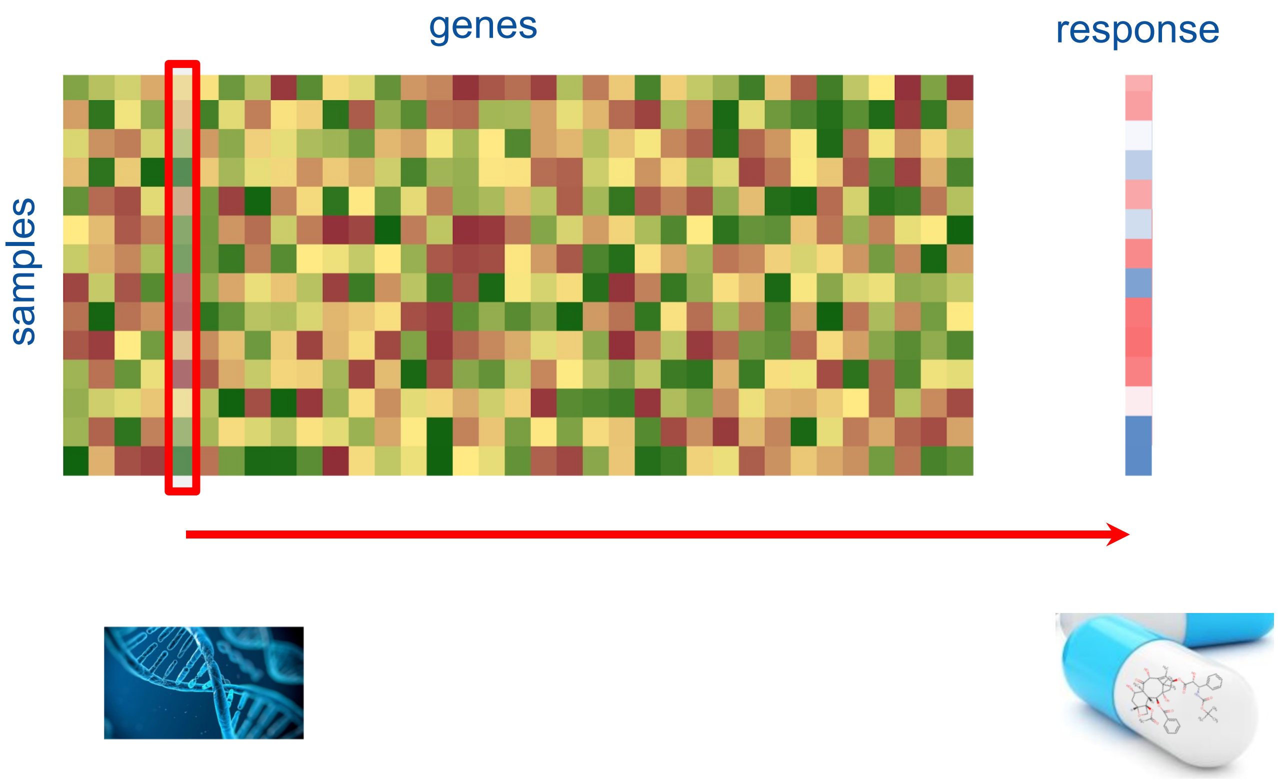
- Score each gene based on the correlation of its expression with drug response



Nat Chem Biol. 2016 Feb;12(2):109-16. doi: 10.1038/nchembio.1986. Epub 2015 Dec 14.

Correlating chemical sensitivity and basal gene expression reveals mechanism of action.

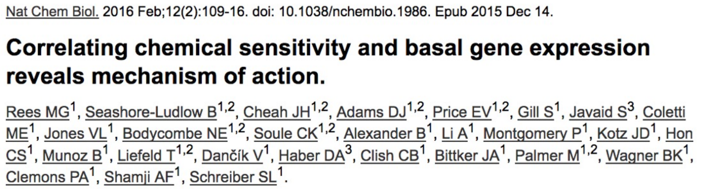
Rees MG[1], Seashore-Ludlow B[1,2], Cheah JH[1,2], Adams DJ[1,2], Price EV[1,2], Gill S[1], Javaid S[3], Coletti ME[1], Jones VL[1], Bodycombe NE[1,2], Soule CK[1,2], Alexander B[1], Li A[1], Montgomery P[1], Kotz JD[1], Hon CS[1], Munoz B[1], Liefeld T[1,2], Dančík V[1], Haber DA[3], Clish CB[1], Bittker JA[1], Palmer M[1,2], Wagner BK[1], Clemons PA[1], Shamji AF[1], Schreiber SL[1].



Standard Gene Prioritization

Examples of current methods:

- Score each gene based on the correlation of its expression with drug response

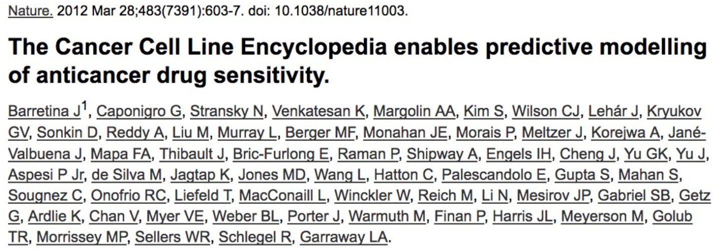


Nat Chem Biol. 2016 Feb;12(2):109-16. doi: 10.1038/nchembio.1986. Epub 2015 Dec 14.

Correlating chemical sensitivity and basal gene expression reveals mechanism of action.

Rees MG[1], Seashore-Ludlow B[1,2], Cheah JH[1,2], Adams DJ[1,2], Price EV[1,2], Gill S[1], Javaid S[3], Coletti ME[1], Jones VL[1], Bodycombe NE[1,2], Soule CK[1,2], Alexander B[1], Li A[1], Montgomery P[1], Kotz JD[1], Hon CS[1], Munoz B[1], Liefeld T[1,2], Dančík V[1], Haber DA[3], Clish CB[1], Bittker JA[1], Palmer M[1,2], Wagner BK[1], Clemons PA[1], Shamji AF[1], Schreiber SL[1].

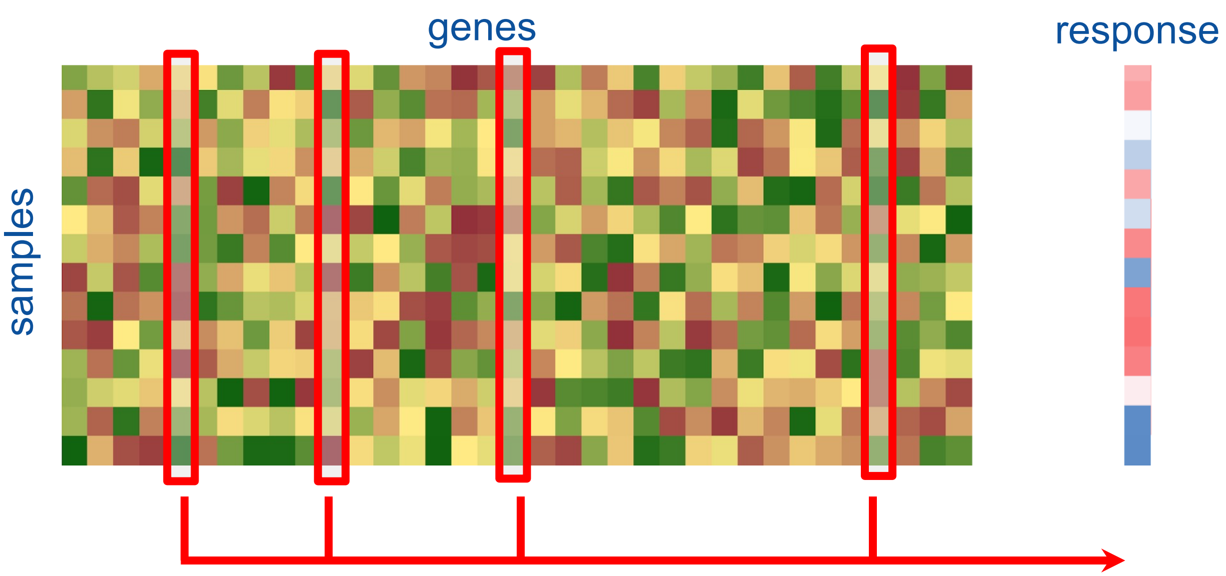
- Use multivariable regression algorithms such as Elastic Net to relate multiple genes’ expression values to drug response



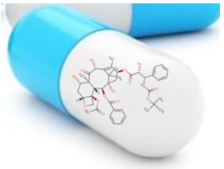
Nature. 2012 Mar 28;483(7391):603-7. doi: 10.1038/nature11003.

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.

Barretina J[1], Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P Jr, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA.



$$\sum w_i \mathbf{x}_i$$

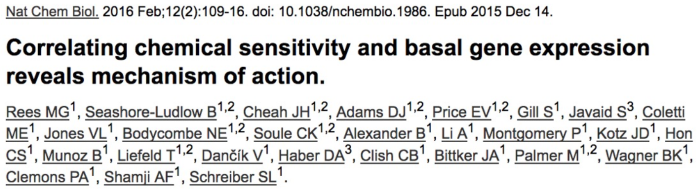


Augmenting Gene Prioritization

Examples of current methods:

- Score each gene based on the correlation of its expression with drug response

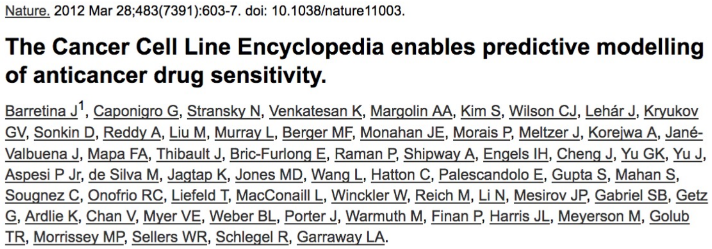
- Use multivariable regression algorithms such as Elastic Net to relate multiple genes' expression values to drug response



Nat Chem Biol. 2016 Feb;12(2):109-16. doi: 10.1038/nchembio.1986. Epub 2015 Dec 14.

Correlating chemical sensitivity and basal gene expression reveals mechanism of action.

Rees MG[1], Seashore-Ludlow B[1,2], Cheah JH[1,2], Adams DJ[1,2], Price EV[1,2], Gill S[1], Javaid S[3], Coletti ME[1], Jones VL[1], Bodycombe NE[1,2], Soule CK[1,2], Alexander B[1], Li A[1], Montgomery P[1], Kotz JD[1], Hon CS[1], Munoz B[1], Liefeld T[1,2], Dančík V[1], Haber DA[3], Clish CB[1], Bittker JA[1], Palmer M[1,2], Wagner BK[1], Clemons PA[1], Shamji AF[1], Schreiber SL[1].



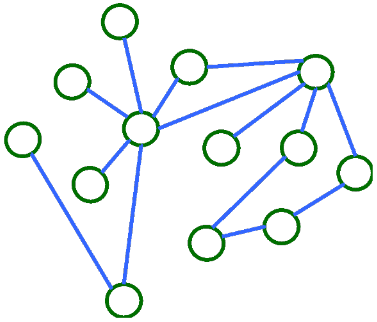
Nature. 2012 Mar 28;483(7391):603-7. doi: 10.1038/nature11003.

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.

Barretina J[1], Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P Jr, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA.

Shortcoming:

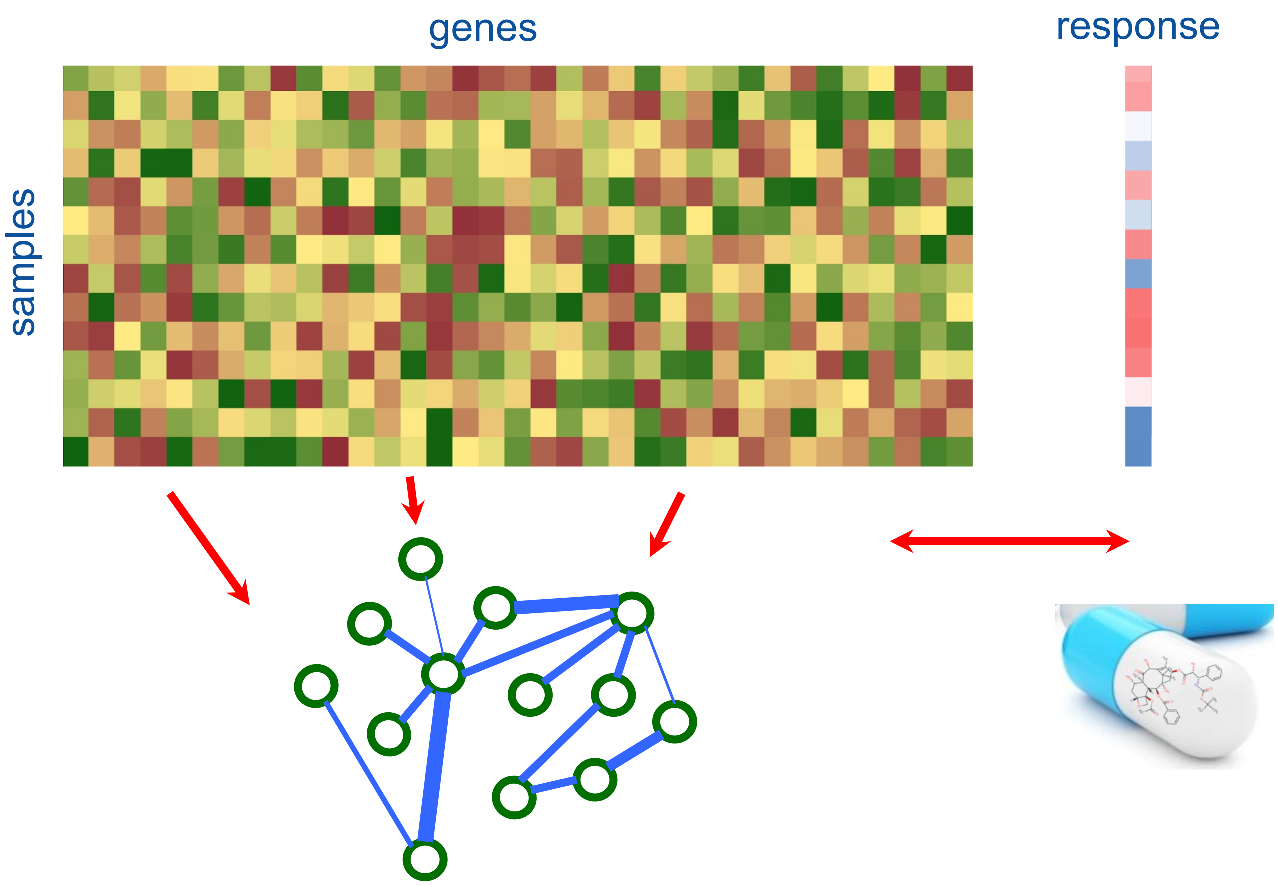
- These methods do not incorporate prior information about the interaction of the genes



Network-Guided Gene Prioritization

Hypothesis:

- Since genes and proteins involved in drug MoA are functionally related, prior knowledge in the form of gene interaction network (e.g. PPI) can improve accuracy of the prioritization task



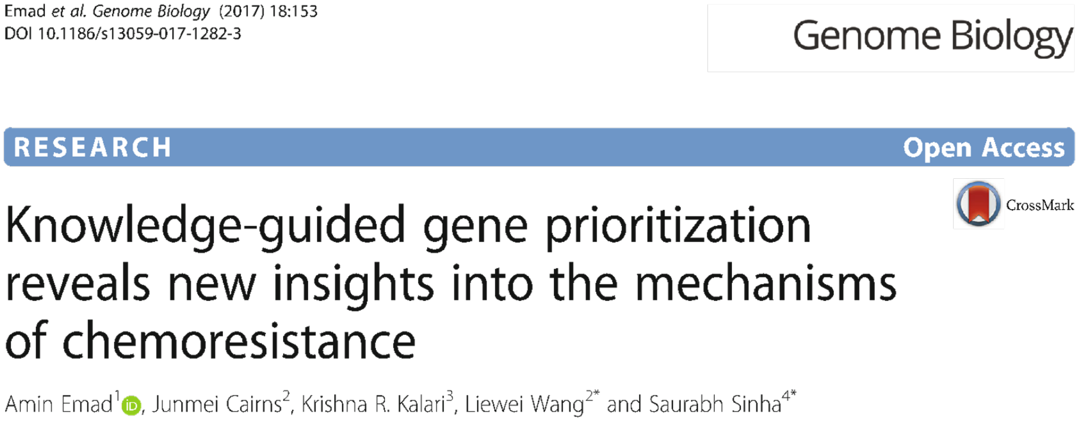
ProGENI

ProGENI: Network-guided gene prioritization

- An algorithm that incorporates gene network information to improve prioritization accuracy

Genome Biology

Featured article: new insights into mechanisms of chemoresistance



Emad *et al. Genome Biology* (2017) 18:153
DOI 10.1186/s13059-017-1282-3

Genome Biology

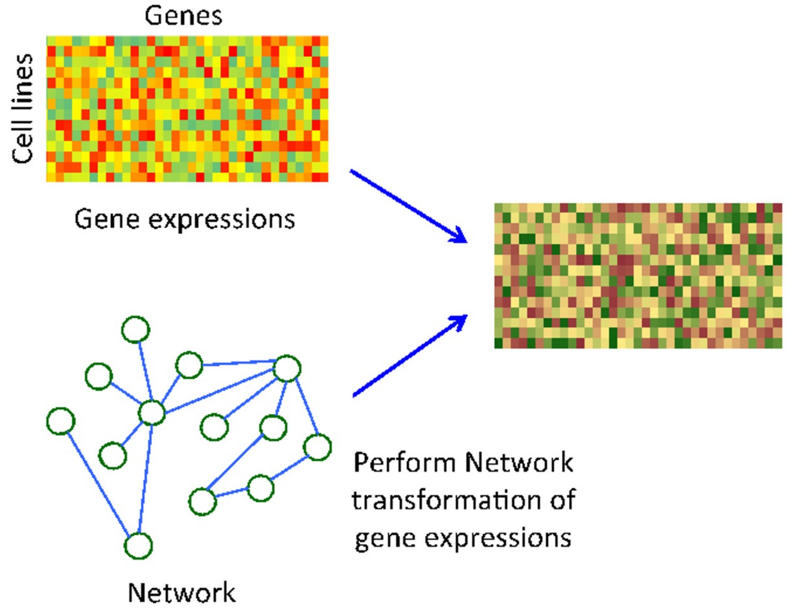
RESEARCH Open Access

Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance

Amin Emad[1], Junmei Cairns[2], Krishna R. Kalari[3], Liewei Wang[2*] and Saurabh Sinha[4*]

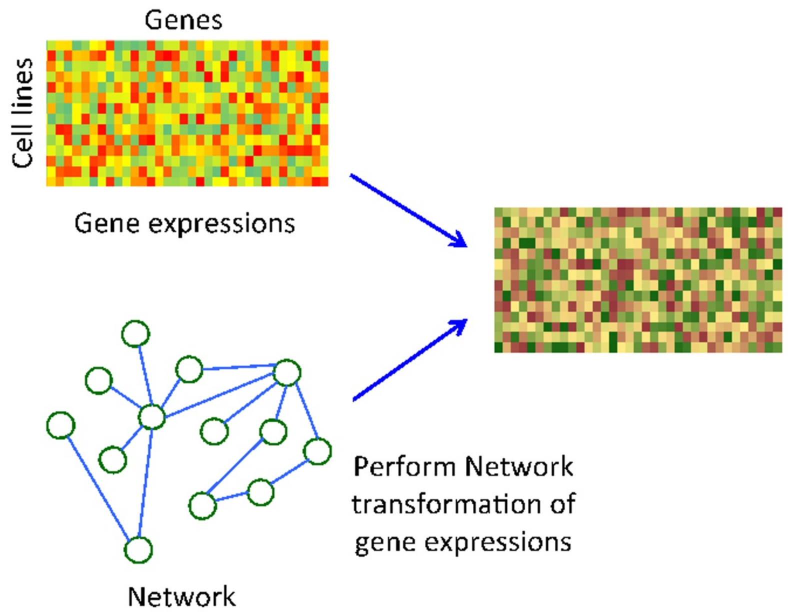
ProGENI Method

Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance



ProGENI Method

Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance



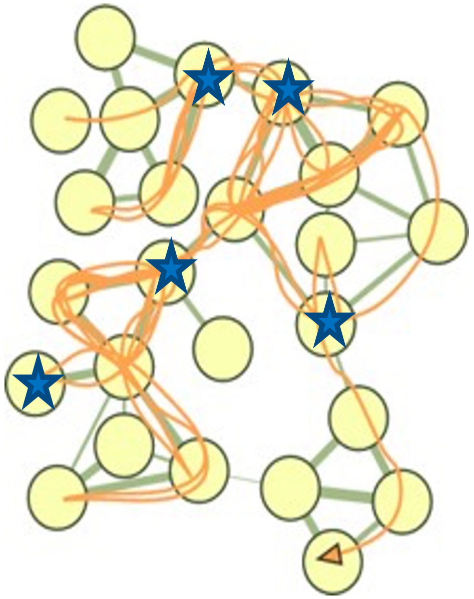
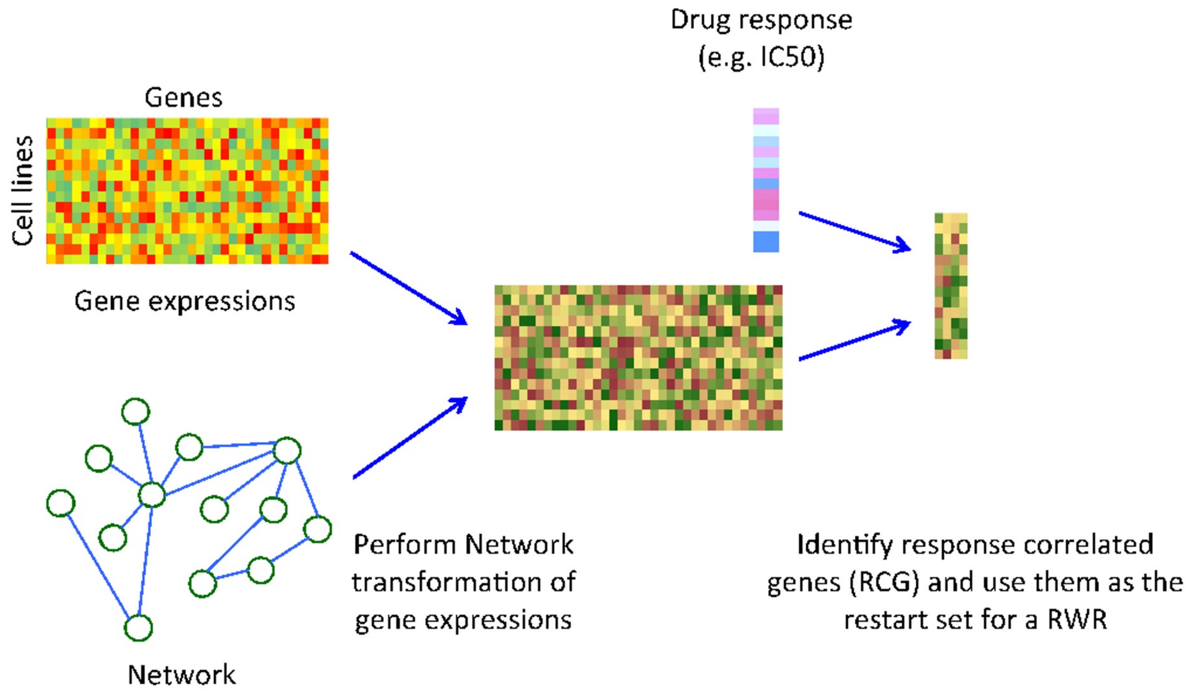


Figure from Rosvall and Bergstrom. "Maps of random walks on complex networks reveal community structure." *Proceedings of the national academy of sciences* 105.4 (2008).

ProGENI Method

Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance

Step 2: Find genes most correlated with drug response (RCG set)

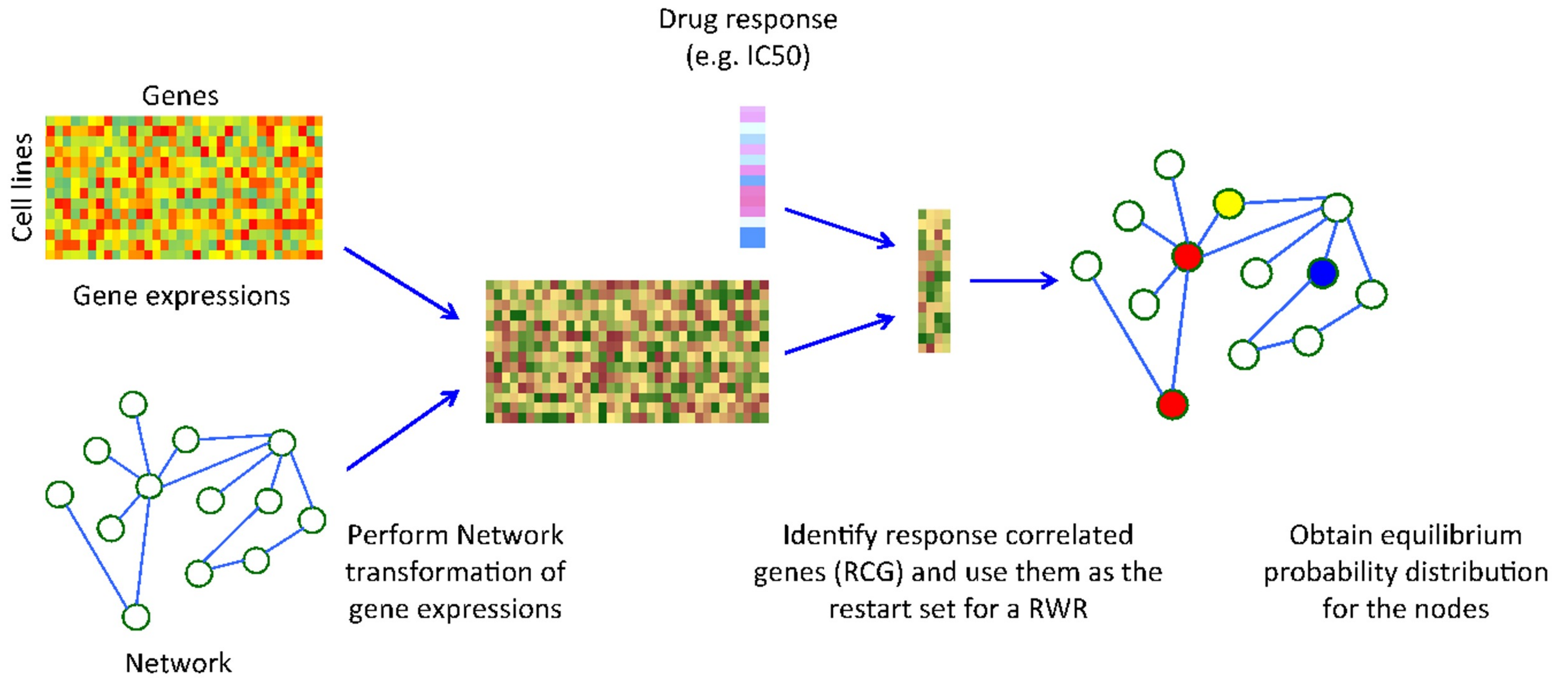

Drug response
(e.g. IC50)
Genes
Cell lines
Gene expressions
Network
Perform Network transformation of gene expressions
Identify response correlated genes (RCG) and use them as the restart set for a RWR

ProGENI Method

Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance

Step 2: Find genes most correlated with drug response (RCG set)

Step 3: Score genes based on their relevance to the RCG set



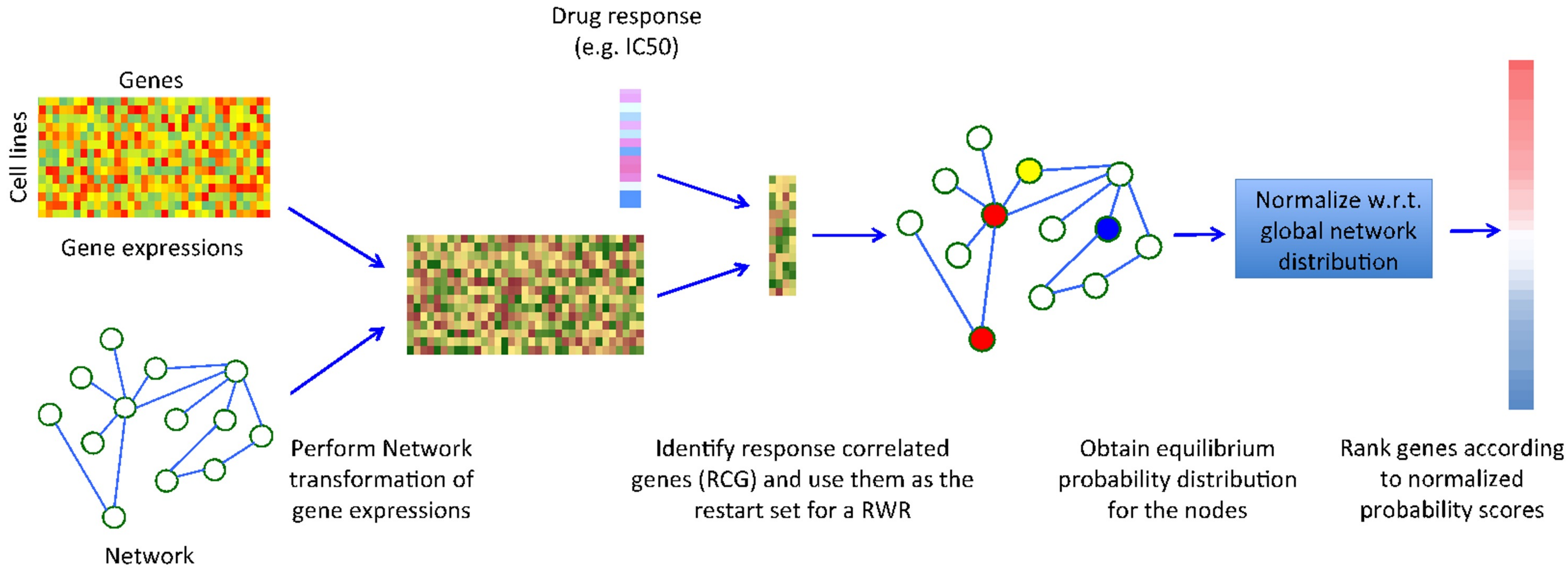
ProGENI Method

Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance

Step 2: Find genes most correlated with drug response (RCG set)

Step 3: Score genes based on their relevance to the RCG set

Step 4: Remove network bias by normalizing scores w.r.t. scores corresponding to global network topology



Perform Network transformation of gene expressions

Identify response correlated genes (RCG) and use them as the restart set for a RWR

Obtain equilibrium probability distribution for the nodes

Rank genes according to normalized probability scores

ProGENI Analysis Datasets

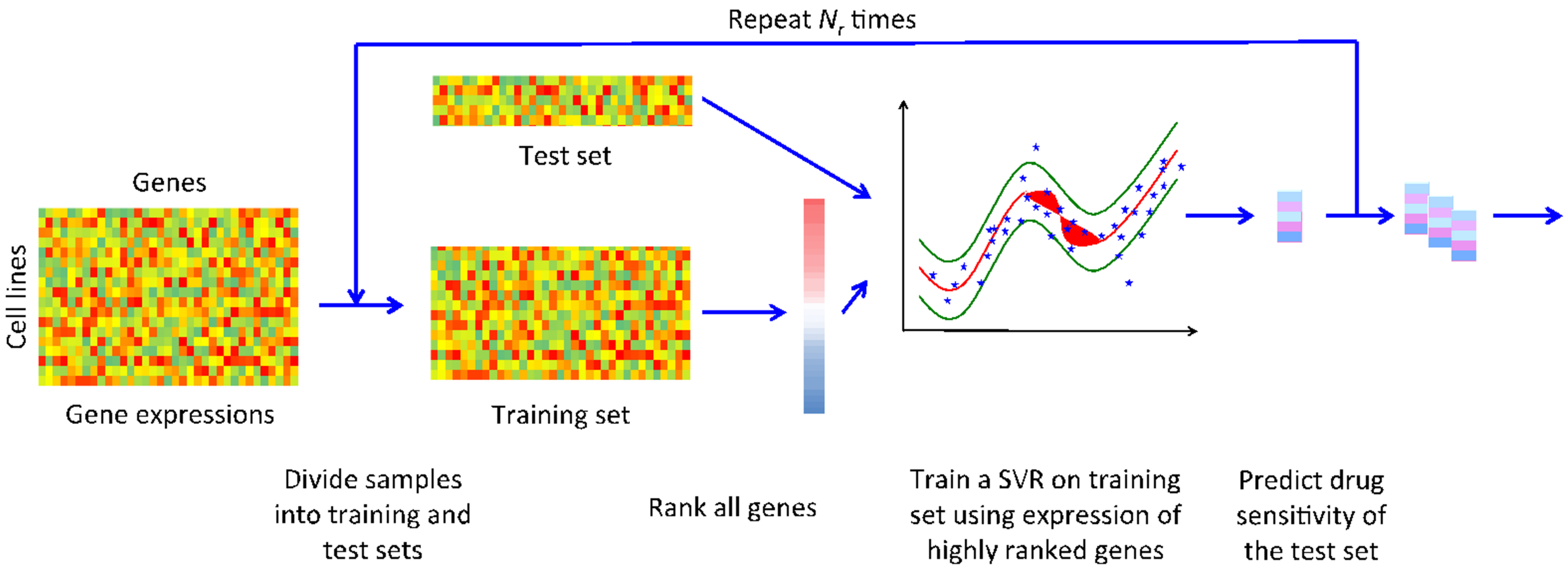
- Human lymphoblastoid cell lines (LCL)
  - Gene expression (~17K genes of ~300 cell lines)
  - Drug response of 24 cytotoxic treatments

- Publicly available dataset from GDSC
  - Gene expression (~13K genes of ~600 cell lines from 13 tissues)
  - Drug response of 139 cytotoxic treatments

- Publicly available prior knowledge
  - Network of gene interactions (PPI and genetic interactions) from STRING (~1.5M edges, ~15.5K nodes)

Validation Using Drug Response Prediction

- Genes ranked highly using a good prioritization method are good predictors of drug sensitivity



Divide samples into training and test sets

Rank all genes

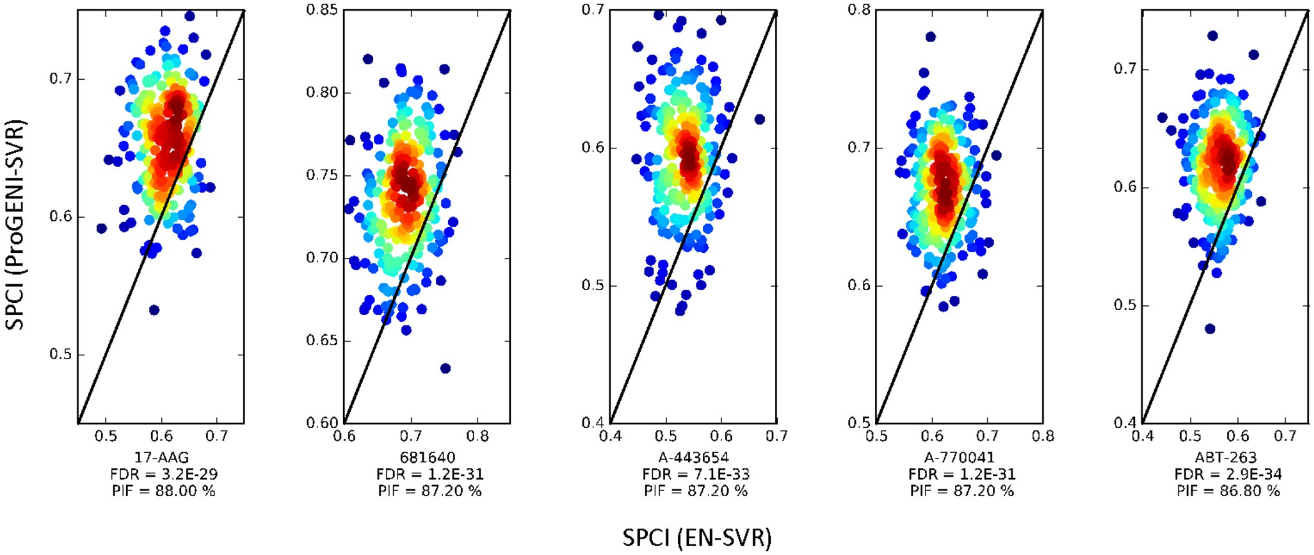
Train a SVR on training set using expression of highly ranked genes

Predict drug sensitivity of the test set

Validation Using Drug Response Prediction

LCL Dataset	Pearson	Elastic Net
Num. Drugs (out of 24) ProGENI > Baseline	14	20
FDR (Wilcoxon signed-rank test)	6.5 E-3	9.6 E-5

GDSC Dataset	Pearson	Elastic Net
Num. Drugs (out of 139) ProGENI > Baseline	66	110
FDR (Wilcoxon signed-rank test)	9.1 E-4	4.0 E-21

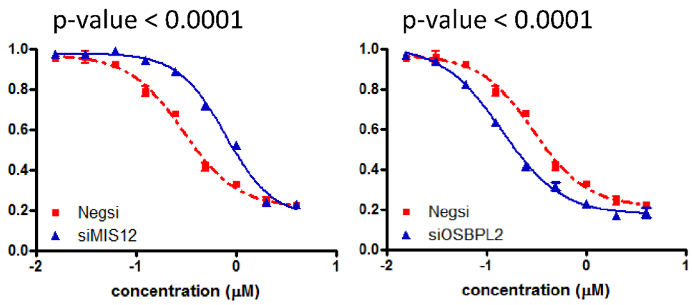


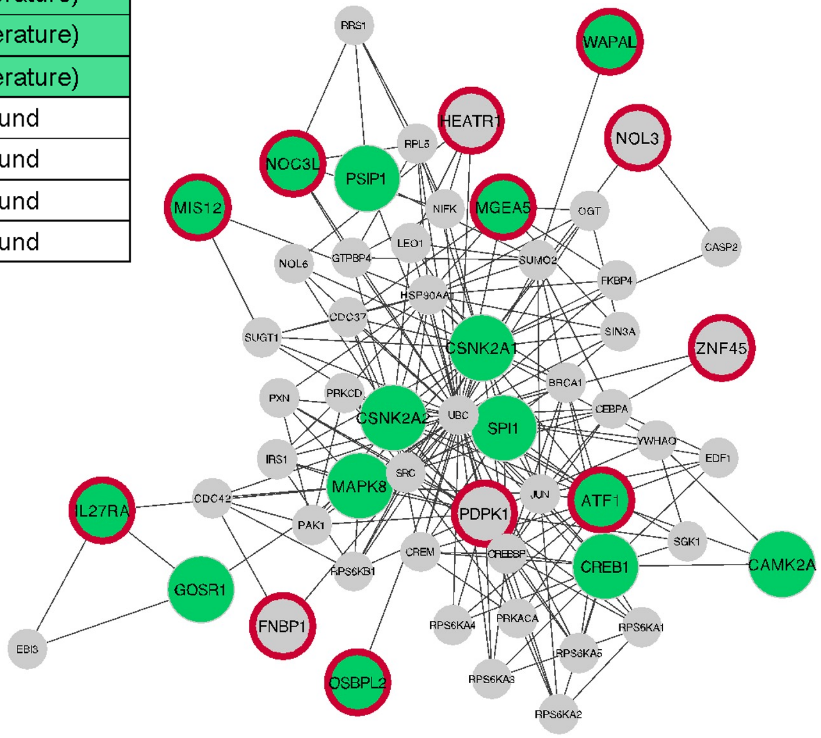
Functional Validation

We validated role of 33 (out of 45) genes (73%) for three drugs.

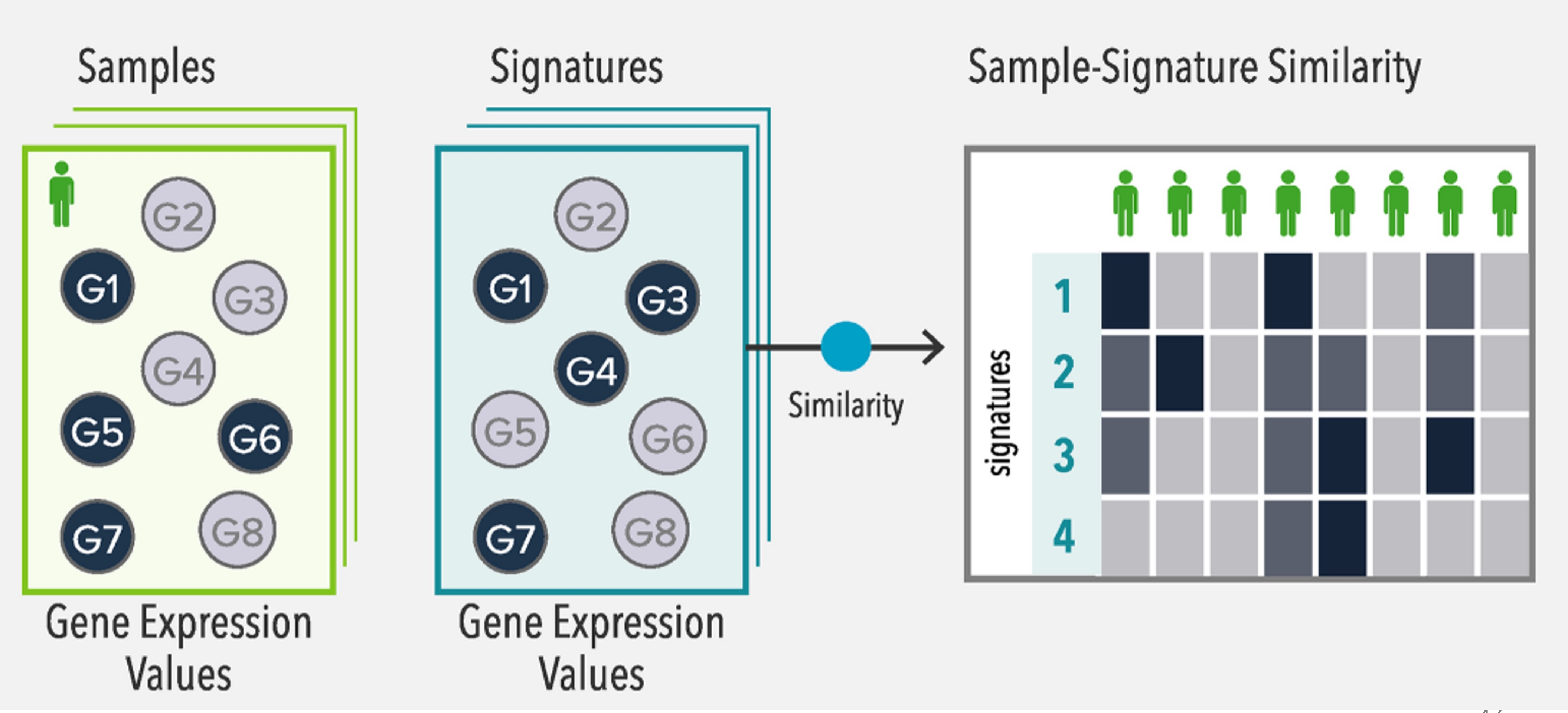
Gene Symbol	Rank (ProGENI)	Rank (Pearson)	Absolute value of Pearson correlation coefficient	Evidence
ATF1	1	1	0.2000	Direct (this study)
MIS12	2	4	0.1887	Direct (this study)
OSBPL2	5	6	0.1865	Direct (this study)
CSNK2A1	7	1587	0.0752	Direct (literature)
PSIP1 (LEDGF)	8	46	0.1537	Direct (literature)
CAMK2A	9	6991	0.0157	Direct (literature)
CSNK2A2	10	4870	0.0347	Direct (literature)
GOSR1	11	6867	0.0167	Direct (this study)
MAPK8	13	7574	0.0112	Direct (literature)
SPI1	14	6287	0.0217	Direct (literature)
CREB1	15	665	0.1000	Direct (literature)
NOC3L	3	3	0.1893	Not found
IL27RA	4	2	0.1911	Not found
MGEA5	6	7	0.1814	Not found
WAPAL	12	8	0.1805	Not found

BT549





Gene Expression Signatures



Gene Expression Signatures

- Massive Transcriptomic Profiling Projects
  - TCGA and ICGC
  - GTEX and CCLE
  - LINCS
- Definitions
  - Projects produce **expression vectors** for samples (e.g. gene expression levels)
  - Scoring the difference in expression between samples of two (or more) conditions produces **differential expression vectors**
- **Signature** (of a biological state):
  - **Gene Set** – differentially, characteristically expressed genes in that state relative to some reference (control or population)
  - **Differential Expression Vector** – the differential expression scores for the subset of genes in the same comparison

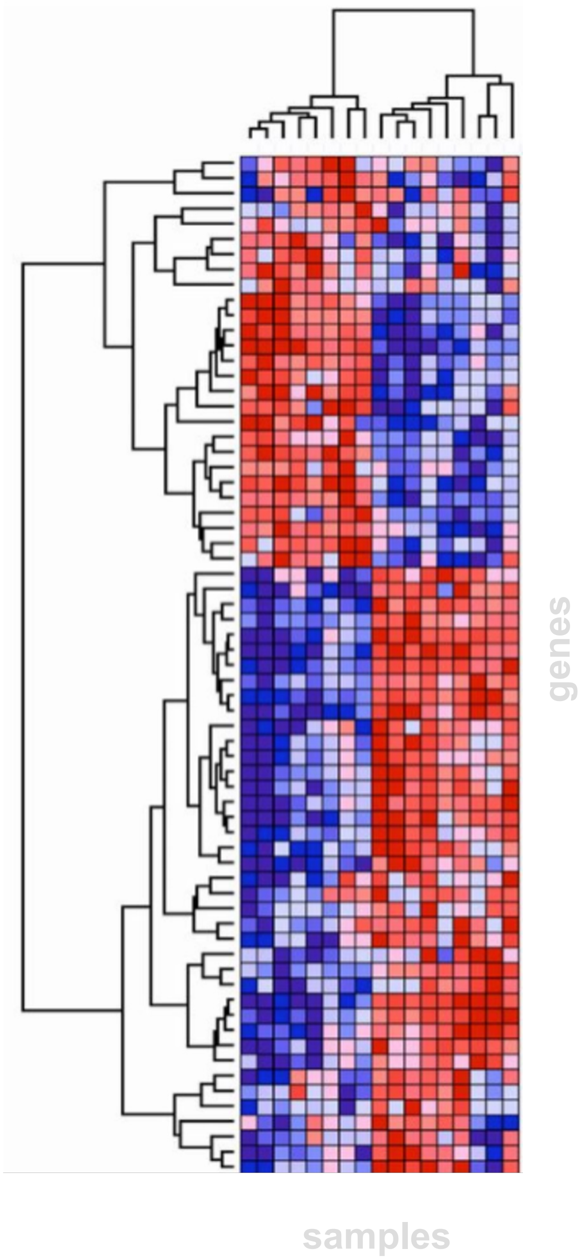
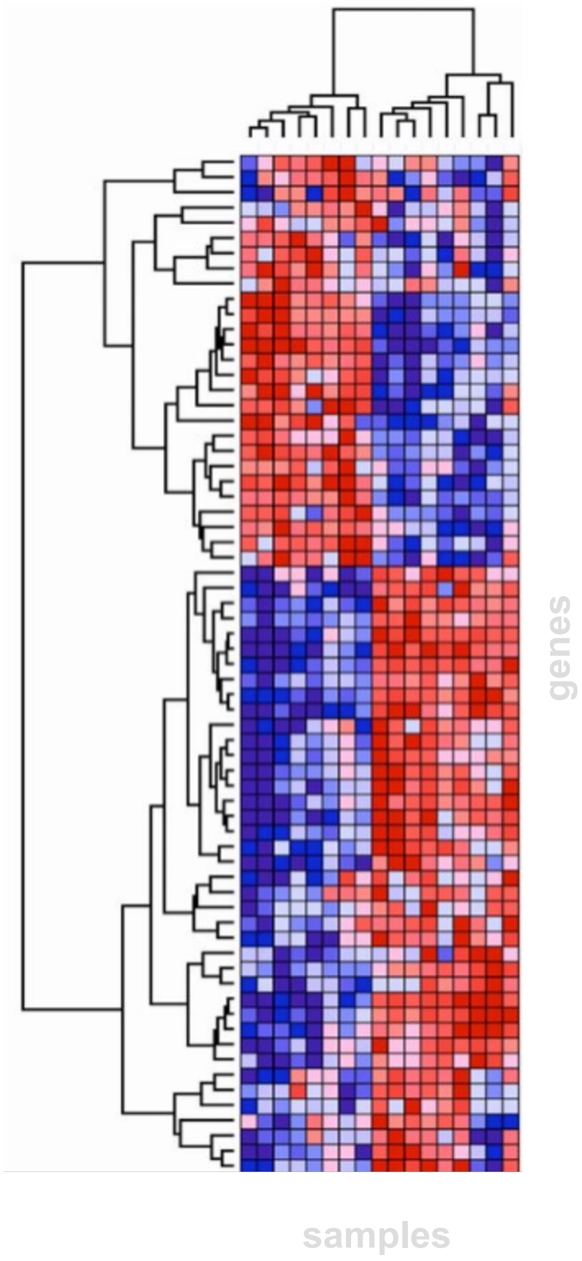


Figure from Greenough, et al. "A gene expression signature that correlates with CD8+ T cell expansion in acute EBV infection." *The Journal of Immunology* 195.9 (2015).

Gene Expression Signatures

- Example Comparisons
  - Mutated vs Wild-Type
  - Metastatic vs Primary
  - Tumor vs Normal
  - Perturbagens
    - Drug Treatment vs Placebo
    - Environmental Stimuli vs Control
- Gene Signatures provide a uniquely characteristic pattern of gene expression that is tied to its studied biological or medical phenomenon
  - Enable researchers to relate samples and other phenomenon by finding the similarity to the gene signatures
  - Focus understanding on underlying mechanism for phenomenon to a subset of gene behaviors



Public Resources for Gene Signatures

- There are many public resources for acquiring gene expression signatures
  - Extracting signatures yourself

    GEO2Enrichr

    GEO2R

  - Libraries of Curated Signatures

    GEN3VA

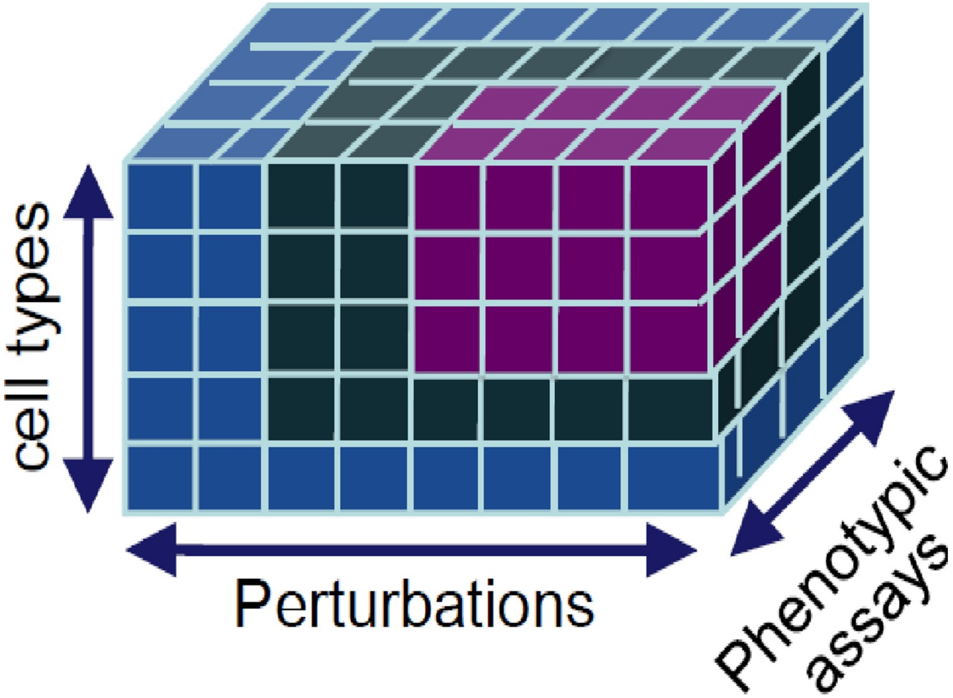
    MSigDB Molecular Signatures Database

- Lab will use signatures from the Library of Integrated Network-Based Cellular Signatures (LINCS)

NIH LINCS PROGRAM

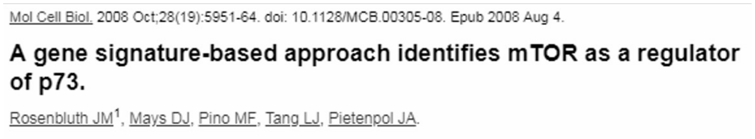
The LINCS DataCube of Signatures

- Gathering a data cube of gene signatures
- Using many different:
  - Cell Types
    - Dozens of cell lines
    - Induced pluripotent stem cells
    - Primary Cells
  - Perturbagens
    - Small molecules / Drugs
    - CRISPR overexpression and
    - shRNA knockdown
    - Microenvironments
    - Ligands
  - Experimental Assays
    - Gene expression: microarray, RNA-seq, L1000
    - Protein expression: RPPA, P100 mass spectrometry
    - Morphological and Proliferation: biochemical and imaging assays



Signature Similarity Analysis

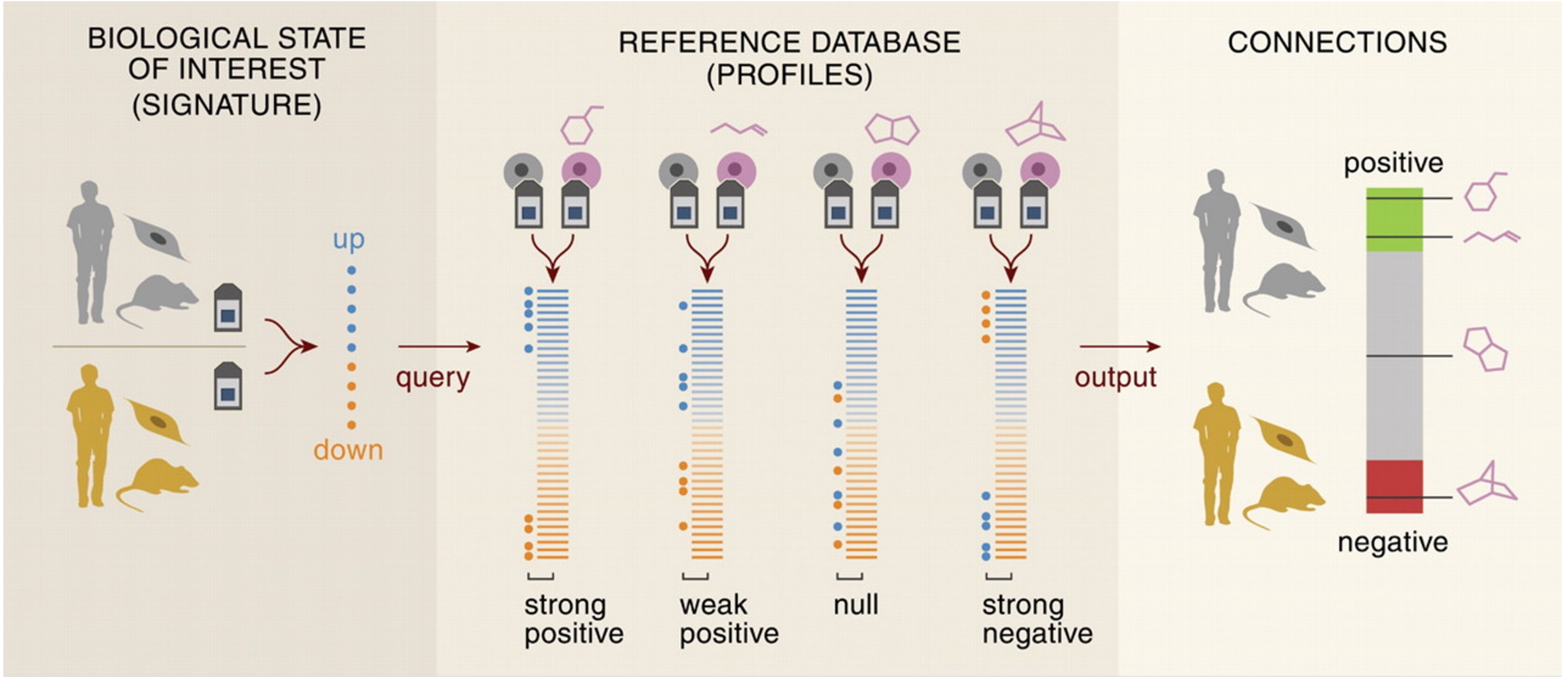
- Given a query signature and a library of reference signatures, how do you find the similar signatures?



Mol Cell Biol. 2008 Oct;28(19):5951-64. doi: 10.1128/MCB.00305-08. Epub 2008 Aug 4.

A gene signature-based approach identifies mTOR as a regulator of p73.

Rosenbluth JM[1], Mays DJ, Pino MF, Tang LJ, Pietenpol JA.



<u>Types of Similarity Comparisons</u>

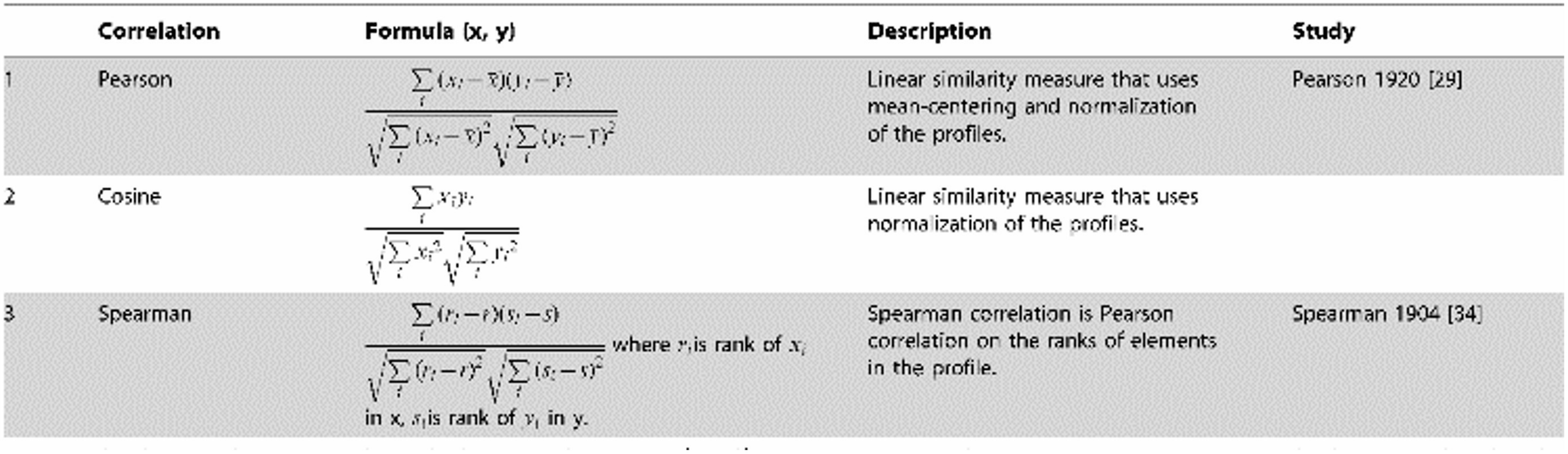
Gene Set & Differential Expression Vector

Differential Expression Vector & Differential Expression Vector

Gene Set & Gene Set

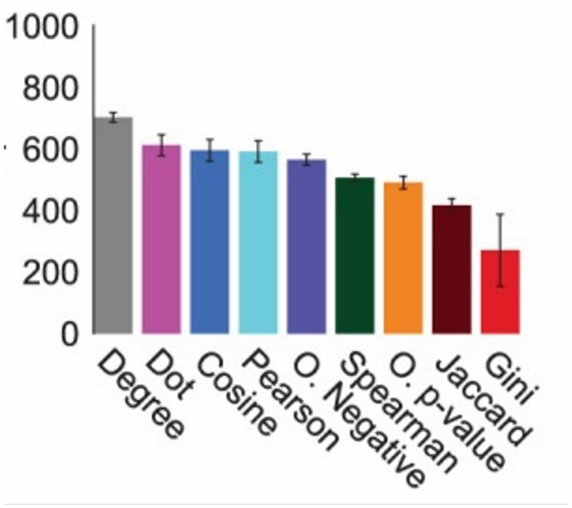
Standard Similarity Measures

- When both signatures are represented as differential expression vectors:



	Correlation	Formula (x, y)	Description	Study
1	Pearson	$\frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2}\sqrt{\sum_i (y_i-\bar{y})^2}}$	Linear similarity measure that uses mean-centering and normalization of the profiles.	Pearson 1920 [29]
2	Cosine	$\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$	Linear similarity measure that uses normalization of the profiles.	
3	Spearman	$\frac{\sum_i (r_i-\bar{r})(s_i-\bar{s})}{\sqrt{\sum_i (r_i-\bar{r})^2}\sqrt{\sum_i (s_i-\bar{s})^2}}$ where r_i is rank of x_i in x, s_i is rank of y_i in y.	Spearman correlation is Pearson correlation on the ranks of elements in the profile.	Spearman 1904 [34]

- In one analysis, they did not observe a large performance difference between the possible measures



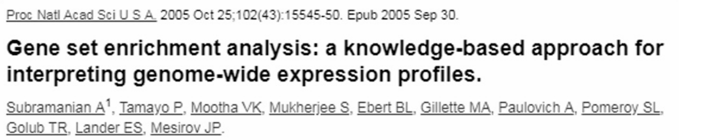
PLoS One. 2013 Jul 10;8(7):e68664. doi: 10.1371/journal.pone.0068664. Print 2013.

Comparison of profile similarity measures for genetic interaction networks.

Deshpande R[1], Vandersluis B, Myers CL.

Gene Set Enrichment Analysis

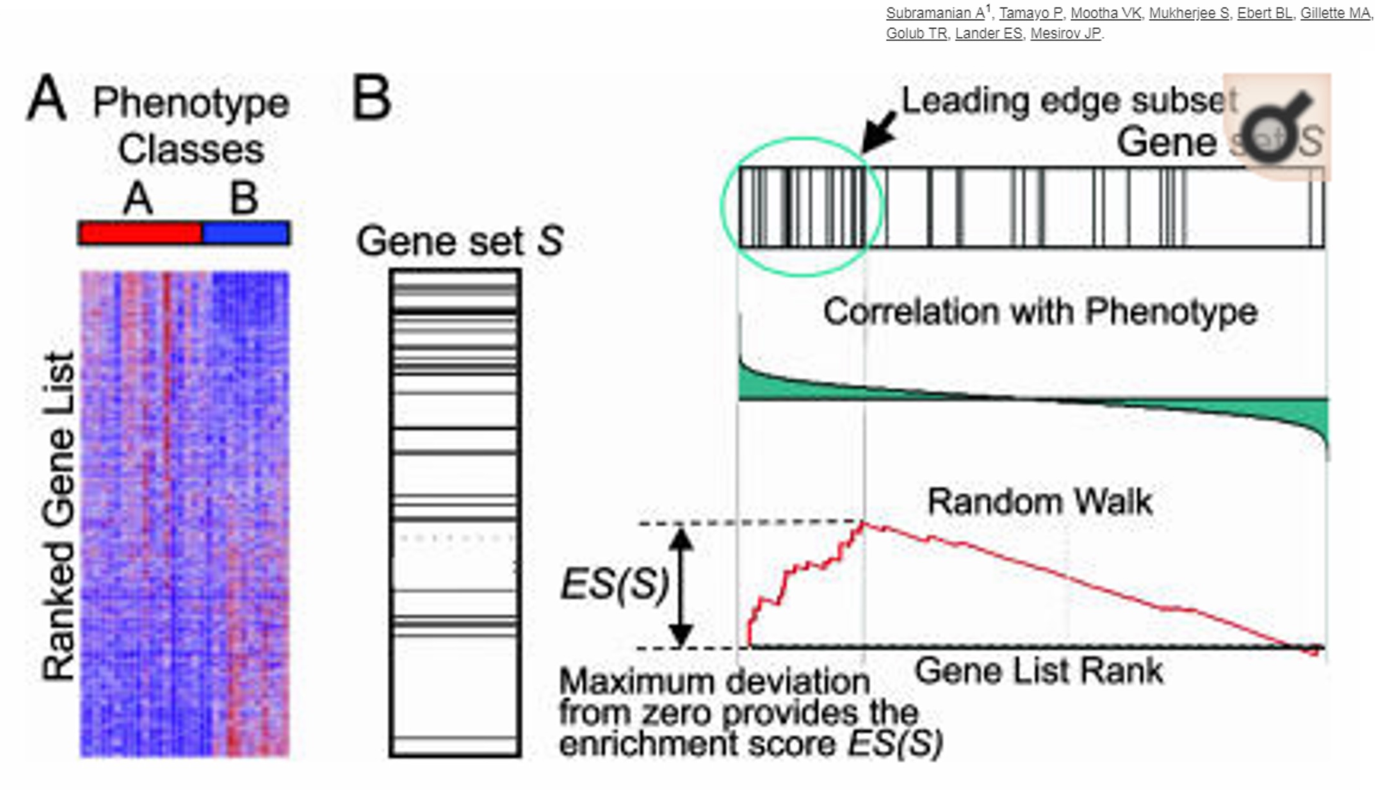
- When sample signature is **vector** and library signature is **gene set**
  - GSEA - http://software.broadinstitute.org/gsea/index.jsp



Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. Epub 2005 Sep 30.

Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.

Subramanian A[1], Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.

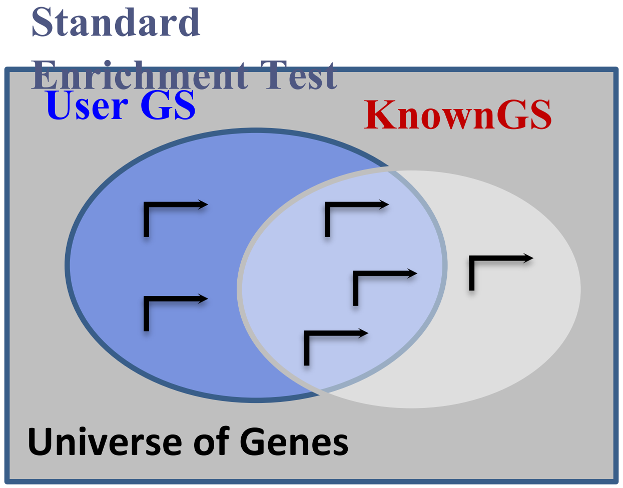


Maximum deviation from zero provides the enrichment score *ES(S)*

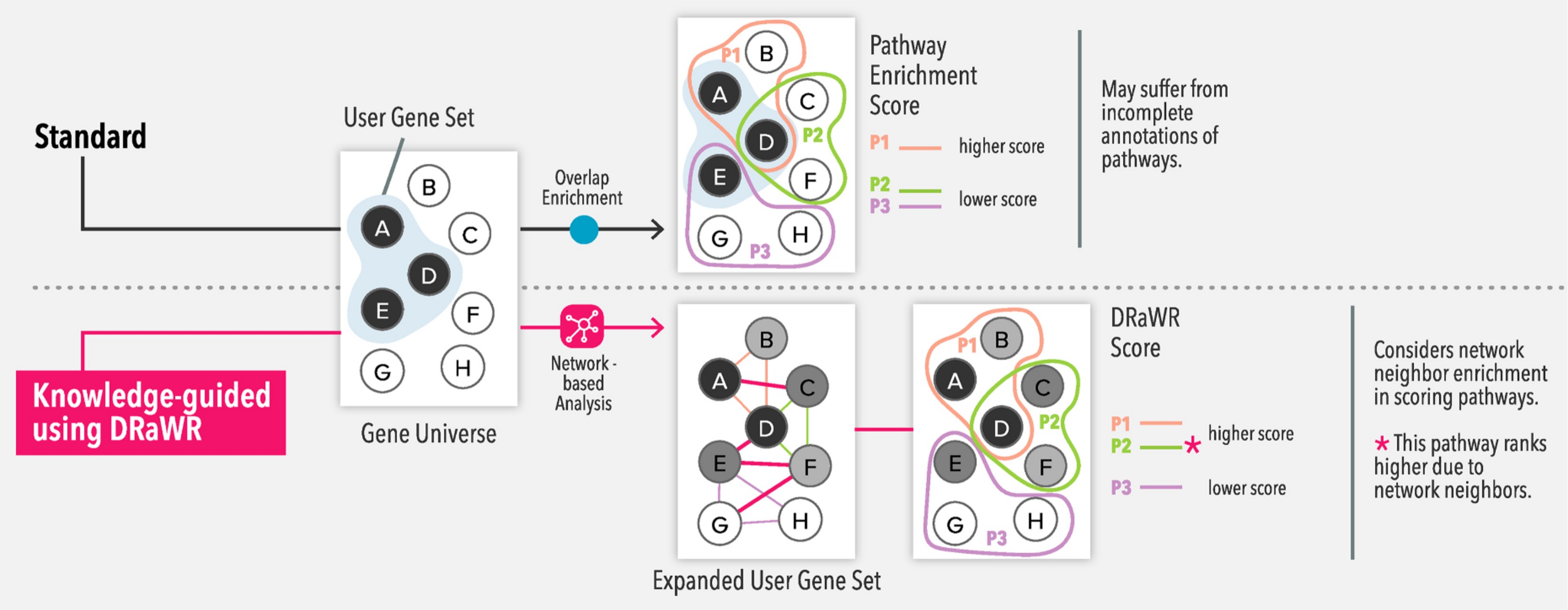
- Modification of the Kolmogorov-Smirnov Statistic
  - Calculate the enrichment score (ES) that represents the amount the genes in the gene set are over-represented in the top or the bottom of the signature vector
  - Estimate statistical significance of the ES by permuting the mappings between the data
  - Adjust for multiple hypothesis testing when analyzing a large number of gene sets

Gene Set Association Tests

- For use when **both** signatures are **gene sets**
  - Also known as Gene Set Characterization
- One-sided exact Fisher / Hypergeometric distribution tests
  - Covered by Saurabh this morning
- Available through tools like:
  - DAVID - https://david.ncifcrf.gov/
  - Enrichr - http://amp.pharm.mssm.edu/Enrichr/
  - Metascape - http://metascape.org/gp/index.html



Network-Guided Gene Set Characterization



Standard

User Gene Set

Overlap Enrichment

Pathway Enrichment Score

P1 higher score

P2 P3 lower score

May suffer from incomplete annotations of pathways.

Knowledge-guided using DRaWR

Gene Universe

Network - based Analysis

Expanded User Gene Set

DRaWR Score

P1 P2 * higher score

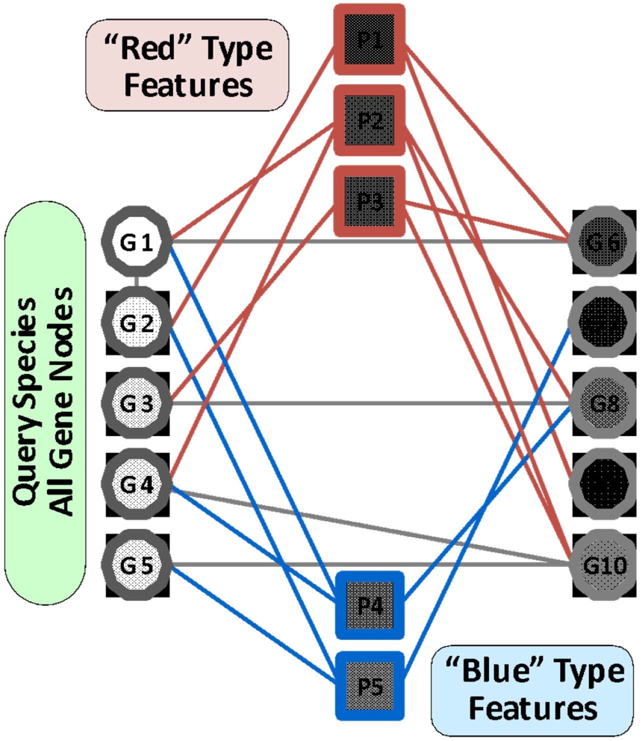
P3 lower score

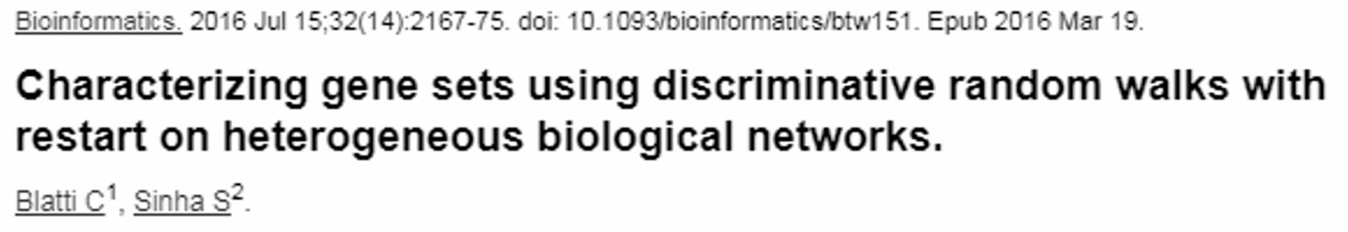
Considers network neighbor enrichment in scoring pathways.

* This pathway ranks higher due to network neighbors.

Idea for a Network-based Method

- Use guilt-by-association principles to find out which annotations are well connected to the query genes in a heterogeneous network.
- These well connected annotations should be specific to the query genes, and not simply hub nodes in the network.
- Developed Discriminative Random Walks with Restart (DRaWR)



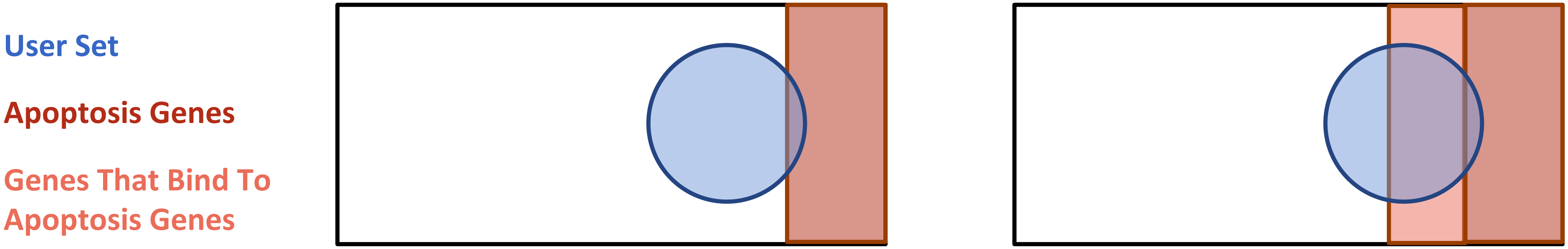

Bioinformatics. 2016 Jul 15;32(14):2167-75. doi: 10.1093/bioinformatics/btw151. Epub 2016 Mar 19.

Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks.

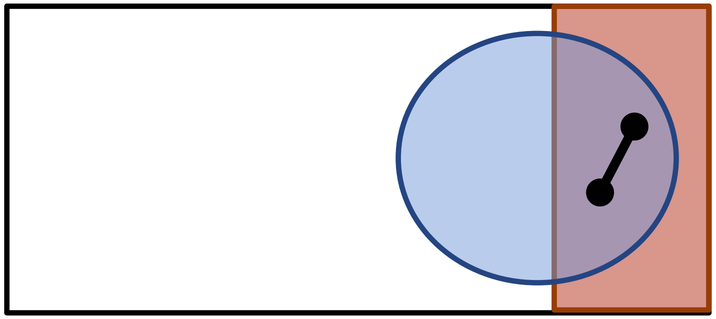
Blatti C[1], Sinha S[2].

Value of Network-Guided Analysis

- Take advantage of gene neighbors

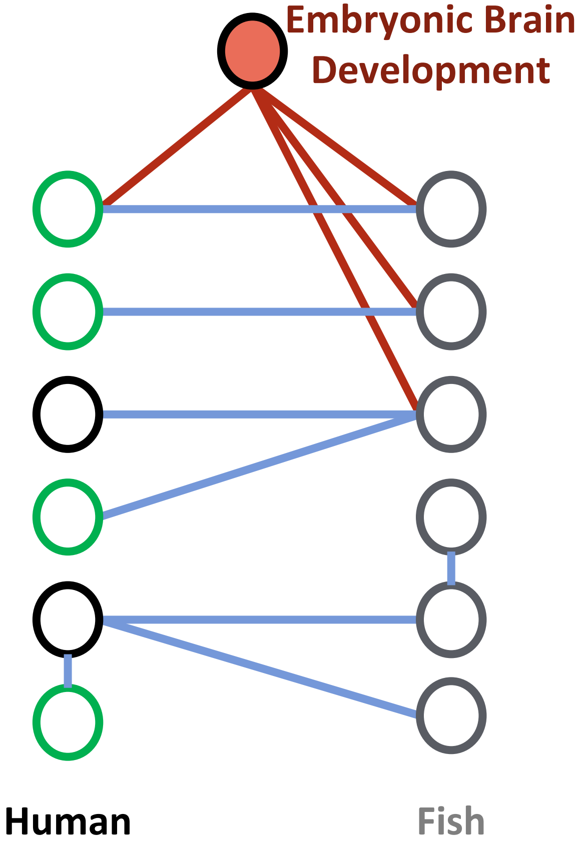


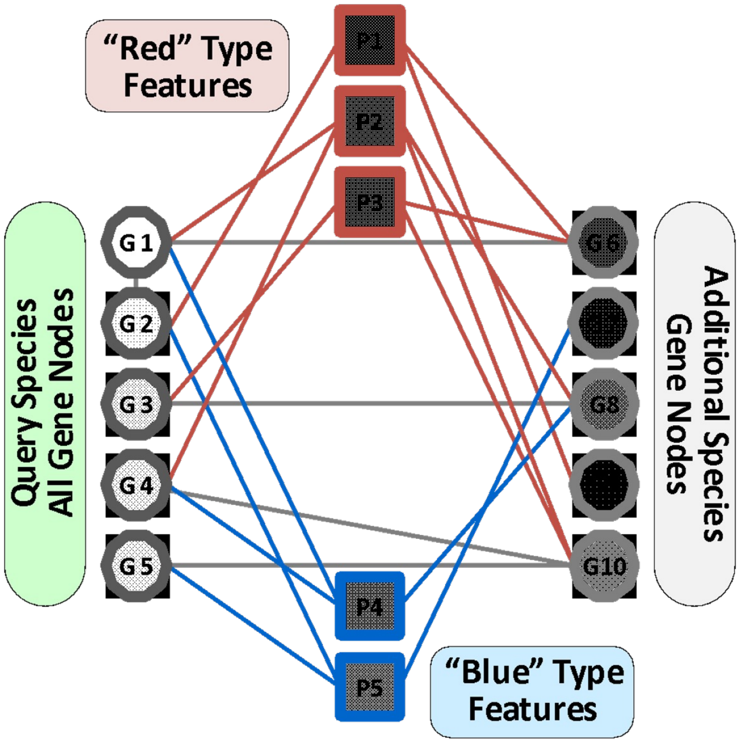
- Incorporate dependencies from separate knowledge in analysis



Value of Network-Guided Analysis

- Extension to poorly annotated domains

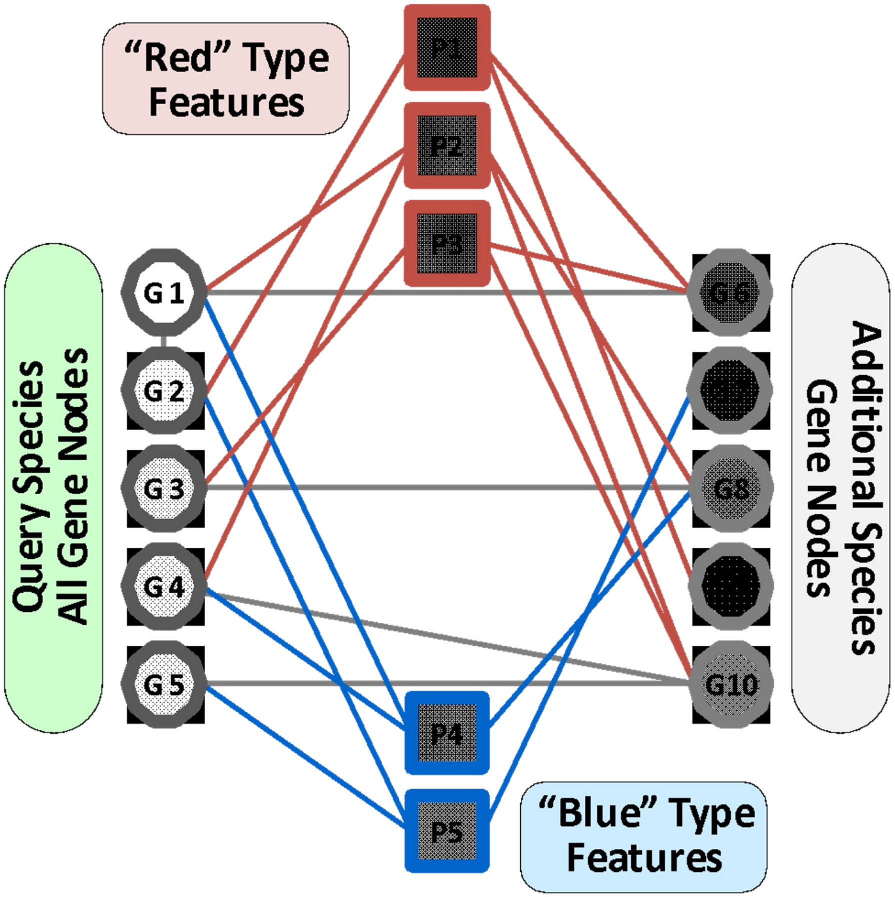




- Integrating multiple data types

Network-based DRaWR Method

- DRaWR – using random walks on a network
    - Construct a heterogeneous network of interest

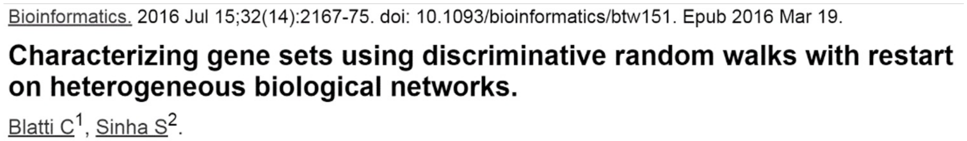


Heterogeneous Edge Types

type_A

type_B

type_C

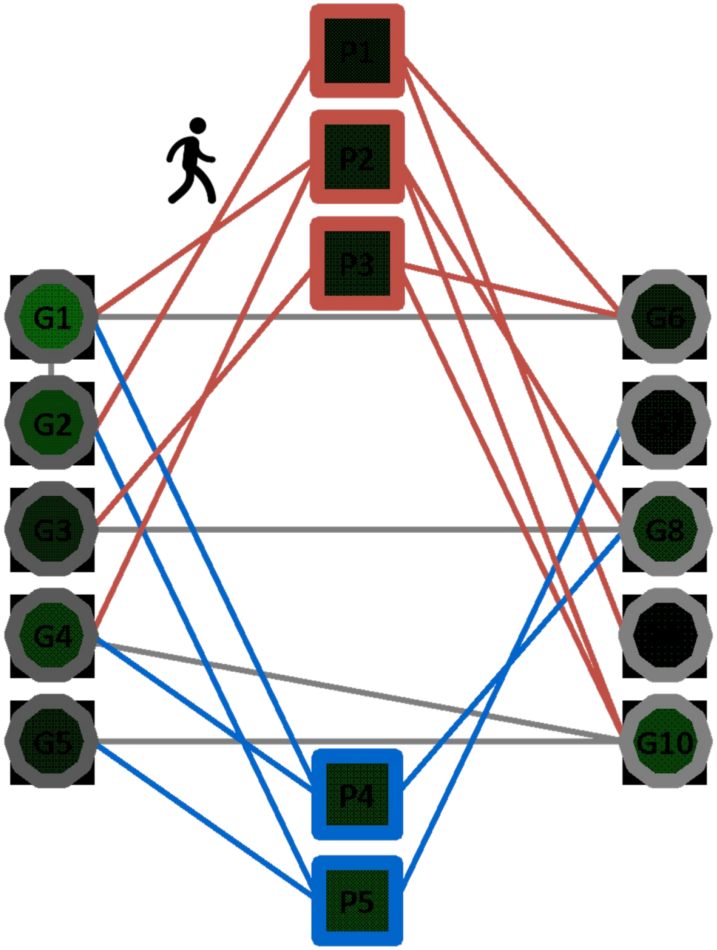

Bioinformatics. 2016 Jul 15;32(14):2167-75. doi: 10.1093/bioinformatics/btw151. Epub 2016 Mar 19.

Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks.

Blatti C[1], Sinha S[2].

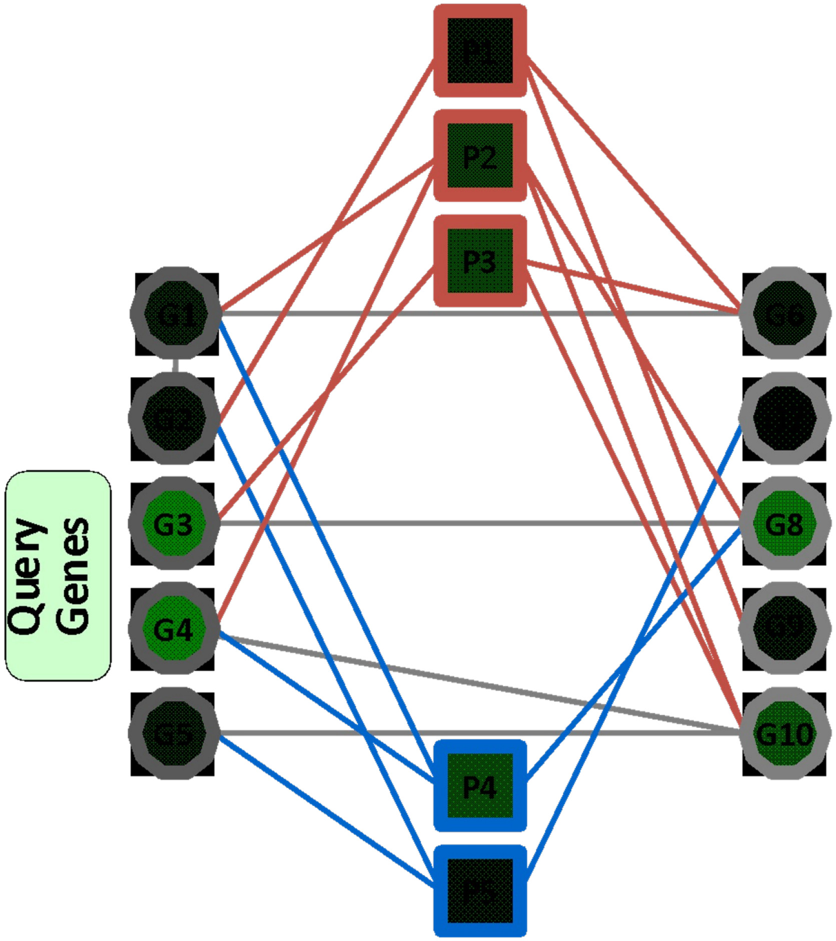
Network Methods for GSC

- DRaWR – using random walks on a network
  - Construct a network of interest
  - Find stationary distribution on network



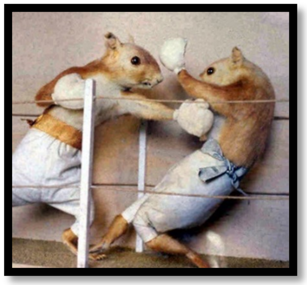
Network Methods for GSC

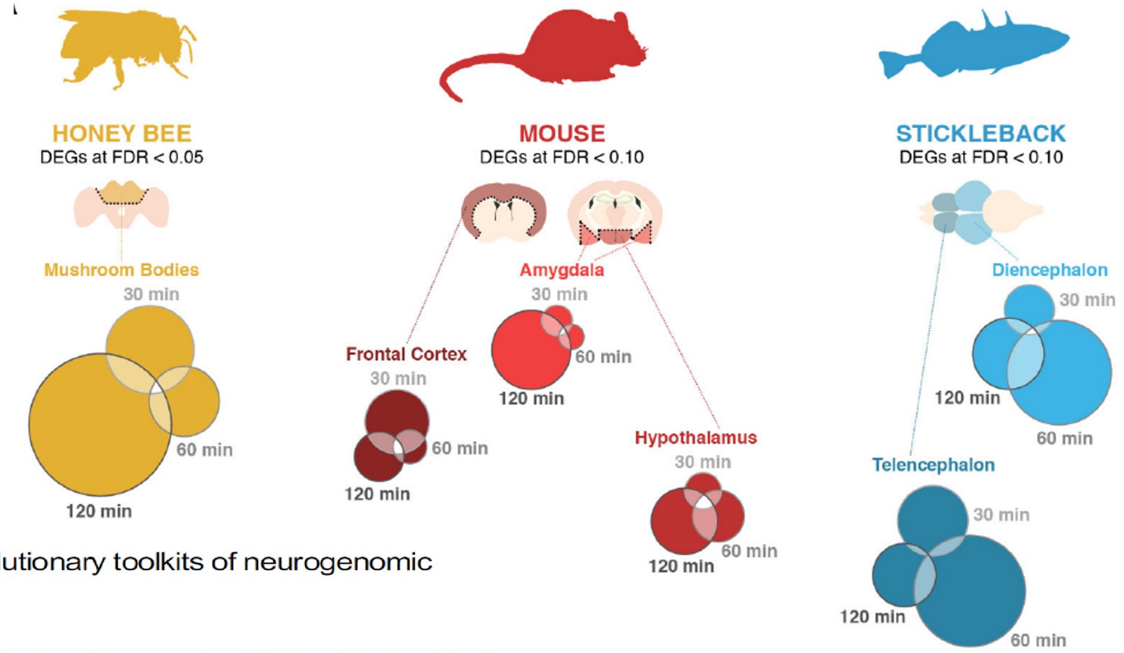
- DRaWR – using random walks on a network
  - Construct a network of interest
  - Find stationary distribution on network
  - Find gene set specific distribution
  - Return annotation nodes that are especially related to the query



Social Aggression Study Application

- Idea: Evolutionary “toolkits” – genes and modules with lineage-specific variations but deep conservation of function
- Questions: Are there toolkits that underlie social behaviors
    - Such as aggressive response to territorial intrusions?
- Study: gather brain transcriptomic responses to social challenge from three social species – honey bees, mice, and stickleback fish
    - With and without exposure to intraspecies intruder
    - From different brain regions and/or durations after event
- Results: sets of differentially expressed genes across three species



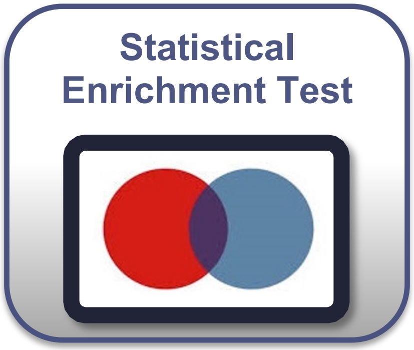


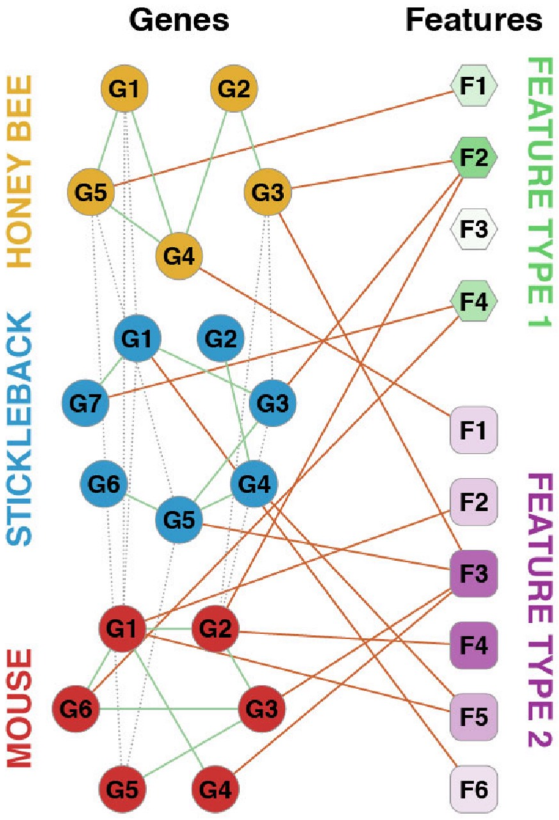
Cross-species systems analysis of evolutionary toolkits of neurogenomic response to social challenge

Michael C. Saul[1], Charles Blatti[1,2], Wei Yang[1,2], Syed Abbas Bukhari[1,3], Hagai Y. Shpigler[1,4], Joseph M. Troy[1,3], Christopher H. Seward[1,5], Laura Sloofman[1,6], Sriram Chandrasekaran[7], Alison M. Bell[1,3,8,9], Lisa Stubbs[1,3,5,9], Gene E. Robinson[1,9,10], Sihai Dave Zhao[1,11,*], and Saurabh Sinha[1,2,10,*].

Failure of Standard Approach

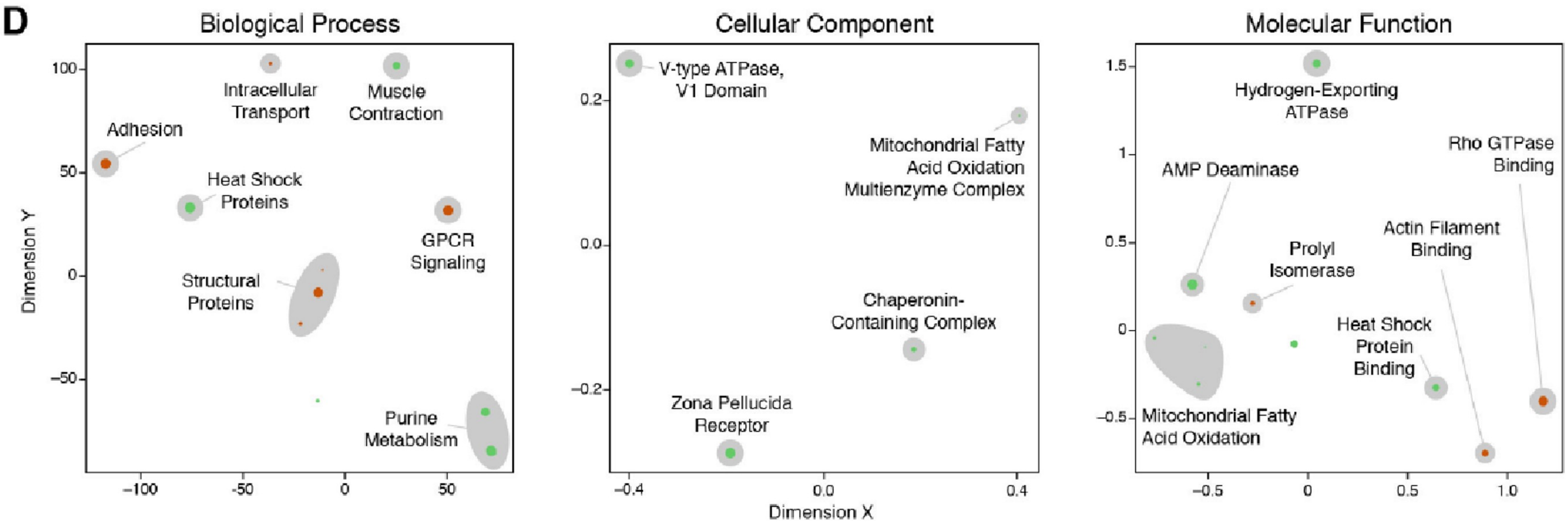
- Would like to find Gene Ontology annotations that:
  - Relate to DE gene sets of all three species
    - However, Gene Ontology annotation quality varies greatly in three species
  - Or relate to DE genes sets of the Mouse
    - However, the corresponding sets from the other species might have greatly different function
- Solution:
  - Integrate Orthology and Gene Ontology information in a three species network
  - Find Gene Ontology terms that are strongly connected to the DE gene sets of all three species simultaneously



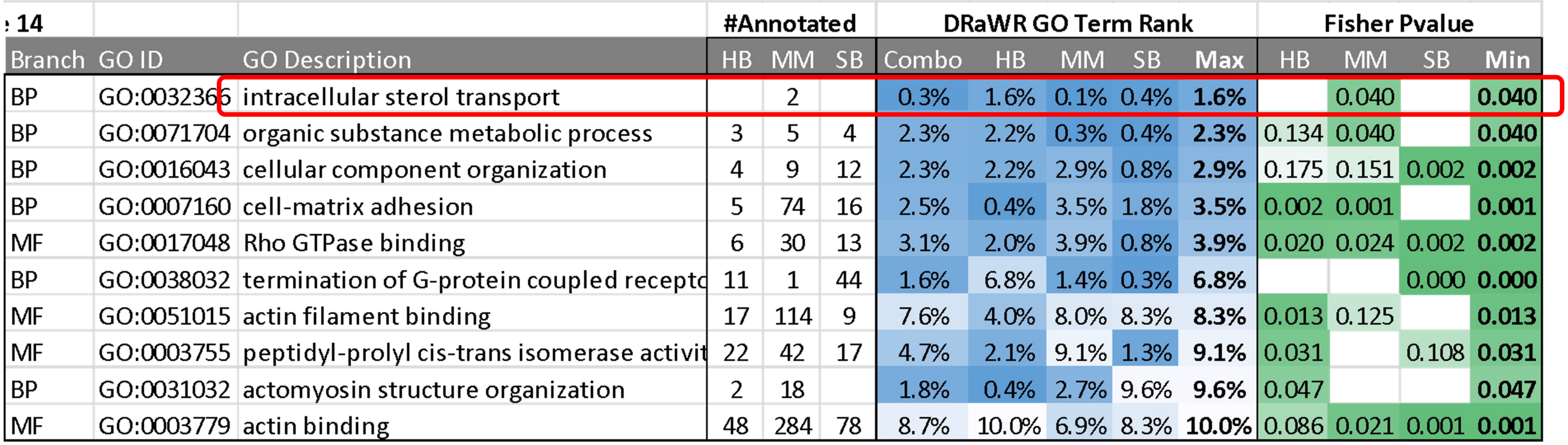


Findings with DRaWR

- Annotations of two (red and green) conserved Gene Modules



- Specific results for red module



14			#Annotated			DRaWR GO Term Rank					Fisher Pvalue			
Branch	GO ID	GO Description	HB	MM	SB	Combo	HB	MM	SB	**Max**	HB	MM	SB	**Min**
BP	GO:0032366	intracellular sterol transport		2		0.3%	1.6%	0.1%	0.4%	**1.6%**		0.040		**0.040**
BP	GO:0071704	organic substance metabolic process	3	5	4	2.3%	2.2%	0.3%	0.4%	**2.3%**	0.134	0.040		**0.040**
BP	GO:0016043	cellular component organization	4	9	12	2.3%	2.2%	2.9%	0.8%	**2.9%**	0.175	0.151	0.002	**0.002**
BP	GO:0007160	cell-matrix adhesion	5	74	16	2.5%	0.4%	3.5%	1.8%	**3.5%**	0.002	0.001		**0.001**
MF	GO:0017048	Rho GTPase binding	6	30	13	3.1%	2.0%	3.9%	0.8%	**3.9%**	0.020	0.024	0.002	**0.002**
BP	GO:0038032	termination of G-protein coupled recepto	11	1	44	1.6%	6.8%	1.4%	0.3%	**6.8%**			0.000	**0.000**
MF	GO:0051015	actin filament binding	17	114	9	7.6%	4.0%	8.0%	8.3%	**8.3%**	0.013	0.125		**0.013**
MF	GO:0003755	peptidyl-prolyl cis-trans isomerase activit	22	42	17	4.7%	2.1%	9.1%	1.3%	**9.1%**	0.031		0.108	**0.031**
BP	GO:0031032	actomyosin structure organization	2	18		1.8%	0.4%	2.7%	9.6%	**9.6%**	0.047			**0.047**
MF	GO:0003779	actin binding	48	284	78	8.7%	10.0%	6.9%	8.3%	**10.0%**	0.086	0.021	0.001	**0.001**

Gene Ranking / Function Prediction

- Given:
  - Novel gene set(s) generated by a genomic researcher
- Task:
  - **Rank** genes for the strength of their relationship to the user’s gene set(s)…
    - … in order to assess the coherence of the genes in the experimental gene set or identify putative related genes

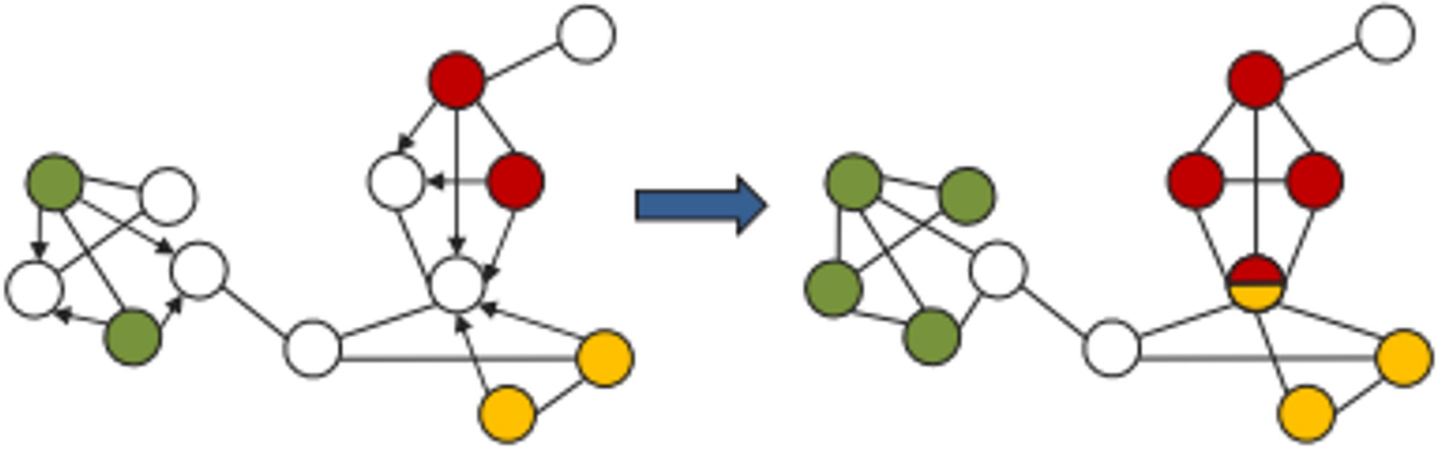
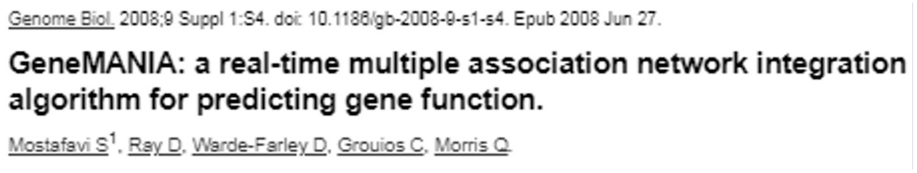


Figure from Arzt, et al. "Pipa: custom integration of protein interactions and pathways." *GI-Jahrestagung*. 2011.

GeneMANIA Approach

- GeneMANIA stands for
  - Multiple Association Network Integration Algorithm
- Main Idea
  - Given a gene set with a known functions
  - And several gene-gene interaction affinity networks
  - Find genes that relate to the functional set through the edges of the given networks
- Approach
  - Find out how well each network predicts the membership of the given set
    - A linear regression-based algorithm that calculates a single composite functional association network from multiple data sources
  - Do label propagation guilt-by-association algorithm on the composite functional association network



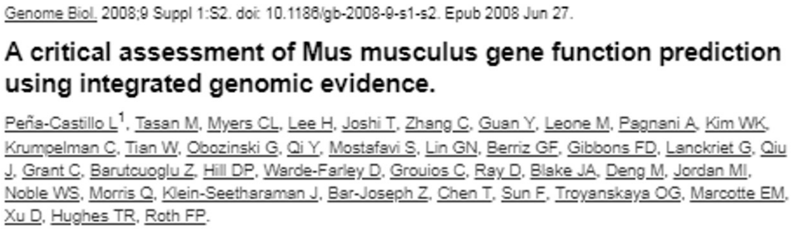
Genome Biol. 2008;9 Suppl 1:S4. doi: 10.1186/gb-2008-9-s1-s4. Epub 2008 Jun 27.

GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function.

Mostafavi S[1], Ray D, Warde-Farley D, Grouios C, Morris Q.

GeneMANIA Performance

- Participated in grand challenge for this function prediction task on Mouse genes

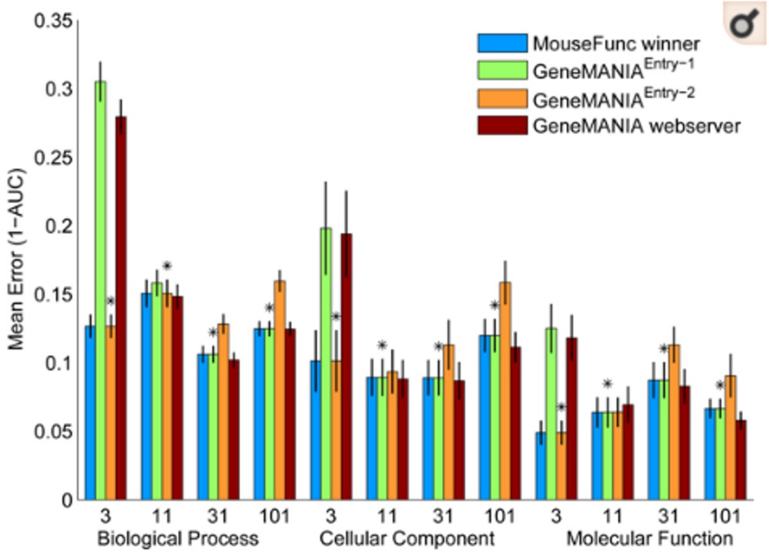


Genome Biol. 2008;9 Suppl 1:S2. doi: 10.1186/gb-2008-9-s1-s2. Epub 2008 Jun 27.

A critical assessment of Mus musculus gene function prediction using integrated genomic evidence.

Peña-Castillo L[1], Tasan M, Myers CL, Lee H, Joshi T, Zhang C, Guan Y, Leone M, Pagnani A, Kim WK, Krumpelman C, Tian W, Obozinski G, Qi Y, Mostafavi S, Lin GN, Berriz GF, Gibbons FD, Lanckriet G, Qiu J, Grant C, Barutcuoglu Z, Hill DP, Warde-Farley D, Grouios C, Ray D, Blake JA, Deng M, Jordan MI, Noble WS, Morris Q, Klein-Seetharaman J, Bar-Joseph Z, Chen T, Sun F, Troyanskaya OG, Marcotte EM, Xu D, Hughes TR, Roth FP.

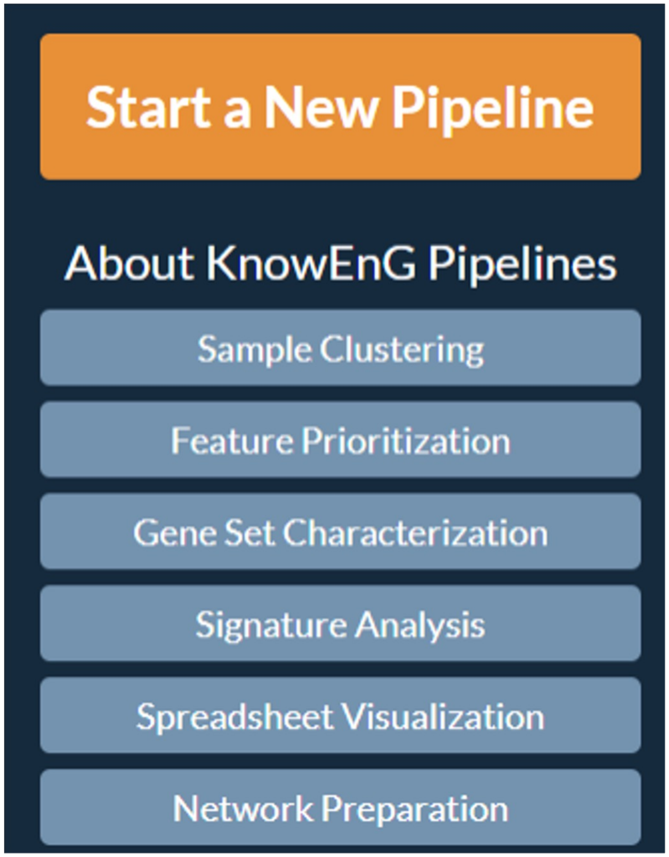
- Did extraordinary well in the competition and has improve method since then



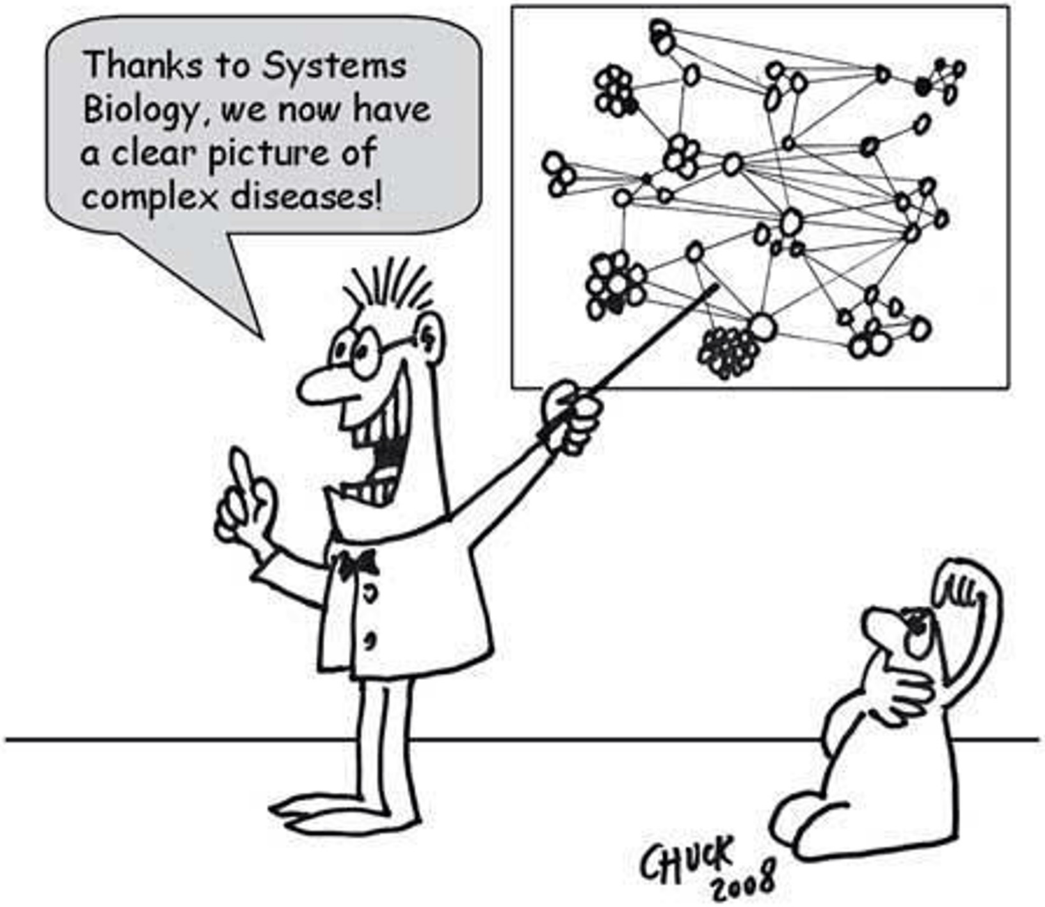
- Has easy to use webserver for running functional prediction with small genesets

In this Lecture and the Lab

- Biological Knowledge Networks
  - KnowEnG Platform
- Network-Guided Sample Clustering
  - Network Based Stratification, COCA
- Network-Guided Gene Prioritization
  - ProGENI
- Gene Signatures and Similarity Methods
  - LINCS, GSEA, Enrichr, DAVID
- Network-based Gene Set Characterization
  - DRaWR
- Network-based Function Prediction
  - GeneMANIA



Thank you, Any Questions?


Thanks to Systems Biology, we now have a clear picture of complex diseases!
CHUCK 2008

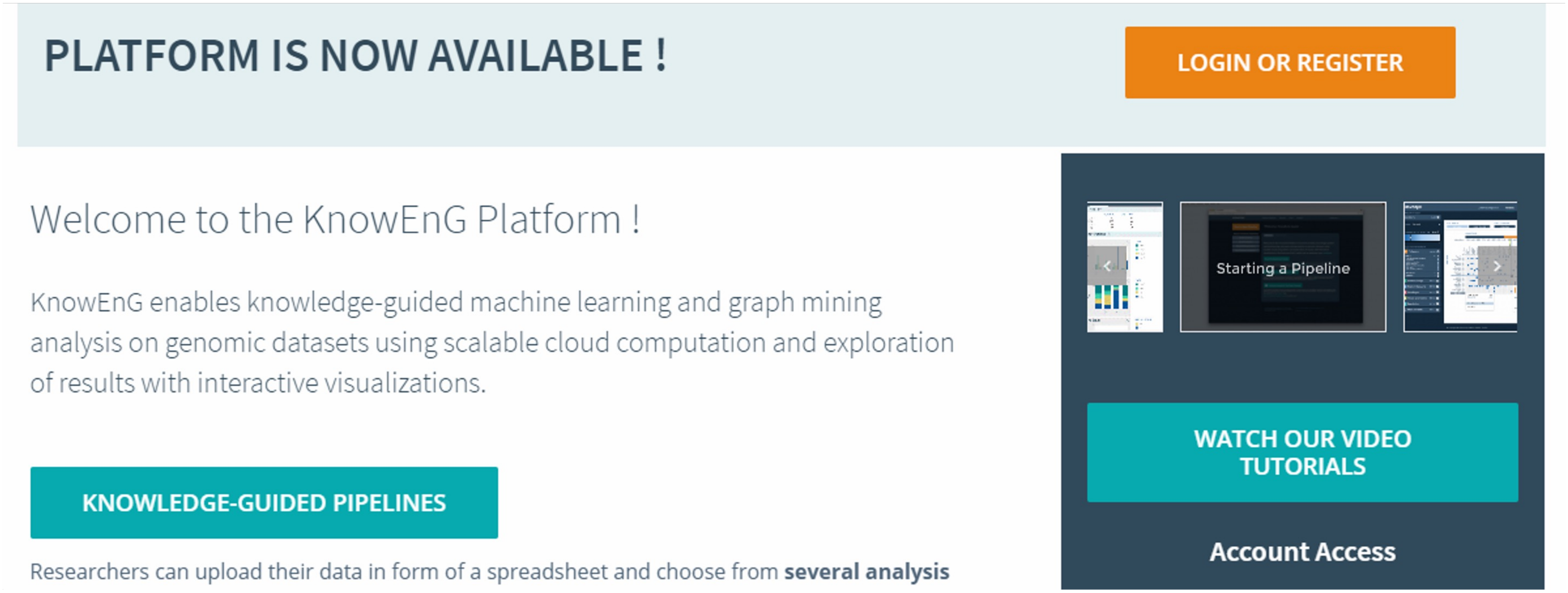
Reproduced by permission of Pierre de Meyts in
Ebrahim. "Metabolomics, nutrition and why epidemiology matters." (2016).

KnowEnG Resources

- Also Check Out:
  - Network Preparation for uploading your custom network to the platform for analysis
  - Signature Analysis for mapping samples to signatures by correlation of omics profiles
- Tutorials:
  - Quickstarts: https://knoweng.org/quick-start/
  - YouTube: https://www.youtube.com/channel/UCjyIIolCaZIGtZC20XLBOyg
- Resources:
  - Data Preparation Guide: https://github.com/KnowEnG/quickstart-demos/blob/master/pipeline_readmes/README-DataPrep.md
  - Knowledge Network Contents:
    - Summary: https://knoweng.org/kn-data-references/
    - Download: https://github.com/KnowEnG/KN_Fetcher/blob/master/Contents.md
- Research
  - Knowledge-guided analysis of omics Data (KnowEng cloud platform paper): https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000583
  - TCGA Analysis Walkthrough: https://github.com/KnowEnG/quickstart-demos/tree/master/publication_data/blatti_et_al_2019
- Source Code:
  - Docker Images: https://hub.docker.com/u/knowengdev/
  - Github Repos: https://knoweng.github.io/
- Other Cloud Platforms
  - https://cgc.sbgenomics.com/public/apps#q?search=knoweng
- Contact Us with Questions and Feedback: knoweng-support@illinois.edu

Using A Permanent KnowEnG Account

- For permanent account:
  - Go to https://knoweng.org/analyze/
    Click on “Create an account”
  - Follow the instructions



PLATFORM IS NOW AVAILABLE !

Welcome to the KnowEnG Platform !

KnowEnG enables knowledge-guided machine learning and graph mining analysis on genomic datasets using scalable cloud computation and exploration of results with interactive visualizations.

Researchers can upload their data in form of a spreadsheet and choose from **several analysis**

Regression algorithms

- Lasso: learns a linear model from the training data using only a few features (sparse linear model)

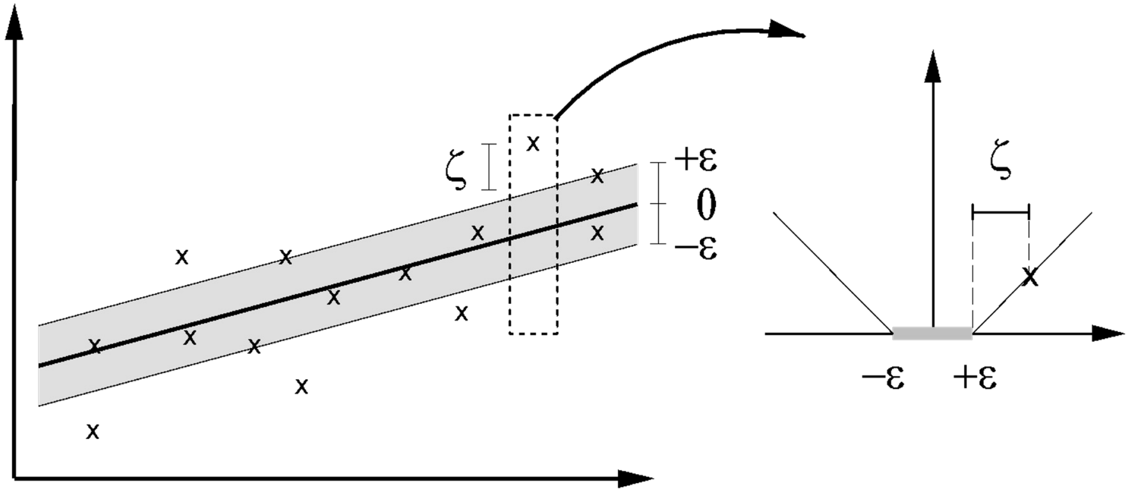
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda_1 ||\boldsymbol{\beta}||_1 \right)$$

- Elastic Net: learns a linear model from the training data by linearly combining ridge and Lasso regression regularization terms (a generalization of both Lasso and ridge regression)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda_2 ||\boldsymbol{\beta}||_2 + \lambda_1 ||\boldsymbol{\beta}||_1 \right)$$

Regression algorithms

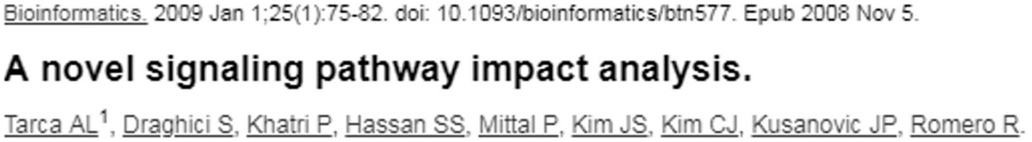
- Kernel-SVR:
  - Linear SVR learns a linear model such that it has at most ε-deviation from the response values and is as flat as possible



(Smola and Schölkopf, 1998)

  - Kernel-SVR generalizes the idea to nonlinear models by mapping the features to a high-dimensional kernel space

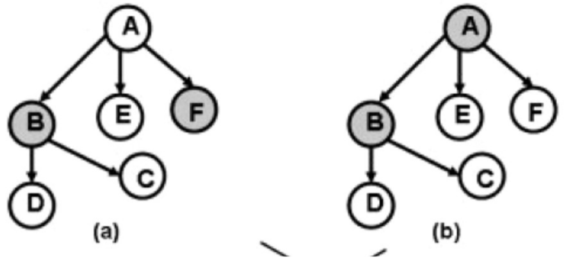
Other Network Based Characterization Methods



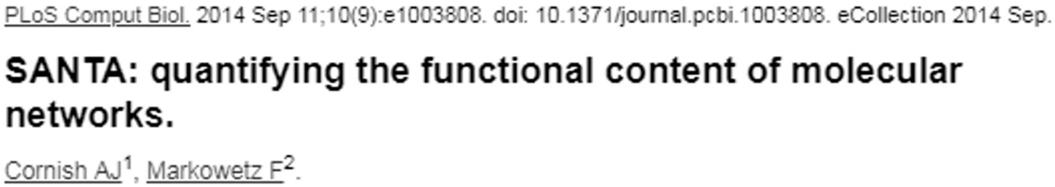
Bioinformatics. 2009 Jan 1;25(1):75-82. doi: 10.1093/bioinformatics/btn577. Epub 2008 Nov 5.

A novel signaling pathway impact analysis.

Tarca AL[1], Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R.



- SPIA Idea:
  - Combine with standard enrichment p-value that asks about the significance of the number of perturbed genes in the pathway
  - Perturbagen p-value, which asks if the amount of total accumulated perturbation after one network propagation step is significant when considering the value it takes with random controls

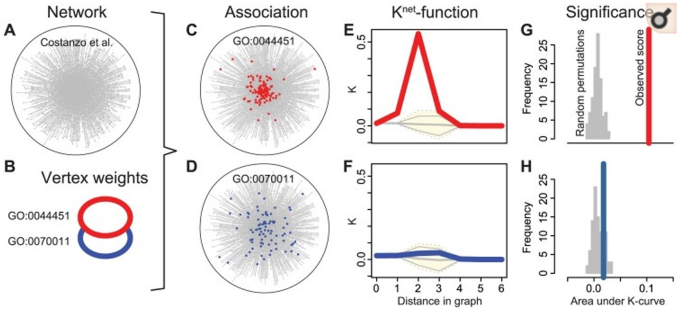


PLoS Comput Biol. 2014 Sep 11;10(9):e1003808. doi: 10.1371/journal.pcbi.1003808. eCollection 2014 Sep.

SANTA: quantifying the functional content of molecular networks.

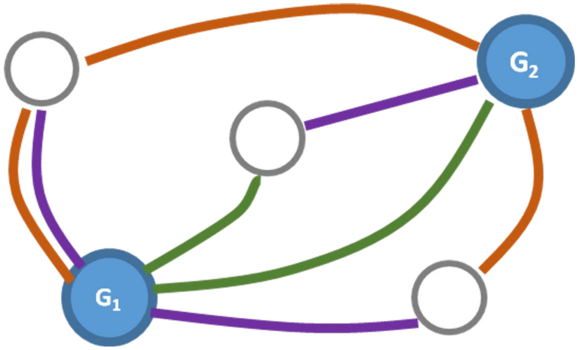
Cornish AJ[1], Markowetz F[2].

Shortest Path Length criteria



Incorporating Meta-Paths

- DRaWR random walks on heterogeneous networks make no consideration / memory of the edge ***types*** they have followed



Paths from G1 -> G2:
type_A
type_A - type_B
type_C- type_C
type_B - type_C (x2)

meta-path:

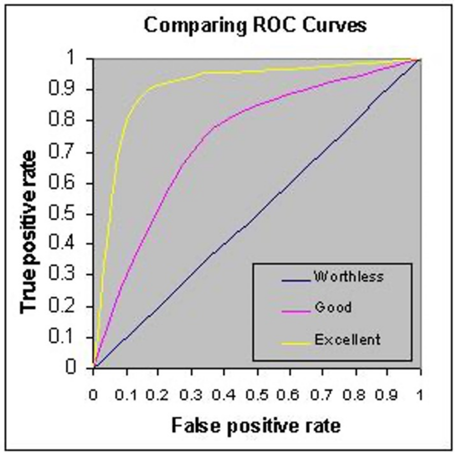
a path defined by sequence of edges types between two nodes

- Explore if similarity in a gene set can best be described by particular ***types of meta-paths*** amongst its genes.

Ranking Genes for Disease

MSigDB Molecular Signatures Database

- Initial Study:
  - 53 MSigDB DE gene sets from separate cancer studies
- Question:
  - If we hide a subset of genes disrupted by the development of cancer, what types of networks are best suited to recover them?
- Evaluation:
  - Partition 75% of DE genes for training, 25% for testing
  - Use DRaWR on KnowNet subnetworks and training data to rank genes
  - Report average AUCs of ranking using test genes as truth



Networks Under Consideration

- Gene-Gene Edge Types
  - H: Homology
  - CoEx: Co-Expression
  - TM: Text Mining
  - Exp: Experimental Interaction
- Gene-Property Edge Types
  - PD: Protein Domains
  - GO: Gene Ontology

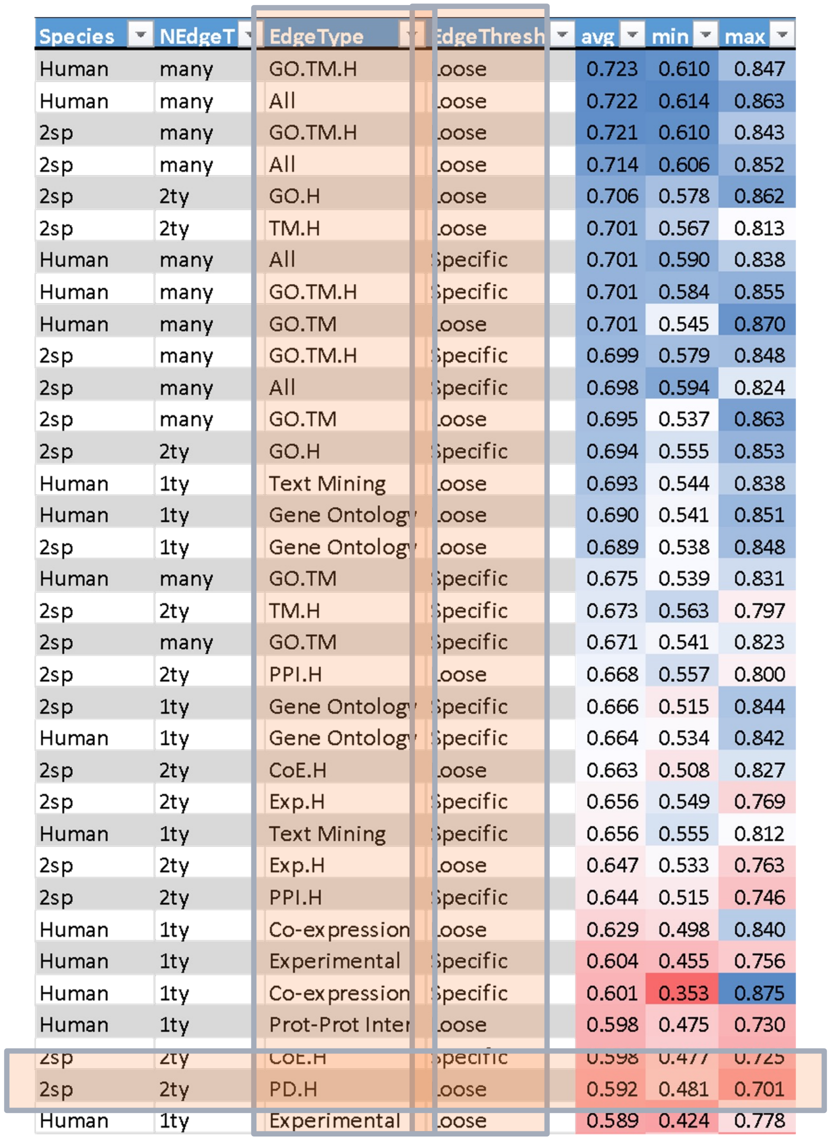
- Number of Species
  - Human: only
  - 2sp: Human and Mouse
- Specificity of the edges
  - Specific: high confidence edges
  - Loose: all edges of that types
- Combinations of Edge Types
  - 1ty: One primary type
  - 2ty: Primary type + homology
  - Many: 3+ edge types

Best Networks

- Gene Ontology annotations and Text Mining relations are the best edge types for recovering cancer set DE genes

- Networks with all edges (Loose) are better at recovering gene than networks with only high confidence edges

- Protein Domain annotations are poor predictors for cancer DE genes, but great for embryonic development



Species	NEdgeT	EdgeType	EdgeThresh	avg	min	max
Human	many	GO.TM.H	Loose	0.723	0.610	0.847
Human	many	All	Loose	0.722	0.614	0.863
2sp	many	GO.TM.H	Loose	0.721	0.610	0.843
2sp	many	All	Loose	0.714	0.606	0.852
2sp	2ty	GO.H	Loose	0.706	0.578	0.862
2sp	2ty	TM.H	Loose	0.701	0.567	0.813
Human	many	All	Specific	0.701	0.590	0.838
Human	many	GO.TM.H	Specific	0.701	0.584	0.855
Human	many	GO.TM	Loose	0.701	0.545	0.870
2sp	many	GO.TM.H	Specific	0.699	0.579	0.848
2sp	many	All	Specific	0.698	0.594	0.824
2sp	many	GO.TM	Loose	0.695	0.537	0.863
2sp	2ty	GO.H	Specific	0.694	0.555	0.853
Human	1ty	Text Mining	Loose	0.693	0.544	0.838
Human	1ty	Gene Ontolog	Loose	0.690	0.541	0.851
2sp	1ty	Gene Ontolog	Loose	0.689	0.538	0.848
Human	many	GO.TM	Specific	0.675	0.539	0.831
2sp	2ty	TM.H	Specific	0.673	0.563	0.797
2sp	many	GO.TM	Specific	0.671	0.541	0.823
2sp	2ty	PPI.H	Loose	0.668	0.557	0.800
2sp	1ty	Gene Ontolog	Specific	0.666	0.515	0.844
Human	1ty	Gene Ontolog	Specific	0.664	0.534	0.842
2sp	2ty	CoE.H	Loose	0.663	0.508	0.827
2sp	2ty	Exp.H	Specific	0.656	0.549	0.769
Human	1ty	Text Mining	Specific	0.656	0.555	0.812
2sp	2ty	Exp.H	Loose	0.647	0.533	0.763
2sp	2ty	PPI.H	Specific	0.644	0.515	0.746
Human	1ty	Co-expression	Loose	0.629	0.498	0.840
Human	1ty	Experimental	Specific	0.604	0.455	0.756
Human	1ty	Co-expression	Specific	0.601	0.353	0.875
Human	1ty	Prot-Prot Inter	Loose	0.598	0.475	0.730
2sp	2ty	CoE.H	Specific	0.598	0.477	0.725
2sp	2ty	PD.H	Loose	0.592	0.481	0.701
Human	1ty	Experimental	Loose	0.589	0.424	0.778