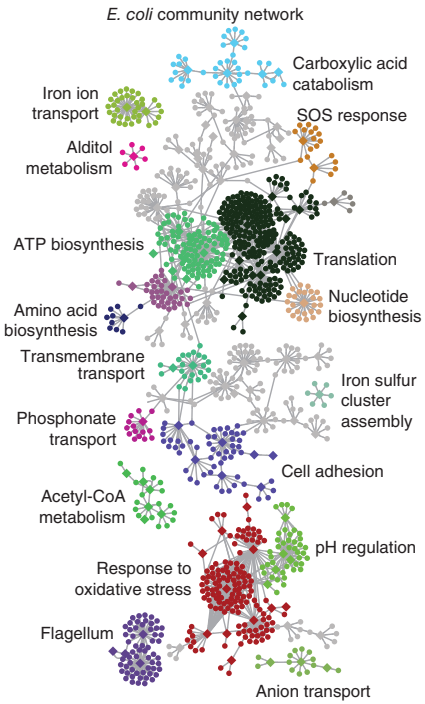


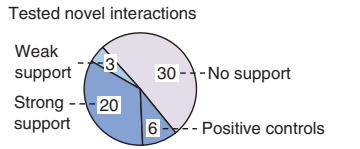
Constructing network models from (and of) expression data

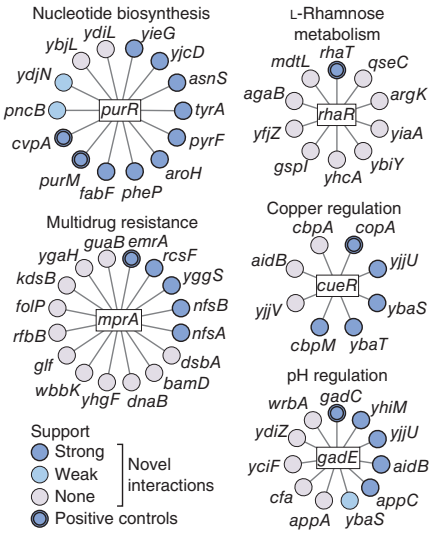
Theodore J. Perkins

Department of Biochemistry, Microbiology and Immunology

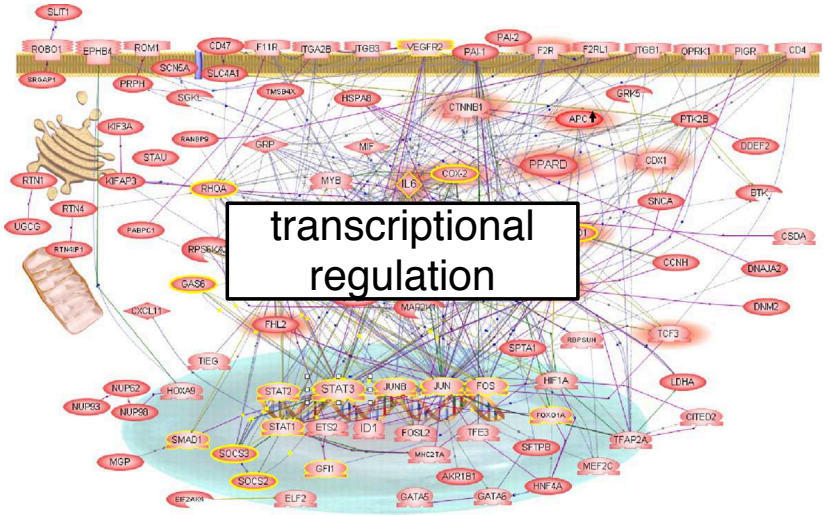
&

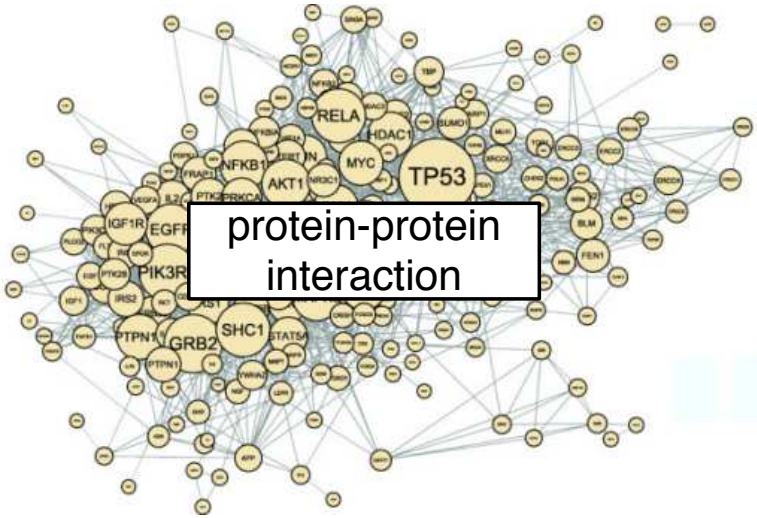
Ottawa Hospital Research Institute

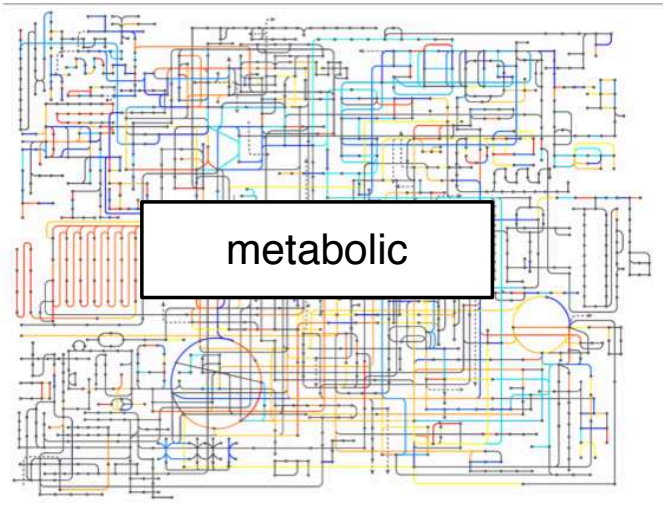


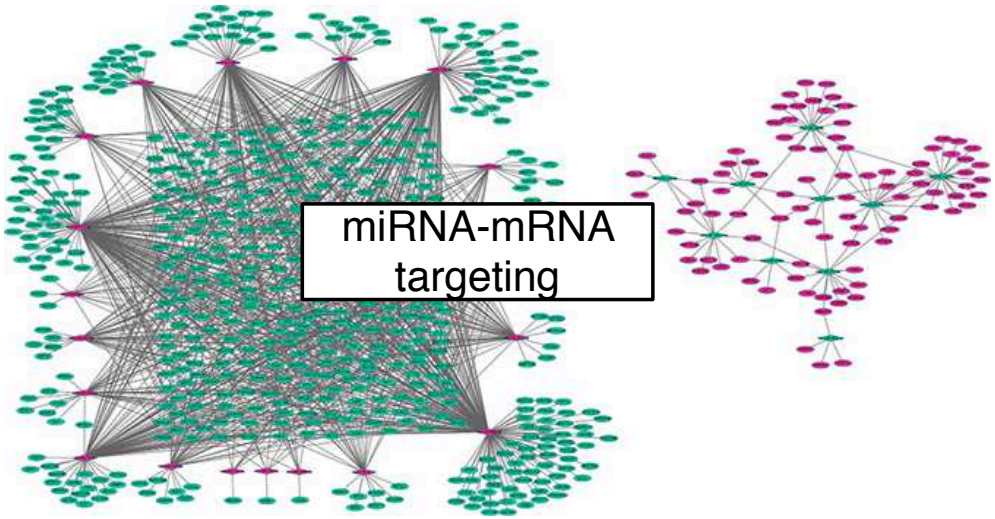


Gene expression is regulated by multiple, overlapping networks

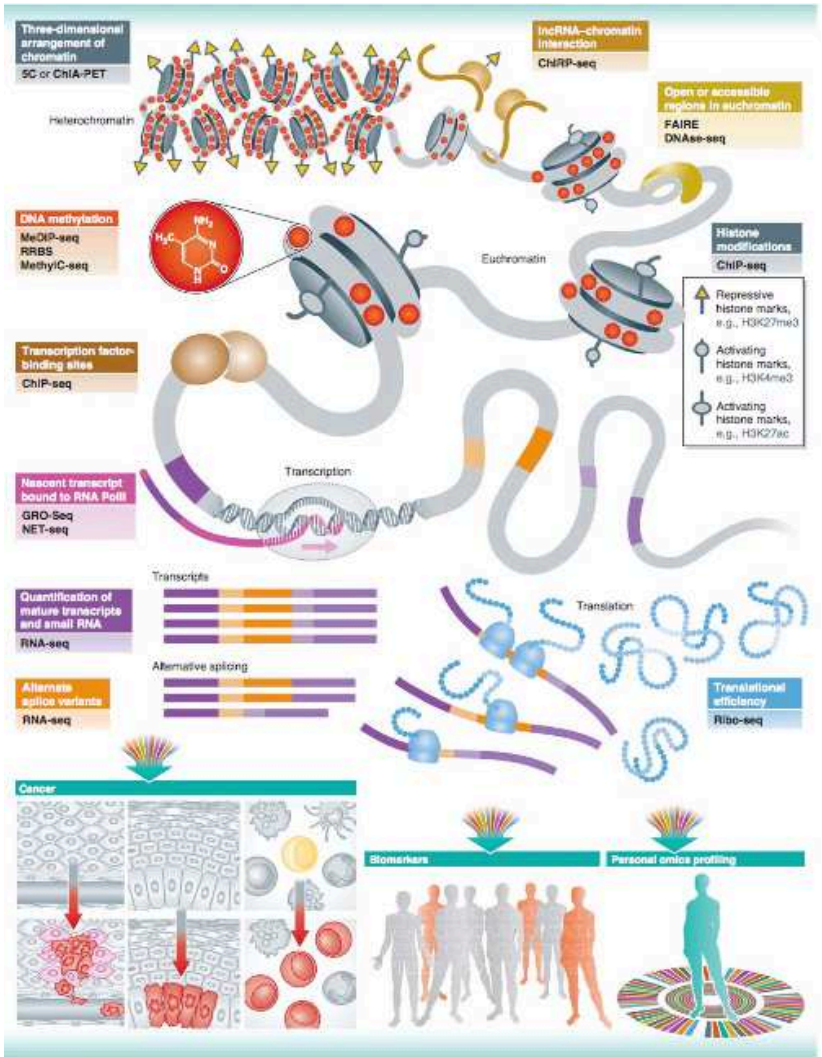




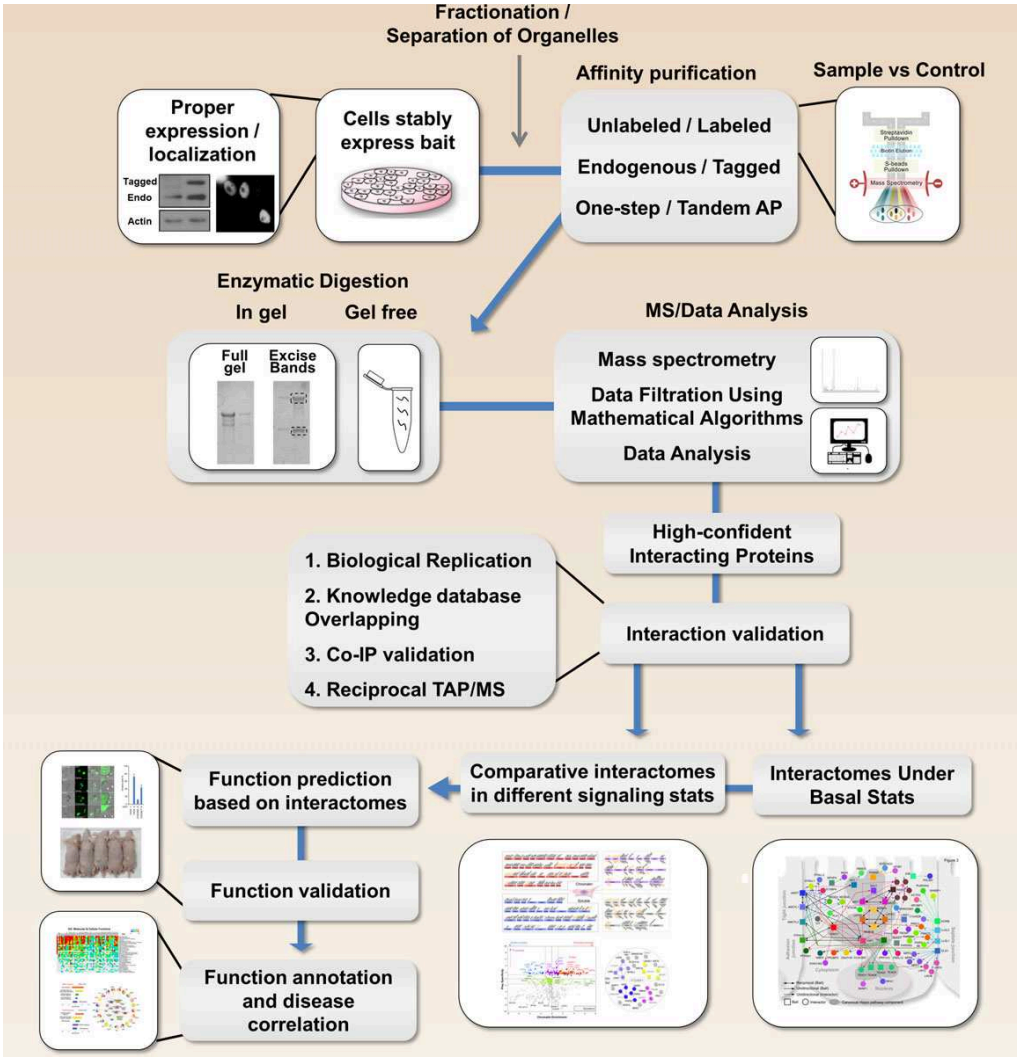




“Direct” experimental measurement of networks

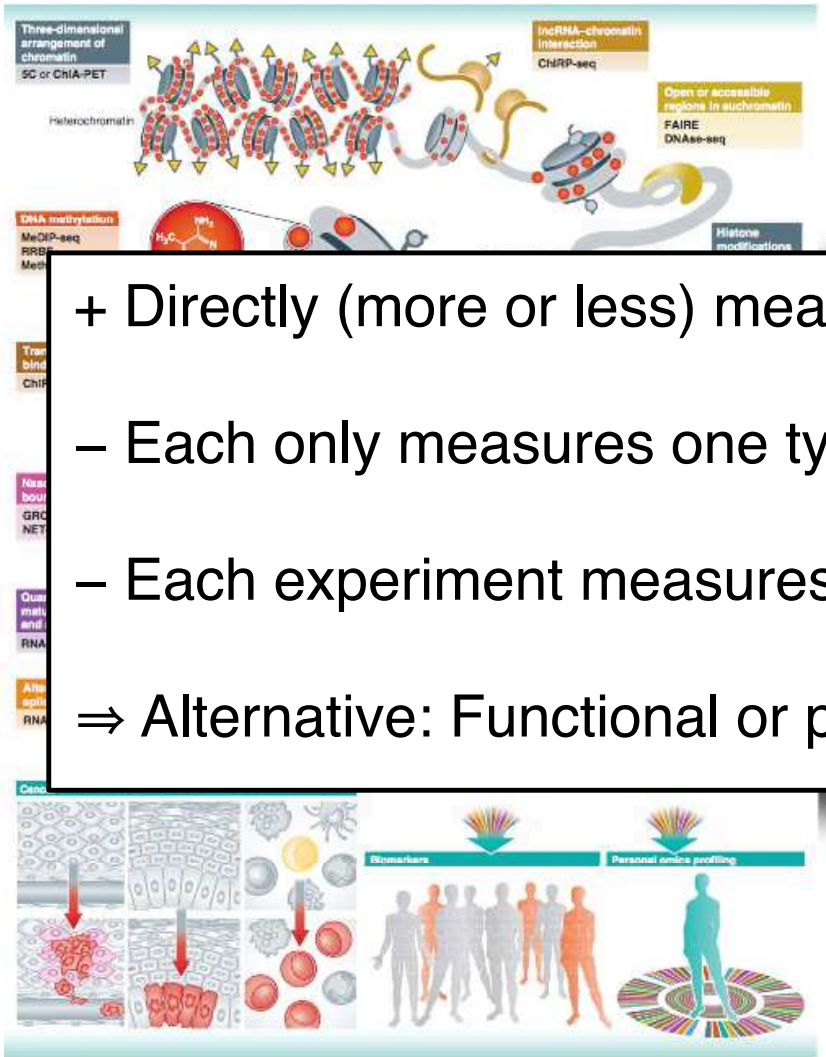


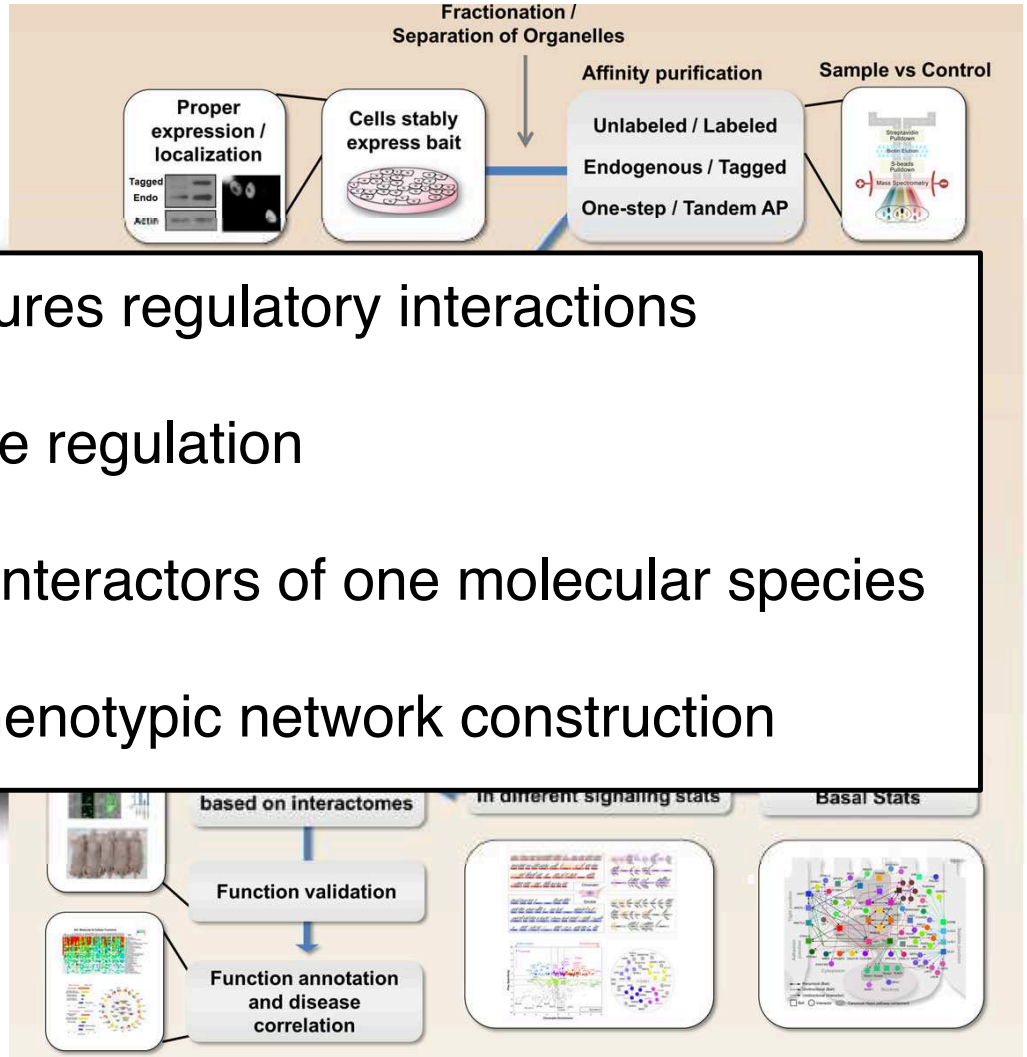
[Soon et al., Molec Sys Biol 2013]



[Li et al., Proteomics 2015]

“Direct” experimental measurement of networks





+ Directly (more or less) measures regulatory interactions

– Each only measures one type regulation

– Each experiment measures interactors of one molecular species

⇒ Alternative: Functional or phenotypic network construction

[Soon et al., Molec Sys Biol 2013]

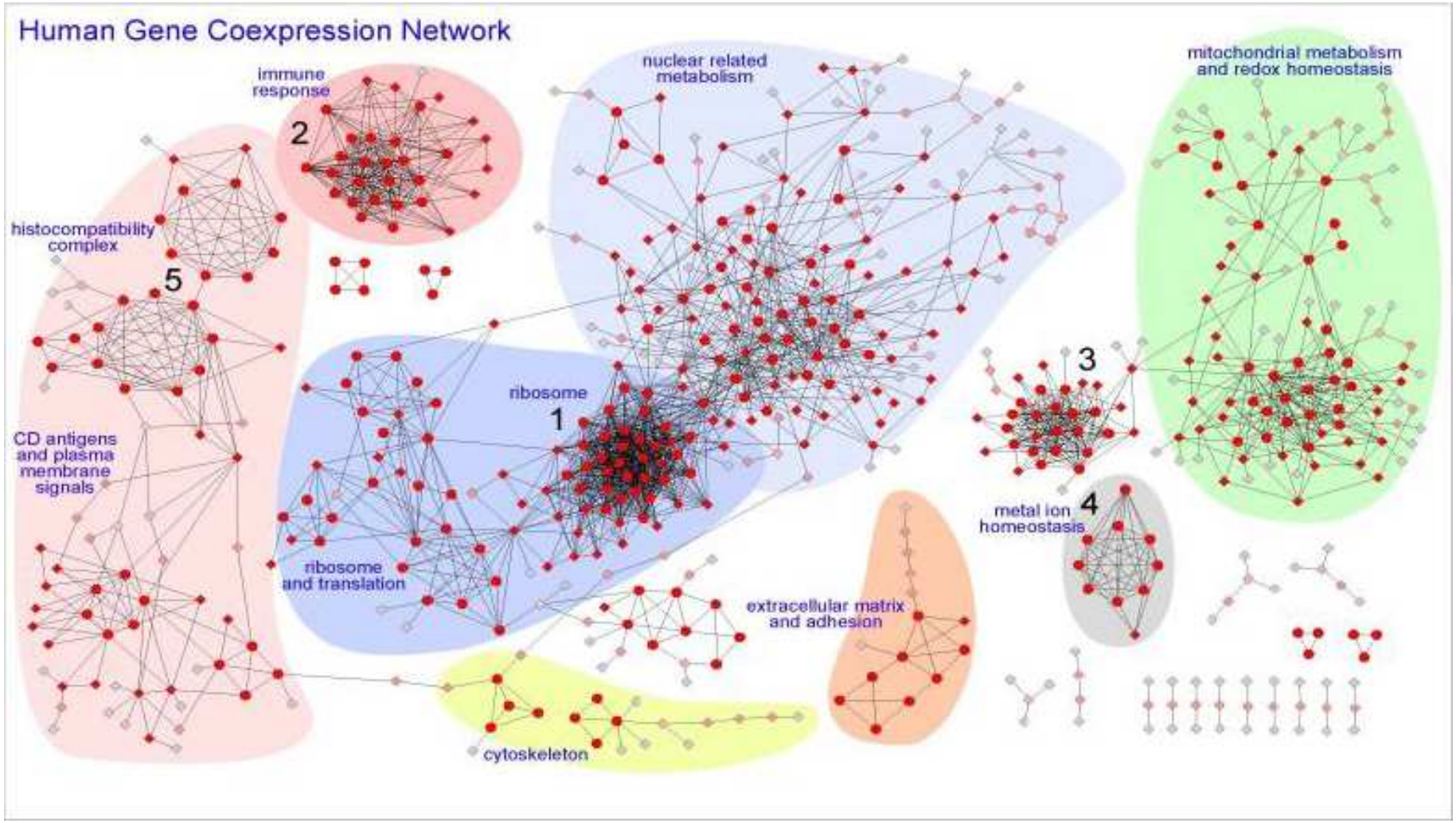
[Li et al., Proteomics 2015]

Outline

- Co-expression networks – linking genes by similarity of expression
  - Examples
  - Relevance Networks
  - ARACNE
  - WCNGA?
  - Bayesian Relevance Networks?
- Epistasis networks – linking genes by interpreting knockout phenotypes
  - Avery & Wasserman's classical theory
  - Data-robust epistasis analysis

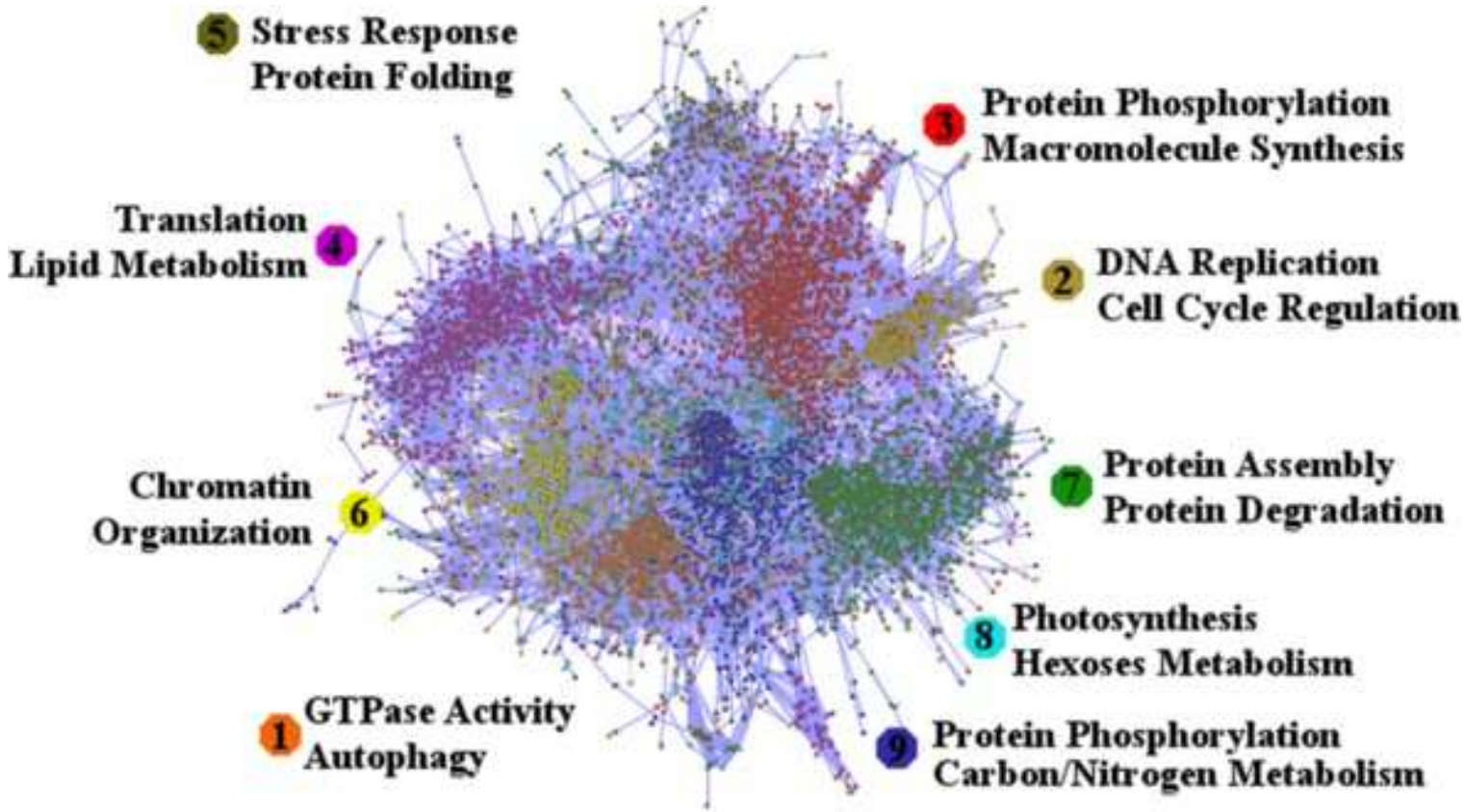
Example co-expression network (Prieto et al., PLoS ONE 2008)

Used microarray expression on 24 human tissues to determine co-expression, finding 15841 high-confidence relationships between 3327 genes.



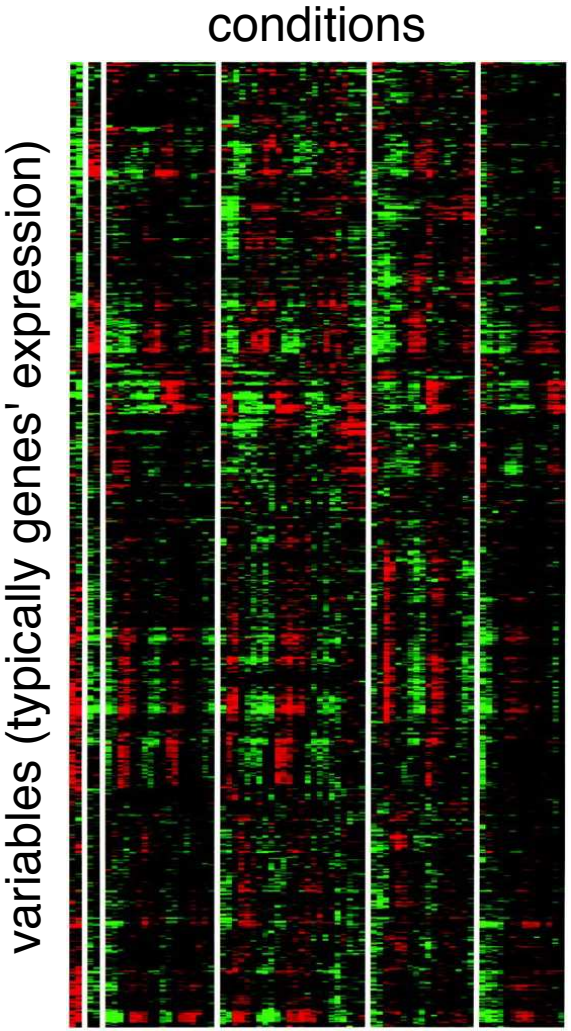
Example (Romero-Campero et al., BMC Genomics 2016)

Based on 50 RNA-seq datasets on 8 genotypes of *Chlamydomonas* under different physiological conditions.



Co-expression networks

Typically, we start with a data matrix measuring the expression of genes under different conditions.



- Main idea: Make a big graph in which "similarly" expressed genes are connected.

- Could represent one TF regulating another, or co-regulated genes in a complex / pathway, or any number of other things . . . .

- The resulting graph can then be inspected / analyzed to extract biological meaning.

- What does "similar" mean?

- When are two variables similar enough?

Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks

Atul J. Butte[†‡], Pablo Tamayo[§], Donna Slonim[§], Todd R. Golub[§¶], and Isaac S. Kohane[†]

[†]Children's Hospital Informatics Program and Division of Endocrinology, Department of Medicine, Children's Hospital, 300 Longwood Avenue, Boston, MA 02115; [§]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142; and [¶]Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115

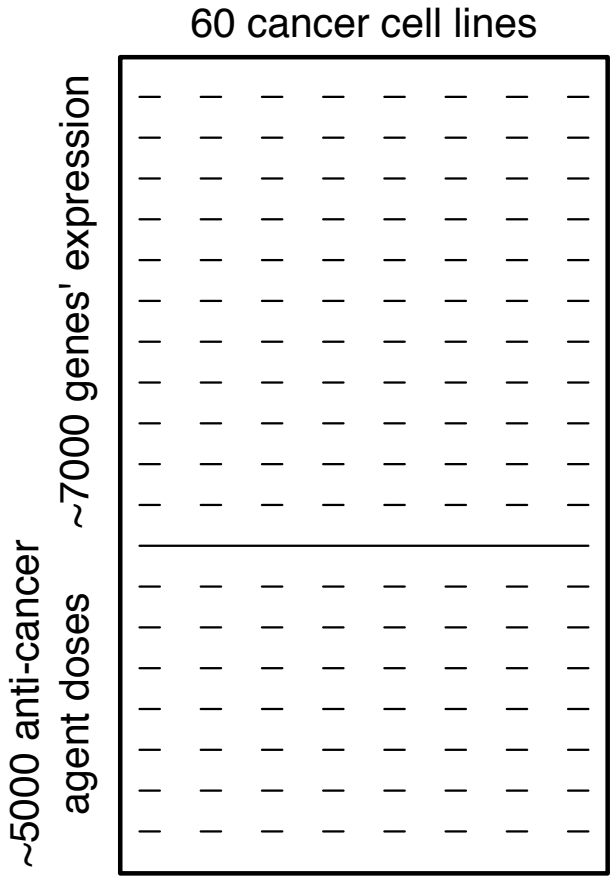
Communicated by Louis M. Kunkel, Harvard Medical School, Boston, MA, August 16, 2000 (received for review May 1, 2000)

In an effort to find gene regulatory networks and clusters of genes that affect cancer susceptibility to anticancer agents, we joined a database with baseline expression levels of 7,245 genes measured by using microarrays in 60 cancer cell lines, to a database with the amounts of 5,084 anticancer agents needed to inhibit growth of those same cell lines. Comprehensive pair-wise correlations were calculated between gene expression and measures of agent susceptibility. Associations weaker than a threshold strength were removed, leaving networks of highly correlated genes and agents called relevance networks. Hypotheses for potential single-gene determinants of anticancer agent susceptibility were constructed. The effect of random chance in the large number of calculations performed was empirically determined by repeated random permutation testing; only associations stronger than those seen in multiply permuted data were used in clustering. We discuss the advantages of this methodology over alternative approaches, such as phylogenetic-type tree clustering and self-organizing maps.

potheses of putative functional relationships between pairs of genes. Specifically, we used baseline RNA expression levels measured from the NCI60, a set of 60 human cancer cell lines used by the National Cancer Institute Developmental Therapeutics Program to screen anticancer agents since 1989 (8). We joined the gene expression levels to a database with measures of cancer susceptibility to anticancer agents, to see how the baseline RNA expression levels in the cell lines correlated with the inhibition of growth of these same cell lines to thousands of anticancer agents. To be clear, RNA expression levels were measured without any exposure to anticancer agents. As shown below, this methodology, termed relevance networks, is able to form clusters without having the problems listed above that are inherent in other methodologies. A feature of a clustering technique such as relevance networks, is that it allows us to find correlations across disparate biological measures, such as RNA expression and susceptibility to pharmaceuticals.

Their data

Butte *et al.* took 60 cancer cell lines as conditions, and used as variables microarray expression data for 6,701 genes and susceptibility of those lines to 4,991 anti-cancer agents.



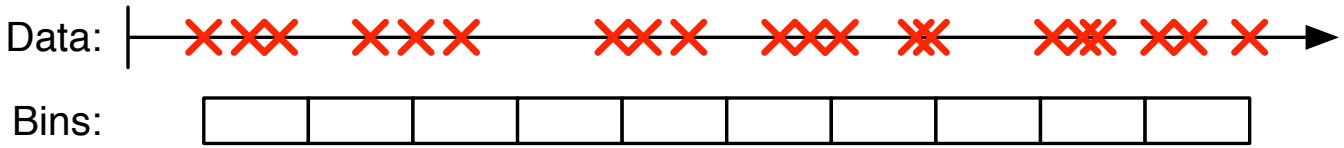
Algorithm outline

1. Remove variables with "low information content" – e.g., genes that are almost always on or always off, or have outliers observations
2. For every pair of variables x and y compute Pearson's (linear) correlation coefficient across conditions r_{xy}
3. Choose a threshold, τ, to determine statistically significant values of r_{xy}
4. Connect nodes x and y with an undirected edge, if $r_{xy}^2 > \tau$

1. Removing low-varability variables

For each variable:

- Discretize min-to-max range into 10 equal-sized bins



- Determine empirical fraction of data in each bin

Fracs:	0.15	0.05	0.10	0.05	0.10	0.10	0.15	0.00	0.15	0.15

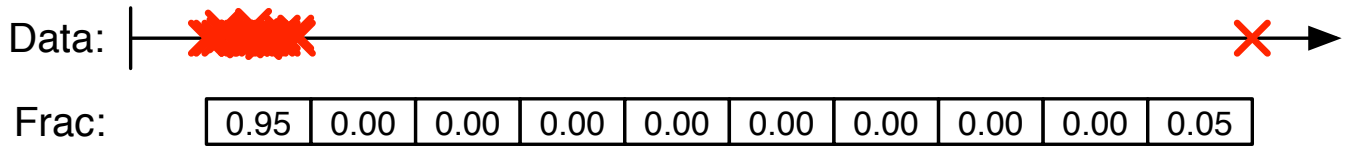
- Compute the (empirical) entropy

$$\begin{aligned} H &= -\sum_{i=1}^{10} f_i \log_2 f_i \\ &= 3.0710 \text{ bits} \end{aligned}$$

Remove from consideration 5% variables with lowest entropy

Targeting outliers?

This especially targets variables with one or a few extreme observations.



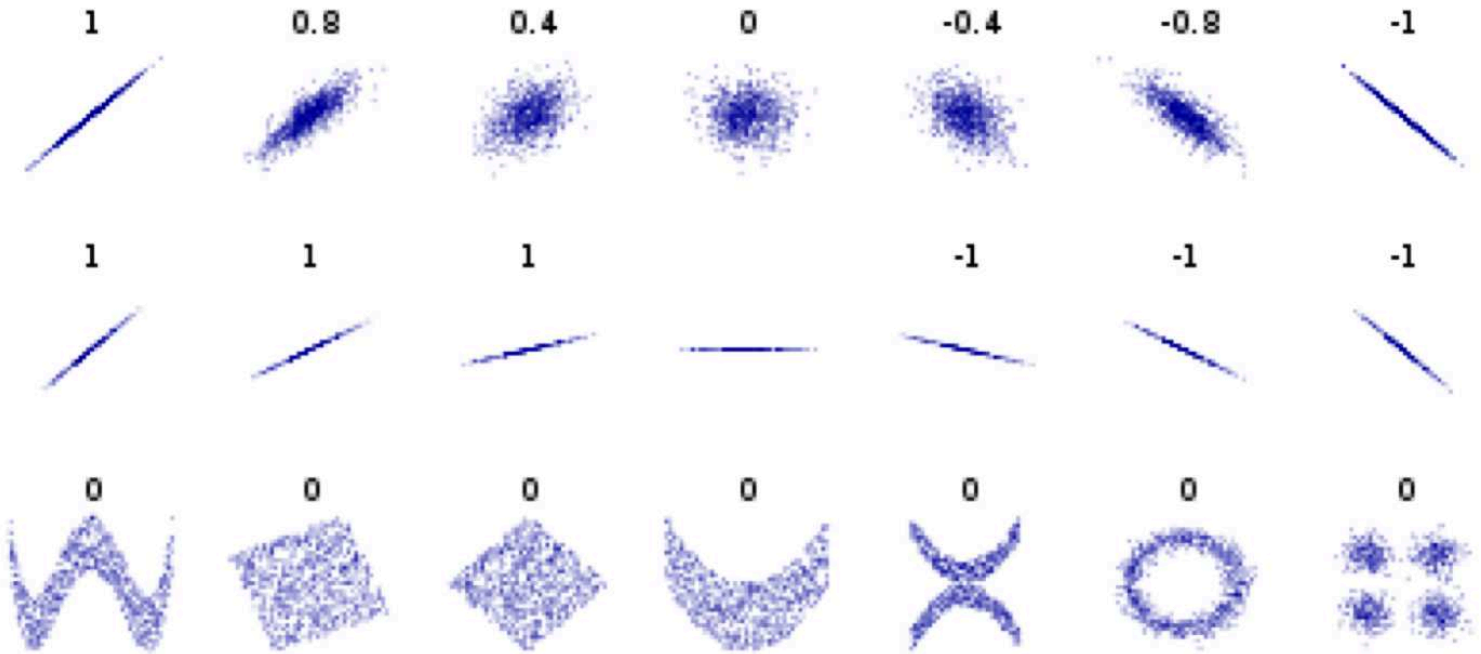
$$H = 0.2864$$

2. Compute pairwise Pearson's correlation coefficients

Linear correlation between $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$:

$$r_{xy} = \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}} = \frac{\sum_{i=1}^{n} \frac{1}{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} \frac{1}{n}(x_i - \bar{x})^2 \sum_{i=1}^{n} \frac{1}{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are sample means.



What constitutes a "significant" correlation?

Here's an idea from computational statitistics – permutation testing...

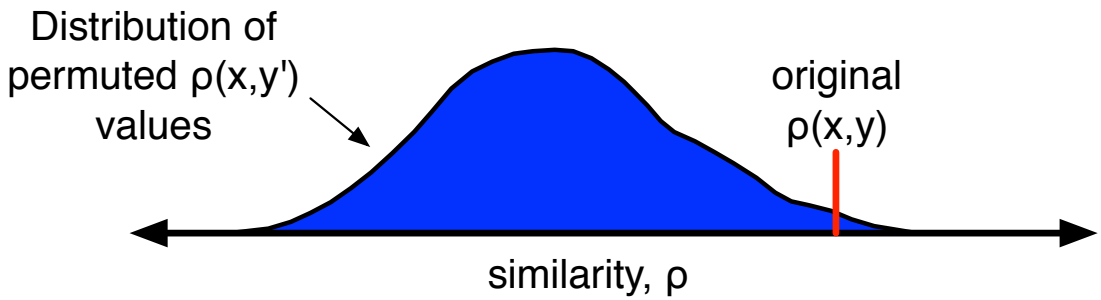
- Suppose you've got paired data on two variables, x and y:

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
y	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8

- Suppose you've got any measure of similarity ρ, which assigns a score to such paired data, $\rho(x, y)$.
- N times, randomly permute the y values and recompute ρ. E.g.:

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
y'	y_3	y_7	y_6	y_1	y_5	y_4	y_2	y_8

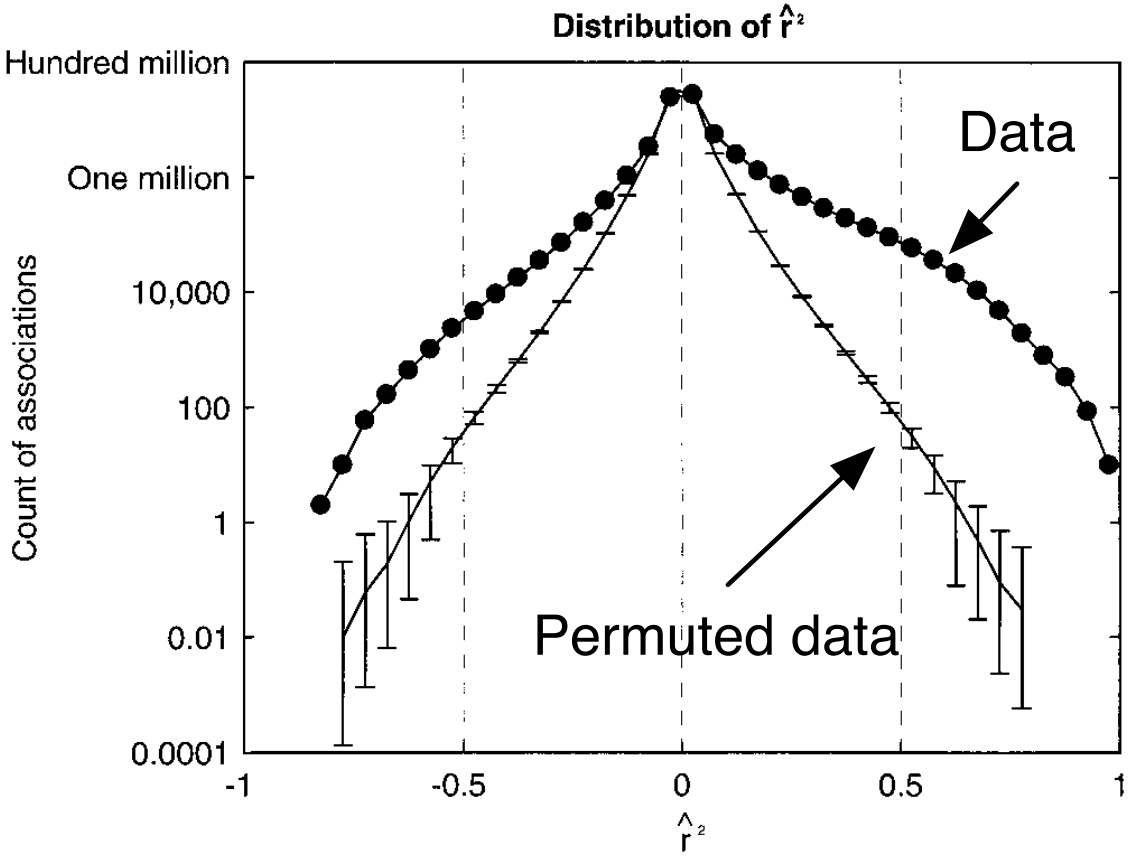
- The location of the original $\rho(x, y)$ with respect to the permuted ρ values gives a p-value.



- Approach is agnostic to the data distribution and similarity measure! Still need to choose a p-value threshold (or FDR)...

3. Permutation testing to choose a correlation cutoff

Permute all rows of data matrix 100 times, recomputing pairwise correlations, and building an empirical distribution.



Choose threshold based on acceptable balance of expected true and false positives.

4. Link variables with correlations exceeding threshold

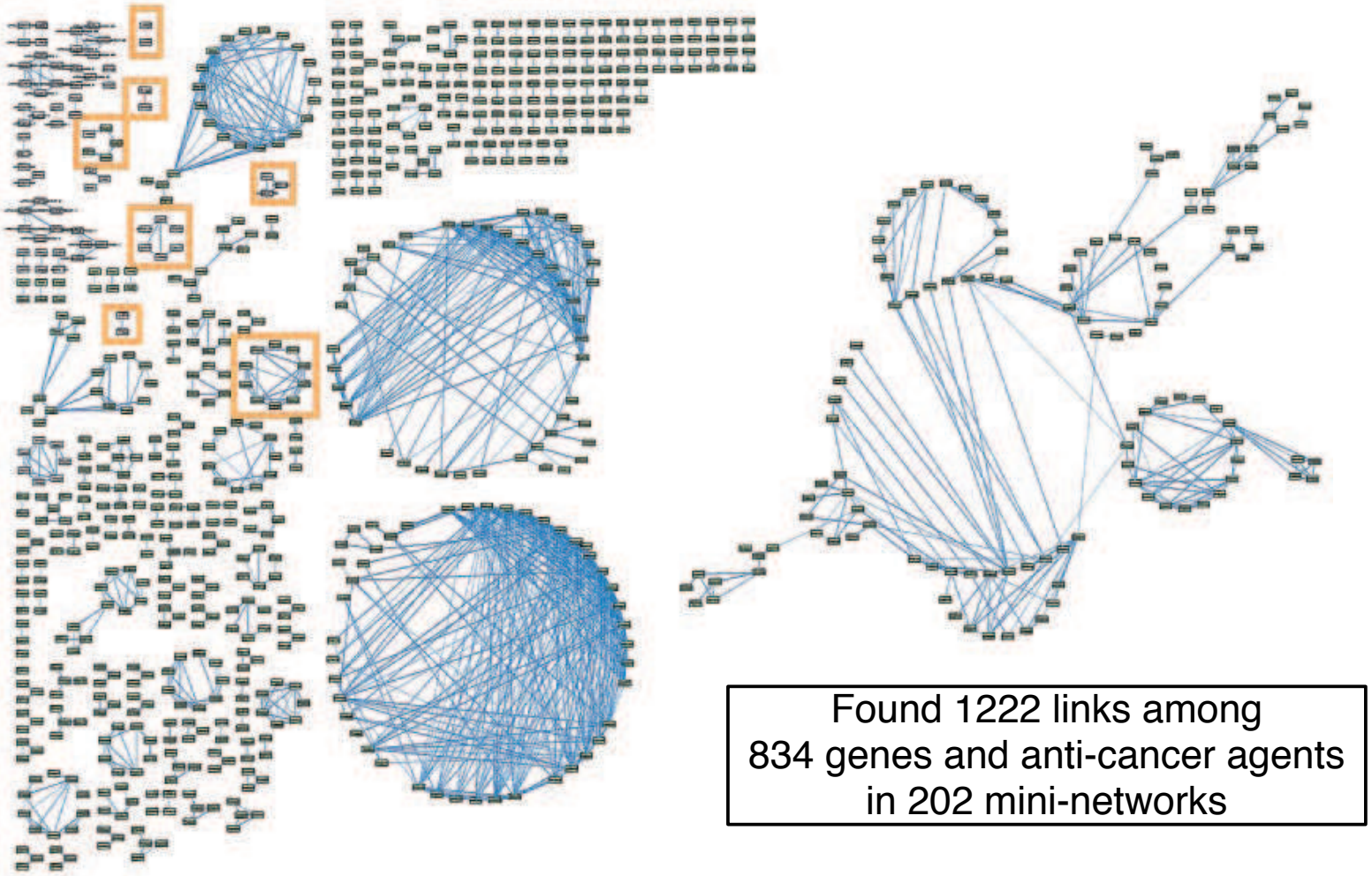


Fig. 2. Relevance networks constructed from the joined databases of baseline gene expression in 60 cancer cell lines and measures of susceptibility of the same cell lines to anticancer agents. The pairs of features (anticancer agents in green boxes, genes in white boxes) with r^2 at or greater than $\pm$ 0.80 were drawn with line thickness proportional to r^2. Features without an association at $\pm$ 0.80 were removed. Associations with negative r^2 are in red. Seven networks are highlighted in orange and are in Table 1. Large versions of all figures and descriptions for each accession number may be found at http://www.chip.org/genomics.

ARACNE (Margolin et al., BMC Bioinfo 2006)

BMC Bioinformatics

BioMed Central

Proceedings

Open Access

ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context

Adam A Margolin[1,2], Ilya Nemenman[2], Katia Basso[3], Chris Wiggins[2,4], Gustavo Stolovitzky[5], Riccardo Dalla Favera[3] and Andrea Califano*[1,2]

Address: [1]Department of Biomedical Informatics, Columbia University, New York, NY 10032, [2]Joint Centers for Systems Biology, Columbia University, New York, NY 10032, [3]Institute for Cancer Genetics, Columbia University, New York, NY 10032, [4]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10032 and [5]IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

Email: Adam A Margolin - adam@dbmi.columbia.edu; Ilya Nemenman - ilya.nemenman@columbia.edu; Katia Basso - kb451@columbia.edu; Chris Wiggins - chw2@columbia.edu; Gustavo Stolovitzky - gustavo@us.ibm.com; Riccardo Dalla Favera - rd10@columbia.edu; Andrea Califano* - califano@c2b2.columbia.edu

* Corresponding author

from NIPS workshop on New Problems and Methods in Computational Biology
Whistler, Canada. 18 December 2004

Published: 20 March 2006

BMC Bioinformatics 2006, **7**(Suppl 1):S7 doi:10.1186/1471-2105-7-S1-S7

The ARACNE algorithm

Attempts to answer two "drawbacks" of Relevance Networks:

- Pearson correlation only captures linear relationships
- "Correlations" between variables may be the result of indirect effects

The agorithm:

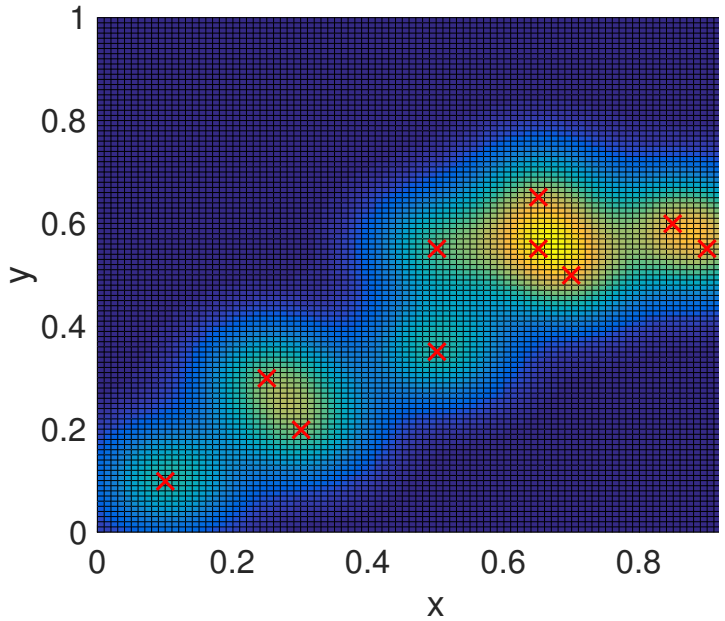
1. Estimate mutual information $MI(x, y)$ between all variables x and y
2. Choose a signficance threshold τ for MI
3. Link variables with mutual information $\geq \tau$
4. Remove $x - y$ link if, for some z, $MI(x, y) < \min(MI(x, z), MI(z, y))$

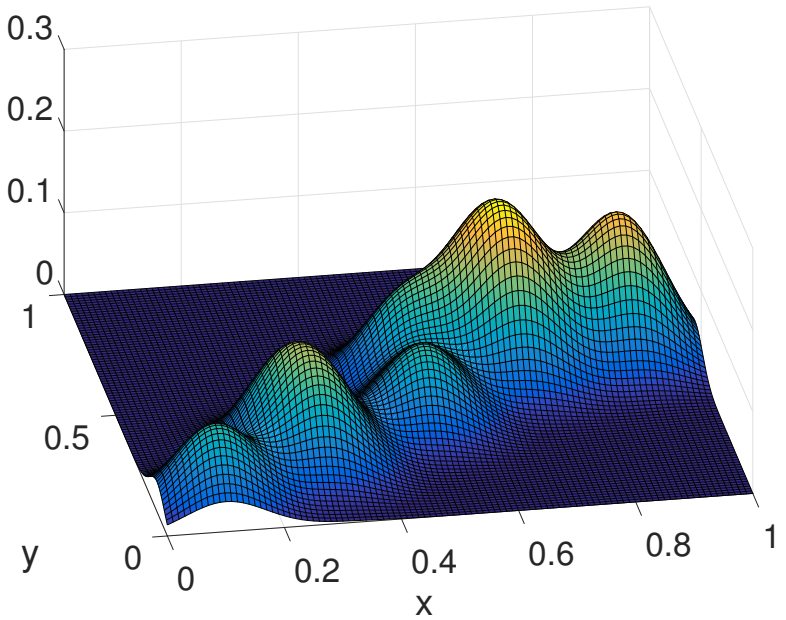
1. Estimate mutual information between variables

- Model joint distribution of x and y with Gaussian mixture model

$$L(x, y) = \sum_{i=1}^{N} (2\pi v)^{-N/2} \exp\left(\frac{(x - x_i)^2 + (y - y_i)^2}{2v^2}\right)$$

where i ranges over the N conditions, v is a "bandwidth" parameter.





1. Estimate mutual information between variables

- Model joint distribution of x and y with Gaussian mixture model

$$L(x,y) = \sum_{i=1}^{N} (2\pi v)^{-N/2} \exp\left(\frac{(x-x_i)^2 + (y-y_i)^2}{2v^2}\right)$$

  where i ranges over the N conditions, v is a "bandwidth" parameter.

- Mutual information estimate is then:

$$MI(x,y) = \frac{1}{N} \sum_{i=1}^{N} \log_2 \frac{L(x,y)}{L(x)L(y)}$$

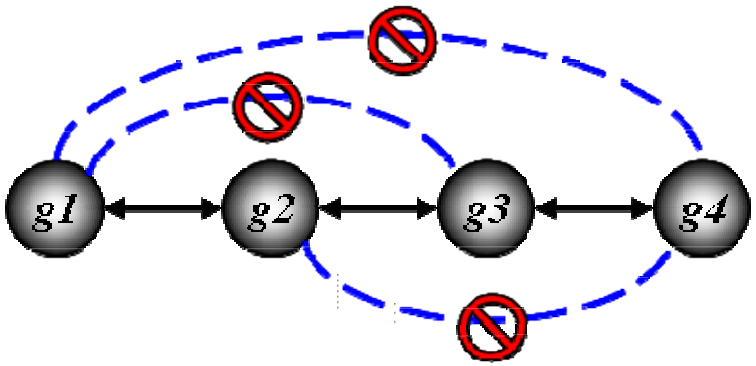
2. Choose a significance threshold τ

- Choose 10^5 random pairs of variables x and y
- For each, permute the y-values, and recompute $MI(x, y)$
- The fraction of these exceeding any threshold τ is an estimate of the p-value for $MI(x, y) = \tau$

  (If $MI(x, y)$ exceeds any of the random pairs, a p-value is assigned by extrapolation.)

3. Remove weak links

The "data processing inequality" says that if we have links $x - y$, $y - z$ and $x - z$, the link with the smallest MI should be removed.

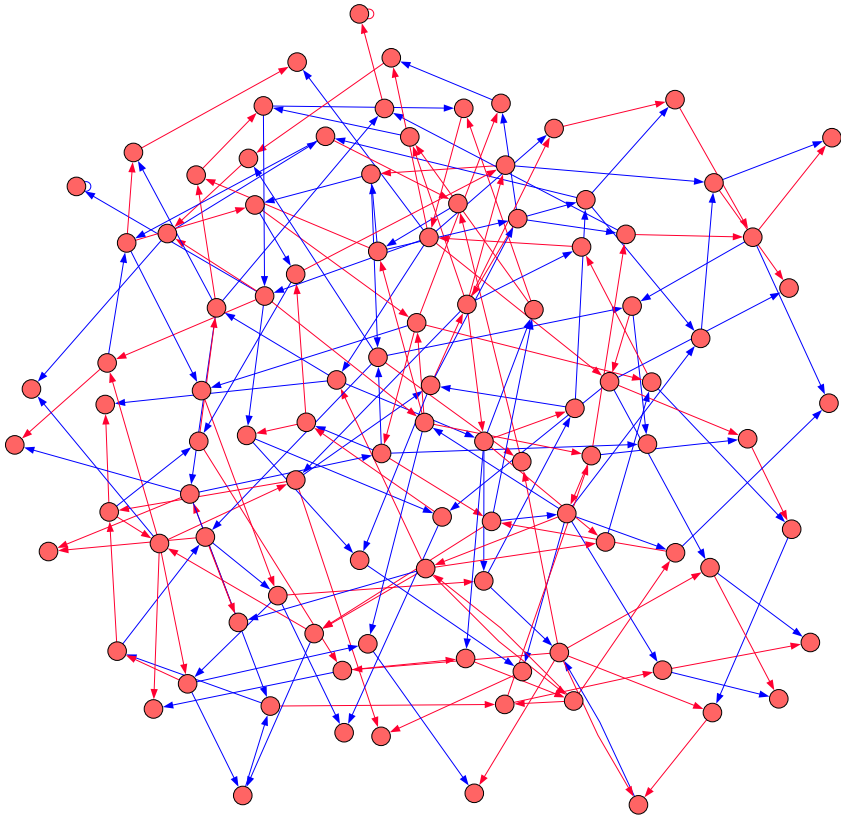


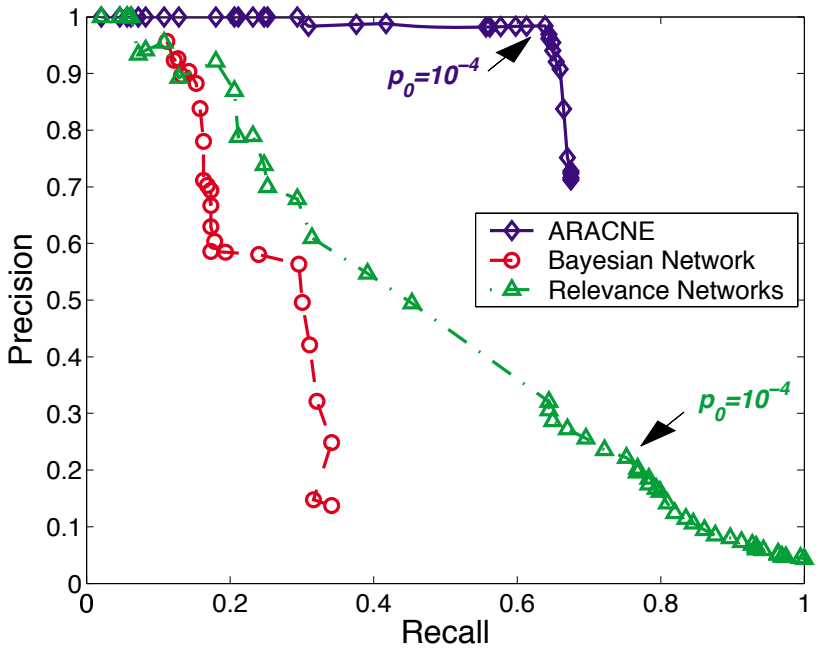
The hope is to remove indirect, correlations in favor of direct, "causal" links.

The resulting network is tree- (really, forest-) structured (unless there are ties for weakest link).

Results on simulated expression data from DREAM

Steady state data with randomly generated network structures, with randomly varied production and decay rates.





BMC Bioinformatics

BioMed Central

Software

Open Access

WGCNA: an R package for weighted correlation network analysis

Peter Langfelder[1] and Steve Horvath*[2]

Address: [1]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA and [2]Department of Human Genetics and Department of Biostatistics, University of California, Los Angeles, CA 90095, USA

Email: Peter Langfelder - Peter.Langfelder@gmail.com; Steve Horvath* - shorvath@mednet.ucla.edu

* Corresponding author

Published: 29 December 2008

BMC Bioinformatics 2008, **9**:559 doi:10.1186/1471-2105-9-559

Received: 24 July 2008
Accepted: 29 December 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/559

Some features

- Uses Pearson (linear) correlation, Spearman rank correlation or biweight midcorrelation to connect genes
- Functions for "module" detection and graph topology analysis
- Functions for correlating genes or modules to a measured "trait"
- Functions for visualization

$\Rightarrow$ It's the most cited of all co-expression papers! (2179 as of Oct 15 2017)

PLOS ONE

RESEARCH ARTICLE

Uncovering robust patterns of microRNA co-expression across cancers using Bayesian Relevance Networks

Parameswaran Ramachandran[1☯¤], Daniel Sánchez-Taltavull[1☯], Theodore J. Perkins[1,2*]

1 Regenerative Medicine Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada K1H8L6, **2** Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, Ontario, Canada K1H8M5

☯ These authors contributed equally to this work.
¤ Current address: The Campbell Family Institute for Breast Cancer Research, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada M5G2M9
* tperkins@ohri.ca

OPEN ACCESS

Citation: Ramachandran P, Sánchez-Taltavull D, Perkins TJ (2017) Uncovering robust patterns of microRNA co-expression across cancers using Bayesian Relevance Networks. PLoS ONE 12(8): e0183103. https://doi.org/10.1371/journal.pone.0183103

Abstract

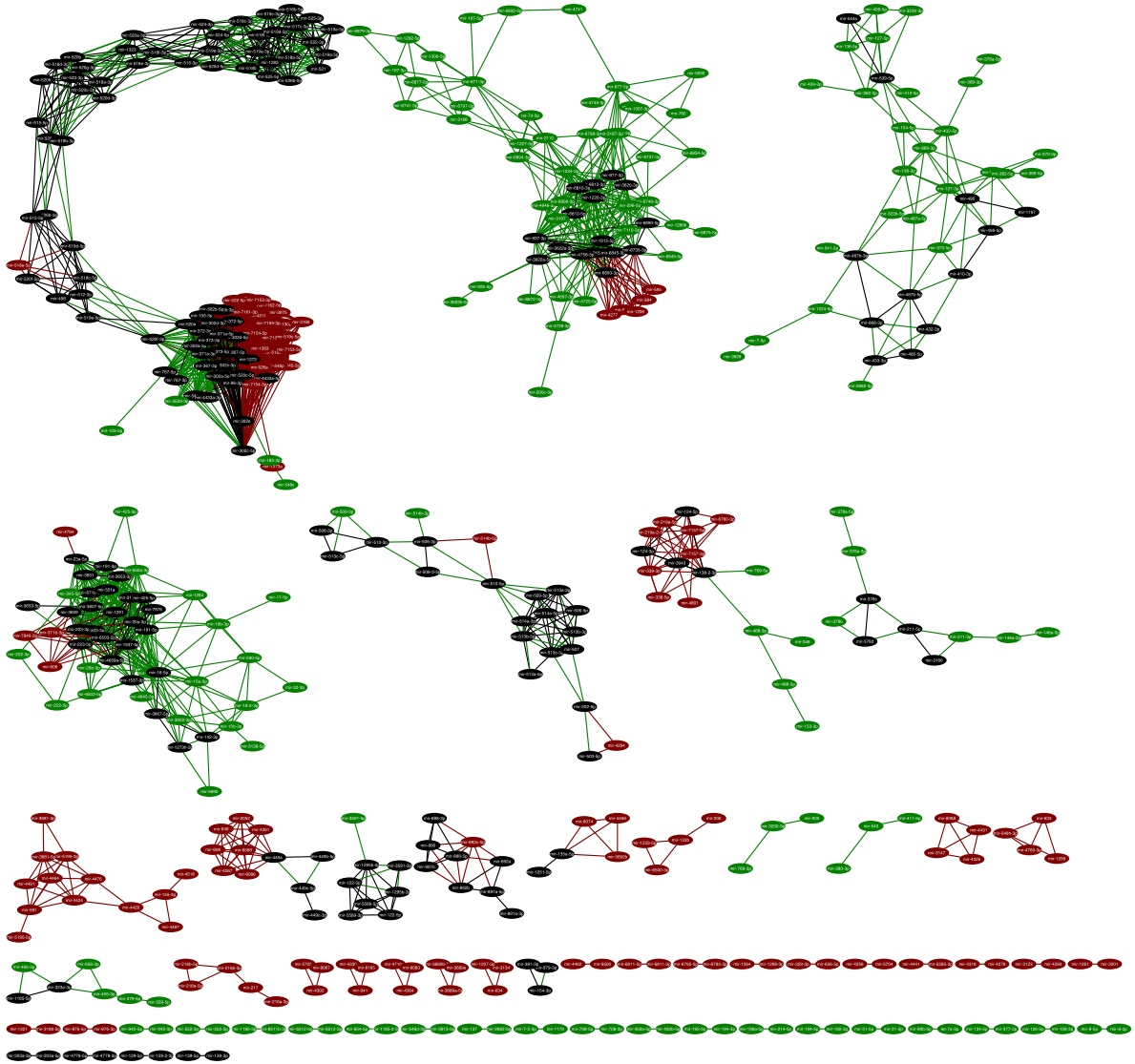
Co-expression networks have long been used as a tool for investigating the molecular circuitry governing biological systems. However, most algorithms for constructing co-expression networks were developed in the microarray era, before high-throughput sequencing—with its unique statistical properties—became the norm for expression measurement. Here we develop Bayesian Relevance Networks, an algorithm that uses Bayesian reasoning about expression levels to account for the differing levels of uncertainty in expression measurements between highly- and lowly-expressed entities, and between samples with

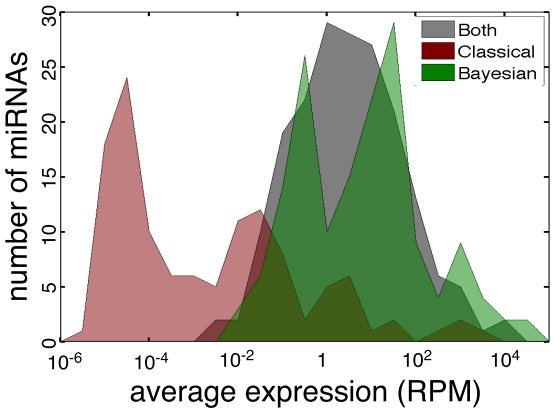
Accounting for varying measurement uncertainty

- With RNA-seq, miRNA-seq, different samples' read depths equate to varying precision of measurement
- With genes', miRNAs' different expression levels, relative precision varies
- Propose Bayesian beliefs over expression levels of gene/miRNA in each sample (Dirichlet, weak non-uniform priors)
- Quantify co-expression by correlation across conditions and beliefs

$$r_{xy}^{B} = \frac{\mathsf{Cov}_{c,u}(x_c, y_c)}{\sqrt{\mathsf{Var}_{c,u}(x_c)\mathsf{Var}_{c,u}(y_c)}}$$

Results on 10,999 miRNA-seq samples from TCGA





- Relevance and Bayesian Relevance Networks differ
- Nodes included in the Bayesian network have higher average expression / greater confidence and replicate better in cross-validation

Co-expression networks summary

- In co-expression networks, the genes whose expression is most similar over a set of conditions are linked
- Similarity can be assessed in several different ways
- Points in favor:
  - \+ Correlation networks can be computed efficiently
  - \+ Readily visualized
  - \+ Subnetwork inspection leads to new hypotheses
  - \+ Can find true / known relationships, as well as many new ones
- Points against:
  - – Links are directionless, and of unclear meaning
    (Though some directional proposals have been made.)
  - – Links are established pairwise only
  - – The networks are not predictive. What if gene x were deleted?

Questions?

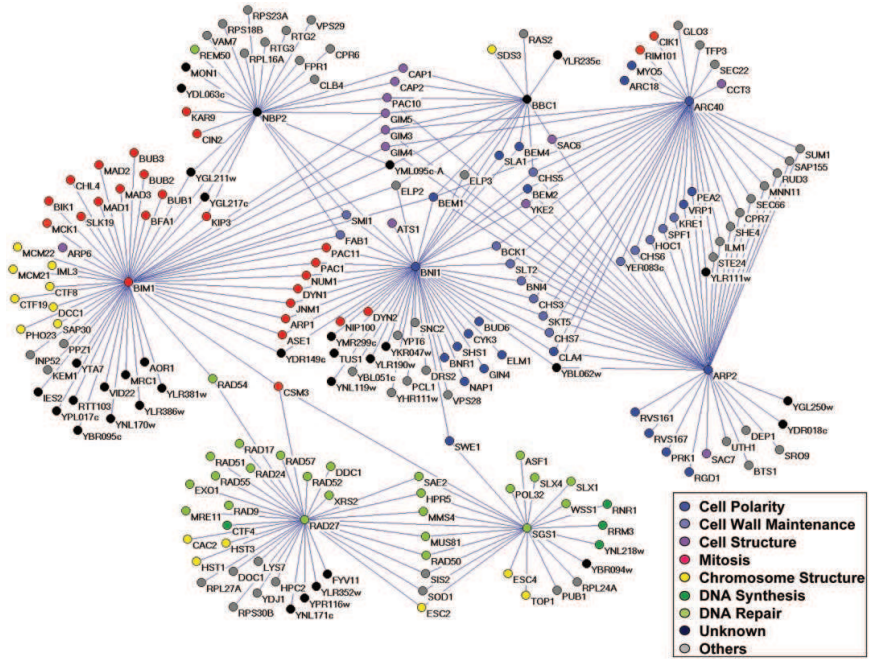
Outline

- Co-expression networks – linking genes by similarity of expression
  - Examples
  - Relevance Networks
  - ARACNE
  - WCNGA?
  - Bayesian Relevance Networks?
- **Epistasis networks – linking genes by interpreting knockout phenotypes**
  - Avery & Wasserman's classical theory
  - Data-robust epistasis analysis

What is epistasis?

Broadly: Epistasis is when deleting/mutating two genes/loci leads to a "surprising" outcome, when compared each deletion/mutation individually.

E.g., **Synthetic lethality**, where deleting two genes kills an organism, even though each individual deletion is harmless (Tong et al., Science 2001)



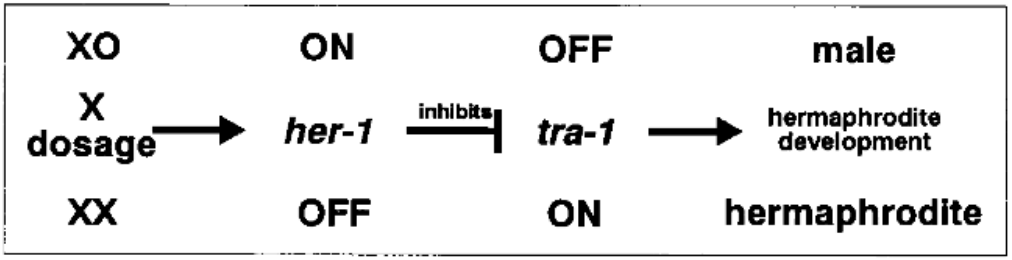
A subtler form is epistasis as **masking**, where the deletion of one gene masks the effect of deleting another.

Ordering gene function: the interpretation of epistasis in regulatory hierarchies

LEON AVERY AND STEVEN WASSERMAN

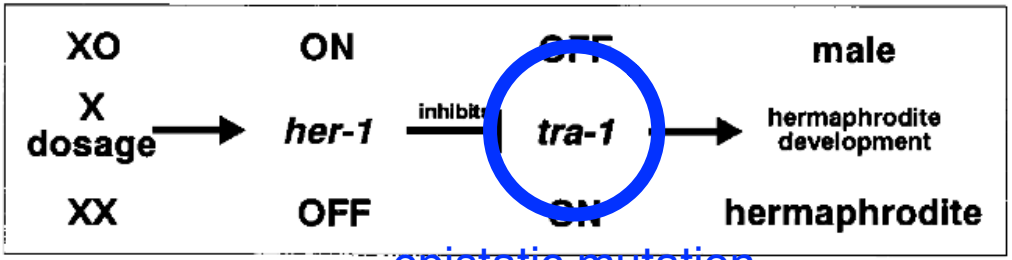
The order of action of genes in a regulatory hierarchy that is governed by a signal can often be determined by the method of epistasis analysis, in which the phenotype of a double mutant is compared with that of single mutants. The epistatic mutation may be in either the upstream or the downstream gene, depending on the nature of the two mutations and the type of regulation. Nevertheless, when the regulatory hierarchy satisfies certain conditions, simple rules allow the position of the epistatic locus in the pathway to be determined without detailed knowledge of the nature of the mutations, the pathway, or the molecular mechanism of regulation.

Example 1 from A & W – Sex determination in C. elegans



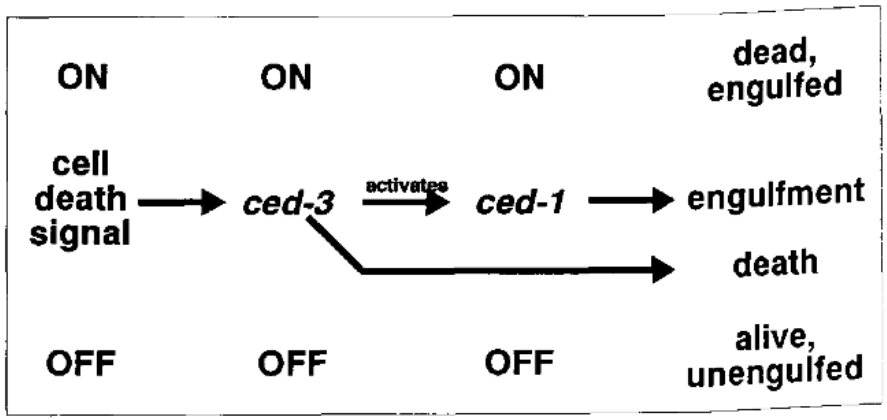
genotype	X dose	her-1	tra-1	phenotype
WT	XO	ON	OFF	♂
	XX	OFF	ON	♀
Δher-1	XO	OFF	ON	♀
	XX	OFF	ON	♀
Δtra-1	XO	ON	OFF	♂
	XX	OFF	OFF	♂
Δher-1 Δtra-1	XO	OFF	OFF	♂
	XX	OFF	OFF	♂

Example 1 from A & W – Sex determination in C. elegans



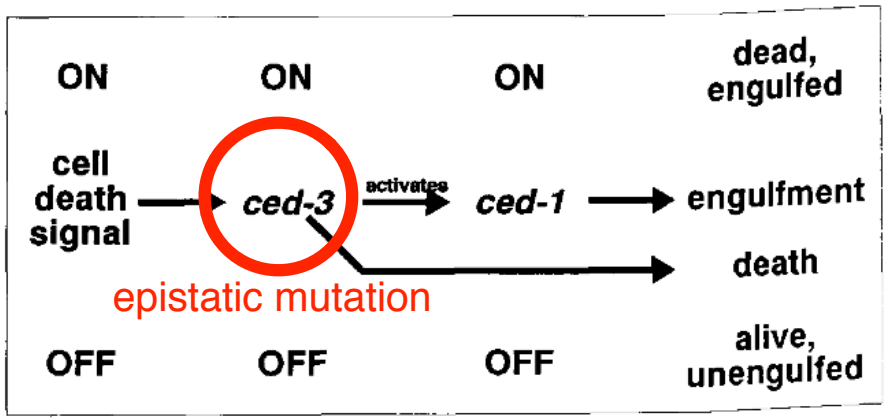
genotype	X dose	her-1	tra-1	phenotype
WT	XO	ON	OFF	♂
	XX	OFF	ON	♀
Δher-1	XO	OFF	ON	♀
	XX	OFF	ON	♀
Δtra-1	XO	ON	OFF	♂
	XX	OFF	OFF	♂
Δher-1 Δtra-1	XO	OFF	OFF	♂
	XX	OFF	OFF	♂

Example 2 from A & W – Apoptosis in C. elegans



genotype	c.d.s	ced-3	ced-1	phenotype
WT	ON	ON	ON	
	OFF	OFF	OFF	
Δced-3	ON	OFF	OFF	
	OFF	OFF	OFF	
Δced-1	ON	ON	OFF	
	OFF	OFF	OFF	
Δced-3 Δced-1	ON	OFF	OFF	
	OFF	OFF	OFF	

Example 2 from A & W – Apoptosis in C. elegans



genotype	c.d.s	ced-3	ced-1	phenotype
WT	ON	ON	ON	
	OFF	OFF	OFF	
Δced-3	ON	OFF	OFF	
	OFF	OFF	OFF	
Δced-1	ON	ON	OFF	
	OFF	OFF	OFF	
Δced-3 Δced-1	ON	OFF	OFF	
	OFF	OFF	OFF	

The mystery of epistasis

Sometimes the epistatic gene is upstream, sometimes downstream.

One gene may activate the other, or may repress the other.

How can we figure this out, based on observing the mutant phenotypes – and *not* the expression of the intermediate genes?

Assumptions and inference rules

Box 1. Assumptions underlying the interpretation of epistasis

Observable signal

(1) There is a signal that affects phenotype. The experimenter can find out the state of the signal, independently of genotype or phenotype.

No cross-talk

(2) The signal and the two genes under study are the sole determinants of phenotype under the conditions of the experiment.

Signal & Genes Boolean

(3) The signal and the two genes are either on or off; there are no intermediate levels of activity. (For instance, partial loss-of-function mutations should be avoided.)

Sequential pathway

(4) In the wild type the signal determines whether one of the genes (the upstream gene) is on or off; this in turn determines whether the second (downstream) gene is on or off.

An analysis of all possible combinations of null and constitutive mutations in both types of models, summarized in Table 1, allows three important generalizations:

(1) A given mutation only affects phenotype either when the signal is on, or when the signal is off, but not both.

For example, *her-1*$^-$ null mutations have a phenotype only in XO worms, and *tra-1*$^-$ null mutations only in XX worms.

(2) If two mutations have phenotypic effects in opposite signal states and one is epistatic to the other, it is the downstream mutation that is epistatic to the upstream mutation.

In sex determination, *tra-1*$^-$ lies downstream of *her-1*$^-$, and is epistatic to it.

(3) If two mutations have phenotypic effects in the same signal state and one is epistatic to the other, it is the upstream mutation that is epistatic to the downstream mutation.

In the cell death pathway, *ced-3*$^-$ lies upstream of *ced-1*$^-$, and is epistatic to it.

Rules in tabular form

TABLE 1. Determining gene order in regulatory hierarchies by epistasis analysis

Which signal states display mutant phenotypes?	Epistatic mutation	Type of mutation		Sign of regulation
		Upstream gene	Downstream gene	
Same	Upstream	Null	Null	+
		Constitutive	Constitutive	+
		Null	Constitutive	−
		Constitutive	Null	−
Opposite	Downstream	Null	Null	−
		Constitutive	Constitutive	−
		Null	Constitutive	+
		Constitutive	Null	+

The results of analysis of the eight possible cases in which regulation may be positive or negative and mutations either null or constitutive. In all eight cases, if there is simple epistasis, the epistatic mutation can be uniquely predicted to be in either the upstream or the downstream gene.

More examples

TABLE 3. Examples of epistasis analysis from yeast, worms, flies and plants

Organism	Pathway	Signal	Upstream mutation	Downstream mutation	Ref.
Same states affected (upstream gene epistatic):					
S. cerevisiae	Cell cycle progression	Time	*cdc28*$^-$	*cdc4*$^-$	3
C. elegans	Programmed cell death	Lineage	*ced-3*$^-$	*ced-1*$^-$	6
C. elegans	Vulval development	Position	*n300*	*lin-15*$^-$	5
D. melanogaster	Sex determination	X : autosome ratio	*tra*$^-$	*ix*$^-$	12
Opposite states affected (downstream gene epistatic):					
S. cerevisiae	Mating type interconversion	Mother/daughter cell	*swi5*$^-$	*sdi1*$^-$	13
S. cerevisiae	Invertase expression	Glucose	*snf1*$^-$	*cid1*$^-$	14
S. cerevisiae	Cell cycle progression	α mating pheromone	*cdc39*$^-$	*ste4* $^-$	9
S. cerevisiae	Biosynthetic enzyme expression	Amino acid starvation	*GCN2*^{c}	*gcn3*$^-$	15
C. elegans	Larval/adult development	Time	*lin-4*$^-$	*lin-14*$^-$	16
C. elegans	Vulval development	Gonadal signal	*let-23*$^-$	*let-60*^{c}	17
C. elegans	Dauer larva formation	Pheromone	*daf-1*$^-$	*daf-12*$^-$	18
D. melanogaster	Segment identity	Position	*esc*$^-$	BX-C$^-$	19
D. melanogaster	Dorsoventral development	Position	*fs(1)K10*$^-$	*gurken*$^-$	4
D. melanogaster	Development of termini	Position	*torso*^{c}	*tailless*$^-$	20
A. thaliana	Flower development	Position	*superman*$^-$	*pistillata*$^-$	21

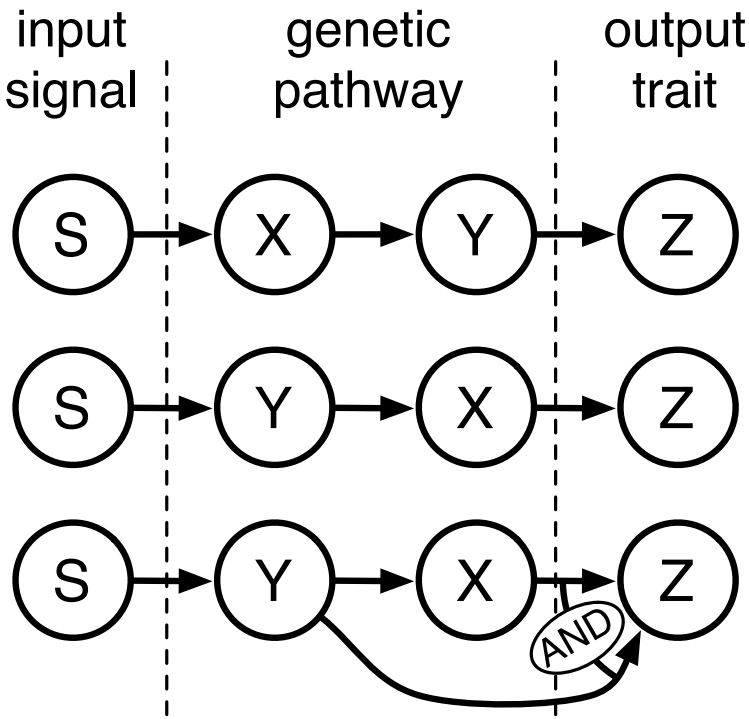
Superscript c denotes constitutive mutations.

Avery & Wasserman key conclusion

Simple logical rules can tell us when observations of epistasis reveal something about pathway structure!

- Includes a test for applicability (single deletions influence in just one signal state)
- *Can* reveal upstream/downstream & activation/repression between the two genes

Epistasis analysis cannot distinguish certain pathways



All arrows indicate activation, and we observed the trait under all possible wild type and knockout conditions, we always see the same thing!

(See Phenix et al. (Chaos, 2013) for thorough analysis of identifiability.)

Quantitative Epistasis Analysis and Pathway Inference from Genetic Interaction Data

Hilary Phenix[1,2], Katy Morin[1,3], Cory Batenchuk[1,2], Jacob Parker[4,5], Vida Abedi[1,2], Liu Yang[1,2,5], Lioudmila Tepliakova[1,2], Theodore J. Perkins[3,4], Mads Kærn[1,2,6]*

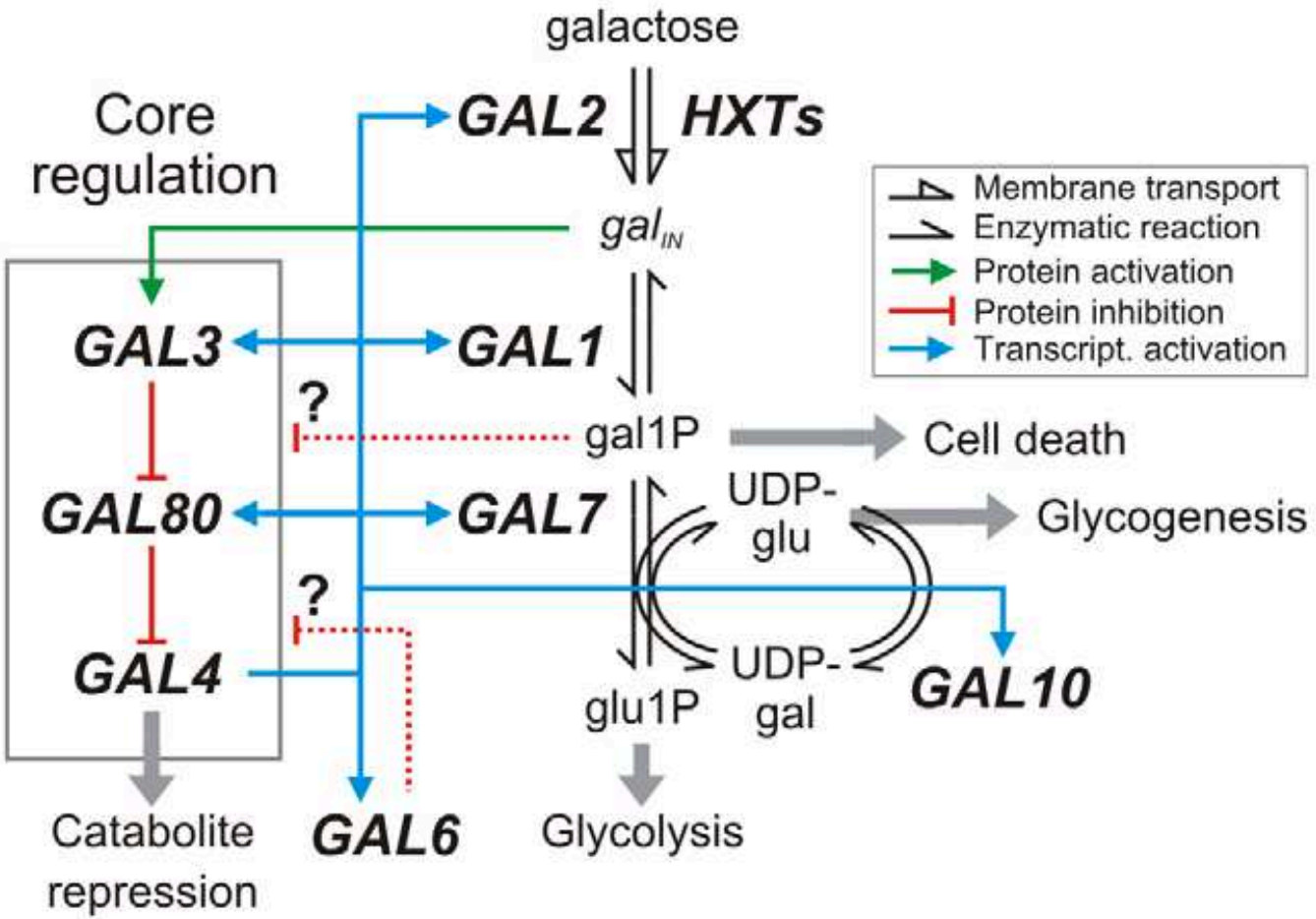
1 Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, Ontario, Canada, **2** Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada, **3** Department of Biochemistry, Immunology and Microbiology, University of Ottawa, Ottawa, Ontario, Canada, **4** Ottawa Hospital Research Institute, Ottawa, Ontario, Canada, **5** Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada, **6** Department of Physics, University of Ottawa, Ottawa, Ontario, Canada

Abstract

Inferring regulatory and metabolic network models from quantitative genetic interaction data remains a major challenge in systems biology. Here, we present a novel quantitative model for interpreting epistasis within pathways responding to an external signal. The model provides the basis of an experimental method to determine the architecture of such pathways, and establishes a new set of rules to infer the order of genes within them. The method also allows the extraction of quantitative parameters enabling a new level of information to be added to genetic network models. It is applicable to any system where the impact of combinatorial loss-of-function mutations can be quantified with sufficient accuracy. We test the method by conducting a systematic analysis of a thoroughly characterized eukaryotic gene network, the galactose utilization pathway in *Saccharomyces cerevisiae*. For this purpose, we quantify the effects of single and double gene deletions on two phenotypic traits, fitness and reporter gene expression. We show that applying our method to fitness traits reveals the order of metabolic enzymes and the effects of accumulating metabolic intermediates. Conversely, the analysis of expression traits reveals the order of transcriptional regulatory genes, secondary regulatory signals and their relative strength. Strikingly, when the analyses of the two traits are combined, the method correctly infers ~80% of the known relationships without any false positives.

Citation: Phenix H, Morin K, Batenchuk C, Parker J, Abedi V, et al. (2011) Quantitative Epistasis Analysis and Pathway Inference from Genetic Interaction Data. PLoS Comput Biol 7(5): e1002048. doi:10.1371/journal.pcbi.1002048

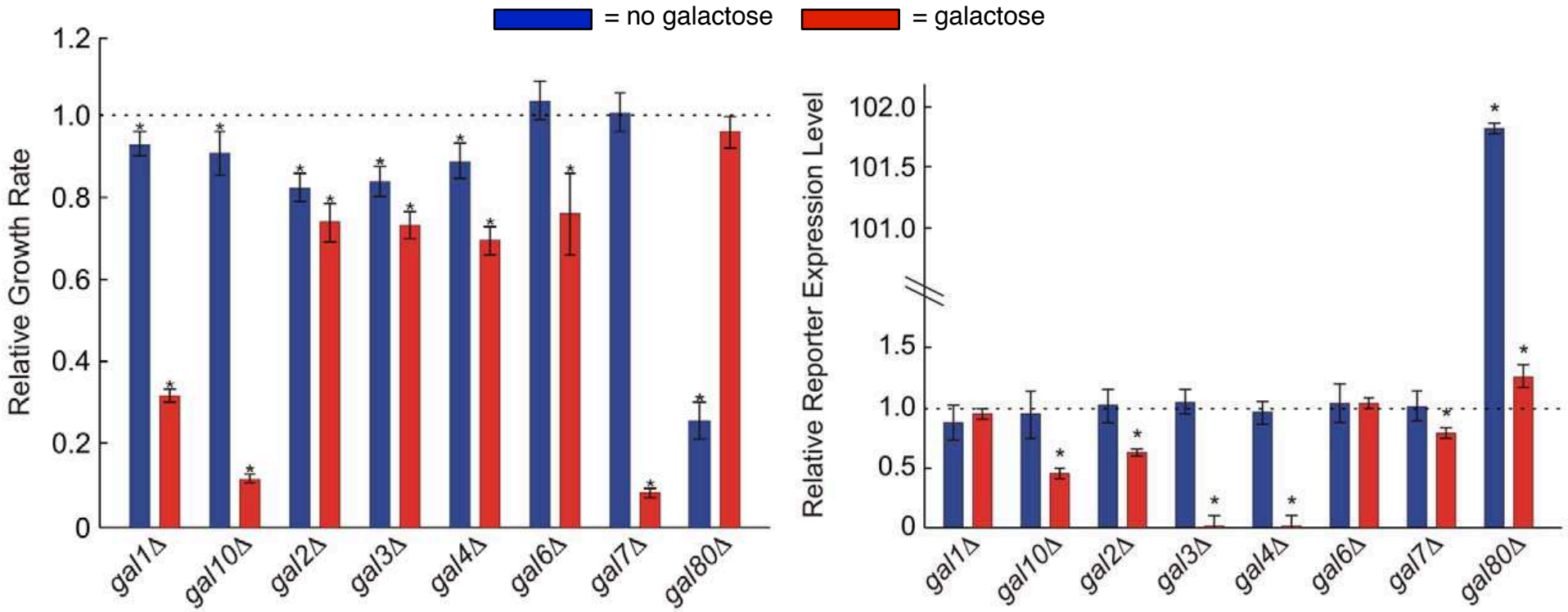
The galactose pathway in yeast – a test case



Experiment

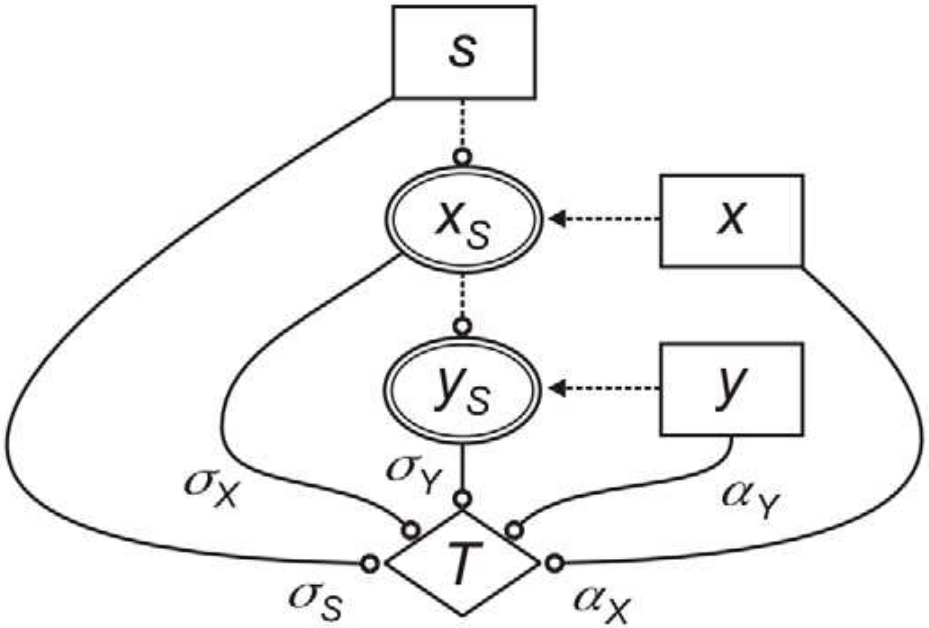
- We considered two signal states: galactose absent (but another sugar in the media) and galactose present (at a certain concentration)
- We considered two different quantitative phenotypes: growth rate relative to wild type, and fluorescence reporter expression driven by Gal10 promoter
- We observed the network in wild type state, and single and pairwise knockout of every gene

Results of single knockout experiments



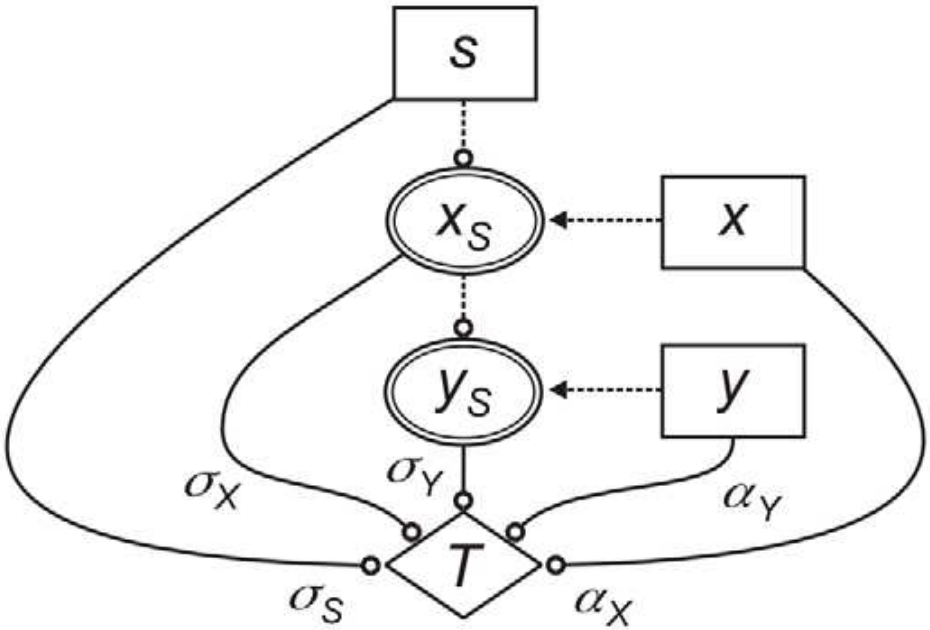
- With experimental noise, when are two trait measurements identical?
- A & W's model predicts that a single knockout can only affect the trait in one of the two signal states

A statistical model accounting for off-pathway, signal-independent effects



- Model distinguishes gene absence/presence (X,Y) from their signal-dependent activities (X_S,Y_S)
- Trait depends linearly on Boolean variables S, X, Y, X_S, Y_S
- We use linear regression to fit parameters σ_S, σ_X, σ_Y, α_X, α_Y across all experimental conditions
- Signs of parameters statistically significantly $\neq 0$ constitute patterns for pathway inference

More concretely . . .



$$T(x,y,s) = T_0 + \alpha_X(1-x) + \alpha_Y(1-y) + \alpha_I(1-x)(1-y) + \sigma_S s + \sigma_X(x_S(x,s)) + \sigma_Y(y_S(x,y,s)),$$

$$s = \begin{cases} 1 & \text{if the signal is ON} \\ 0 & \text{otherwise} \end{cases},$$

$$x = \begin{cases} 1 & \text{if } X \text{ is deleted} \\ 0 & \text{otherwise} \end{cases},$$

$$y = \begin{cases} 1 & \text{if } Y \text{ is deleted} \\ 0 & \text{otherwise} \end{cases}.$$

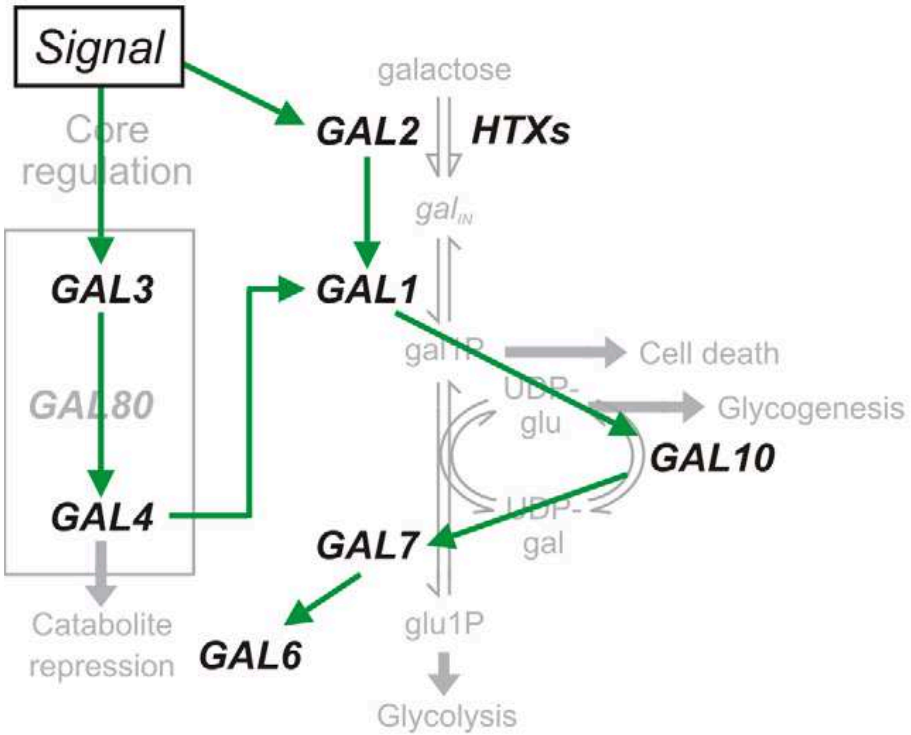
$$x_S(x,s) = \begin{cases} (1-x)s & \text{if } S \text{ activates } X \\ (1-x)(1-s) & \text{if } S \text{ represses } X \end{cases}$$

$$y_S(x,y,s) = \begin{cases} (1-y)x_S & \text{if } X \text{ activates } Y \\ (1-y)(1-x_S) & \text{if } X \text{ represses } Y \end{cases}$$

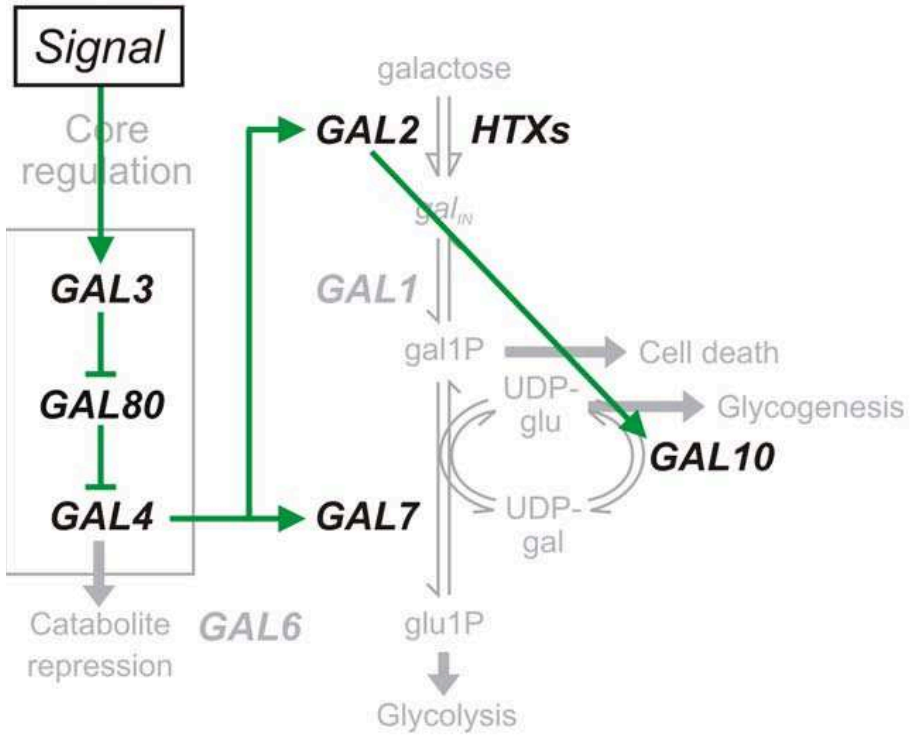
- Eight choices of "core" pathway structure: X, Y upstream/downstream, activation/repression for the first two links
- Core pathways that predict T within experimental error are considered possible explanations
- Extra links depend on whether parameters are significantly $+$ or $-$
- Predicted X, Y relationships assembled to reconstruct whole network

Results on the galactose network

from growth data



from expression data



- We are largely able to correctly reconstruct the galactose regulatory and metabolic pathways

Key conclusions of Phenix et al. 2011

Epistasis analysis can be successful on "noisy" quantitative data, even with there are off-pathway effects, using a more general model

- Standard linear regression techniques account for noise and tell us which parameters are significant
- Our model separately quantifies signal-dependent and signal-independent effects of the genes in the pathways

Epistasis analysis conclusions

- Avery & Wasserman laid the foundation for reasoning about patterns of traits observed upon gene perturbations, and how pathways can be reconstructed — but sometimes we cannot infer pathway stucture by this approach
- Much recent work (by us and many others) has aimed at loosening their assumptions, to account for noisy, quantitative data, off-pathway effects, complex phenotypes (e.g. probabilities of activation), etc.
- The hunt is on for ways to overcome the limits of classical (& modern) approaches to epistasis — e.g. our own work on dynamical epistasis analysis (Awdeh et al. 2015, 2017)

Questions?