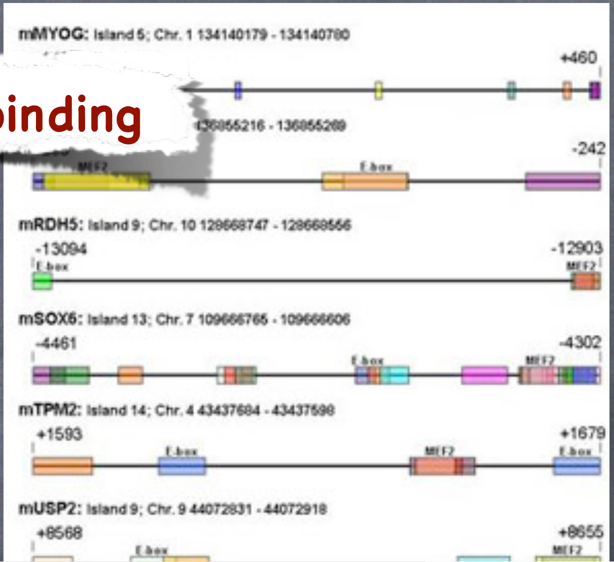
7. Transcriptional regulation from microarray data

Warning: Statistical physics.
It only works on average.

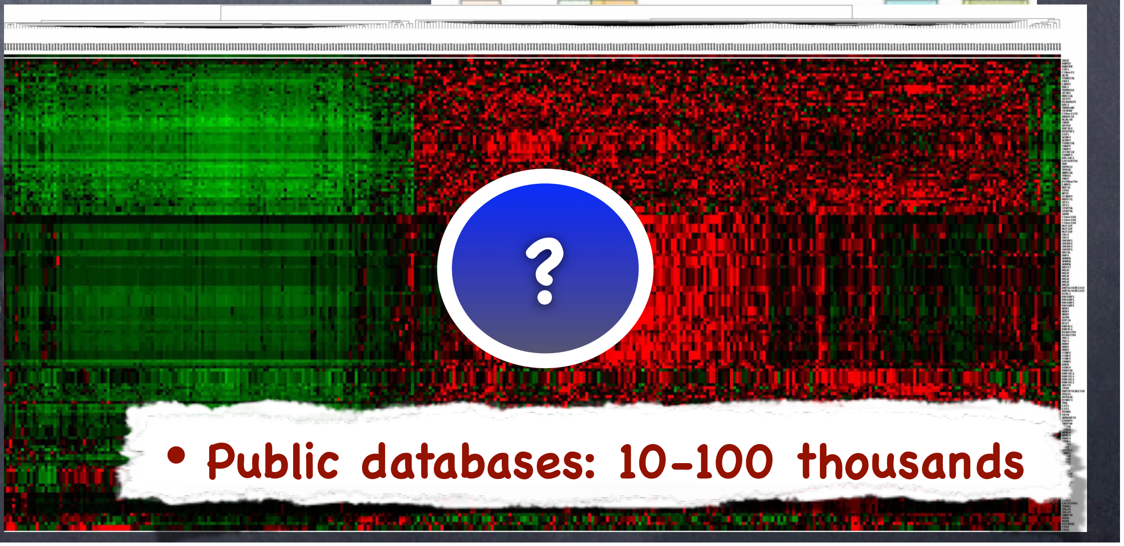
http://regan.med.harvard.edu/CVBR-course.php

How about the entire genetic regulatory network?

- Put together known literature
  - TRANSFAC, TRANSPATH
  - INGENUITY
  - InterAct



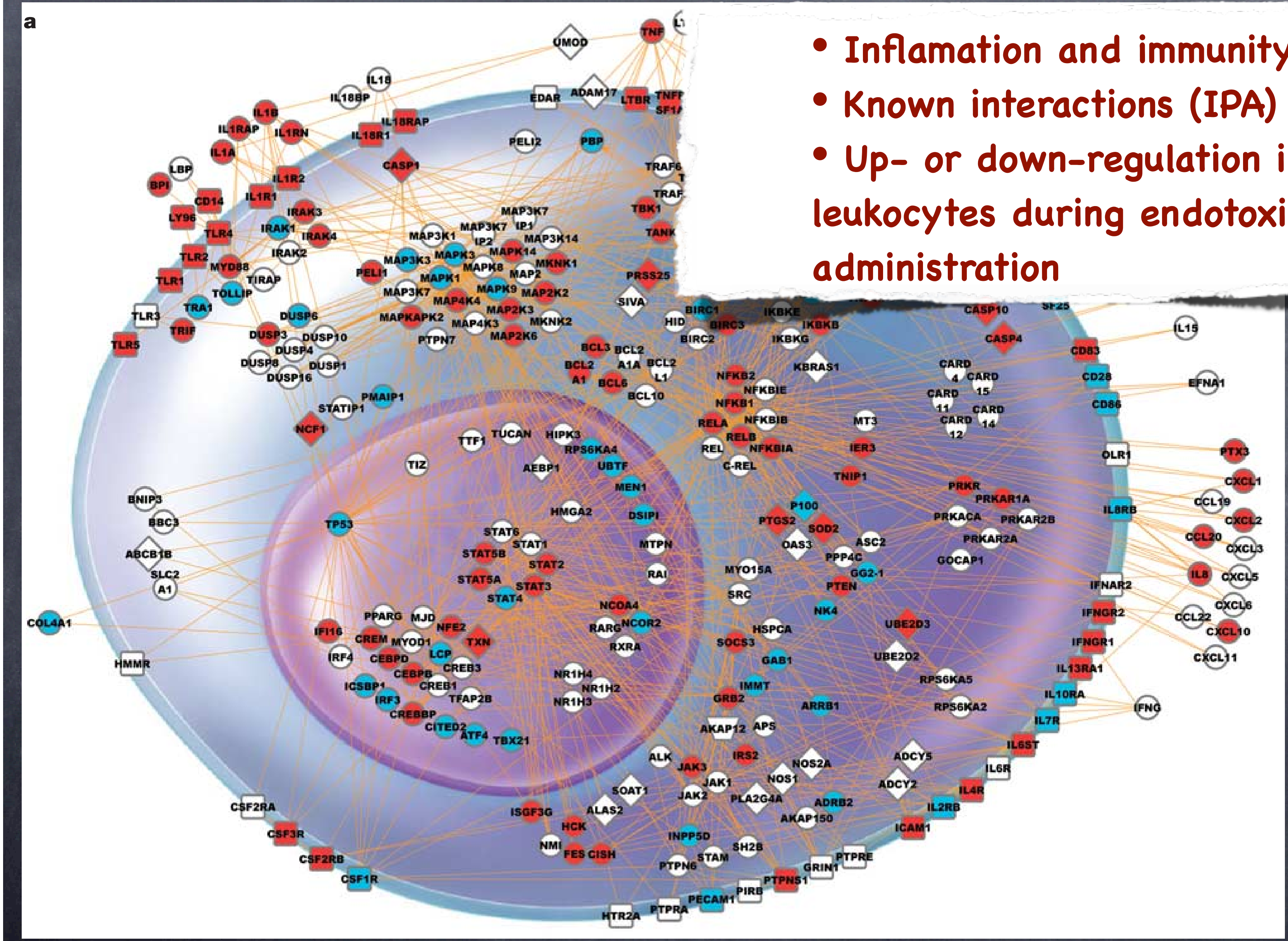
- Predict TF binding



- Public databases: 10-100 thousands

If we know enough interactions:

- Inflamation and immunity genes
- Known interactions (IPA)
- Up- or down-regulation in blood leukocytes during endotoxin administration



An overview of approaches

Co-expression networks

Boolean networks

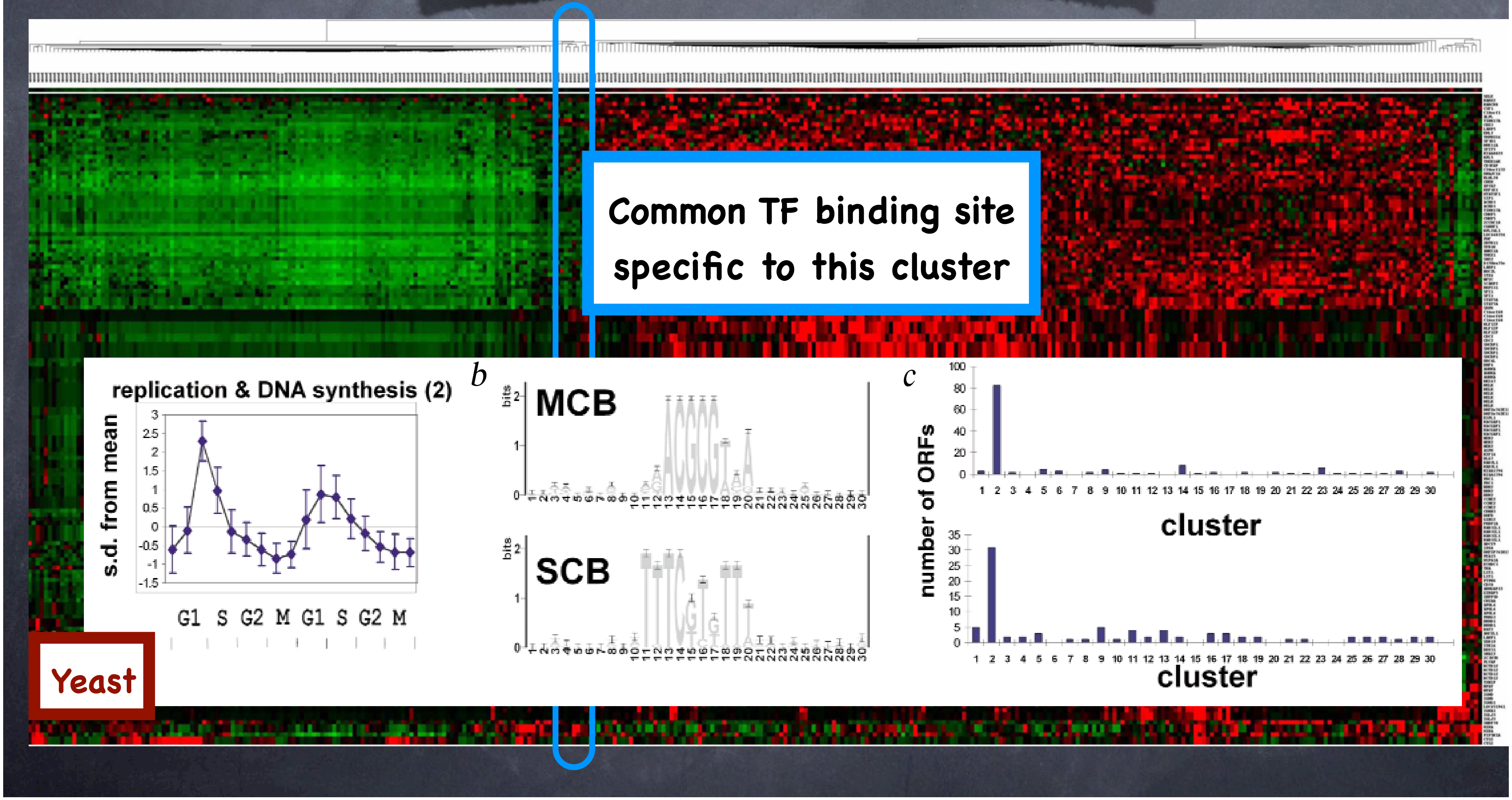
Bayesian networks

ODEs and regression

- Known literature
- Transcription factors
- Binding site information
- Perturbation experimental design

How far does correlation take us?

- Hypothesis: co-expression means
  - common upstream regulator
  - shared function



Common TF binding site specific to this cluster

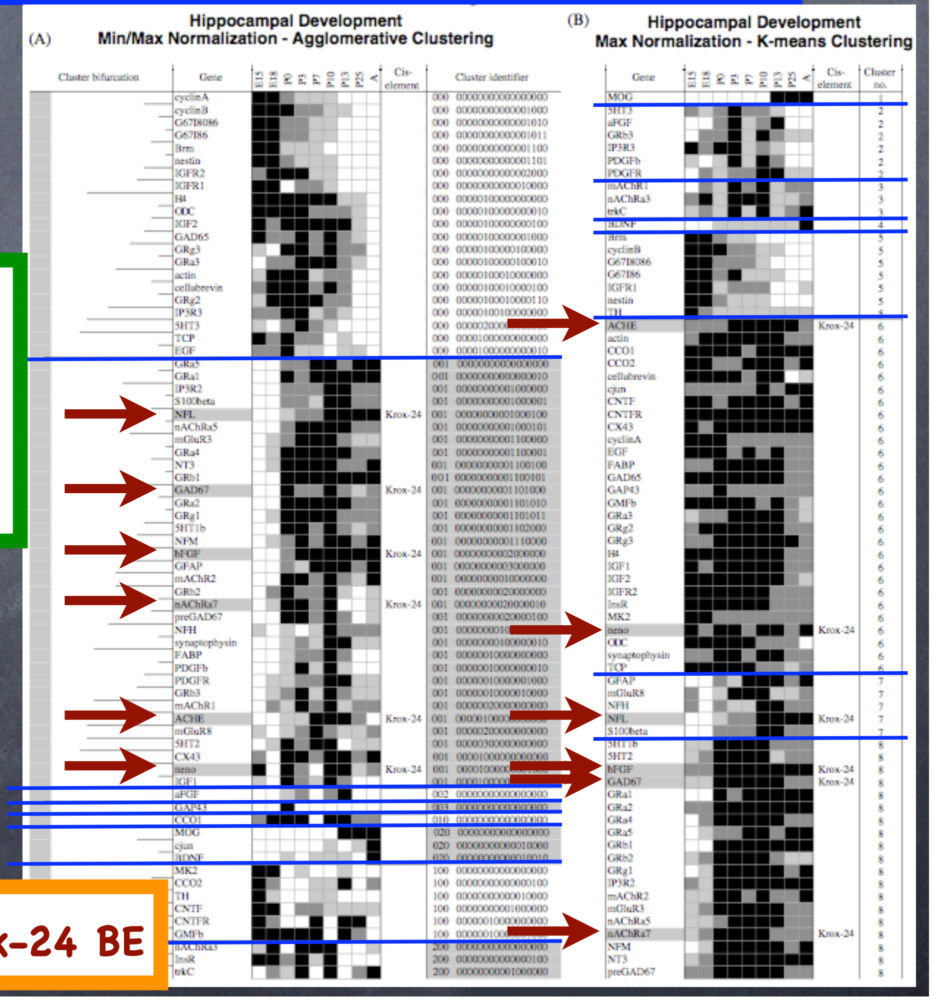
Yeast

Sounds nice! How does clustering go?

- Data preparation & normalization
- A distance metric: what does "similar" mean?
    - Pearson correlation
    - Euklidean distance
    - Mutual Information
    - ...
- Clustering
    - Hier
    - Non-hierarchical
        - K-means
        - SOM
        - PCA
    - Fuzzy clustering
    - ...

All capture functionally related gene groups

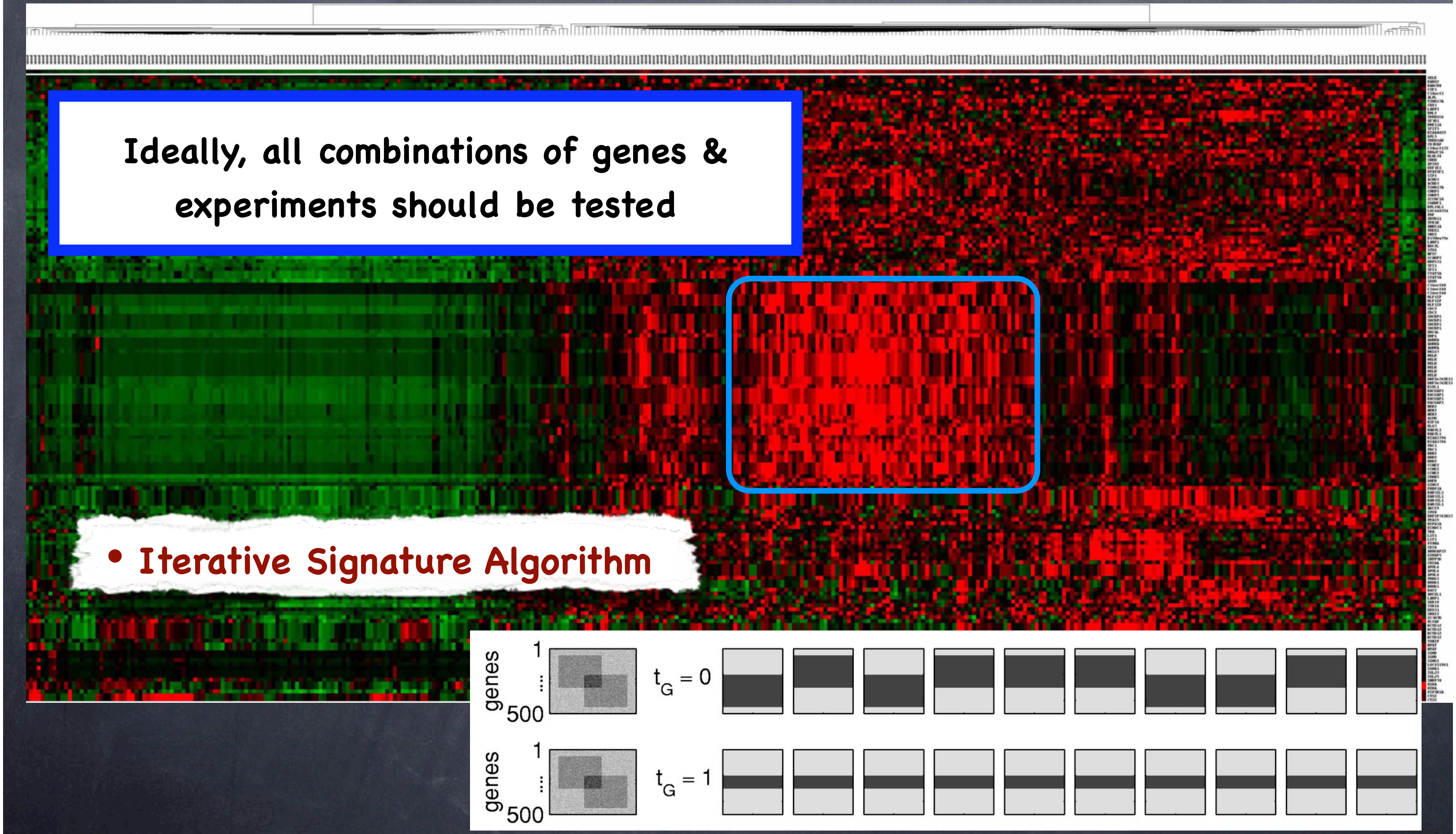
Different results on the same dataset



No consensus on which one is better (no metric for "better")

Rat, Krox-24 BE

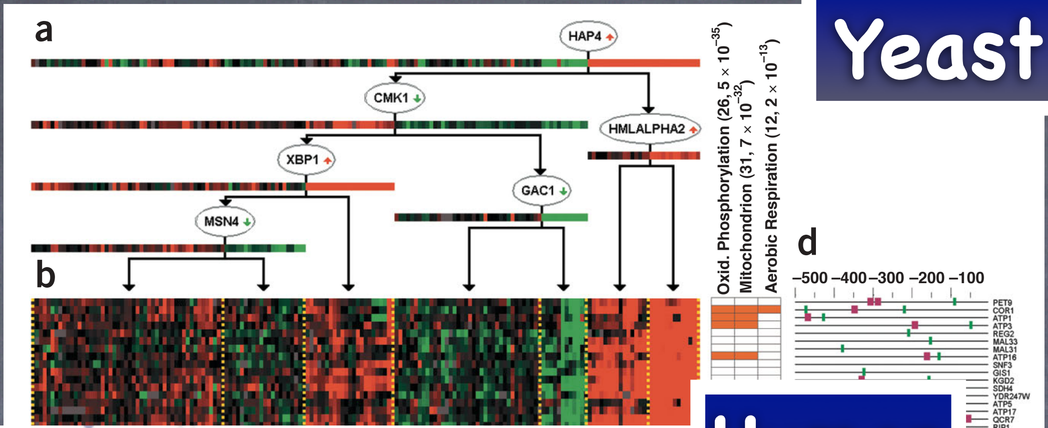
Coherent groups of genes & experiments



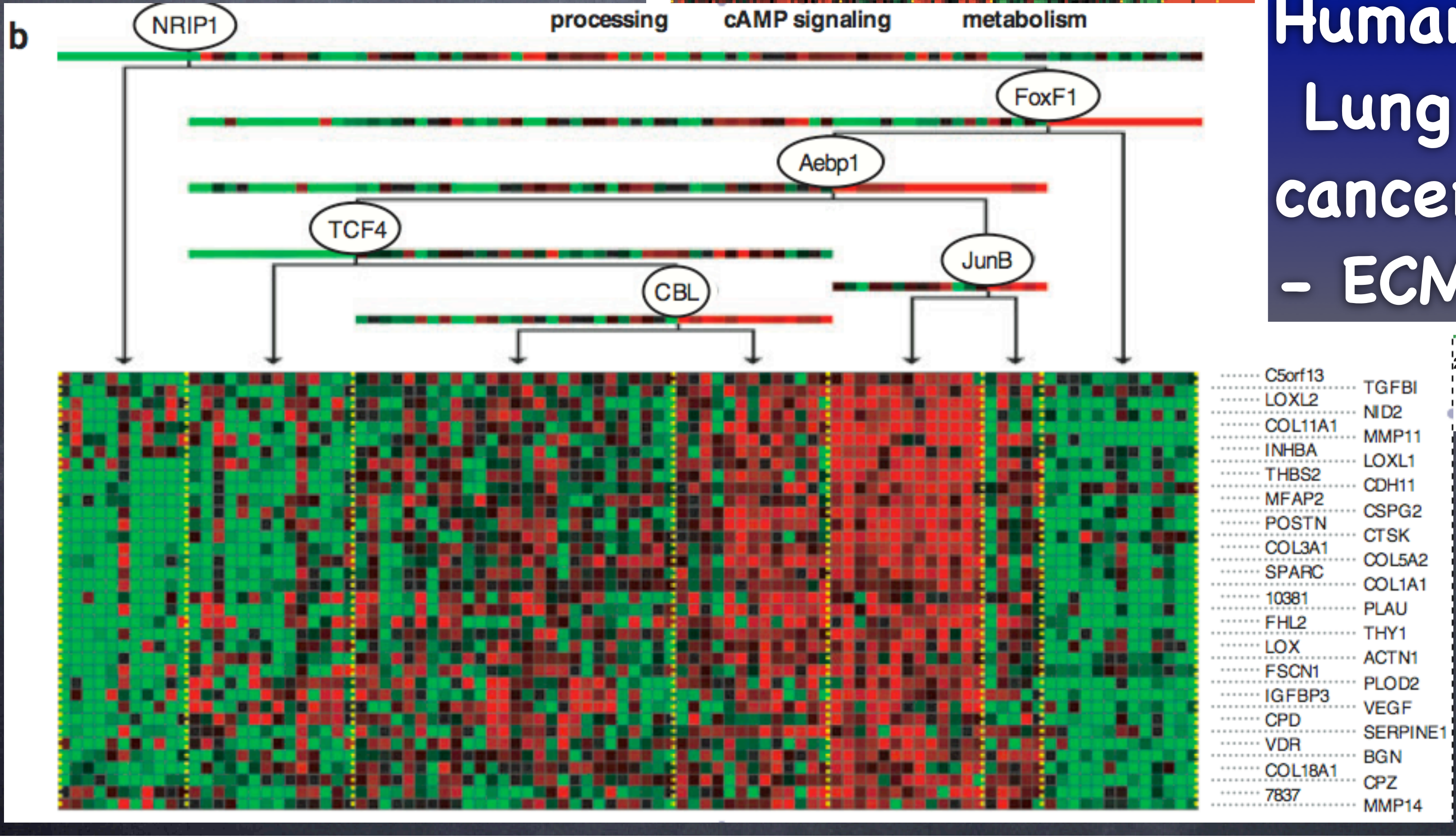
Ideally, all combinations of genes & experiments should be tested

- Iterative Signature Algorithm

Regulators with best predicting power



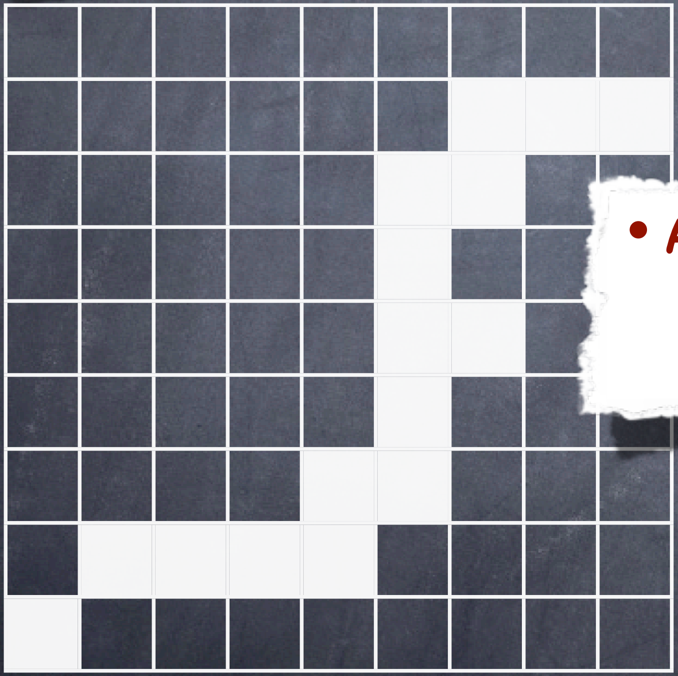
Yeast



Human Lung cancer - ECM

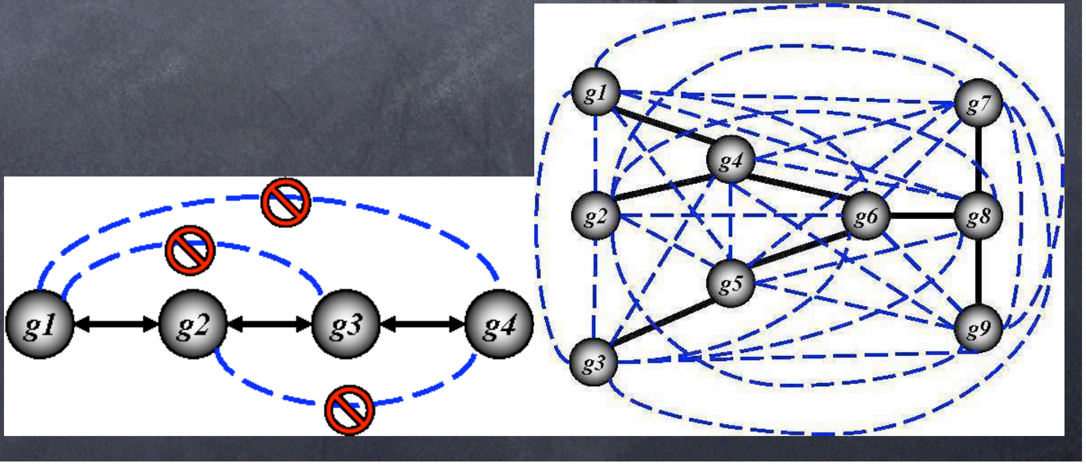
Does correlated behavior translate to direct regulation?

- **Mutual Information**
  - **non-linear measure of correlated behavior**

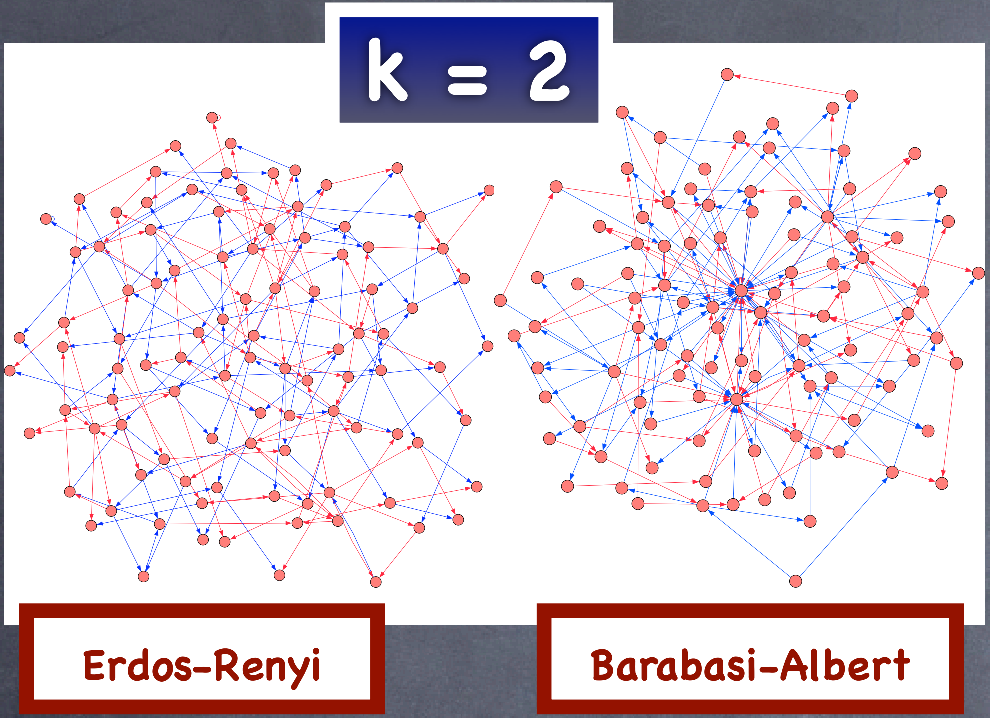


$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right),$$

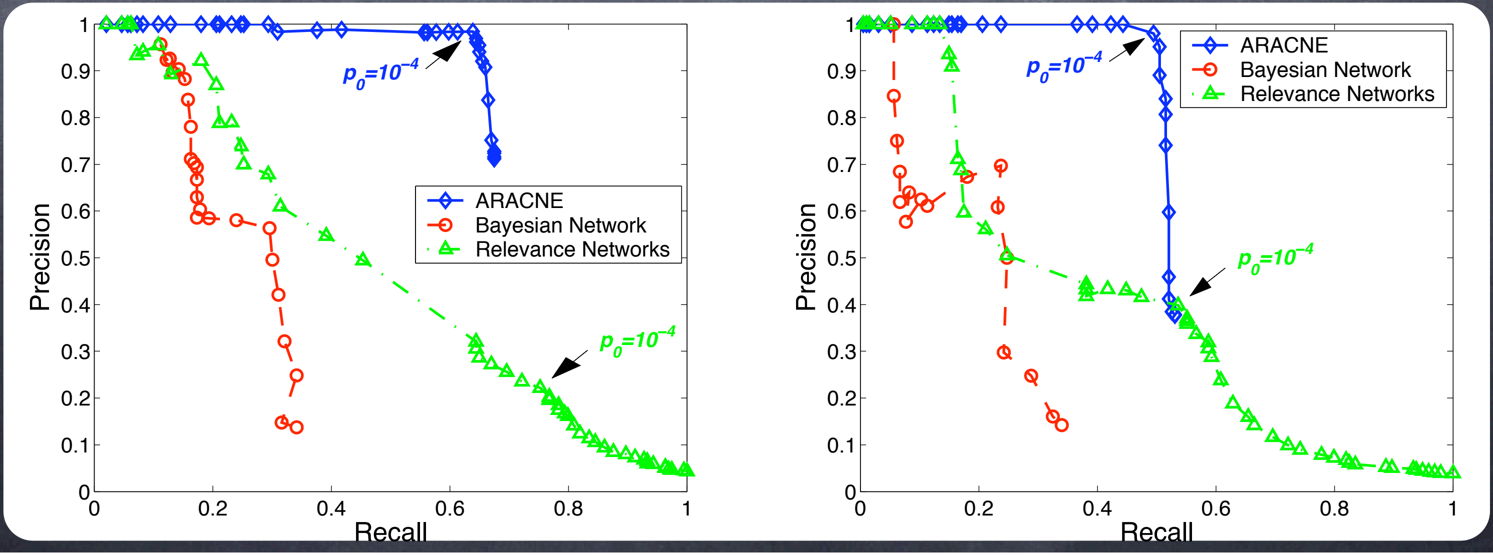
- **ARACNE**
  - **Mutual Information**
  - **Data processing inequality**



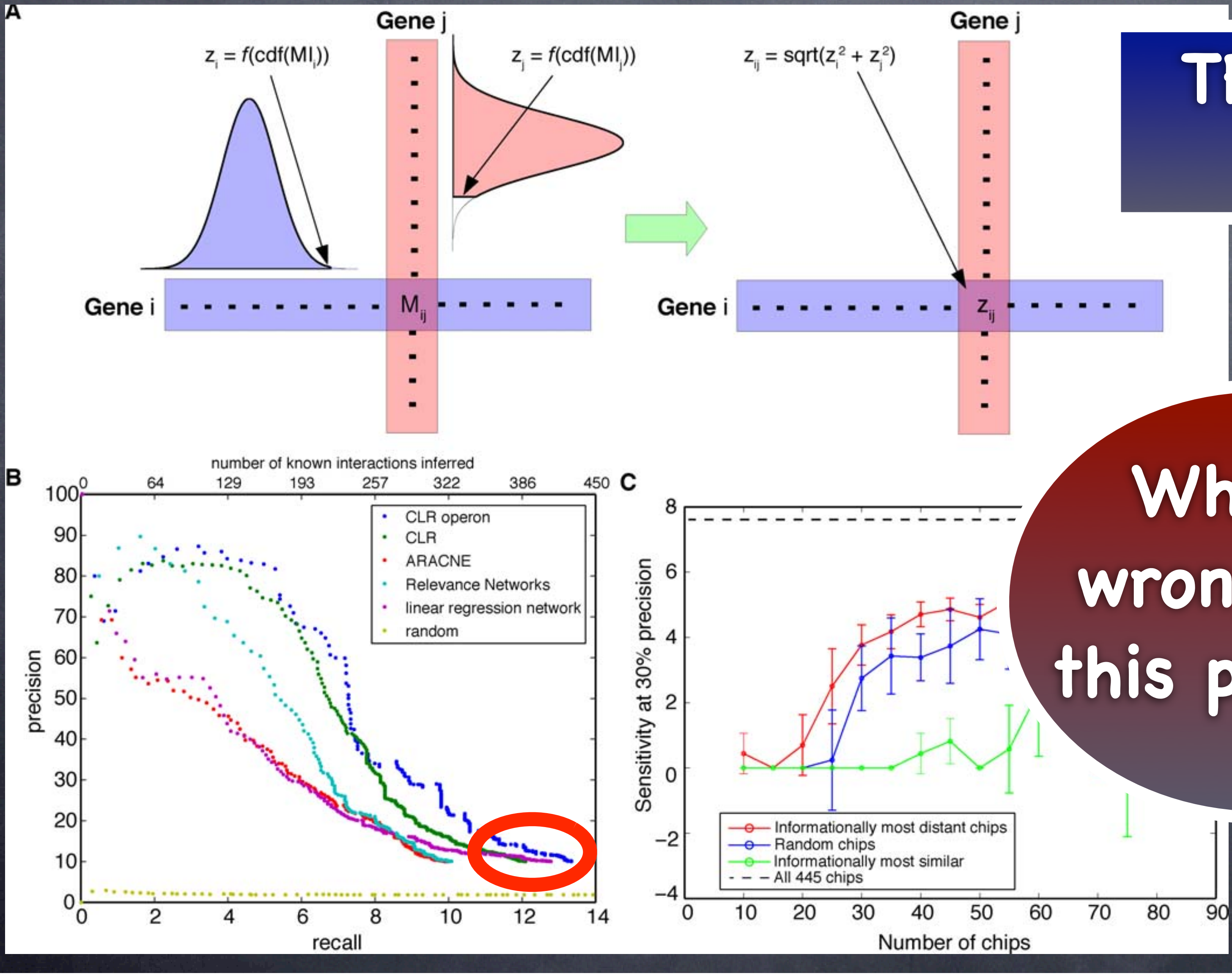
Test of ARACNE



How about real microarray data?



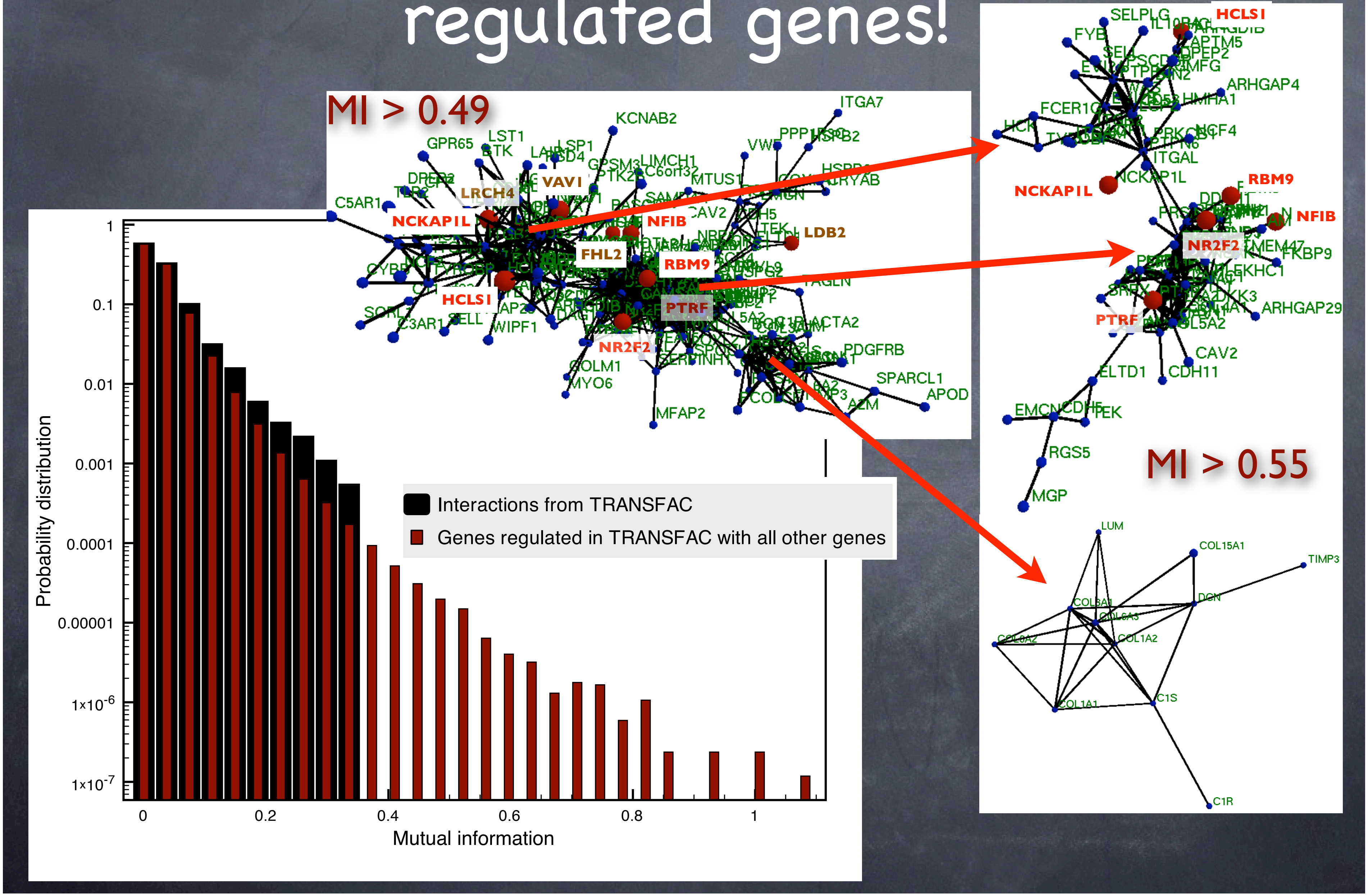
CLR - Context Likelihood of Relatedness



TF -> G only

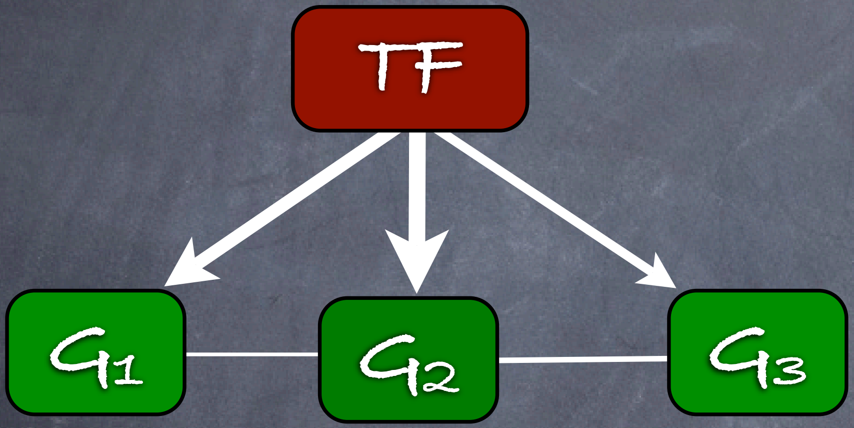
What is wrong with this picture?

Co-expression is stongest between co-regulated genes!

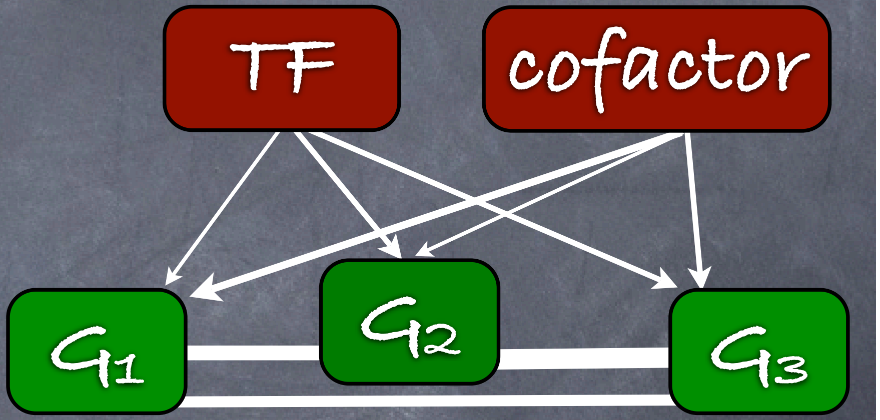


Reasons behind the trouble

- ARACNe IFF 1 input/ gene!



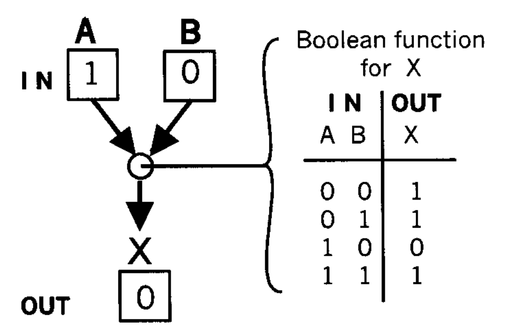
- Fails for combinatorial regulation



Is direct regulatory information contained in MA data?

Boolean network approaches

- State of the system, or gene activity profile: (0,1,1,0,0,1,1,1,0,...,0)
- State changes in time: trajectory in nD space
- Structure of state space determines all possible dynamics



- Only a small fraction of all cell states are stable
- The system does not visit all possible states
- Attractors:

ints: a state in which

are satisfied

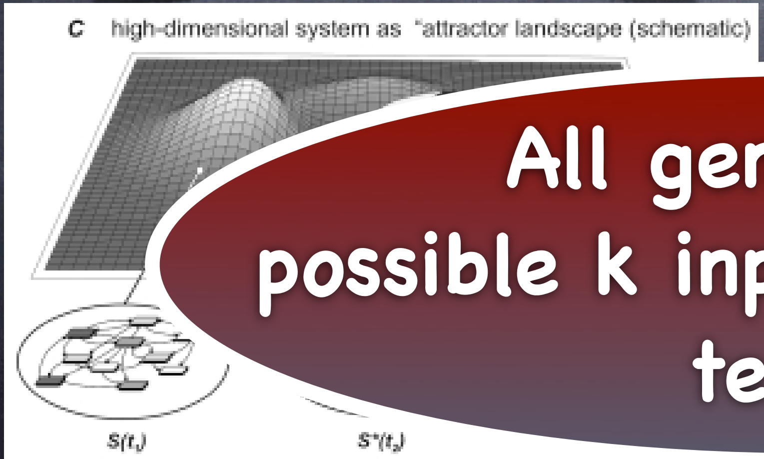
e number

which the

es

Attractor basins

✳ Ordered systems


C high-dimensional system as "attractor landscape (schematic)

All genes <- all possible k inputs need to be tested

Finding expression dependencies

- Bayesian networks
    - Directed Acyclic Graph
    - Its structure describes the conditional probabilities that best fit the data
    - Discrete or continuous
    - Extremely time-consuming on large networks

- Regression based methods

$$X_i^{(k)} = \sum_j h_{j\to i}\cdot X_j^{(k)} + p_i^{(k)}$$

$N_{exp} \ll N_{gene}$

1: Maximum k inputs / gene

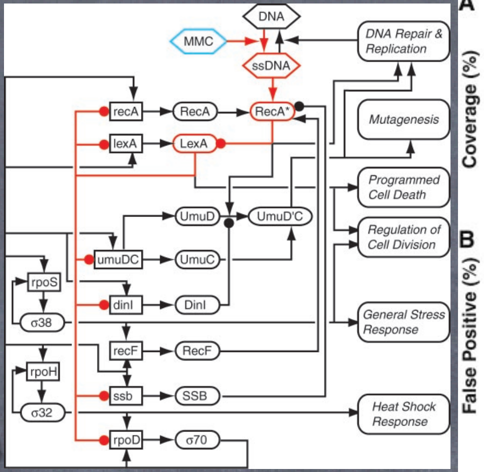
2: Find relevant part of gene space

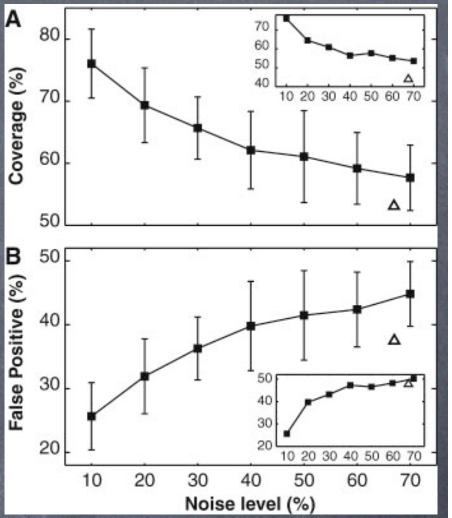
NIR - Network identification by Multiple regression

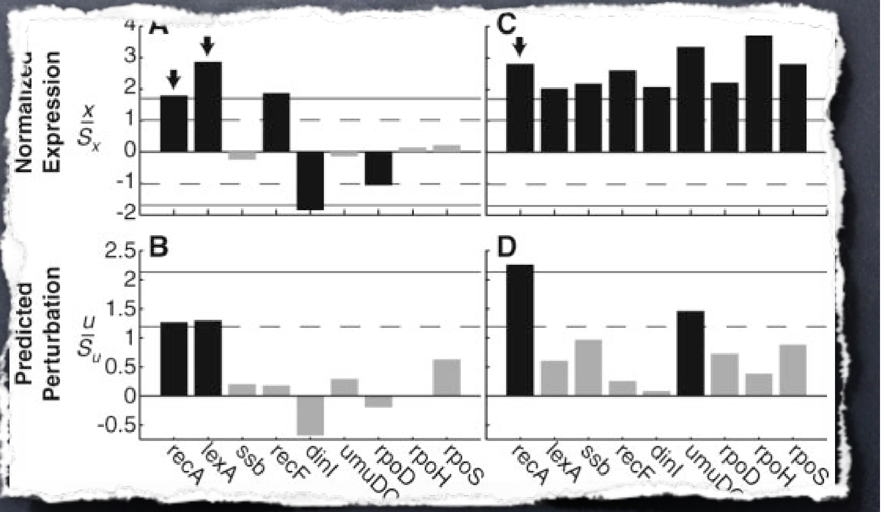
$$X_i^{(k)} = \sum_j h_{j\to i} \cdot X_j^{(k)} + p_i^{(k)}$$

$$\sum z_j^{(k)} \cdot A'_{ji} = -p_i^{(k)}$$

- Iterative approximation
  - 0th approximation:
  - calculate $p_i^{(k)}$
  - for each gene
    - use $n \in (Q, M)$ equations where $p_i < p_c$
    - solve







Summer break!

Next lecture:

Sep 14
12 PM