# Evaluation of inferred networks

**Sushmita Roy**

sroy@biostat.wisc.edu

**Computational Network Biology**
Biostatistics & Medical Informatics 826

https://compnetbiocourse.discovery.wisc.edu

Sep 27th 2018

# Evaluating the network

- Assessing confidence
- Area under the precision recall curve
- Do modules or target sets of genes participate in coherent function?
- Can the network predict expression in a new condition?

# Assessing confidence in the learned network

- Typically the number of training samples is not sufficient to reliably determine the "right" network
- One can however estimate the confidence of specific features of the network
  - Graph features $f(G)$
- Examples of $f(G)$
  - An edge between two random variables
  - Order relations: Is $X$, $Y$'s ancestor?

# How to assess confidence in graph features?

- What we want is $P(f(G)/D)$, which is

$$\Sigma_G f(G) P(G|D)$$

- But it is not feasible to compute this sum

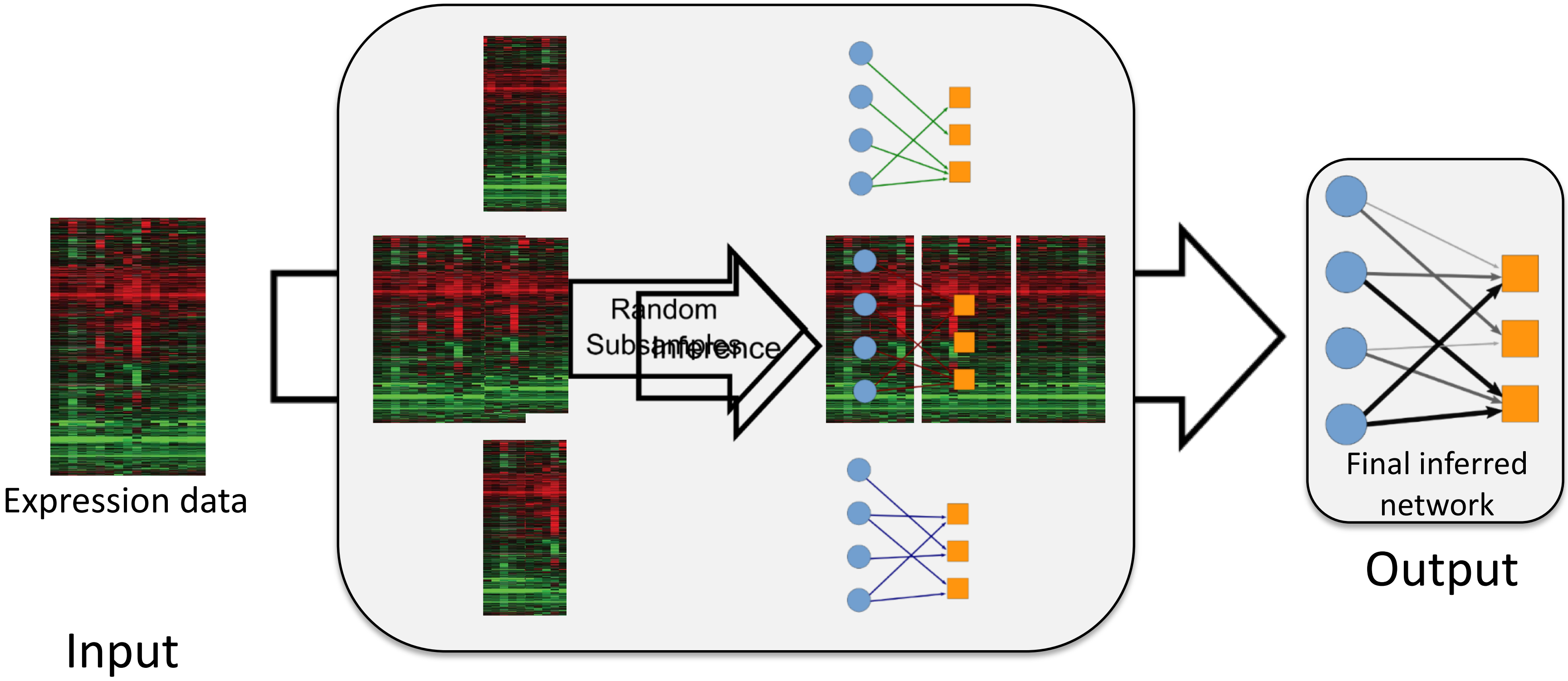
- Instead we will use a “bootstrap” procedure

# Bootstrap to assess graph feature confidence

- For $i=1$ to $m$
  - Construct dataset $\mathbf{D}_i$ by sampling with replacement $N$ samples from dataset $\mathbf{D}$, where $N$ is the size of the original $\mathbf{D}$
  - Learn a graphical model $\{G_{i,}\Theta_i\}$
- For each feature of interest $f$, calculate confidence

$$\mathrm{Conf}(f) = \frac{1}{m}\sum_{i=1}^{m} f(G_i)$$

# Bootstrap/stability selection



Expression data

Random Subsamples

Inference

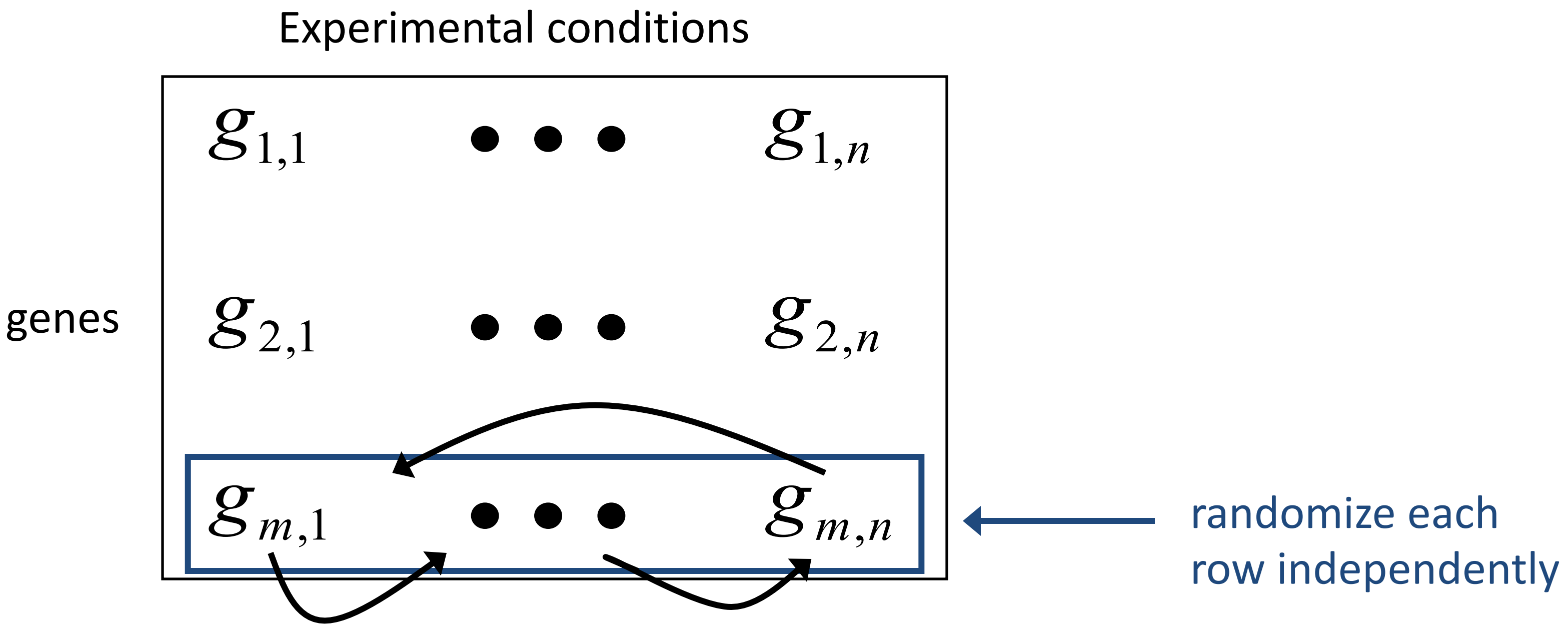
Final inferred network

Output

Input

# Does the bootstrap confidence represent real relationships?

- Compare the confidence distribution to that obtained from randomized data
- Shuffle the columns of each row (gene) separately
- Repeat the bootstrap procedure



Experimental conditions

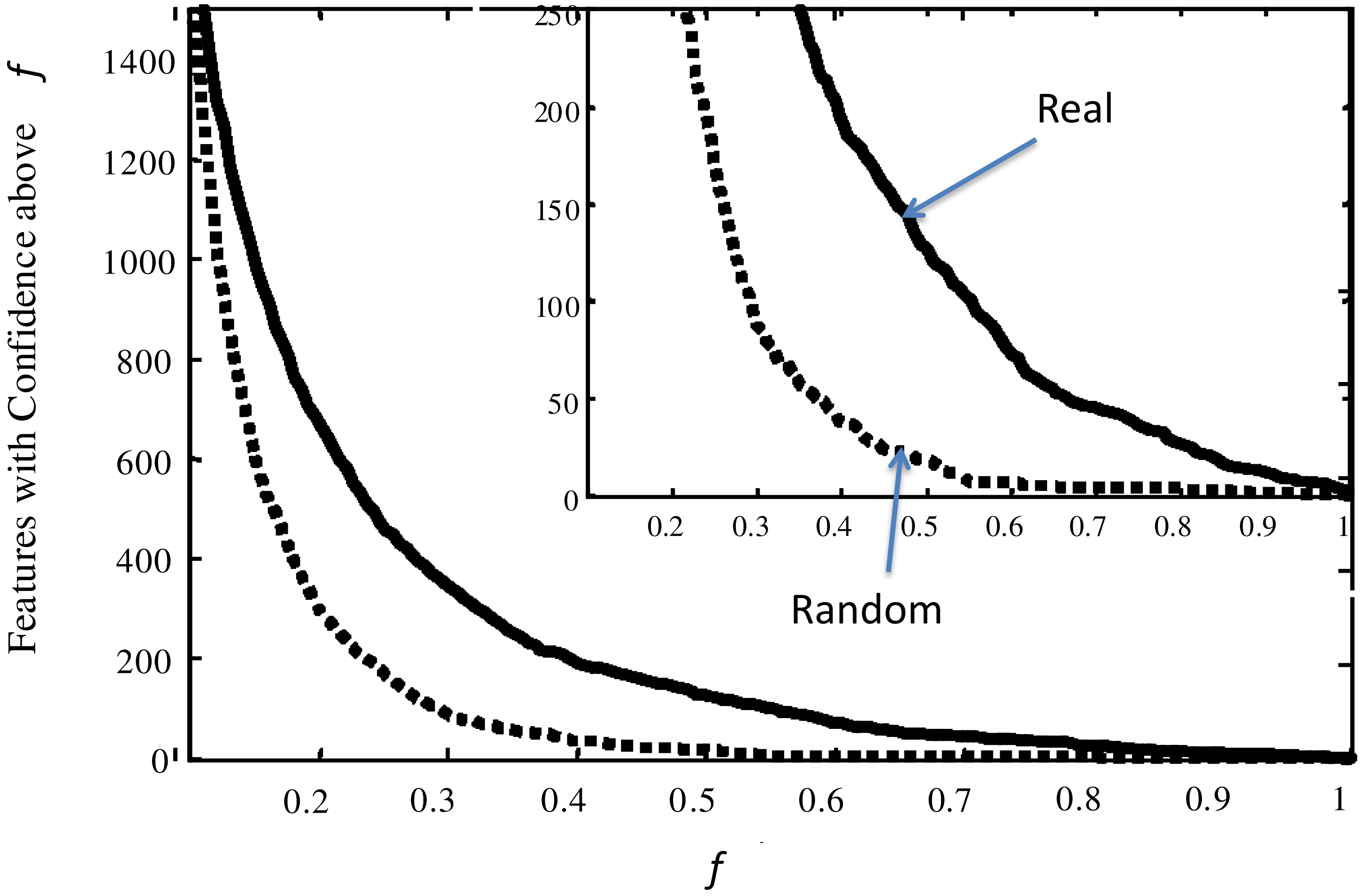
genes

$$\begin{array}{ccc} g_{1,1} & \bullet\bullet\bullet & g_{1,n} \\ g_{2,1} & \bullet\bullet\bullet & g_{2,n} \\ g_{m,1} & \bullet\bullet\bullet & g_{m,n} \end{array}$$

randomize each row independently

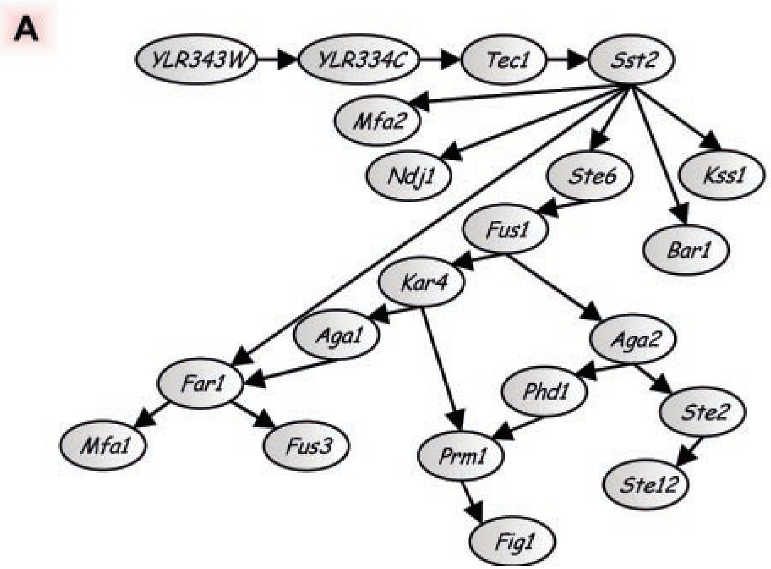
Slide credit Prof. Mark Craven

# Bootstrap-based confidence differs between real and actual data

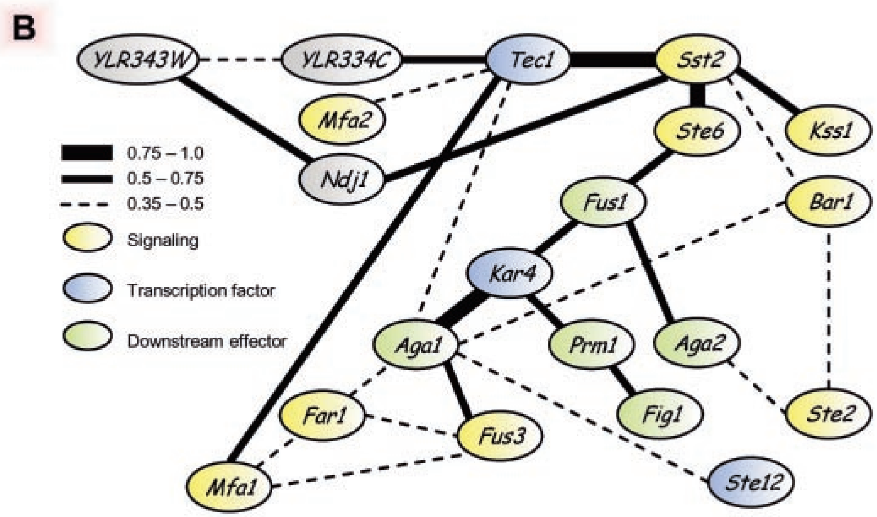


Friedman et al 2000

# Example of a high confidence sub-network



One learned Bayesian network



Bootstrapped confidence Bayesian network: highlights a subnetwork associated with yeast mating pathway. Colors indicate genes with known functions.

Nir Friedman, Science 2004

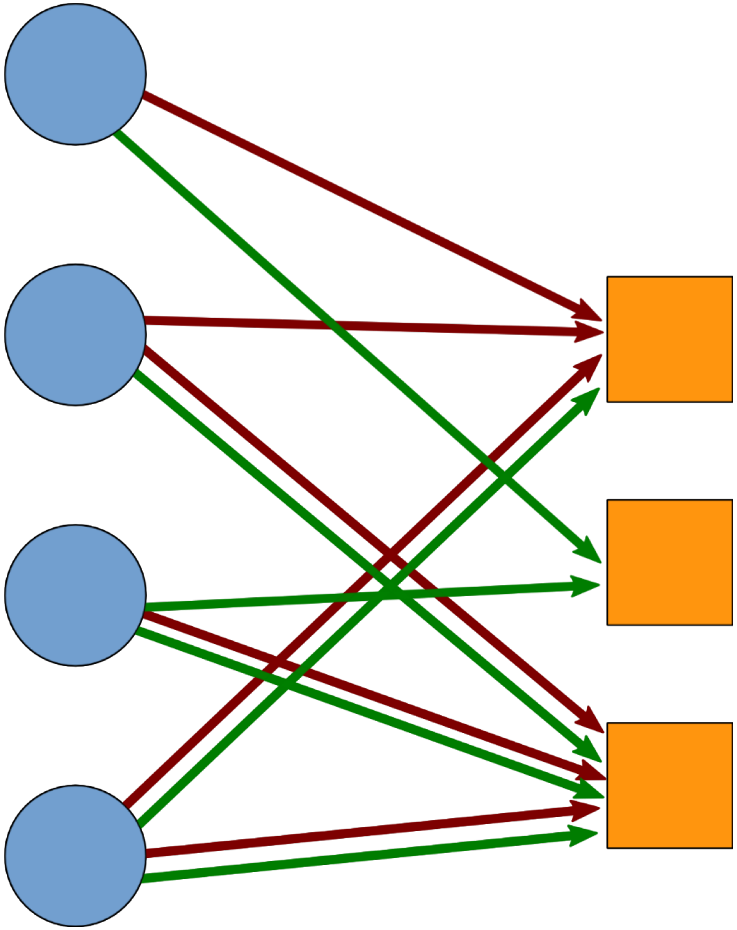
# Area under the precision recall curve (AUPR)

- Assume we know what the “right” network is
- One can use Precision-Recall curves to evaluate the predicted network
- Area under the PR curve (AUPR) curve quantifies performance

$$\text{Precision} = \frac{\text{\# of correct edges}}{\text{\# of predicted edges}}$$

$$\text{Recall} = \frac{\text{\# of correct edges}}{\text{\# of true edges}}$$

# Edge based comparison (AUPR)



Inferred network

Both networks

True positives

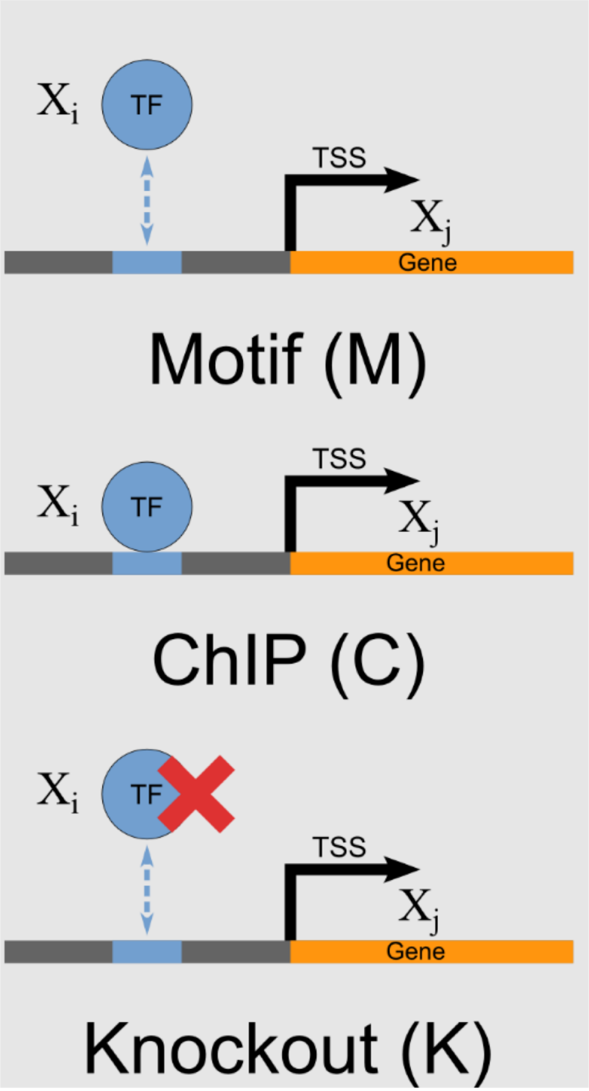
Slide credit: Alireza Fotuhi Siahpirani

# Experimental datasets to assess network structure for gene regulatory networks

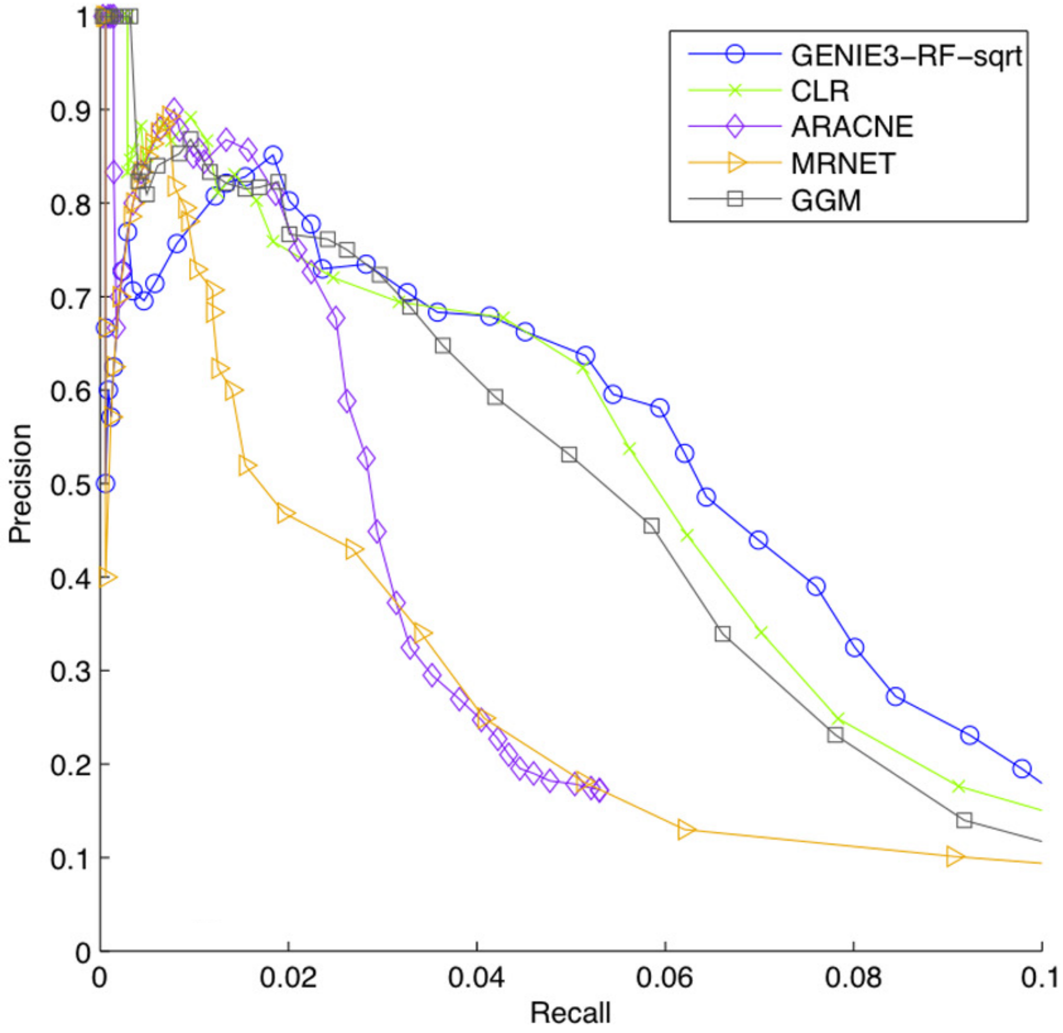
- Sequence specific motifs

- ChIP-chip and ChIP-seq

- Factor knockout followed by whole-transcriptome profiling

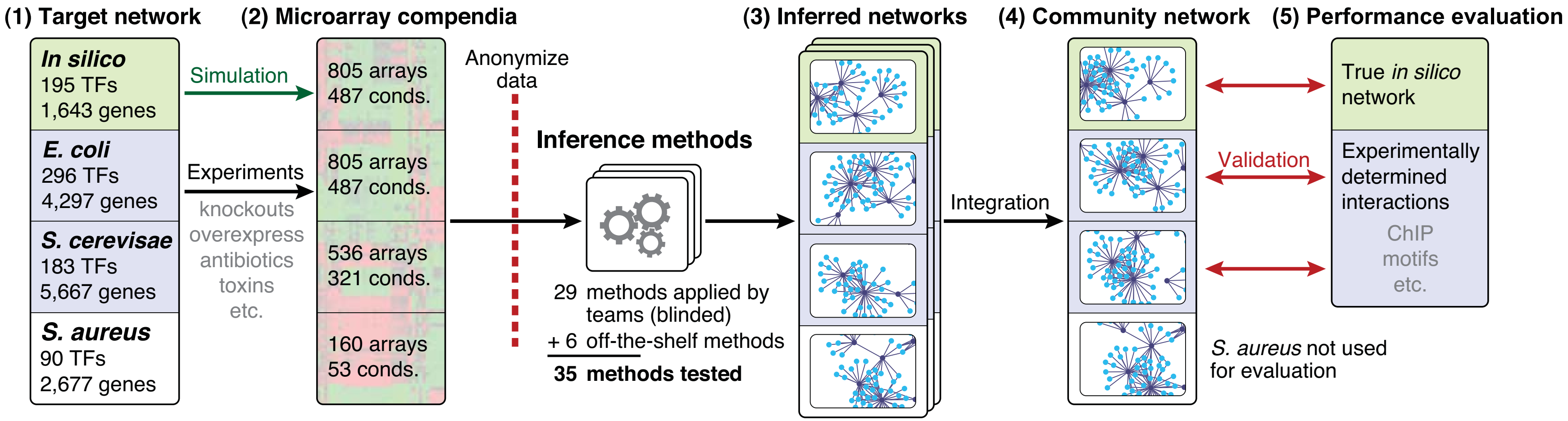


# AUPR based performance comparison



# DREAM: Dialogue for reverse engineeting assessments and methods

## Community effort to assess regulatory network inference

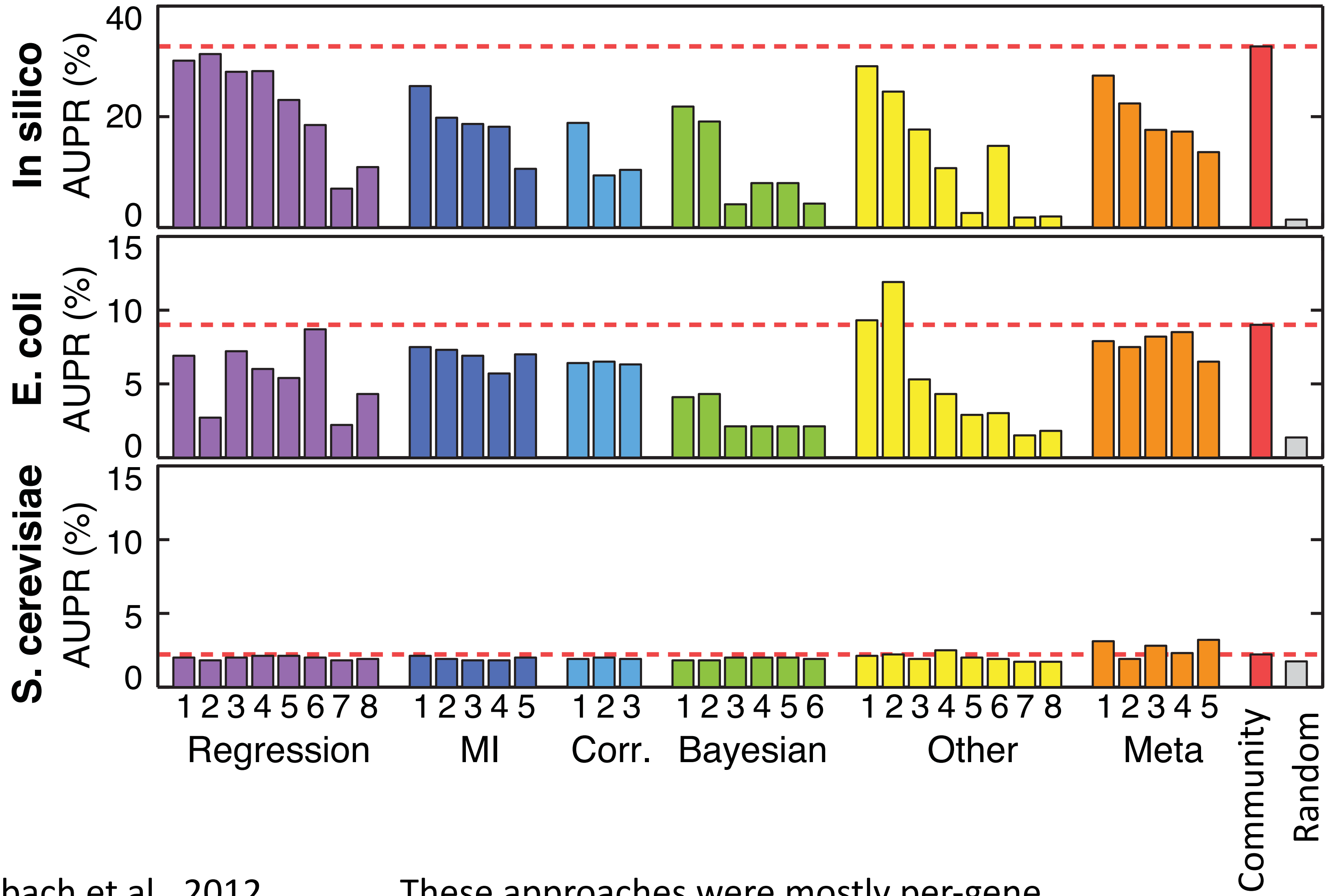


## DREAM 5 challenge

Previous challenges: 2006, 2007, 2008, 2009, 2010

Marbach *et al.* 2012, Nature Methods

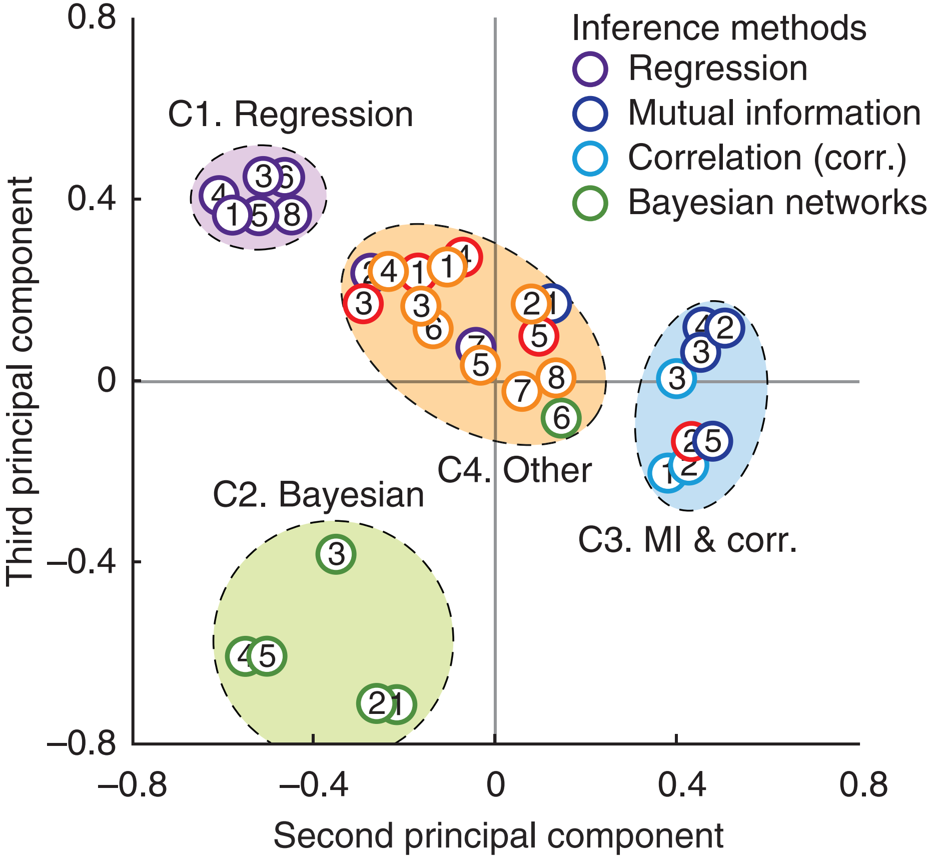
# Where do different methods rank?



Marbach et al., 2012

These approaches were mostly per-gene

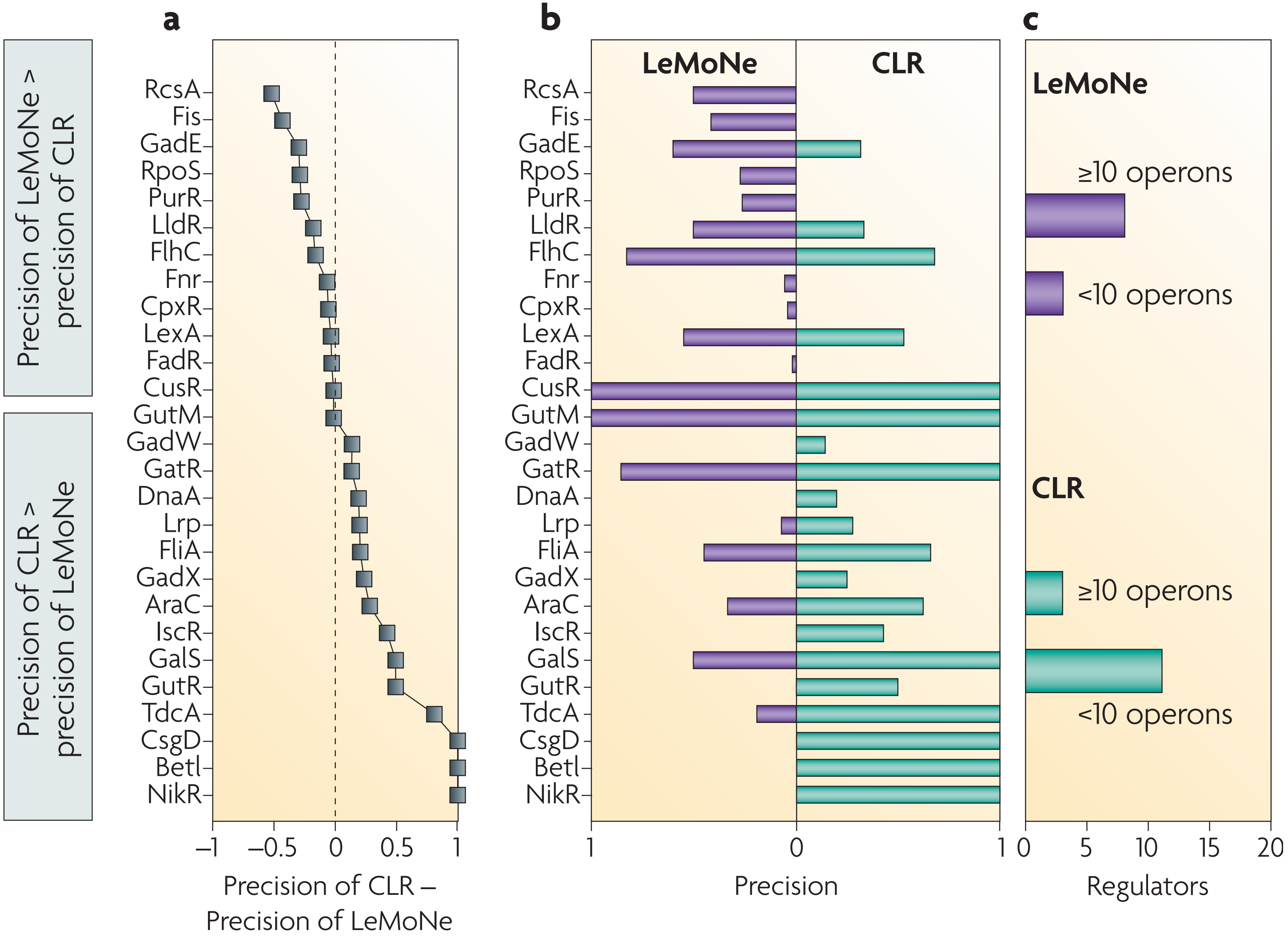
# Methods tend to cluster together



These approaches were mostly per-gene

Marbach et al., 2012

# Comparing per-module (LeMoNe) and per-gene (CLR) methods



Marchal & De Smet, Nature Reviews Microbiology, 2010

# Some comments about expression-based network inference methods

- We have seen multiple types of algorithms to learn these networks
  - Per-gene methods (learn regulators for individual genes)
    - Sparse candidate, GENIE3, ARACNE, CLR
  - Per-module methods
    - Module networks: learn regulators for sets of genes/modules
    - Other implementations of module networks exist
      - LIRNET: Learning a Prior on Regulatory Potential from eQTL Data (Su In Lee et al, Plos genetics 2009, http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000358)
      - LeMoNe: Learning Module Networks (Michoel et al 2007, http://www.biomedcentral.com/1471-2105/8/S2/S5)
  - Methods that combine per-gene and per-module (MERLIN)
- Methods differ in
  - how they quantify dependence between genes
  - Higher-order or pairwise
  - Focus on structure or structure & parameters
- Expression alone is not enough to infer the structure of the network
- Integrative approaches that combine expression with other types of data are likely more successful (next lectures)

# References

- Markowetz, Florian and Rainer Spang. "Inferring cellular networks-a review.." *BMC bioinformatics* 8 Suppl 6 (2007): S5+.
- N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601-620, Aug. 2000. [Online]. Available: http://dx.doi.org/10.1089/106652700750050961
- Dependency Networks for Inference, Collaborative Filtering and Data visualization  Heckerman, Chickering, Meek, Rounthwaite, Kadie 2000
- Inferring Regulatory Networks from Expression Data Using Tree-Based Methods Van Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, Pierre Geurts, Plos One 2010
- D. Marbach et al., "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, no. 8, pp. 796-804, Jul. 2012. [Online]. Available: http://dx.doi.org/10.1038/nmeth.
- R. De Smet and K. Marchal, "Advantages and limitations of current network inference methods." *Nature reviews. Microbiology*, vol. 8, no. 10, pp. 717-729, Oct. 2010. [Online]. Available: http://dx.doi.org/10.1038/nrmicro2419