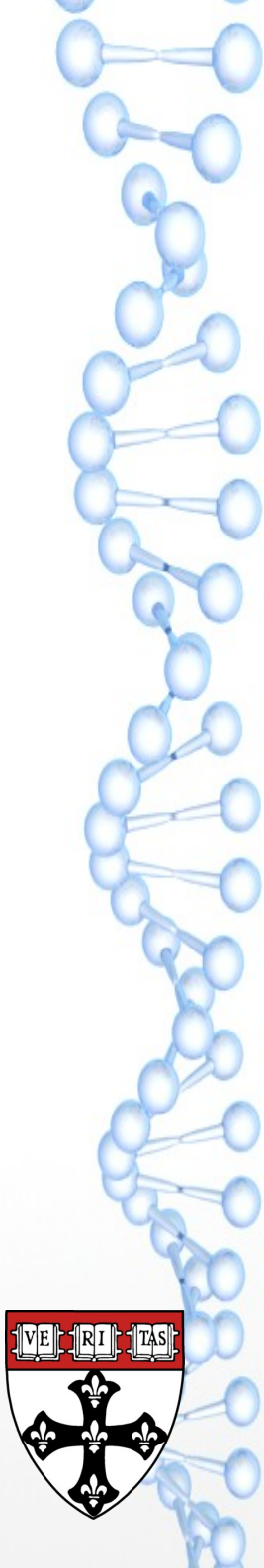


Methods in Case-Control Gene Regulatory Networks

Oral Qualifying Exam

May 18, 2015

Dan Schlauch

Outline

1) Why network inference?

2) The challenges of GRN inference.

3) The challenges of GRN differentiation.

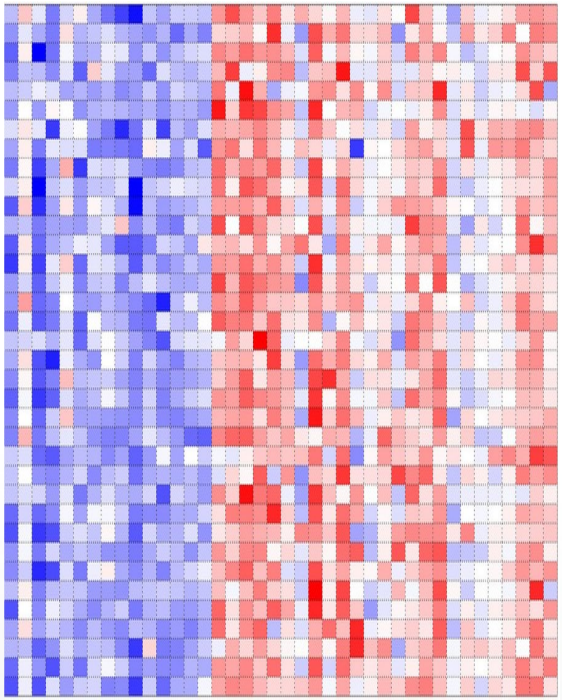
4) BERE, a novel GRN algorithm.

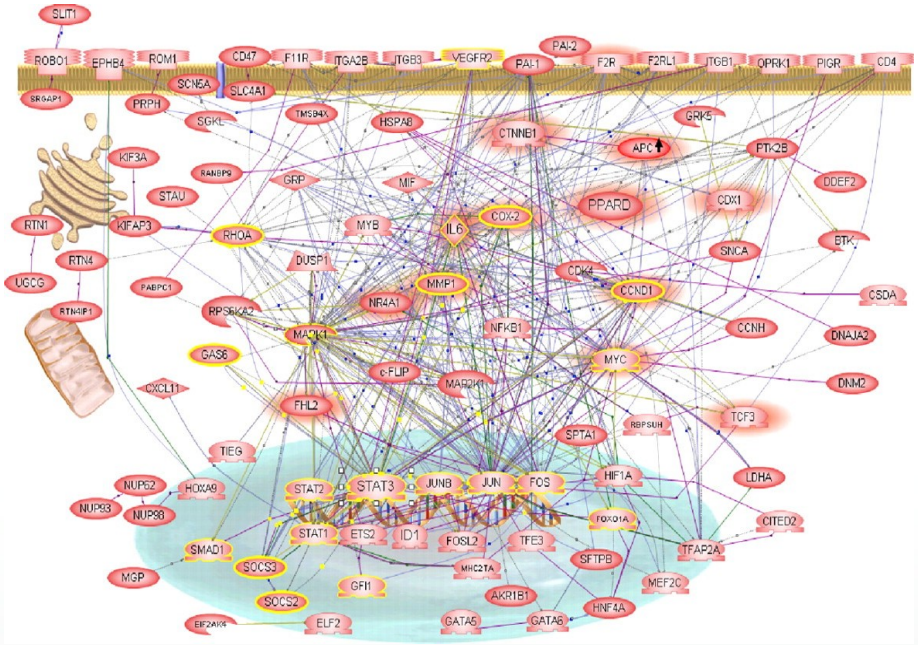
5) A novel method for identifying meaningful structural changes in GRNs in case-control studies.

6) Future work

Why Gene Regulatory Network Inference?

- Genes are not independent objects.
- How are they related?





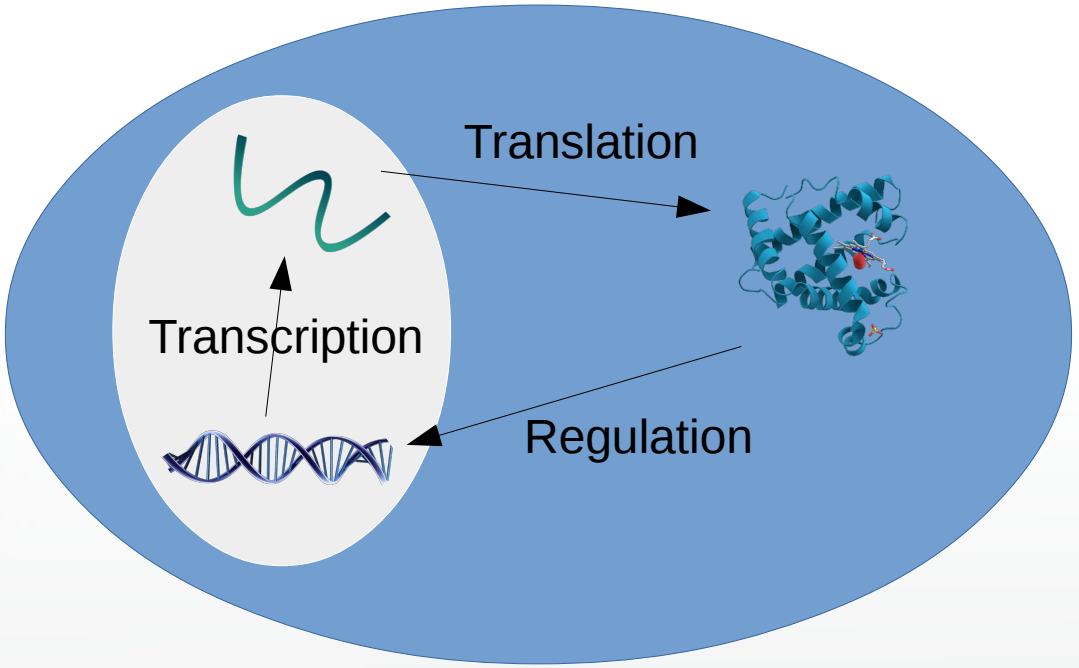
Amir Abdollahi et al. PNAS 2007;104:12890-12895

GRN Inference

- **Goal**: Reverse engineer regulatory mechanisms based on our set of information.
- Information may include
  - Gene expression data
  - DNA sequence information
  - Known protein-protein and protein-DNA interactions.
- **Common approach**: Model GRN as a graph with genes as nodes and edges as molecular interactions.

Biological Challenges

- Measurements of gene expression are at the mRNA level.
- Measurements only consist of mRNA abundance.
- Experimental data is collected as static snapshots.
- Biological variability can be difficult to induce.



Statistical Challenges

- Gene expression measurements are noisy.
- Model complexity may require the estimate of too many model parameters.
  - May be computationally intractable.
  - May be statistically undetermined.

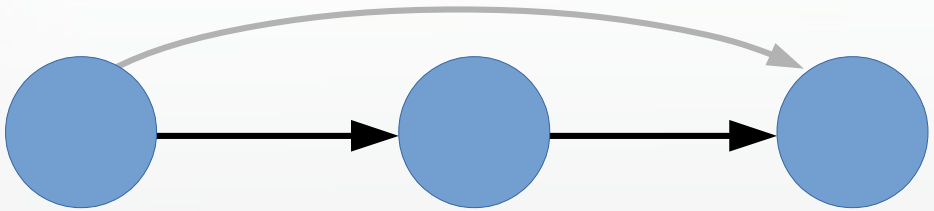
  “The curse of dimensionality”

How to address dimensionality?

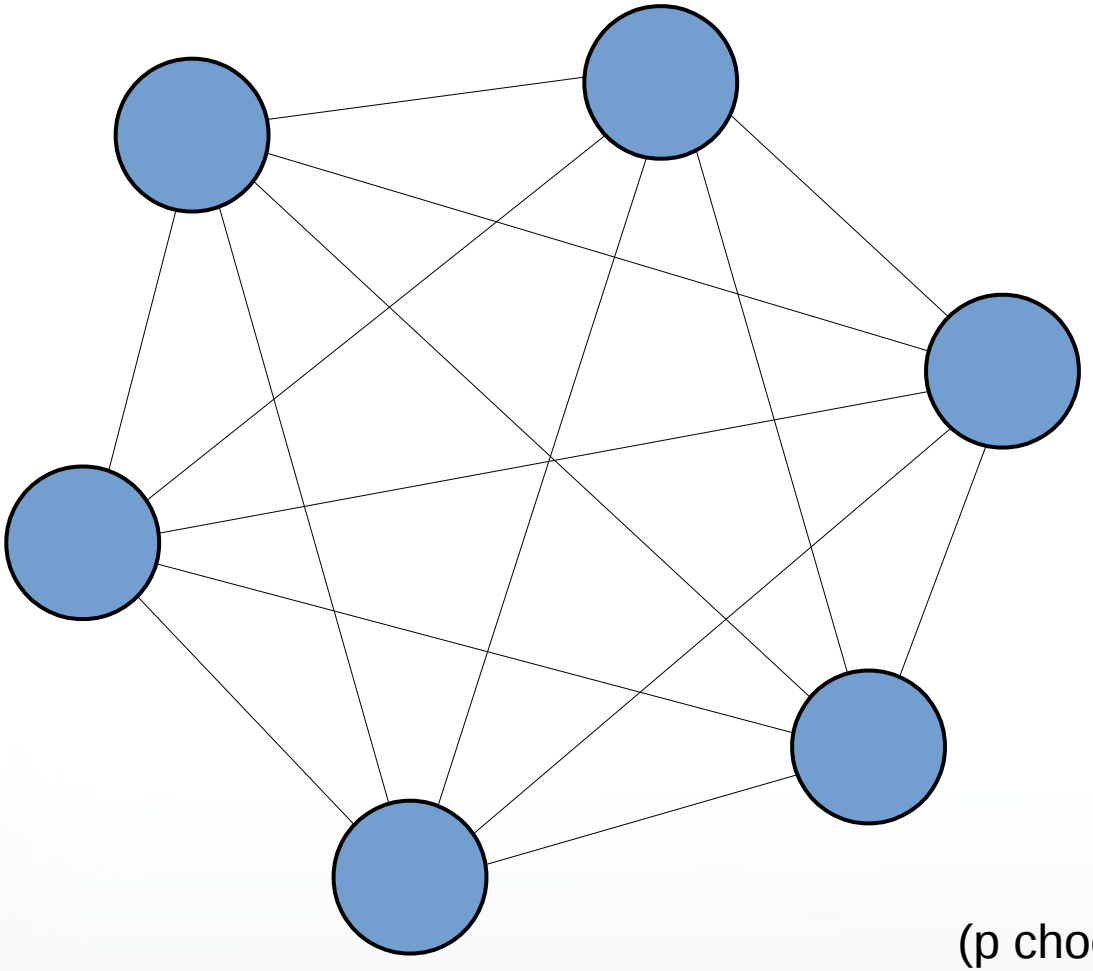
Assume sparsity.

- Define simpler model to reduce parameter space.
- Use *a priori* information to eliminate potential edges.
- Use regularized regression methods to impose sparsity.
- Use heuristic approaches based on priors.

Define model interpretation to allow edges to define “influence”.

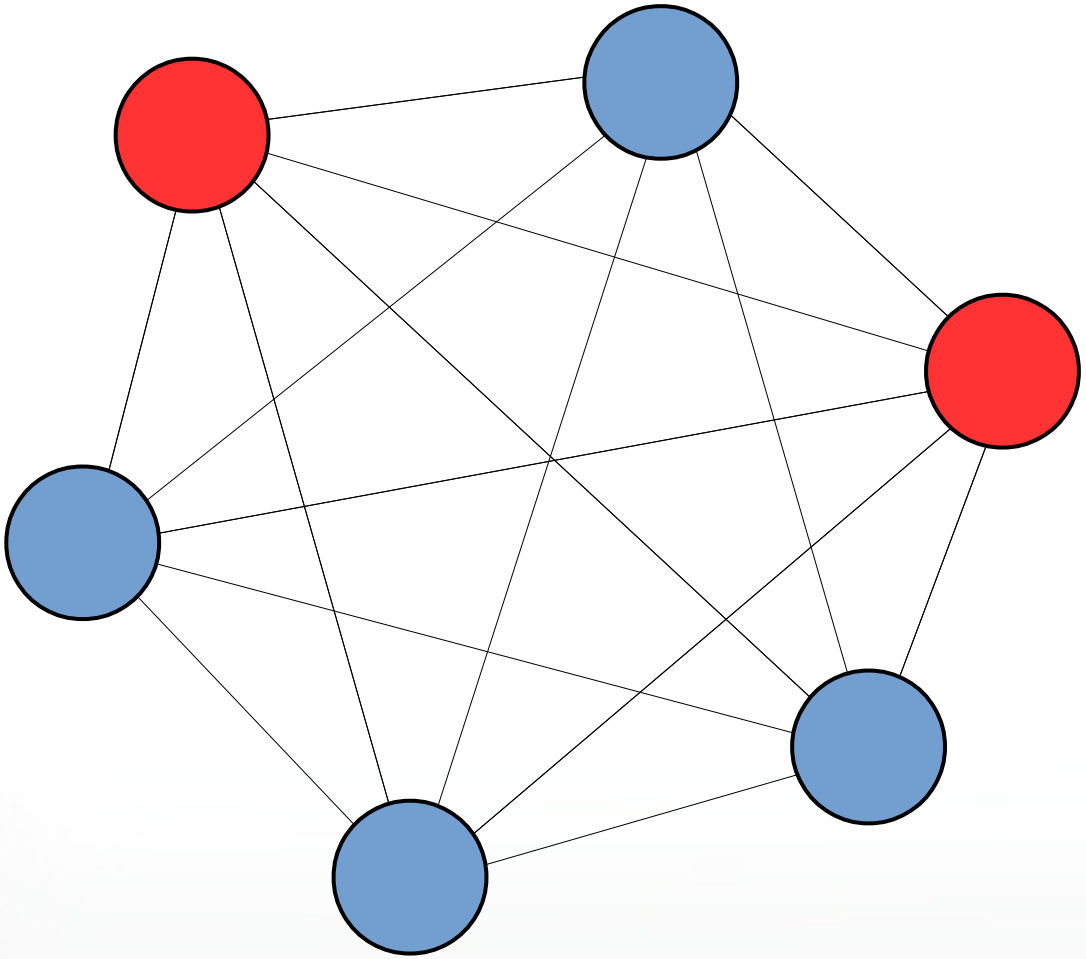


Challenges in GRN inference

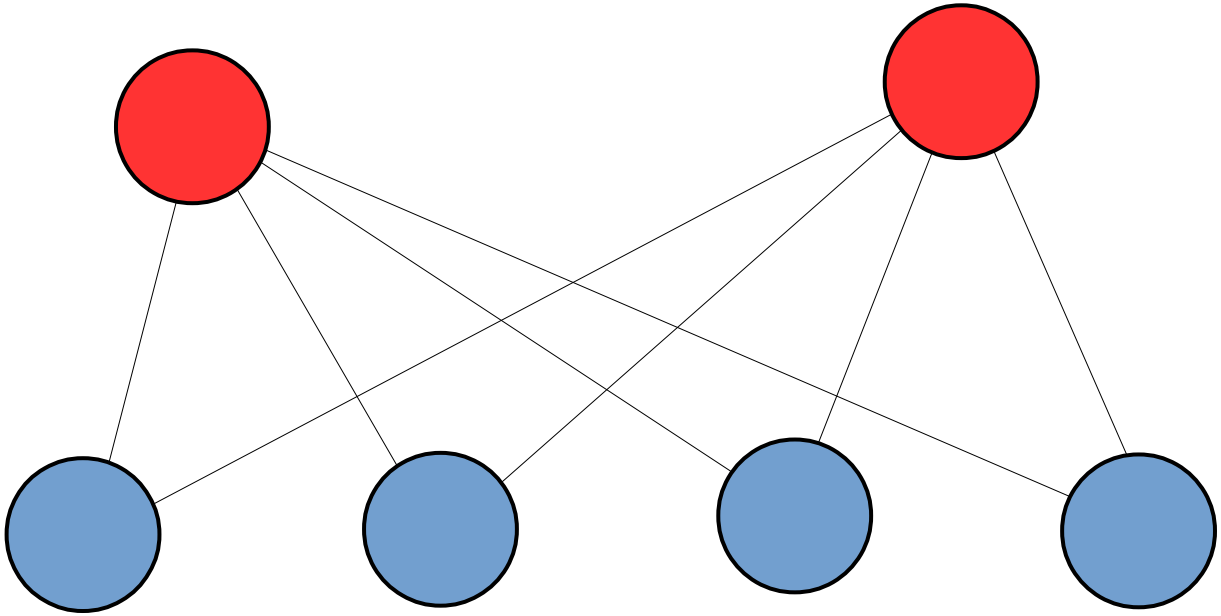


(p choose 2) edges

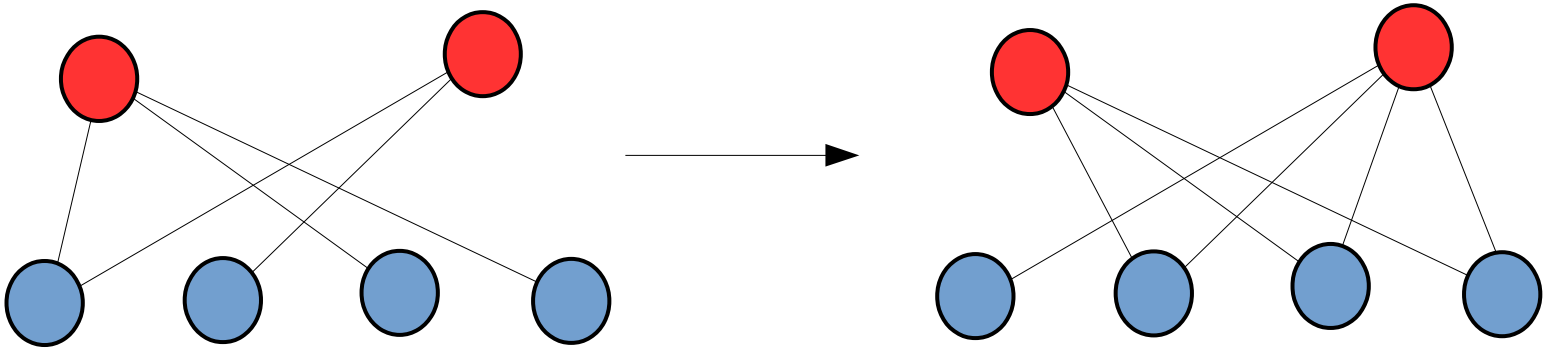
Challenges in GRN inference



Challenges in GRN inference



The network differentiation problem



Background:

- Transcription factors may behave in different ways in different contexts.
- The targeted set of genes are defined by post-translational factors not measured by gene expression.
- These changes in “involvement” may not be readily observed using standard differential gene expression analyses.

Network differentiation challenges

- Current network inference methods yield relatively poorly predictive edgeweights at the individual interaction level.
- Comparison of two networks involves the comparison of millions of noisy edges.
- Best algorithms rely heavily on static information.

Gene expression
Sequence Motifs

Observed Partition
COPD vs SMC

Random Partitions
Null x 100

BERE

Observed COPD Network

Observed SMC Network

Null NetworkA

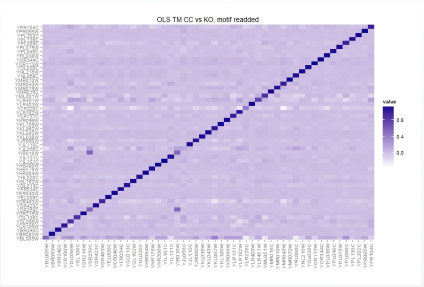
Null NetworkB

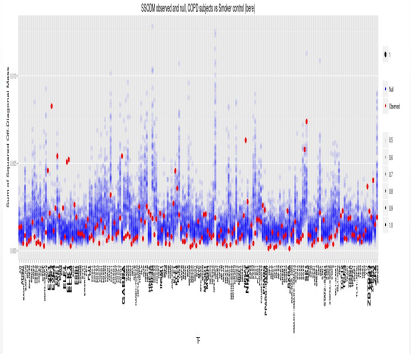
X 1000

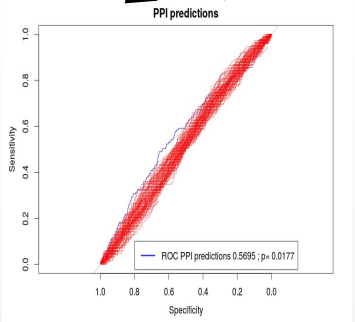
Transition Matrix

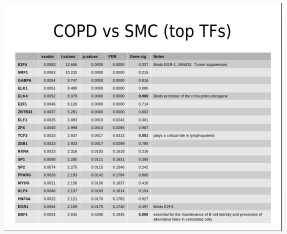
X 1000

Transition Matrix analysis









Gene expression
Sequence Motifs

Observed Partition
COPD vs SMC

Random Partitions
Null x 100

BERE

Observed COPD Network

Observed SMC Network

Null NetworkA

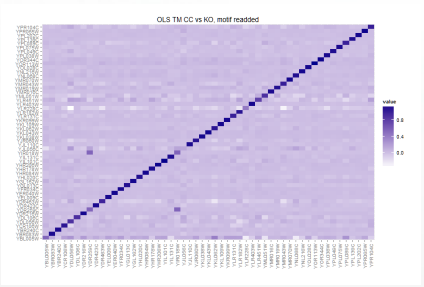
Null NetworkB

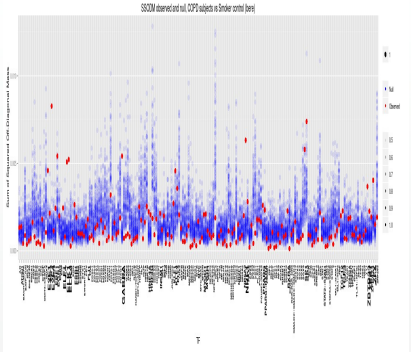
X 1000

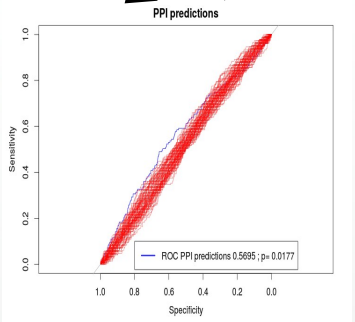
Transition Matrix

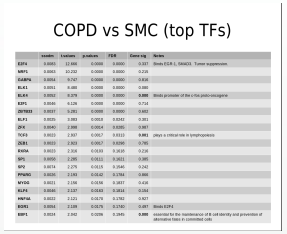
X 1000

Transition Matrix analysis

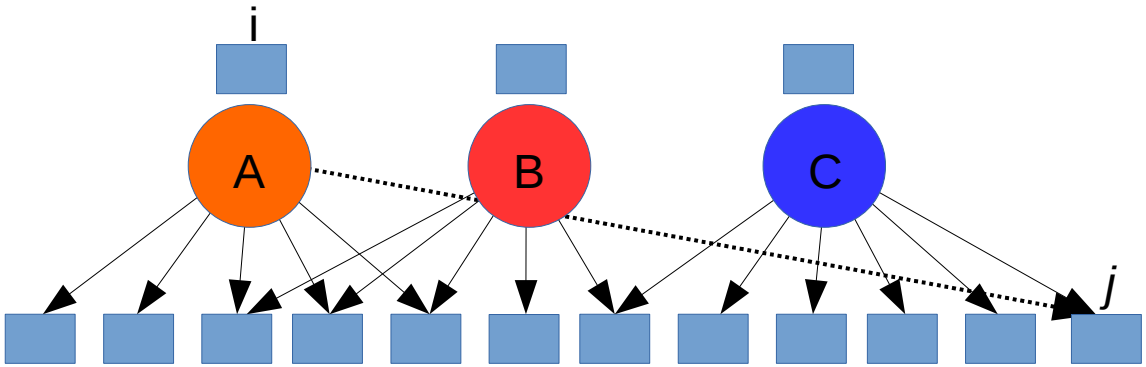


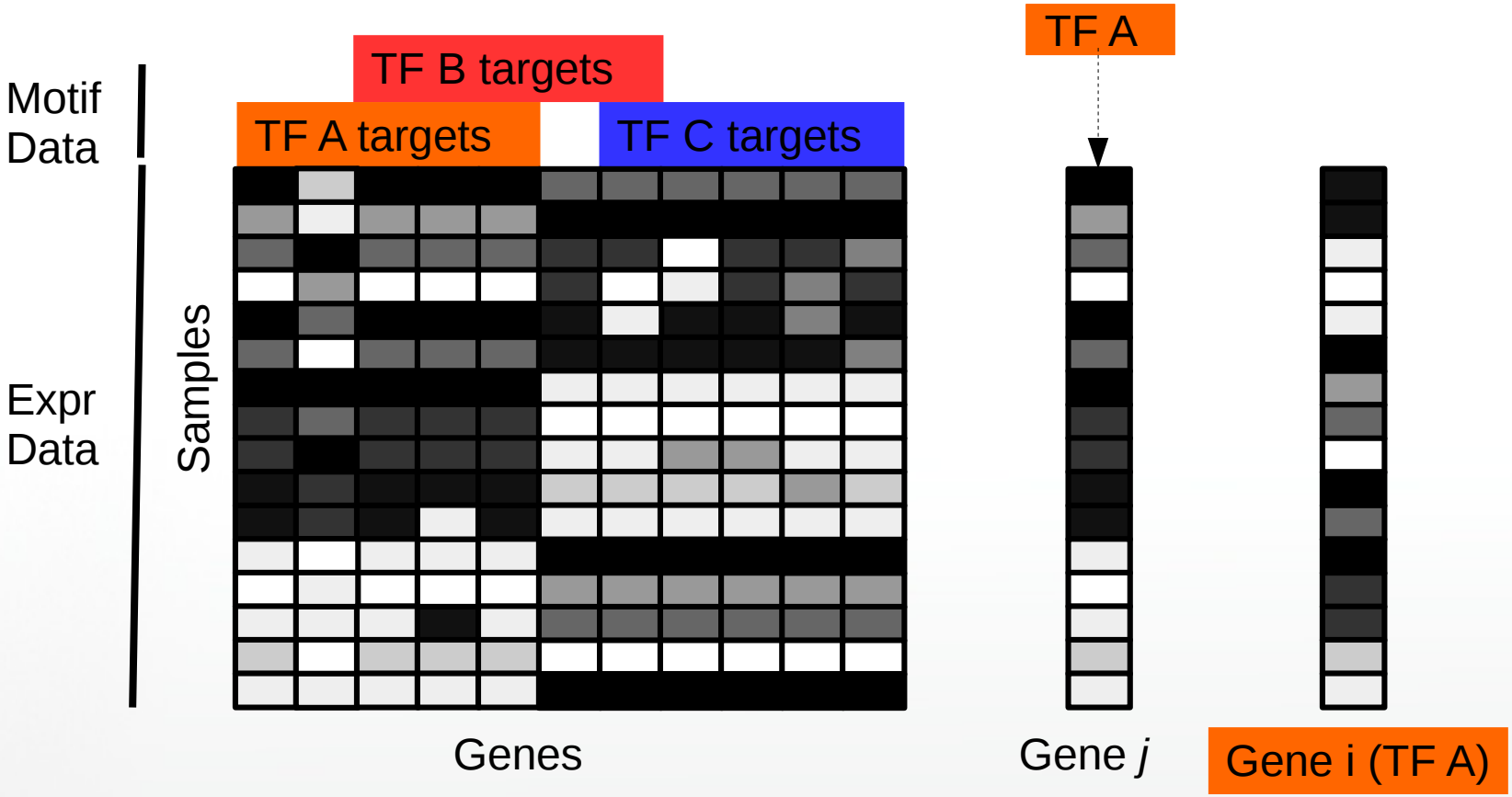




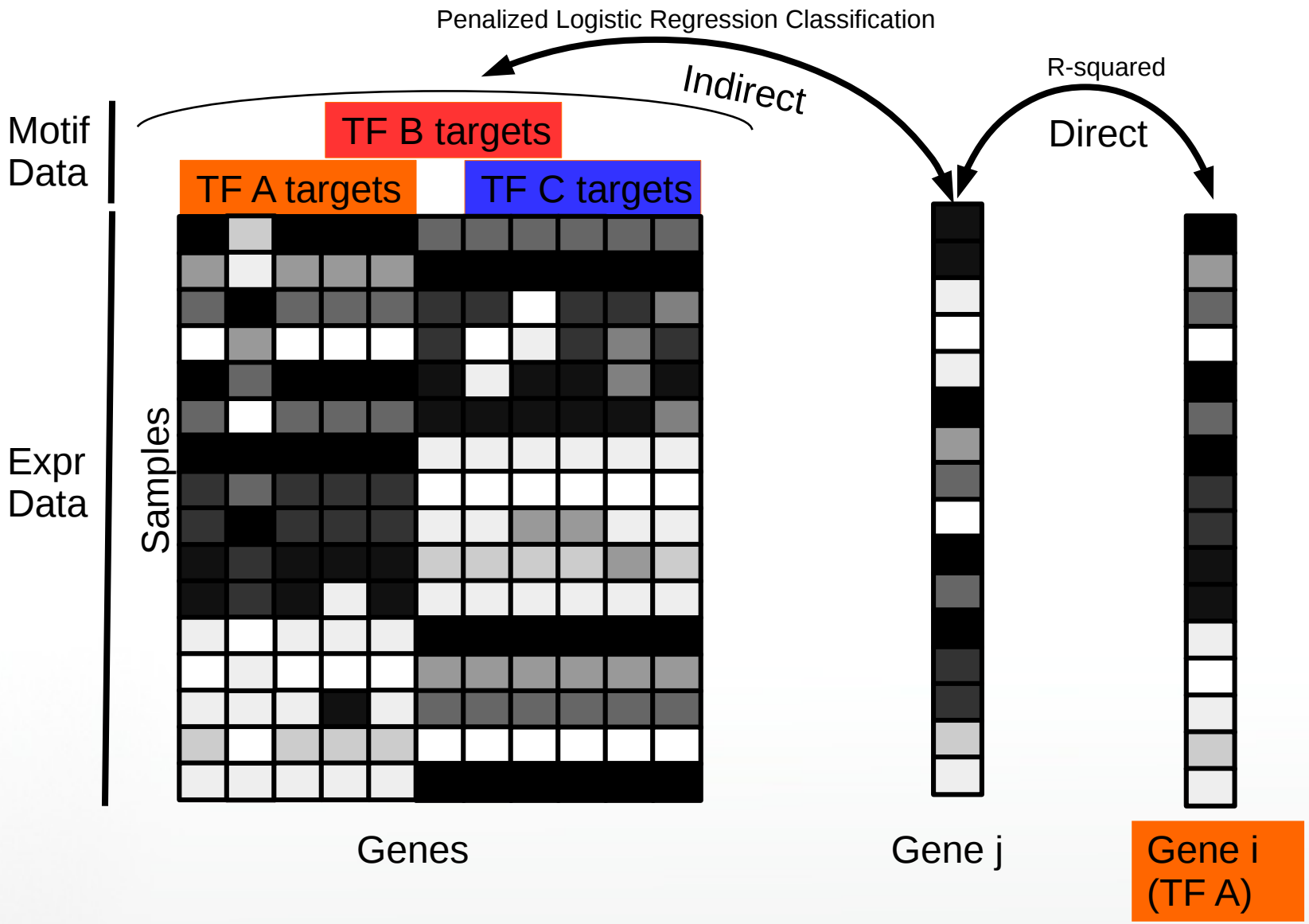


Bipartite **E**dge **R**econstruction from **E**xpression Data





BERE


Penalized Logistic Regression Classification
Indirect
R-squared
Direct
Motif Data
Expr Data
TF B targets
TF A targets
TF C targets
Samples
Genes
Gene j
Gene i (TF A)

BERE - direct

Divide evidence for regulation into 2 parts:

1.) **Direct evidence**

Measured by squared conditional correlation with expression level for transcription factor.

$$d_{i,j} = cor\left(g_i, g_j \middle| \left\{g_{k,-j} : k \neq j, k \in \mathbf{TF}\right\}\right)^2$$

$$\begin{aligned} X_i^* &= X_i - X_{TF}\left(X_{TF}' X_{TF}\right)^{-1} X_{TF}' X_i \\ X_j^* &= X_j - X_{TF}\left(X_{TF}' X_{TF}\right)^{-1} X_{TF}' X_j \\ d_{i,j} &= \frac{X_i^{*\prime} X_j^*}{\sqrt{\left(X_i^{*\prime} X_i^*\right)\left(X_j^{*\prime} X_j^*\right)}} \end{aligned}$$

This results in a limited order partial correlation network. Typically feasible to run with without regularization.

BERE – indirect

2.) **Indirect evidence**

Classification from a regularized logistic regression, with penalty model matrix as inverse TF A expression levels.
Regularization here is across samples. We are not attempting to do feature selection and are using an L_2 penalty.

The goal is to find the maximum of the penalized log likelihood function:

$$\sum_{i=1}^{n} log\left[exp\left(\beta'\mathbf{x_i}\right)^{Y_i}\left\{1-exp\left(\beta'\mathbf{x_i}\right)\right\}^{1-Y_i}\right]-\lambda\beta'\mathbf{Q}\beta$$

Q is diagonal with values equal to the inverse transcription factor expression.

BERE – consensus

How to combine predicted edgeweights?

1.) Rank indirect and direct contributions by TF.
2.) Combine with a weighted sum.

$$\text{edgeweight}_i = (1-\alpha)\left[rank\left(d_i\right)\right] + \alpha\left[rank\left(e_i\right)\right], i \in \{1, \ldots, p\}$$

Greater organism complexity → greater indirect weight.

Optimal indirect weights	
DREAM5 data	**alpha**
In Silico	.33
E. coli	.61
Saccharomyces cerevisiae	.88

BERE - summary

Method overview:

1.) Model gene regulatory network as a bipartite graph between *m* transcription factors and *p* genes.

2.) Consider the direct evidence of regulation.

The squared conditional coexpression of gene *i* and gene *j*, where gene *i* is a transcription factor.

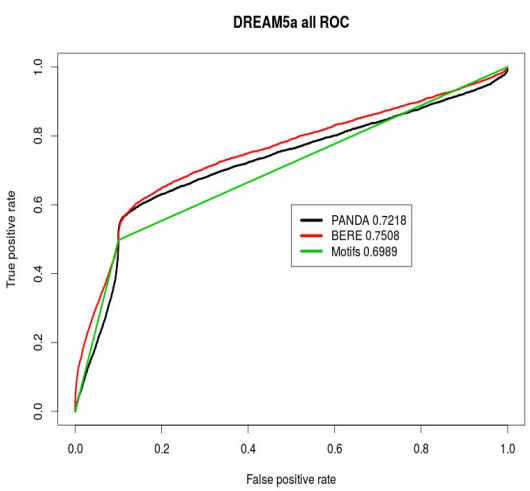
3.) Consider the indirect evidence of regulation.

Use presence of sequence binding motif for TF *i* near gene *j* as a classification label and fit a penalized logistic regression model across all genes.

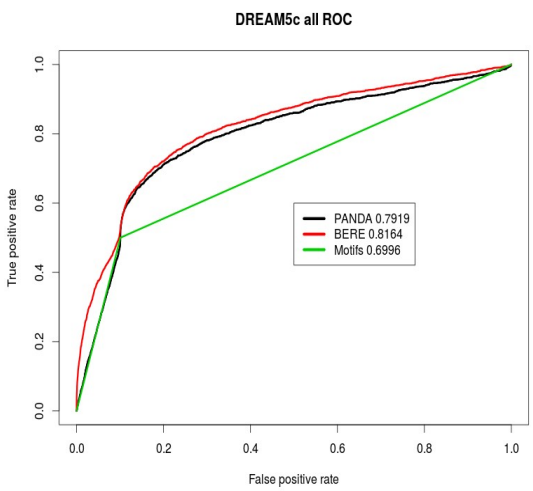
4.) Combine indirect and direct evidence into a score for network edgeweights.

BERE

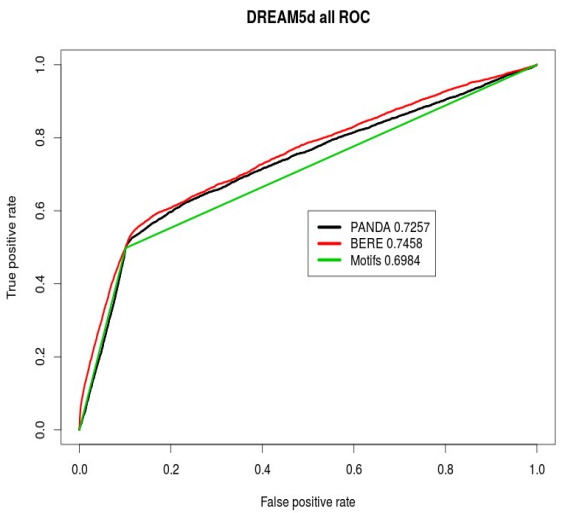
In Silico



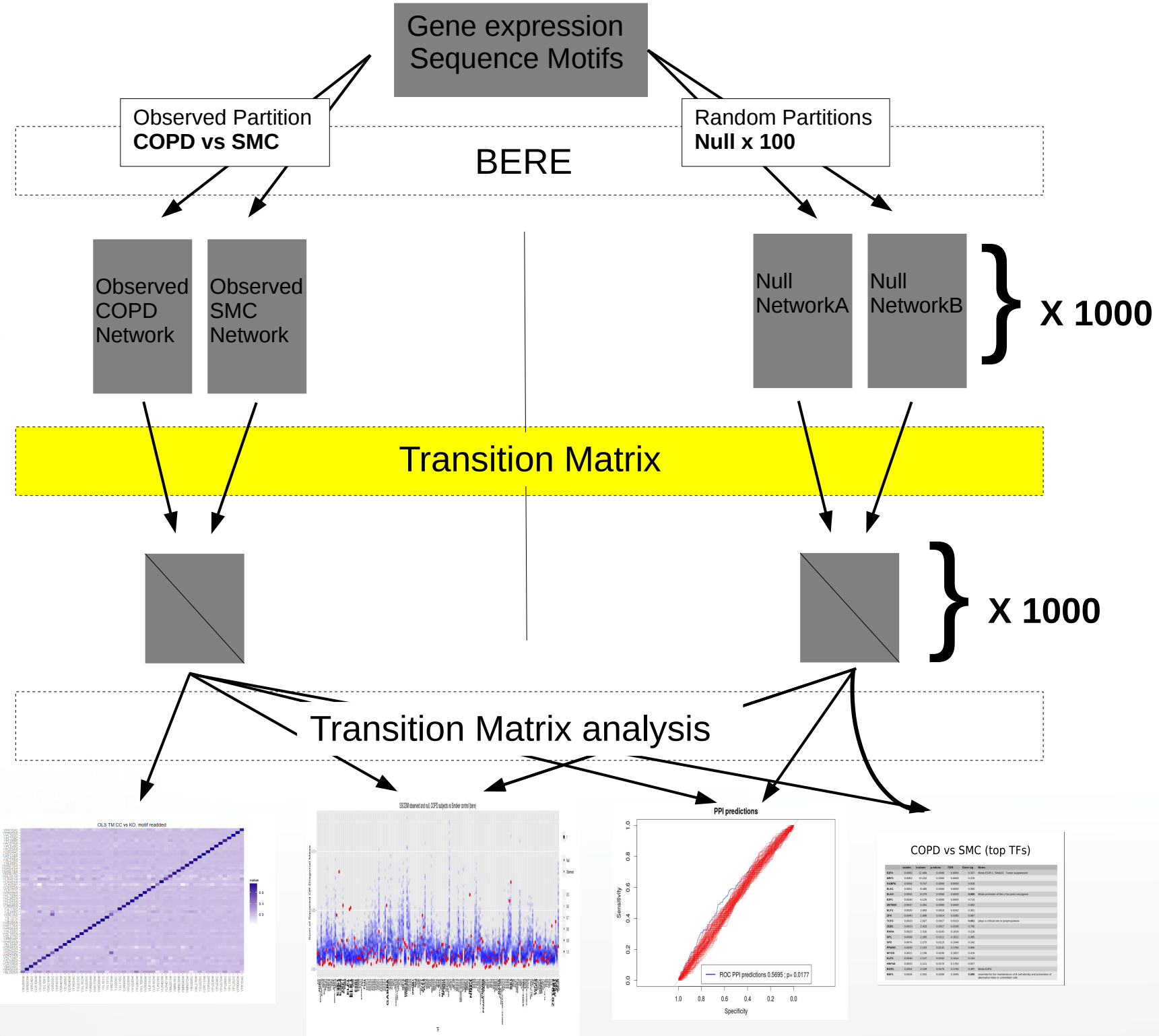
E. coli



Yeast

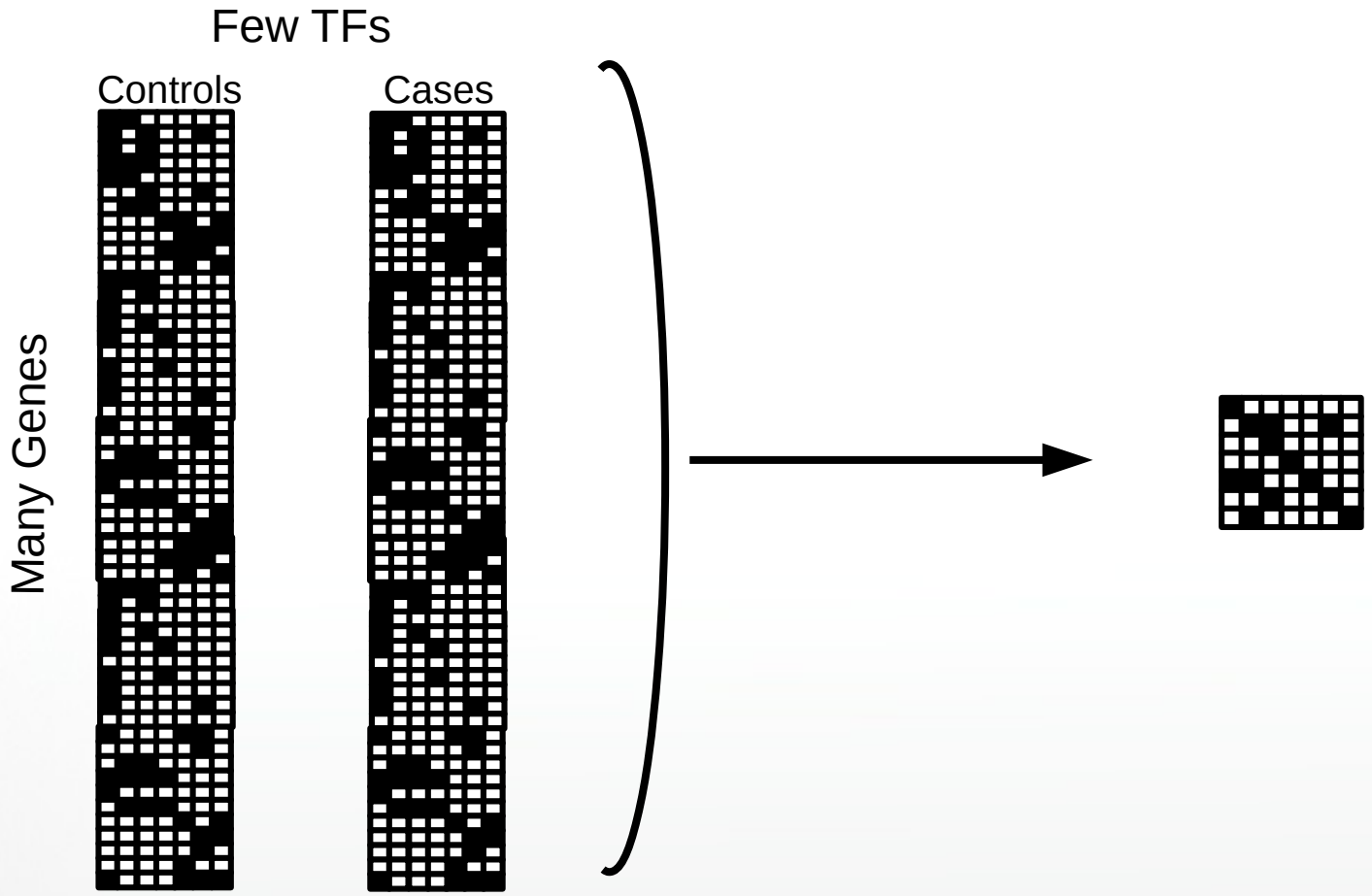


Running R package: 8GB RAM, 2.40Ghz	Time
2555 genes, 53 TF, 106 samples	11s
17342 genes, 189 TF, 226 samples	12m, 20s


Gene expression
Sequence Motifs
Observed Partition
COPD vs SMC
Random Partitions
Null x 100
BERE
Observed
COPD
Network
Observed
SMC
Network
Null
NetworkA
Null
NetworkB
X 1000
Transition Matrix
X 1000
Transition Matrix analysis
OLS TM CC vs KO, motif readded
PPI predictions
ROC PPI predictions 0.5695 ; p= 0.0177
Sensitivity
Specificity
COPD vs SMC (top TFs)

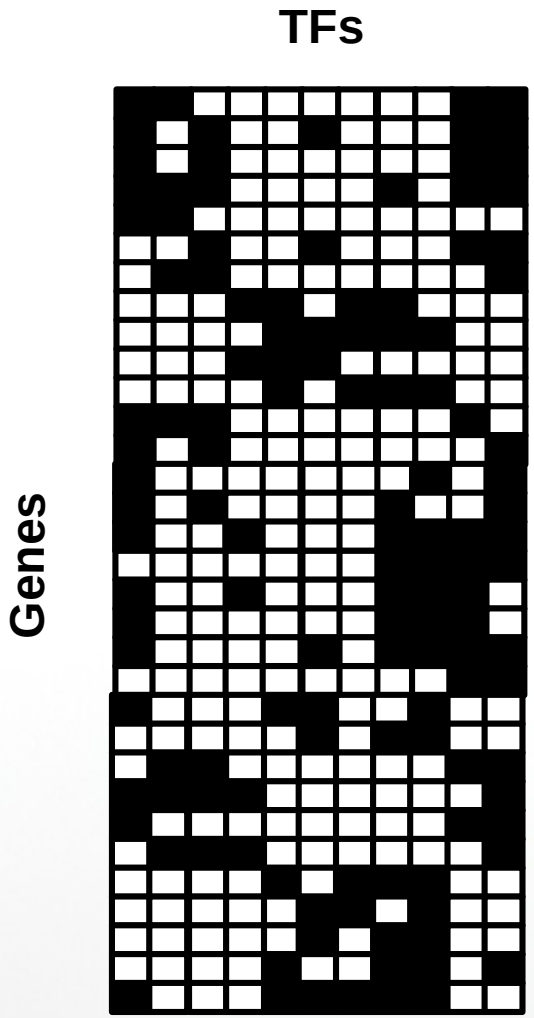
Transition Matrix Approach

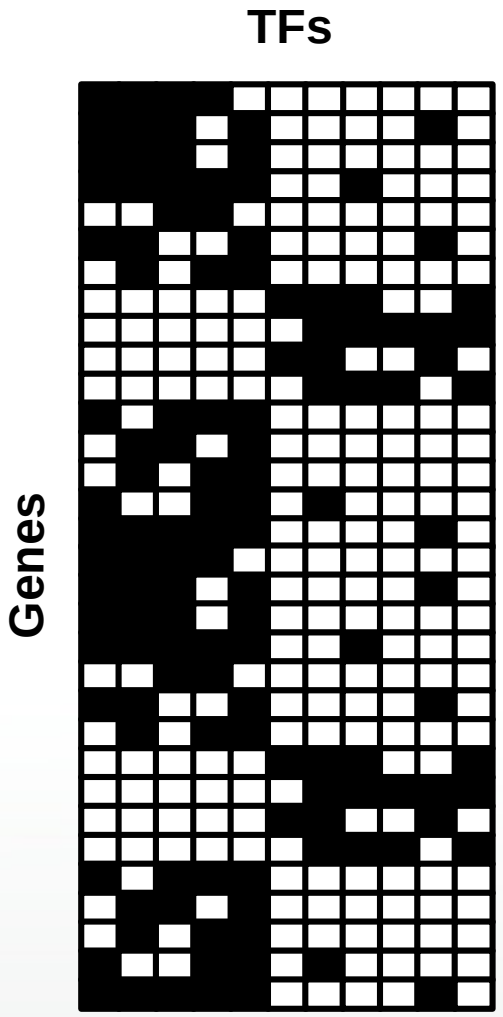
We can view the problem as a dimension reduction problem.



Transition Matrix Approach

Consider two adjacency matrices...

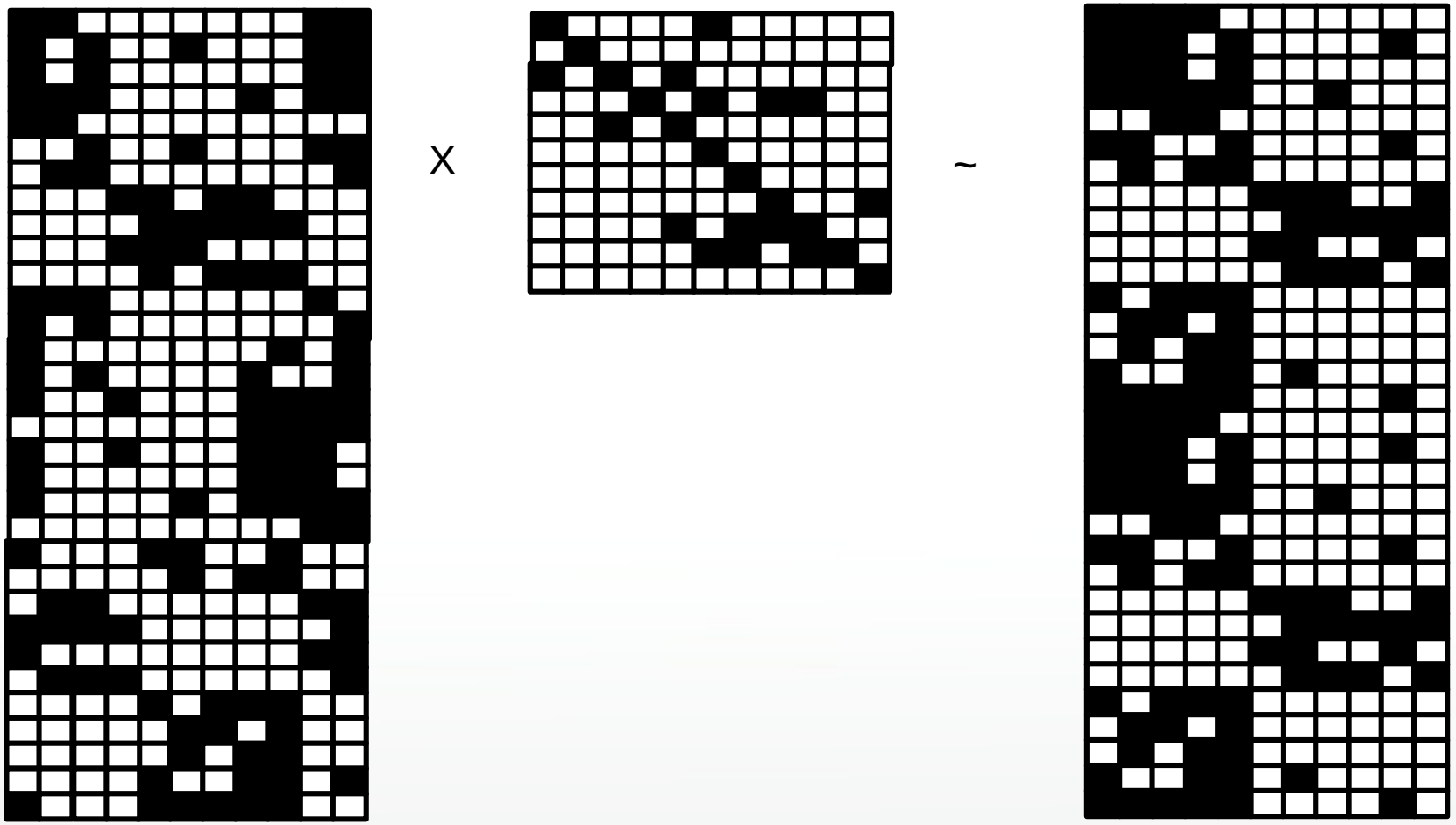




Transition Matrix Approach

Consider two adjacency matrices...

Smoker Control **Tau** **COPD**



The Transition Matrix (Tau) problem

Consider two adjacency matrices, **A** and **B** representing the adjacency matrices for two GRNs estimated from a case-control study. Each matrix has dimensions (*p* x *m*) representing the set of *p* genes targeted by *m* TFs. We seek a matrix, **T**, such that

B = AT + E

$$\begin{bmatrix} \mathbf{b}_{i1} \\ \mathbf{b}_{i2} \\ \vdots \\ \mathbf{b}_{ip} \end{bmatrix} = \tau_{1,i} \begin{bmatrix} \mathbf{a}_{11} \\ \mathbf{a}_{21} \\ \vdots \\ \mathbf{a}_{p1} \end{bmatrix} + \tau_{2,i} \begin{bmatrix} \mathbf{a}_{12} \\ \mathbf{a}_{22} \\ \vdots \\ \mathbf{a}_{p2} \end{bmatrix} + \cdots + \tau_{p,i} \begin{bmatrix} \mathbf{a}_{1p} \\ \mathbf{a}_{2p} \\ \vdots \\ \mathbf{a}_{pp} \end{bmatrix} + \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{bmatrix}$$

The Transition Matrix (Tau)

- Each column in the TM can be thought of as being the best linear combination of columns in the control AM that “create” the columns in the COPD.
- We want to focus on changes in targeting behavior of a TF in terms of biologically recognized alternative targets.
- In reconstructing case-targets for a TF, first account for targets in control for that TF.
- Assume target-transfer is sparse.

The Transition Matrix (Tau)

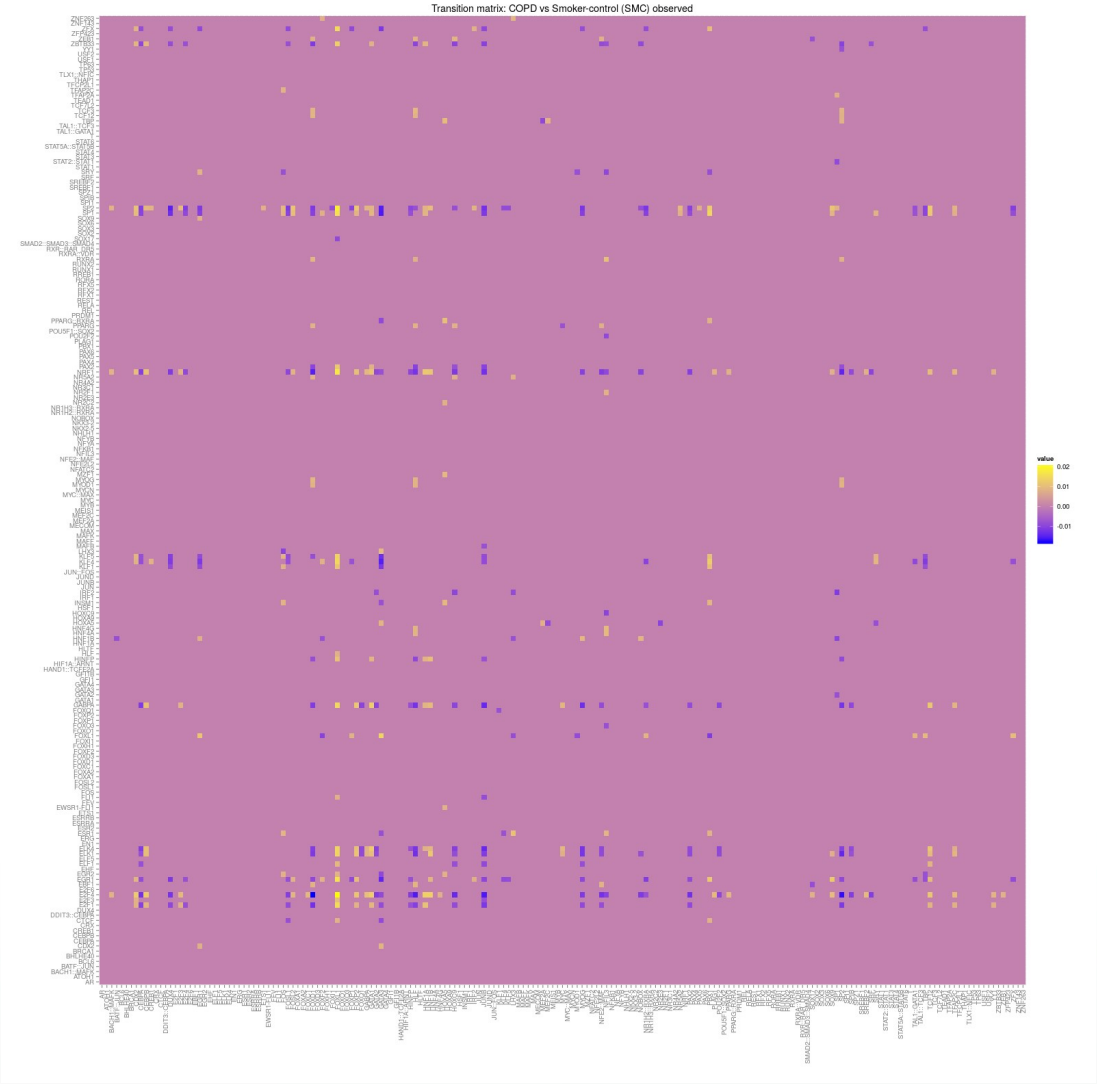
- We can satisfy these properties with an L_1 regularization, aka LASSO.
  - For a column, *k*, we perform the following error minimization.

$$\sum_{i=1}^{p}\left(\mathbf{B}_{i,k}-\sum_{j=1}^{m}A_{i,j}\mathbf{T}_{j,k}\right)^2+\lambda\beta'\mathbf{Q}\beta$$

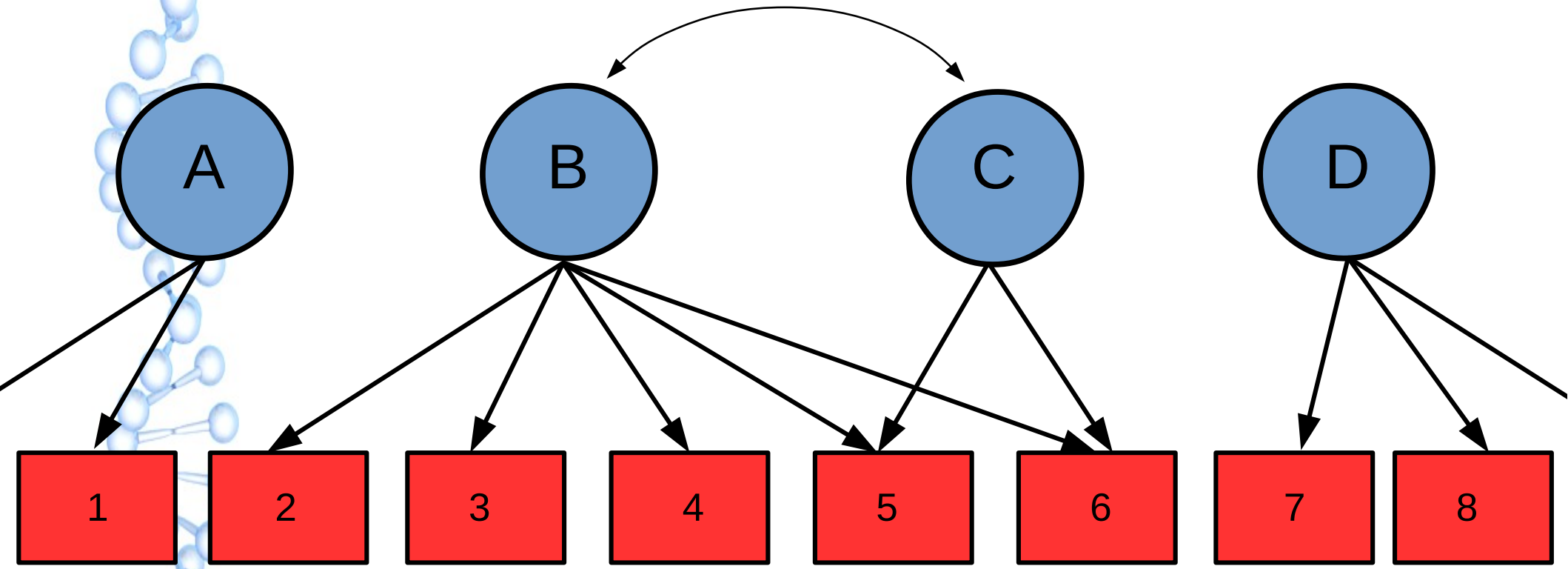
$$\mathbf{Q}_{i,j}=\begin{cases}1 & for\, i=j\neq k\\ 0 & elsewhere\end{cases}$$

- Penalty model matrix is a diagonal matrix with value 0 for it's own TF and 1 for all others.

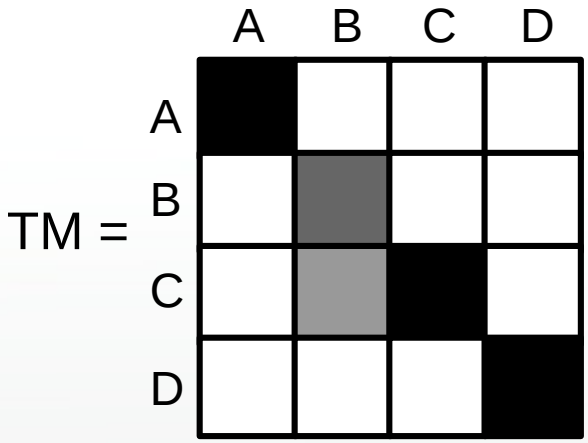
The Transition Matrix (Tau)



An Example

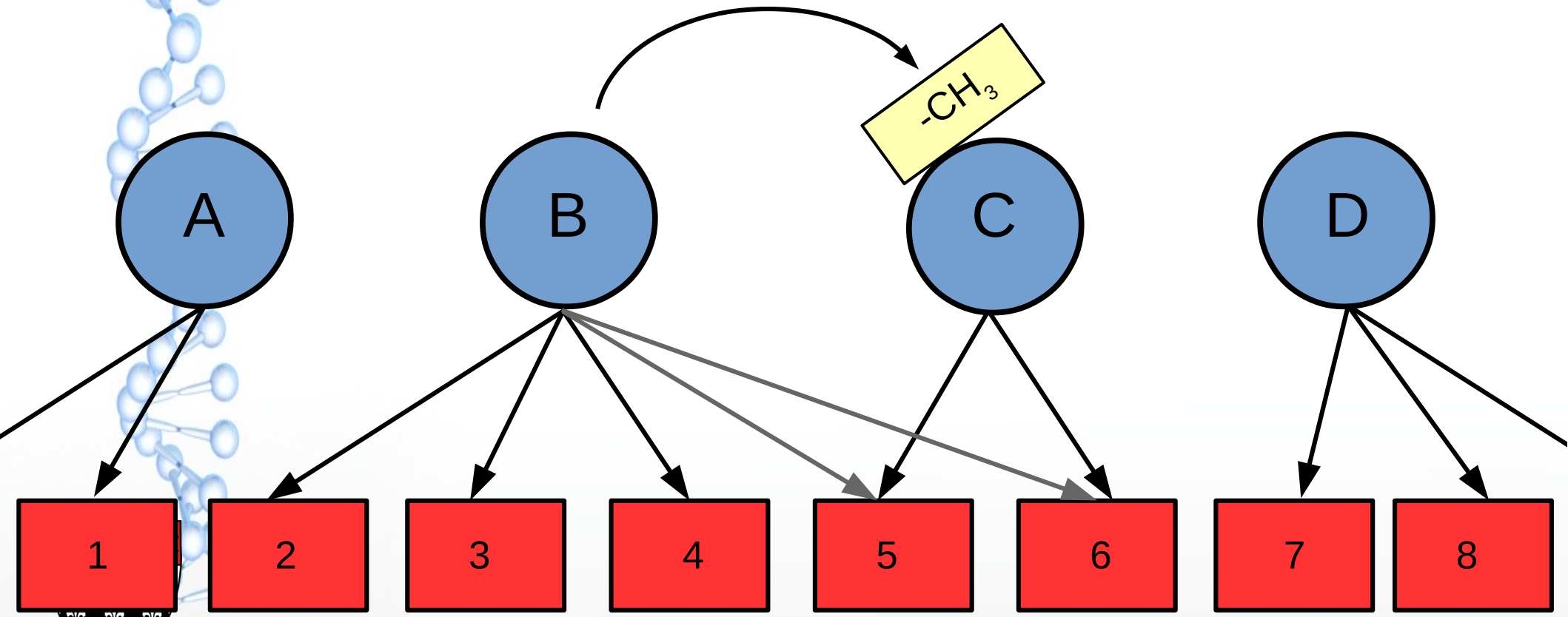


Cases Network



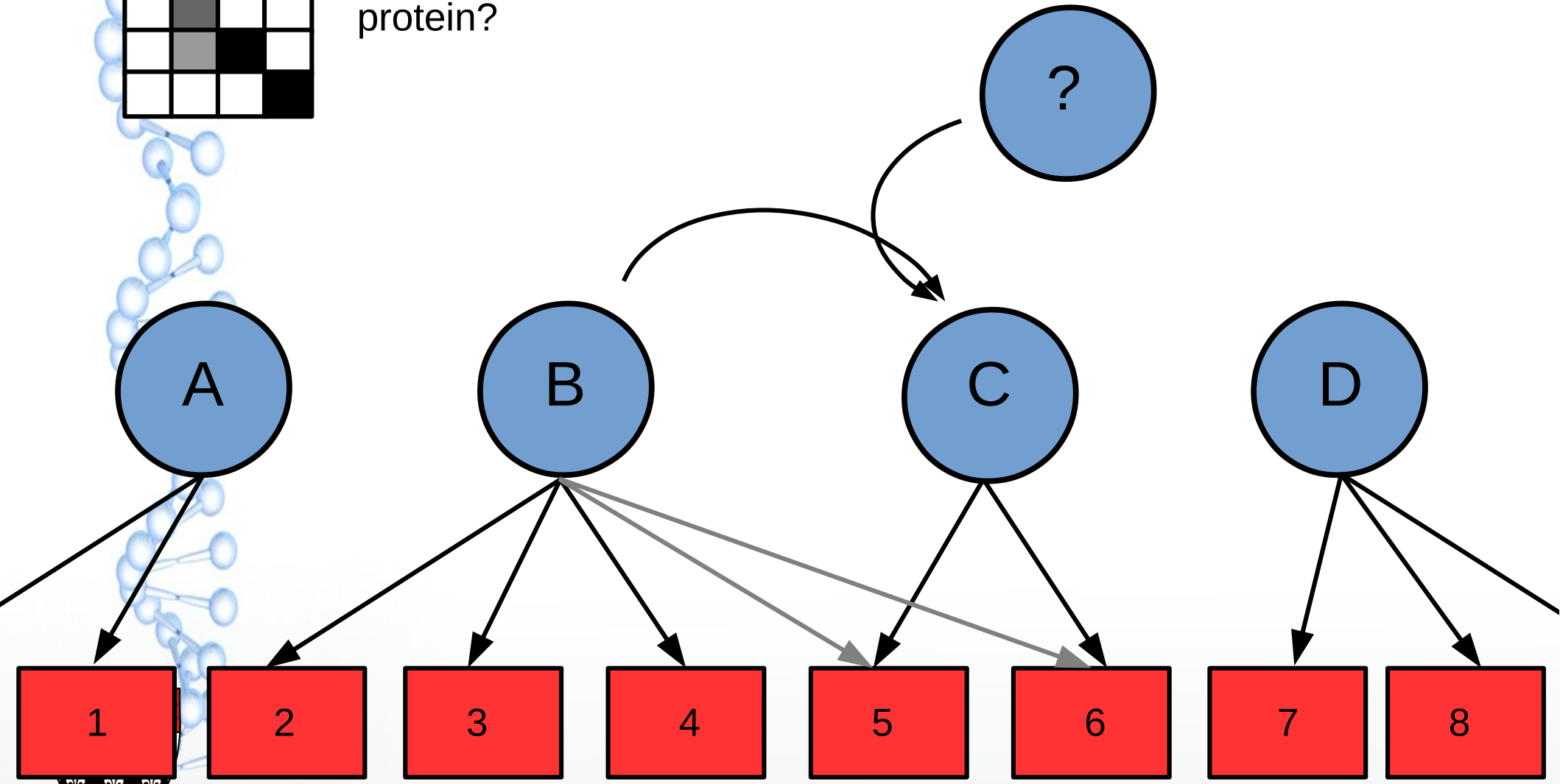
Biological Mechanism

Suggested mechanism #1:
Differential methylation of the gene for TF C?



Biological Mechanism

Suggested mechanism #2:
Protein complex of B-C-? with unknown 3rd protein?



Gene expression
Sequence Motifs

Observed Partition
COPD vs SMC

Random Partitions
Null x 100

BERE

Observed COPD Network

Observed SMC Network

Null NetworkA

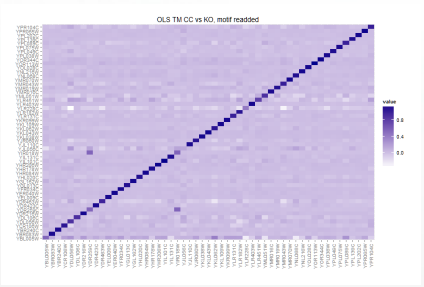
Null NetworkB

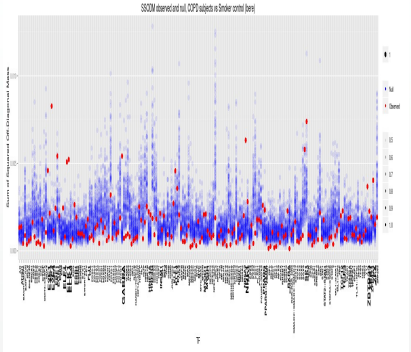
X 1000

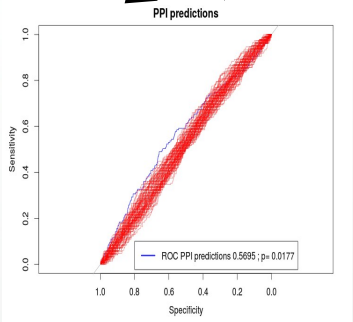
Transition Matrix

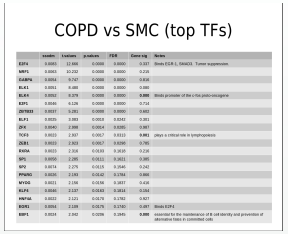
X 1000

Transition Matrix analysis









Evaluating the Transition Matrix

We want to quantify the change in targeting which has a biological basis. The overall TF involvement can be simply measured as

$$s_j = \frac{\sum_{i=1}^{m} I\left(i \neq j\right) \tau_{i,j}^2}{\sum_{i=1}^{m} \tau_{i,j}^2}$$

s_j (differential TF involvement) is the proportion of variability in targeting for TF_j in transitioning from controls to cases which is explained by alternative TF targets.

Null distribution depends on motif structure and can be estimated via resampling on a per-TF basis

Permutation inference on differential TFI statistic

1. Gene expression samples are randomly assigned to case and control forming the null-case and null-control with group sizes preserved.

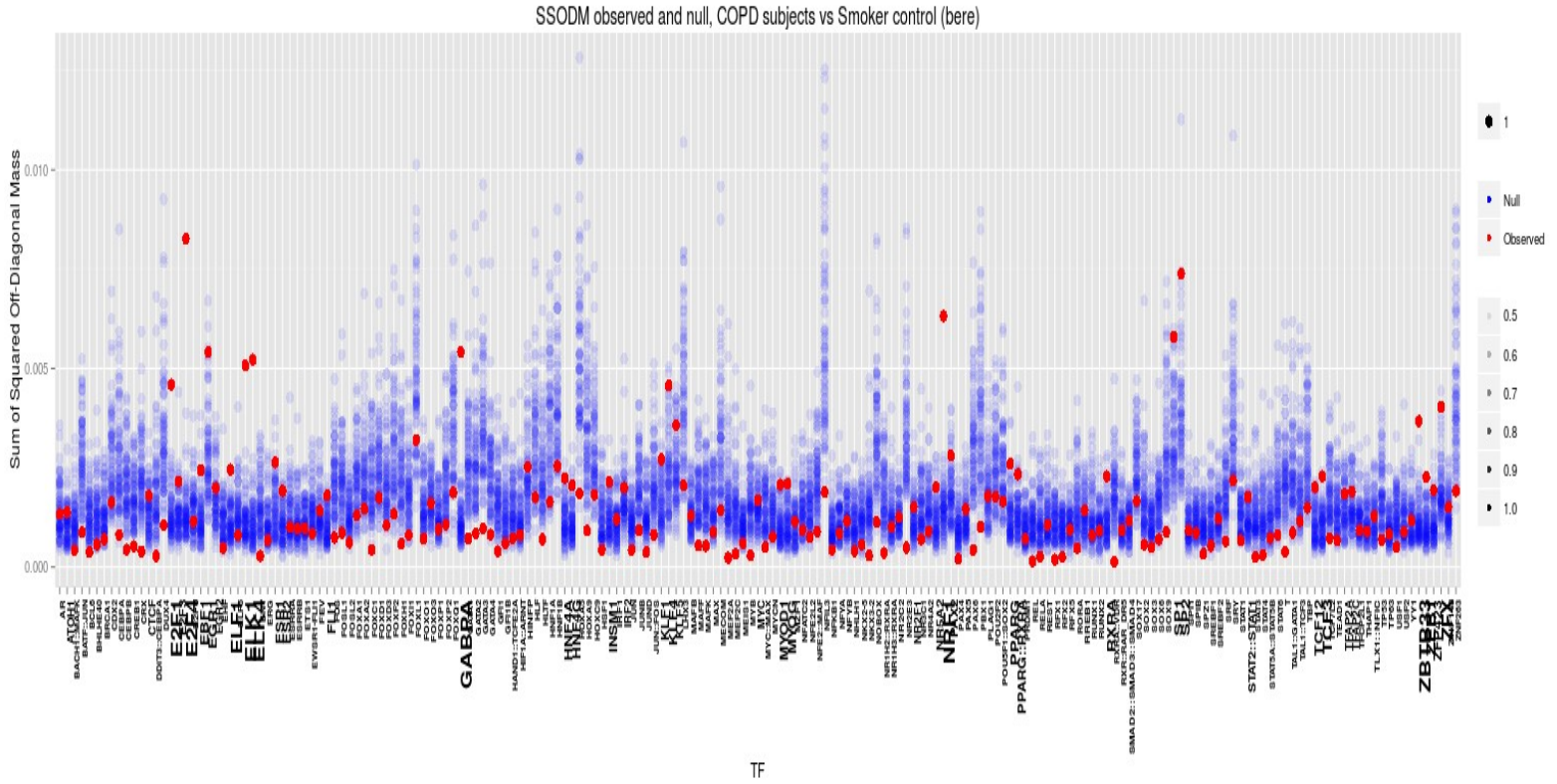
2. GRNs are reconstructed for the null-case and null-control with the same prior regulatory structure.

3. The transition matrix algorithm is applied for the two null networks.

4. The differential TFI is calculated for each TF.

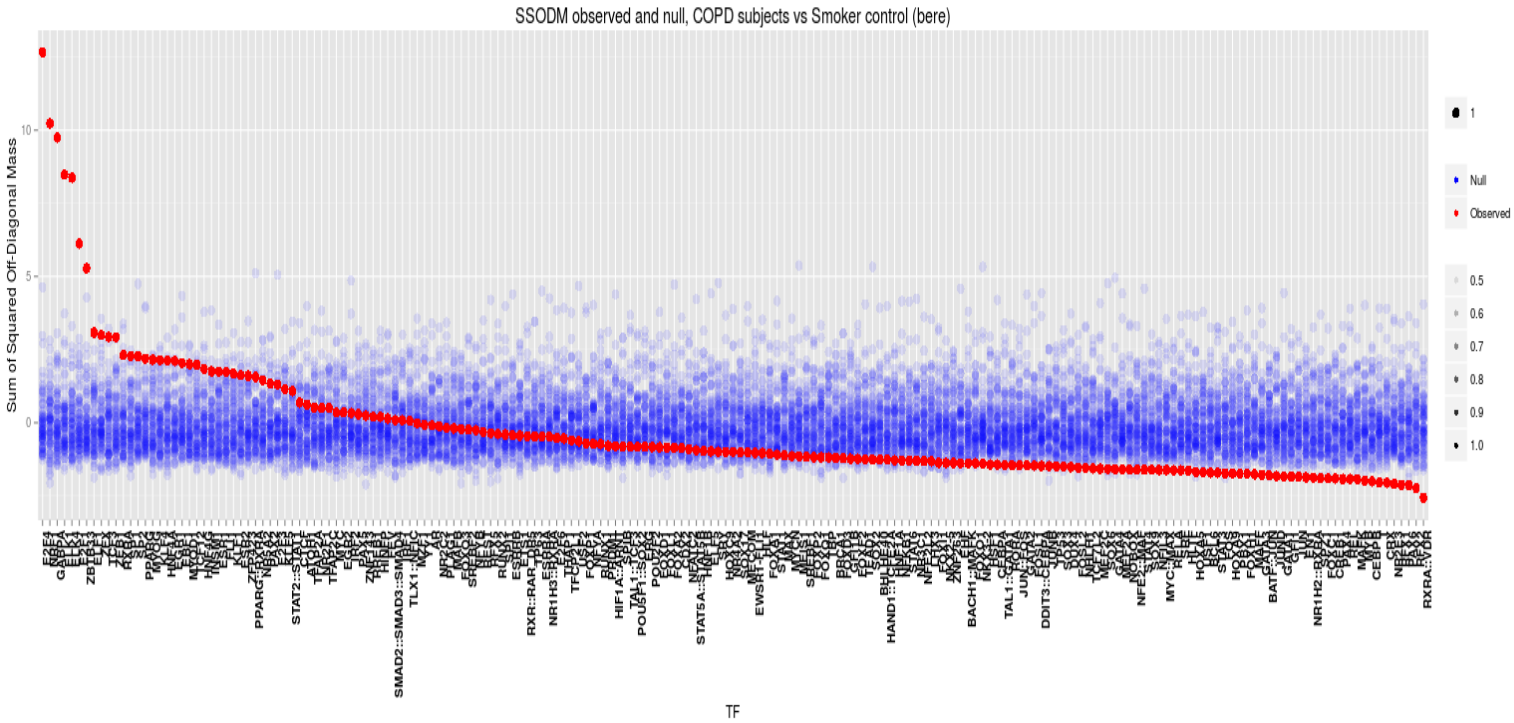
5. Repeat 1-4 1000 times.

Application to a case-control COPD study



Differential transcription factor involvement distribution under the null (blue), with the observed differential TFI (red).

Application to a case-control COPD study



Observed differential TFI (red) standardized by the estimated distribution under the null.

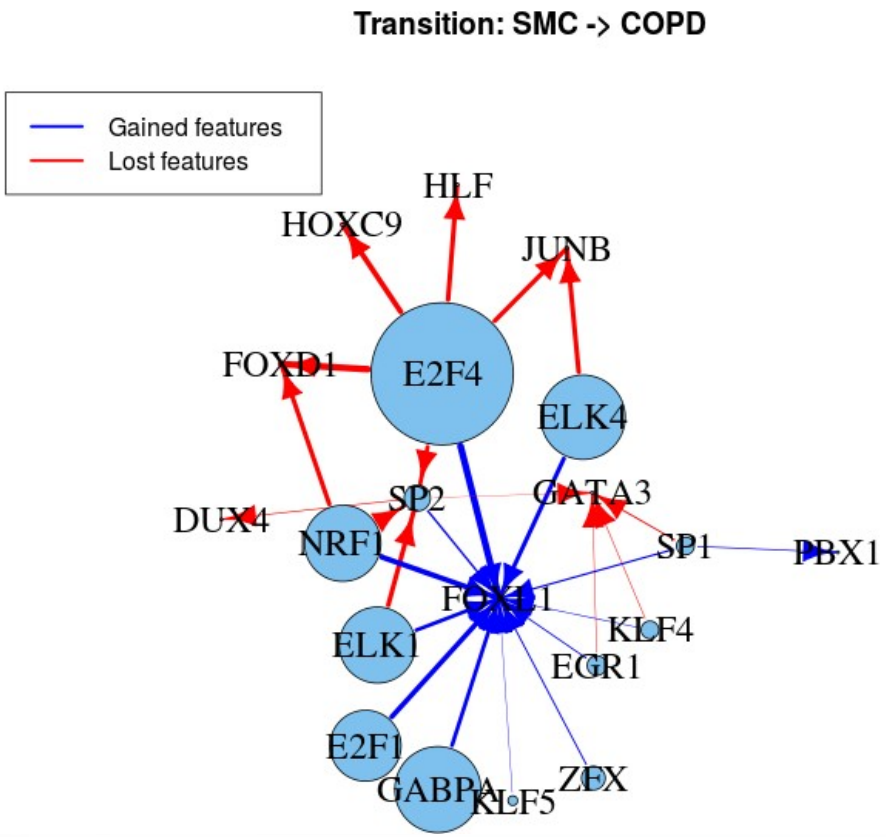
Application to a case-control COPD study

	t-statistic	p-values	FDR	Sig (LIMMA)	Notes
E2F4	12.666	0.0000	0.0000	0.337	Binds EGR-1, SMAD3. Tumor suppression.
NRF1	10.232	0.0000	0.0000	0.215	Acts on nuclear genes encoding respiratory subunits and components of the mitochondrial transcription and replication machinery.
GABPA	9.747	0.0000	0.0000	0.816	Related to NRF1, involved in activation of cytochrome oxidase expression and nuclear control of mitochondrial function
ELK1	8.480	0.0000	0.0000	0.080	Binds to the the serum response factor
ELK4	8.379	0.0000	0.0000	0.000	Binds promoter of the c-fos proto-oncogene
E2F1	6.126	0.0000	0.0000	0.714	E2F family...
ZBTB33	5.281	0.0000	0.0000	0.602	shown to interact with HDAC3, Nuclear receptor co-repressor 1
ELF1	3.083	0.0010	0.0242	0.301	primarily expressed in lymphoid cells
ZFX	2.998	0.0014	0.0285	0.987	gene on the X chromosome

Application to a case-control COPD study

Changing TF	Trainer TF	Gain/Loss	p-value	FDR
GABPA	SPIB	Loss	1.07E-009	3.82E-005
E2F4	PAX2	Loss	1.22E-008	2.17E-004
ELK4	SPIB	Loss	1.83E-008	2.18E-004
E2F4	SPIB	Loss	3.53E-008	3.15E-004
E2F4	ZEB1	Gain	4.70E-008	3.36E-004
E2F4	YY1	Gain	6.76E-008	4.02E-004
E2F4	SREBF2	Gain	1.46E-007	7.46E-004
NRF1	SPIB	Loss	3.64E-007	1.63E-003
E2F4	FOXL1	Gain	4.10E-007	1.63E-003
E2F1	YY1	Gain	4.23E-007	1.51E-003
E2F4	FOXD1	Loss	5.07E-007	1.65E-003
NRF1	BACH1::MAFK	Gain	5.39E-007	1.61E-003
E2F4	BACH1::MAFK	Gain	6.25E-007	1.72E-003
E2F4	PPARG	Gain	8.24E-007	2.10E-003
NRF1	YY1	Gain	1.26E-006	3.00E-003
NRF1	PPARG	Gain	1.46E-006	3.27E-003
E2F4	GABPA	Gain	1.62E-006	3.40E-003
ELK4	MYOG	Loss	2.11E-006	4.19E-003
GABPA	ZEB1	Gain	2.24E-006	4.22E-003
GABPA	MYOG	Loss	3.27E-006	5.83E-003

Application to a case-control COPD study



Limitations
(a non-comprehensive list)

- Lack of ability to make causal inference.
- Lack of attempt to identify specific TF-gene regulatory changes.
- Limited validation metrics available.
- Power to detect differential TFI depends on regulatory prior count and structure.

Future Work

Extend transition matrix by developing methods to adjust for confounding.

Goal:

Remove the effects of known covariate confounders from the Transition matrix.

Anticipated challenges:

Common confounder adjustments, such as ComBat, are not sufficient here because they adjust at individual gene level. It is the differential patterns that need addressing as opposed to relative gene expression levels

Approach:

Our method will involve identification the transition matrix conditional on a set of measured covariates.

Approach will explore a number of methods including an method for identifying regulatory networks of the confounders, predicting regulatory networks of case-control and identifying the residual regulation above what was predicted with the confounders.

Acknowledgements

Oral Qual Committee:
John Quackenbush
Kimbie Glass
JP Onnela

COPD group