



# **Methods in Case-Control Gene Regulatory Networks**

Oral Qualifying Exam

May 18, 2015

Dan Schlauch



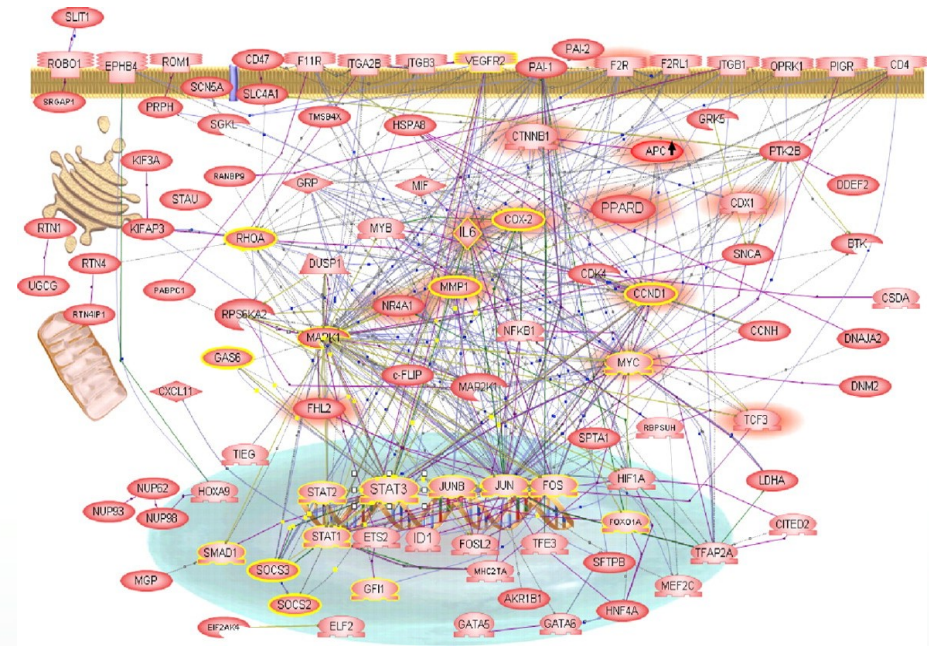
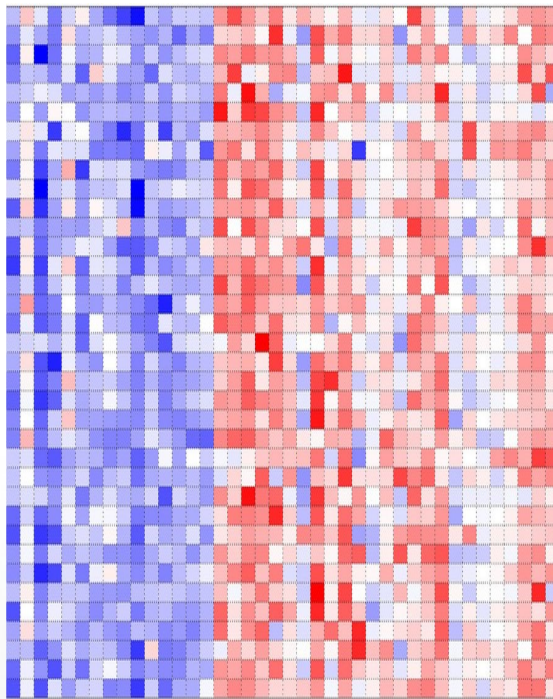
# Outline

- 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?



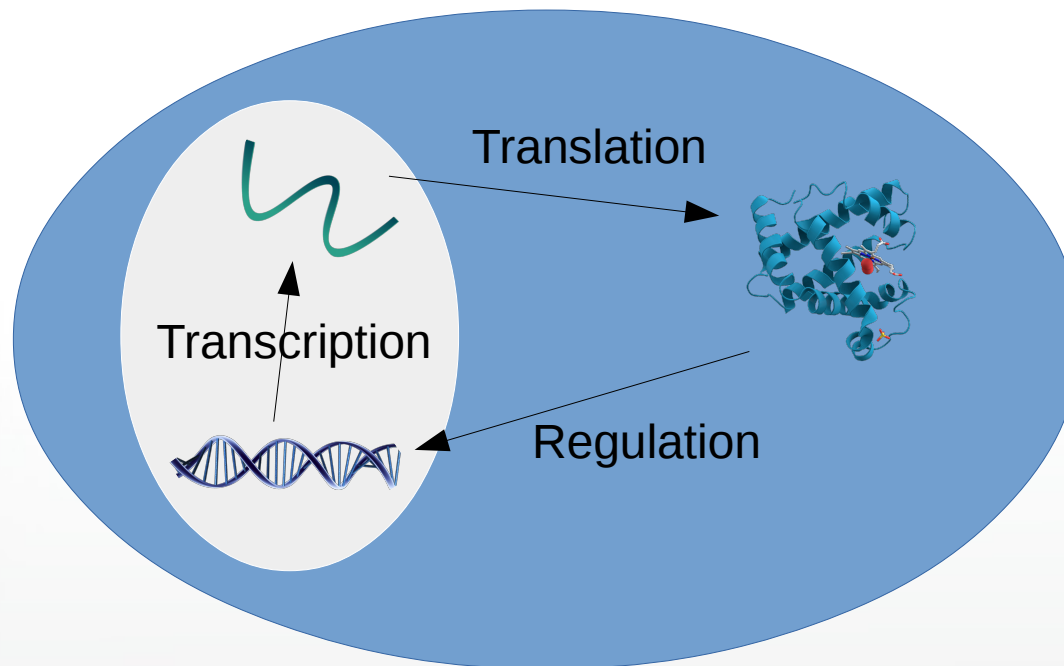
# GRN Inference

- **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”

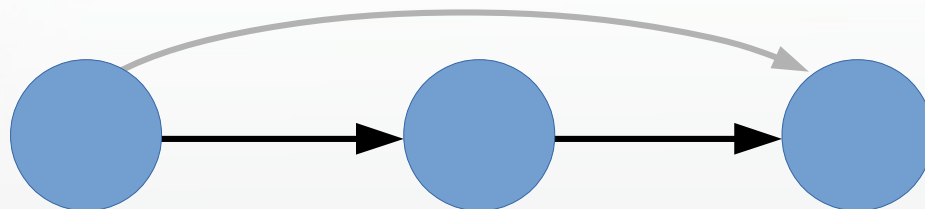


# How to address dimensionality?

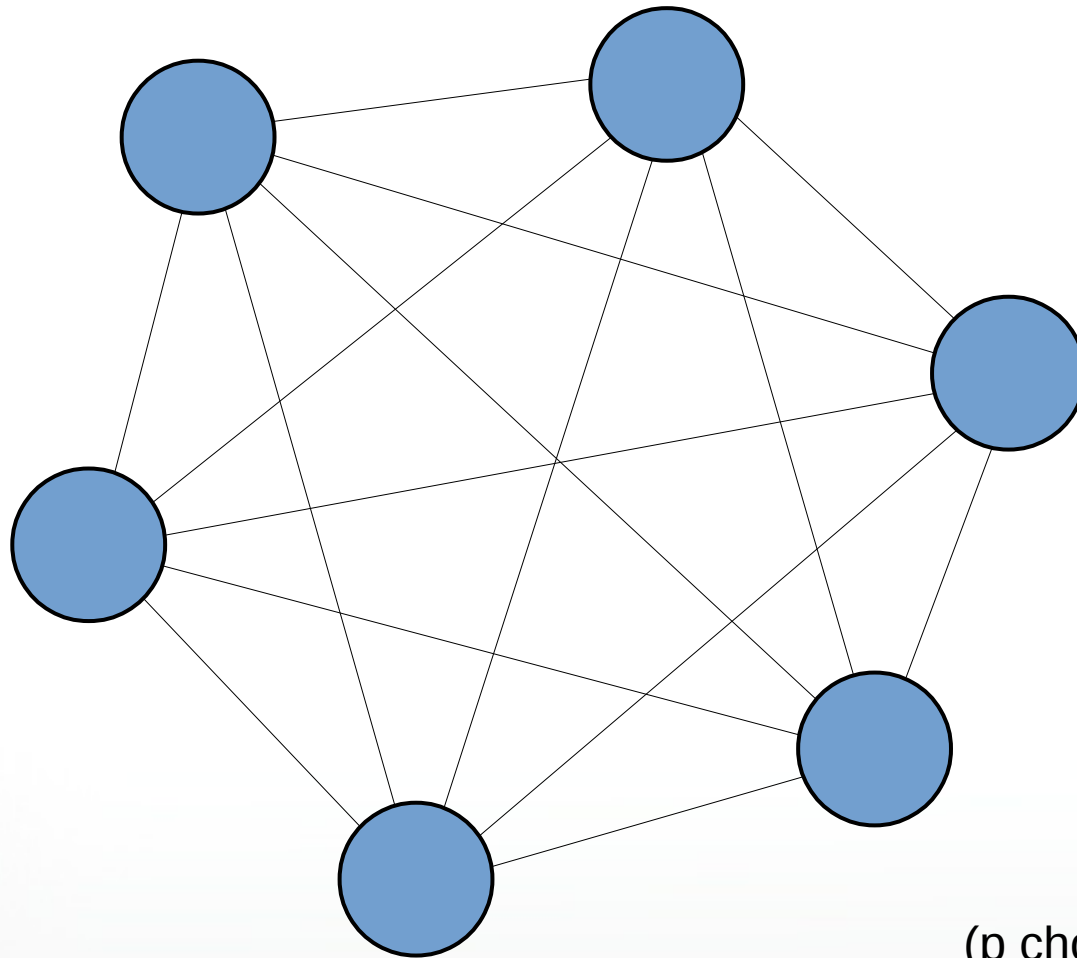
Assume sparsity.

- 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

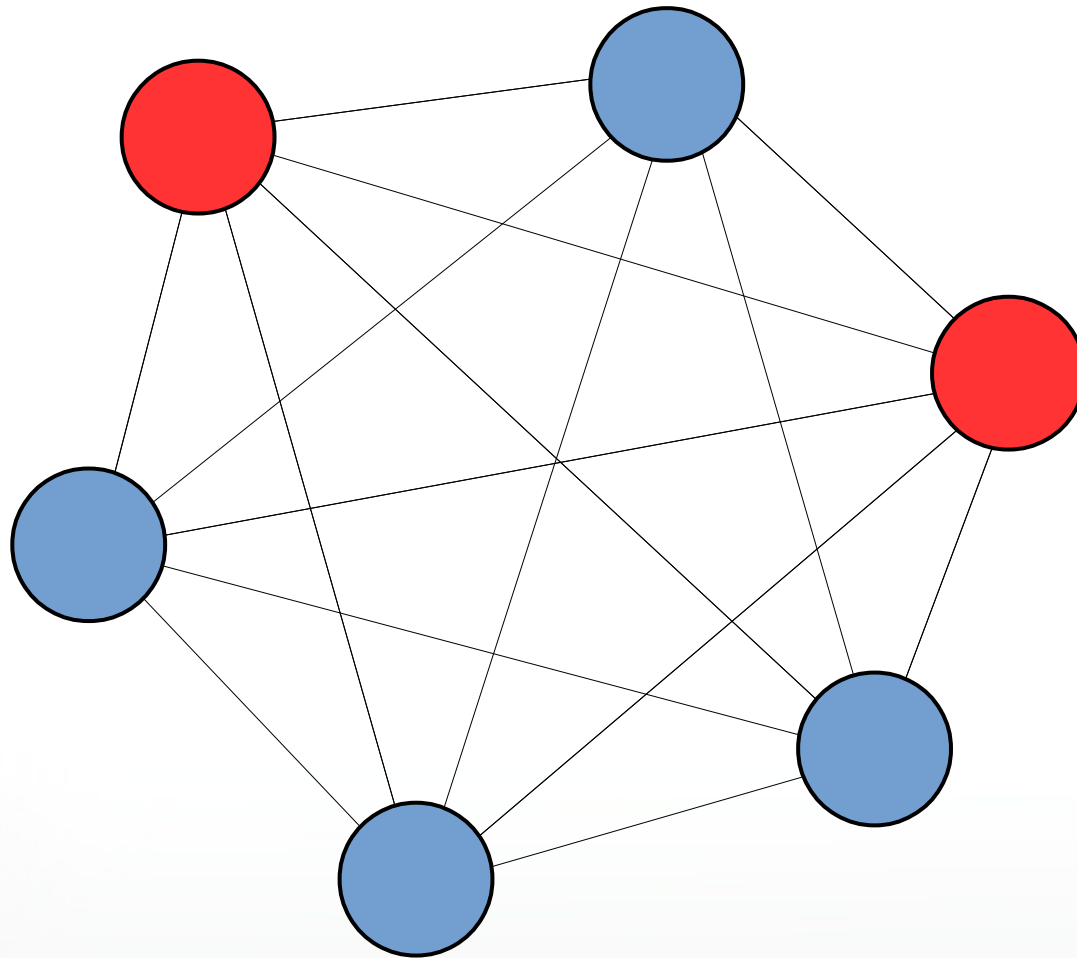


(p choose 2) edges

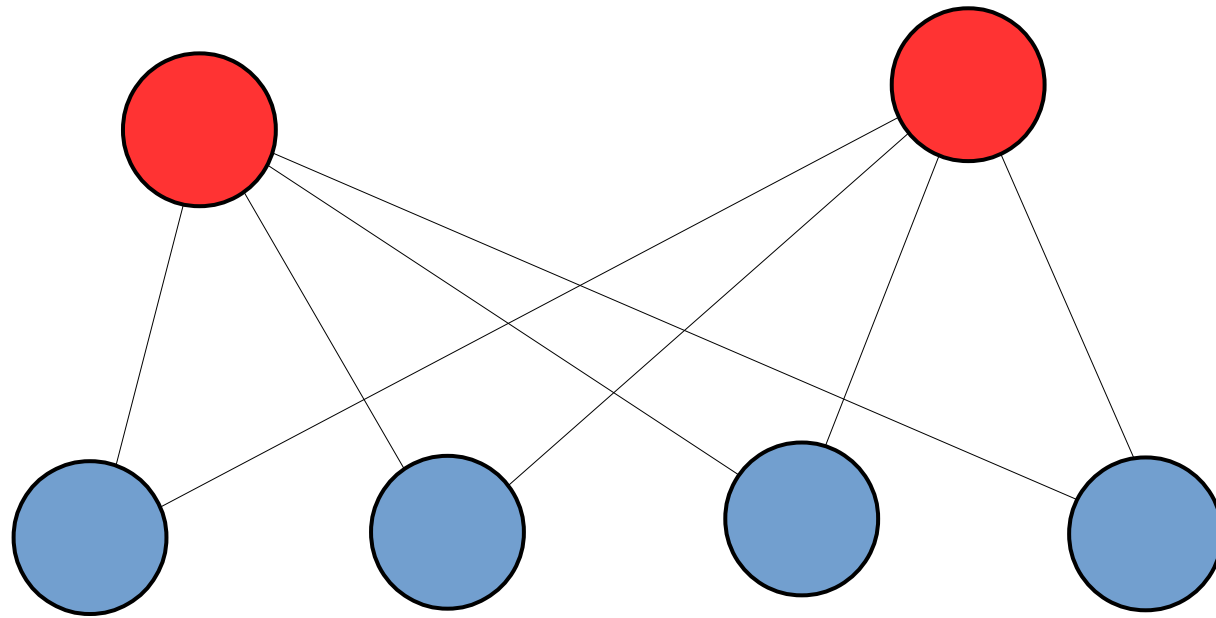




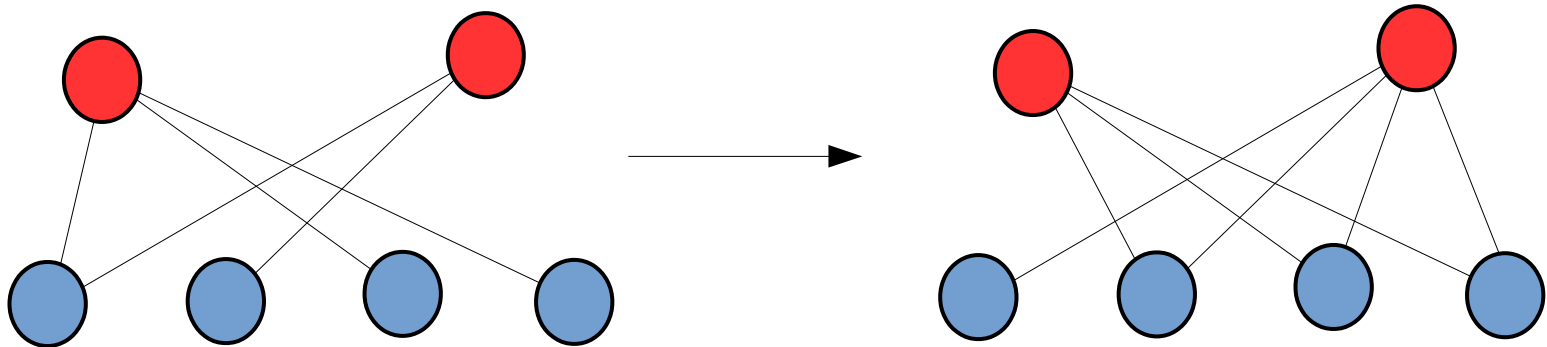
# Challenges in GRN inference



# 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.



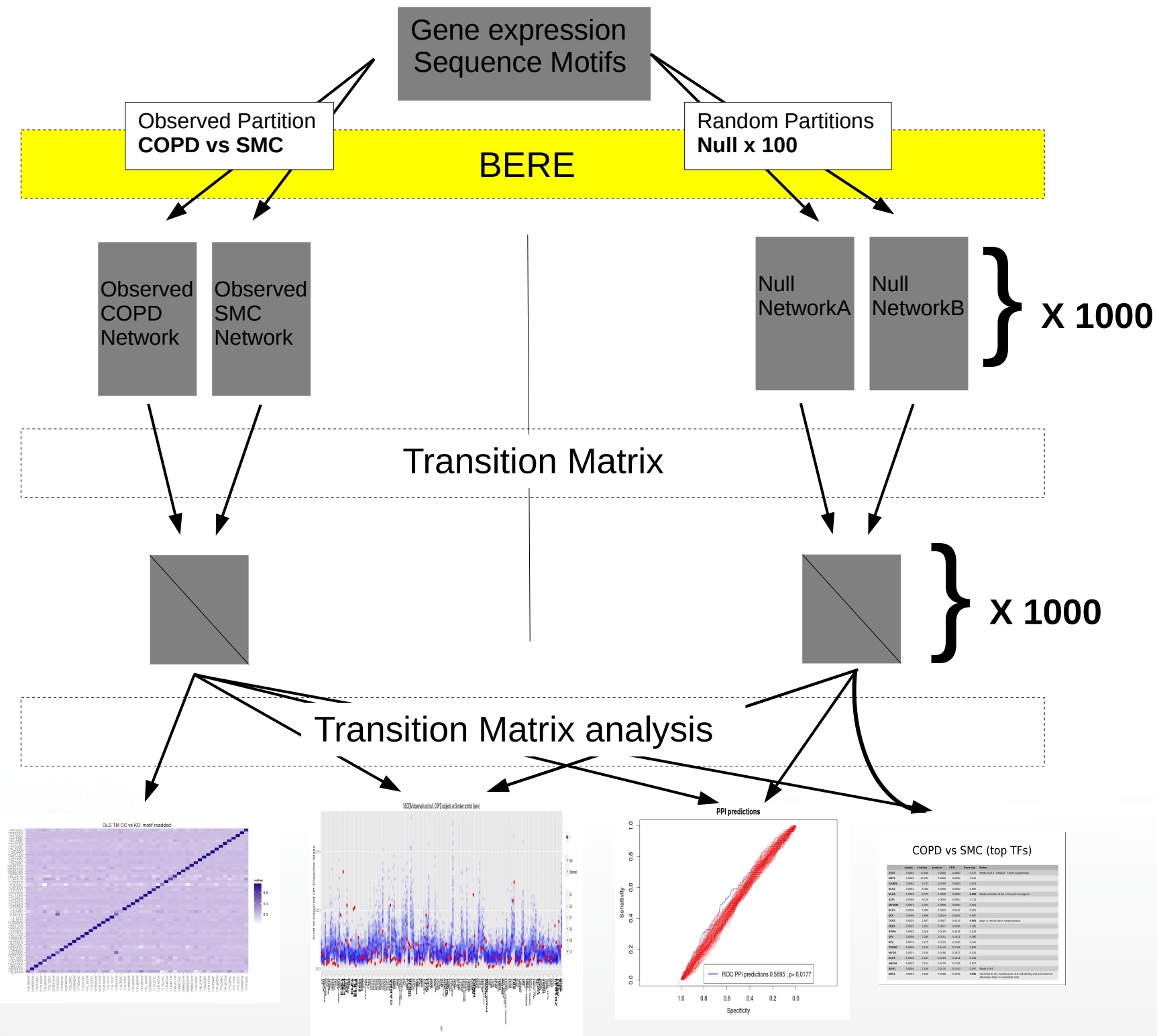
# Network differentiation challenges

- 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.



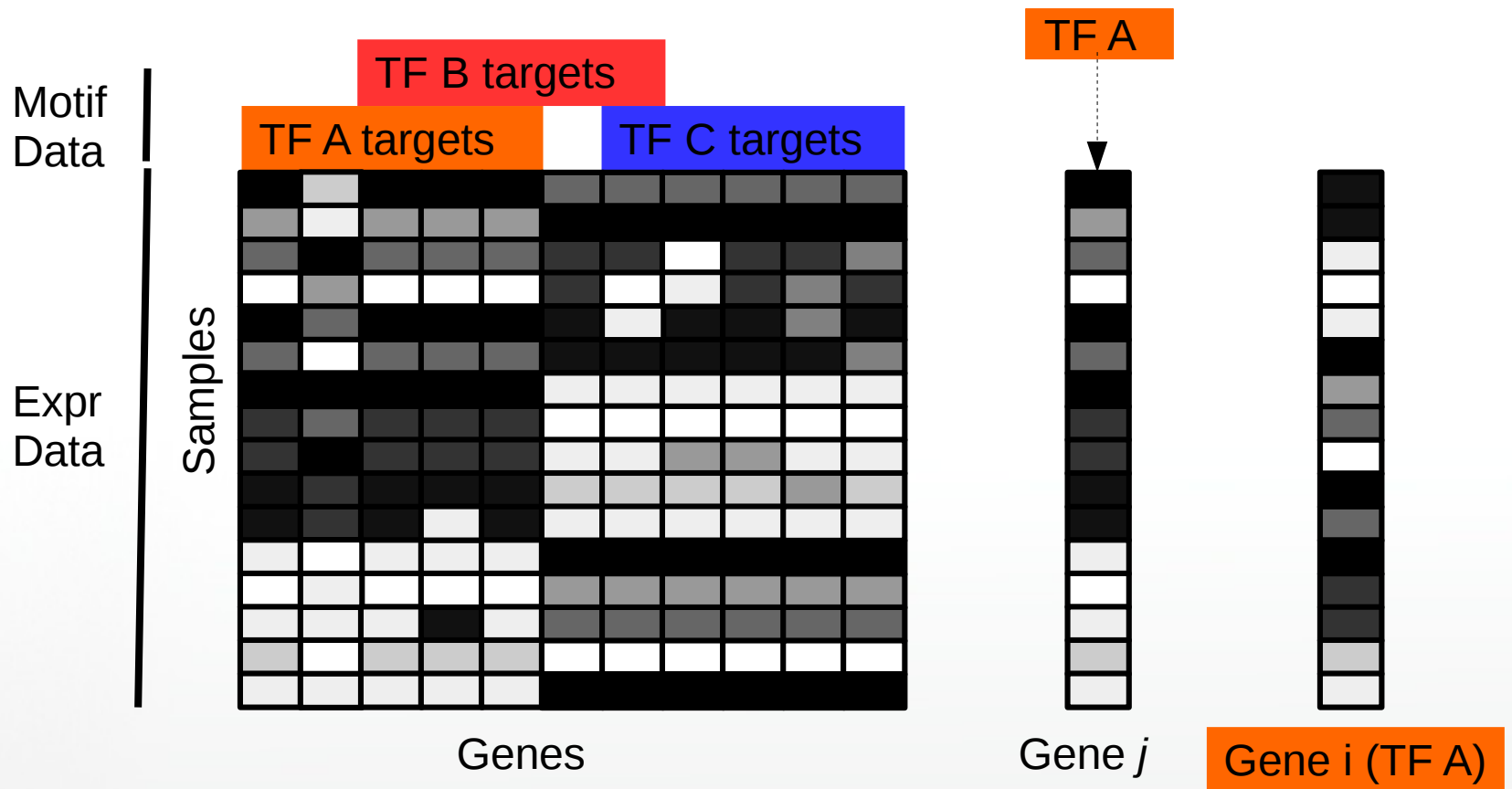
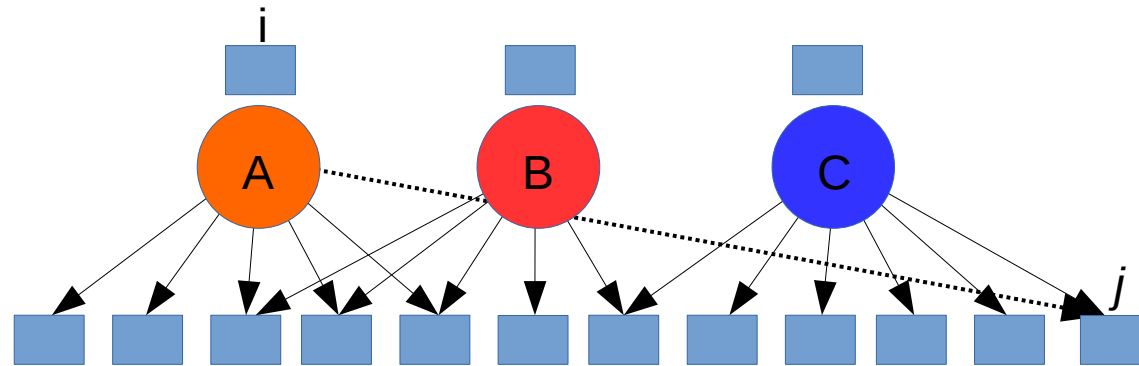




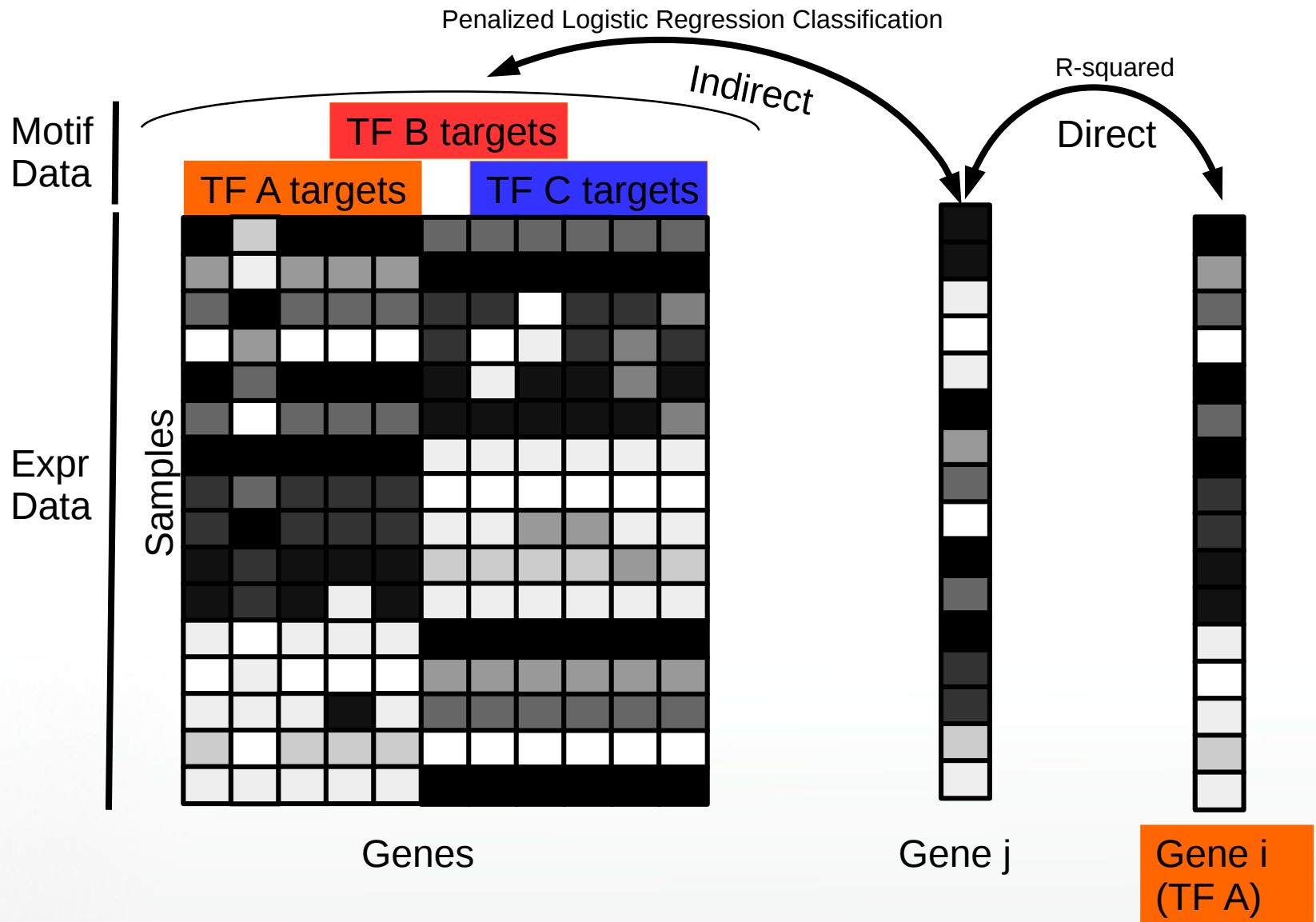


TF	Score	Rank	TF	Score	Rank
TF1	0.9999	1	TF2	0.9999	2
TF3	0.9999	3	TF4	0.9999	4
TF5	0.9999	5	TF6	0.9999	6
TF7	0.9999	7	TF8	0.9999	8
TF9	0.9999	9	TF10	0.9999	10
TF11	0.9999	11	TF12	0.9999	12
TF13	0.9999	13	TF14	0.9999	14
TF15	0.9999	15	TF16	0.9999	16
TF17	0.9999	17	TF18	0.9999	18
TF19	0.9999	19	TF20	0.9999	20
TF21	0.9999	21	TF22	0.9999	22
TF23	0.9999	23	TF24	0.9999	24
TF25	0.9999	25	TF26	0.9999	26
TF27	0.9999	27	TF28	0.9999	28
TF29	0.9999	29	TF30	0.9999	30
TF31	0.9999	31	TF32	0.9999	32
TF33	0.9999	33	TF34	0.9999	34
TF35	0.9999	35	TF36	0.9999	36
TF37	0.9999	37	TF38	0.9999	38
TF39	0.9999	39	TF40	0.9999	40
TF41	0.9999	41	TF42	0.9999	42
TF43	0.9999	43	TF44	0.9999	44
TF45	0.9999	45	TF46	0.9999	46
TF47	0.9999	47	TF48	0.9999	48
TF49	0.9999	49	TF50	0.9999	50

# Bipartite Edge Reconstruction from Expression Data



# BERE



# BERE - direct

Divide evidence for regulation into 2 parts:

## 1.) **Direct evidence**

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

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

$$X_i^* = X_i - X_{TF} (X_{TF}' X_{TF})^{-1} X_{TF}' X_i$$

$$X_j^* = X_j - X_{TF} (X_{TF}' X_{TF})^{-1} X_{TF}' X_j$$

$$d_{i,j} = \frac{X_i^{*'} X_j^*}{\sqrt{(X_i^{*'} X_i^*) (X_j^{*'} X_j^*)}}$$

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:

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

$\mathbf{Q}$  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.

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

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.

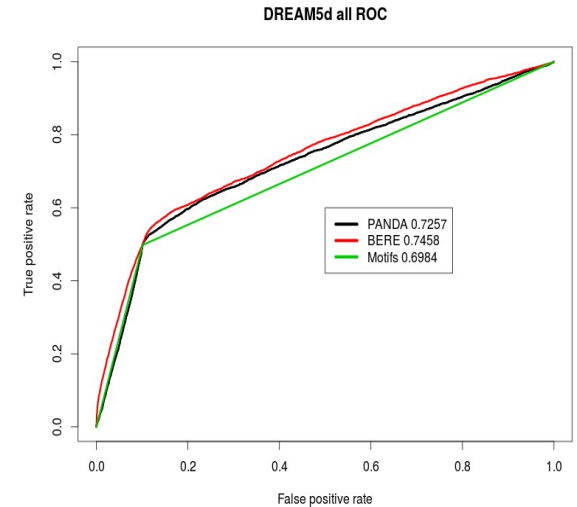
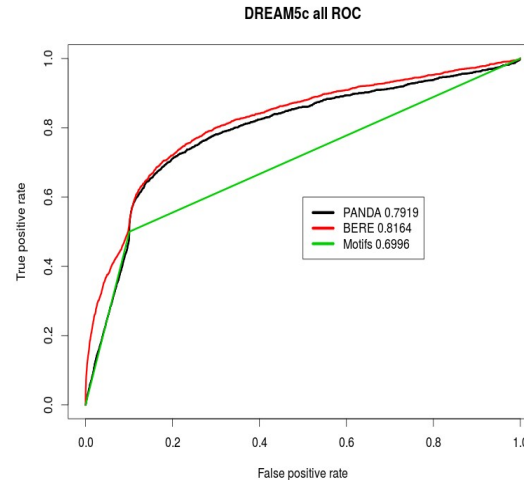
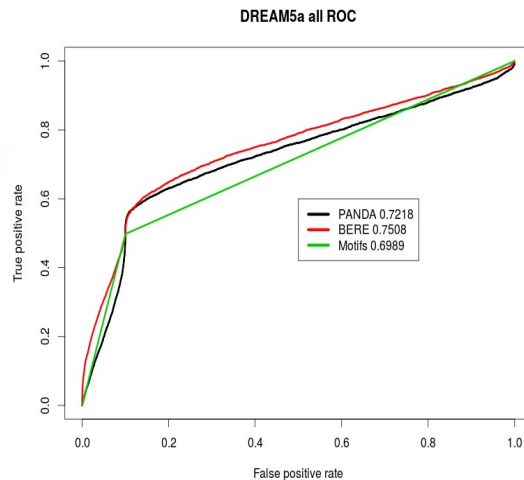


# BERE

In Silico

*E. coli*

Yeast



**Running R package: 8GB RAM, 2.40Ghz**

**Time**

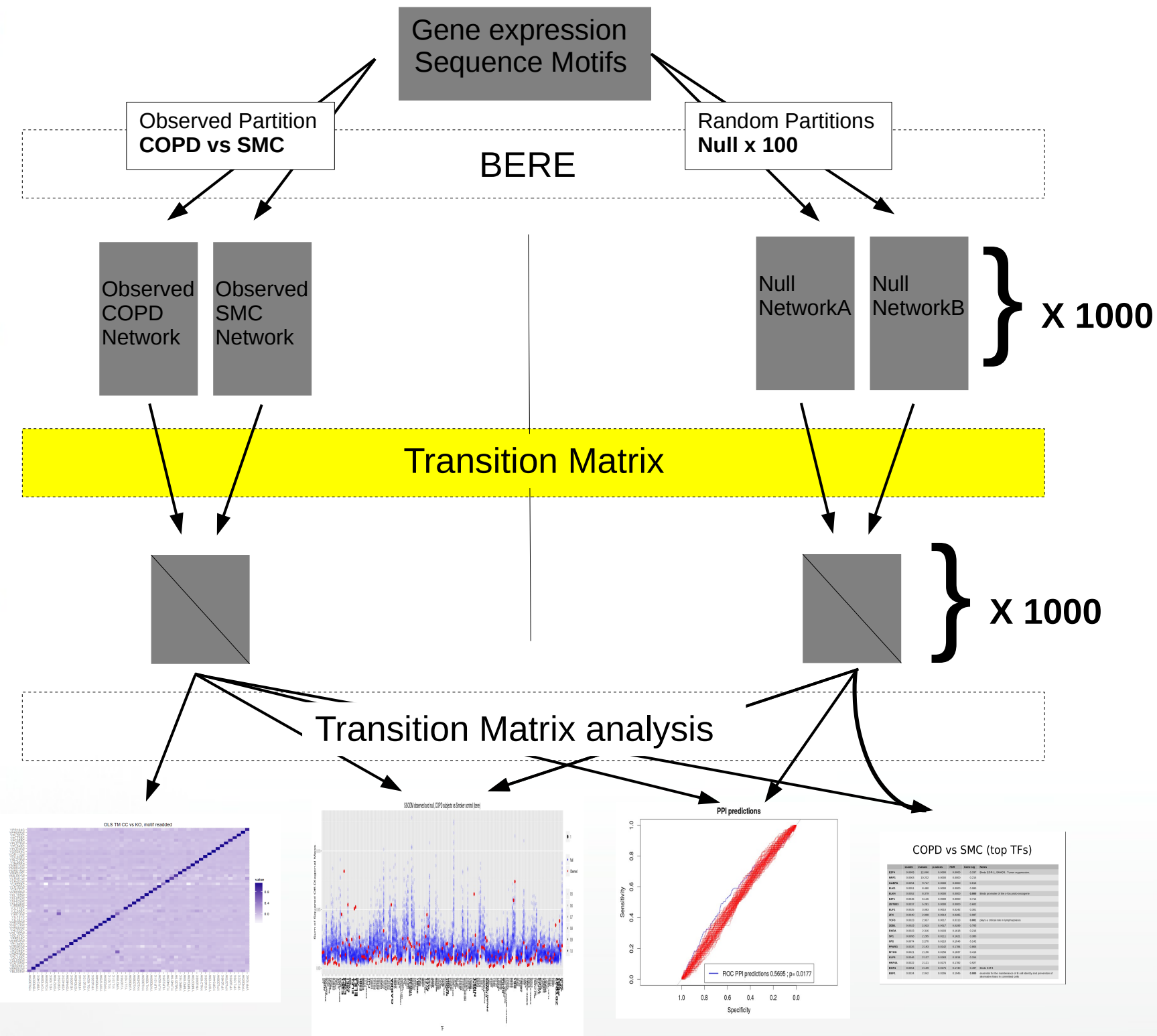
2555 genes, 53 TF, 106 samples

11s

17342 genes, 189 TF, 226 samples

12m, 20s

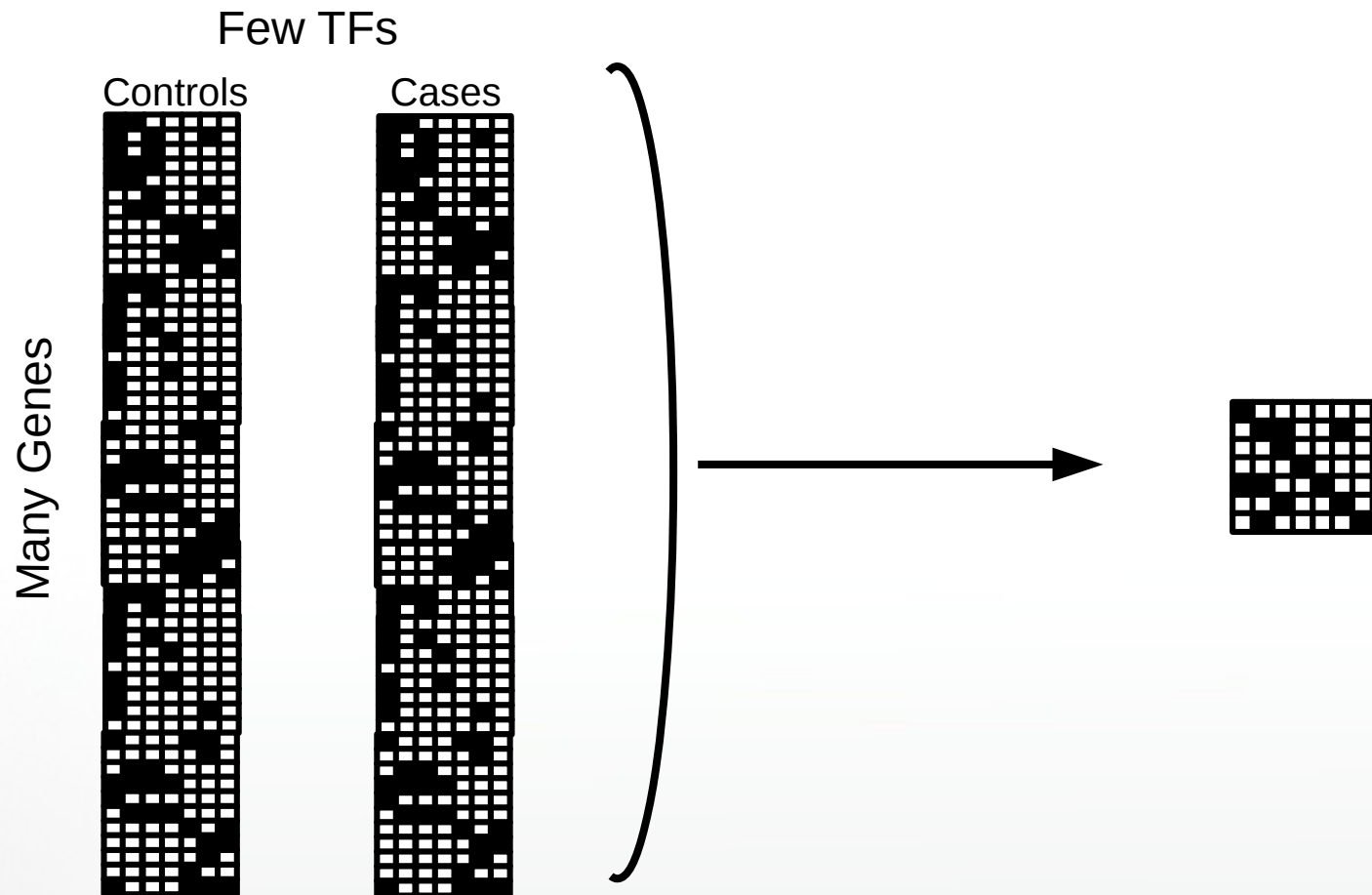




TF	Score	Rank	TF	Score	Rank	TF	Score	Rank
TF1	0.999	1	TF2	0.999	2	TF3	0.999	3
TF4	0.999	4	TF5	0.999	5	TF6	0.999	6
TF7	0.999	7	TF8	0.999	8	TF9	0.999	9
TF10	0.999	10	TF11	0.999	11	TF12	0.999	12
TF13	0.999	13	TF14	0.999	14	TF15	0.999	15
TF16	0.999	16	TF17	0.999	17	TF18	0.999	18
TF19	0.999	19	TF20	0.999	20	TF21	0.999	21
TF22	0.999	22	TF23	0.999	23	TF24	0.999	24
TF25	0.999	25	TF26	0.999	26	TF27	0.999	27
TF28	0.999	28	TF29	0.999	29	TF30	0.999	30
TF31	0.999	31	TF32	0.999	32	TF33	0.999	33
TF34	0.999	34	TF35	0.999	35	TF36	0.999	36
TF37	0.999	37	TF38	0.999	38	TF39	0.999	39
TF40	0.999	40	TF41	0.999	41	TF42	0.999	42
TF43	0.999	43	TF44	0.999	44	TF45	0.999	45
TF46	0.999	46	TF47	0.999	47	TF48	0.999	48
TF49	0.999	49	TF50	0.999	50	TF51	0.999	51
TF52	0.999	52	TF53	0.999	53	TF54	0.999	54
TF55	0.999	55	TF56	0.999	56	TF57	0.999	57
TF58	0.999	58	TF59	0.999	59	TF60	0.999	60
TF61	0.999	61	TF62	0.999	62	TF63	0.999	63
TF64	0.999	64	TF65	0.999	65	TF66	0.999	66
TF67	0.999	67	TF68	0.999	68	TF69	0.999	69
TF70	0.999	70	TF71	0.999	71	TF72	0.999	72
TF73	0.999	73	TF74	0.999	74	TF75	0.999	75
TF76	0.999	76	TF77	0.999	77	TF78	0.999	78
TF79	0.999	79	TF80	0.999	80	TF81	0.999	81
TF82	0.999	82	TF83	0.999	83	TF84	0.999	84
TF85	0.999	85	TF86	0.999	86	TF87	0.999	87
TF88	0.999	88	TF89	0.999	89	TF90	0.999	90
TF91	0.999	91	TF92	0.999	92	TF93	0.999	93
TF94	0.999	94	TF95	0.999	95	TF96	0.999	96
TF97	0.999	97	TF98	0.999	98	TF99	0.999	99
TF100	0.999	100	TF101	0.999	101	TF102	0.999	102
TF103	0.999	103	TF104	0.999	104	TF105	0.999	105
TF106	0.999	106	TF107	0.999	107	TF108	0.999	108
TF109	0.999	109	TF110	0.999	110	TF111	0.999	111
TF112	0.999	112	TF113	0.999	113	TF114	0.999	114
TF115	0.999	115	TF116	0.999	116	TF117	0.999	117
TF118	0.999	118	TF119	0.999	119	TF120	0.999	120
TF121	0.999	121	TF122	0.999	122	TF123	0.999	123
TF124	0.999	124	TF125	0.999	125	TF126	0.999	126
TF127	0.999	127	TF128	0.999	128	TF129	0.999	129
TF130	0.999	130	TF131	0.999	131	TF132	0.999	132
TF133	0.999	133	TF134	0.999	134	TF135	0.999	135
TF136	0.999	136	TF137	0.999	137	TF138	0.999	138
TF139	0.999	139	TF140	0.999	140	TF141	0.999	141
TF142	0.999	142	TF143	0.999	143	TF144	0.999	144
TF145	0.999	145	TF146	0.999	146	TF147	0.999	147
TF148	0.999	148	TF149	0.999	149	TF150	0.999	150
TF151	0.999	151	TF152	0.999	152	TF153	0.999	153
TF154	0.999	154	TF155	0.999	155	TF156	0.999	156
TF157	0.999	157	TF158	0.999	158	TF159	0.999	159
TF160	0.999	160	TF161	0.999	161	TF162	0.999	162
TF163	0.999	163	TF164	0.999	164	TF165	0.999	165
TF166	0.999	166	TF167	0.999	167	TF168	0.999	168
TF169	0.999	169	TF170	0.999	170	TF171	0.999	171
TF172	0.999	172	TF173	0.999	173	TF174	0.999	174
TF175	0.999	175	TF176	0.999	176	TF177	0.999	177
TF178	0.999	178	TF179	0.999	179	TF180	0.999	180
TF181	0.999	181	TF182	0.999	182	TF183	0.999	183
TF184	0.999	184	TF185	0.999	185	TF186	0.999	186
TF187	0.999	187	TF188	0.999	188	TF189	0.999	189
TF190	0.999	190	TF191	0.999	191	TF192	0.999	192
TF193	0.999	193	TF194	0.999	194	TF195	0.999	195
TF196	0.999	196	TF197	0.999	197	TF198	0.999	198
TF199	0.999	199	TF200	0.999	200	TF201	0.999	201
TF202	0.999	202	TF203	0.999	203	TF204	0.999	204
TF205	0.999	205	TF206	0.999	206	TF207	0.999	207
TF208	0.999	208	TF209	0.999	209	TF210	0.999	210
TF211	0.999	211	TF212	0.999	212	TF213	0.999	213
TF214	0.999	214	TF215	0.999	215	TF216	0.999	216
TF217	0.999	217	TF218	0.999	218	TF219	0.999	219
TF220	0.999	220	TF221	0.999	221	TF222	0.999	222
TF223	0.999	223	TF224	0.999	224	TF225	0.999	225
TF226	0.999	226	TF227	0.999	227	TF228	0.999	228
TF229	0.999	229	TF230	0.999	230	TF231	0.999	231
TF232	0.999	232	TF233	0.999	233	TF234	0.999	234
TF235	0.999	235	TF236	0.999	236	TF237	0.999	237
TF238	0.999	238	TF239	0.999	239	TF240	0.999	240
TF241	0.999	241	TF242	0.999	242	TF243	0.999	243
TF244	0.999	244	TF245	0.999	245	TF246	0.999	246
TF247	0.999	247	TF248	0.999	248	TF249	0.999	249
TF250	0.999	250	TF251	0.999	251	TF252	0.999	252
TF253	0.999	253	TF254	0.999	254	TF255	0.999	255
TF256	0.999	256	TF257	0.999	257	TF258	0.999	258
TF259	0.999	259	TF260	0.999	260	TF261	0.999	261
TF262	0.999	262	TF263	0.999	263	TF264	0.999	264
TF265	0.999	265	TF266	0.999	266	TF267	0.999	267
TF268	0.999	268	TF269	0.999	269	TF270	0.999	270
TF271	0.999	271	TF272	0.999	272	TF273	0.999	273
TF274	0.999	274	TF275	0.999	275	TF276	0.999	276
TF277	0.999	277	TF278	0.999	278	TF279	0.999	279
TF280	0.999	280	TF281	0.999	281	TF282	0.999	282
TF283	0.999	283	TF284	0.999	284	TF285	0.999	285
TF286	0.999	286	TF287	0.999	287	TF288	0.999	288
TF289	0.999	289	TF290	0.999	290	TF291	0.999	291
TF292	0.999	292	TF293	0.999	293	TF294	0.999	294
TF295	0.999	295	TF296	0.999	296	TF297	0.999	297
TF298	0.999	298	TF299	0.999	299	TF300	0.999	300

# Transition Matrix Approach

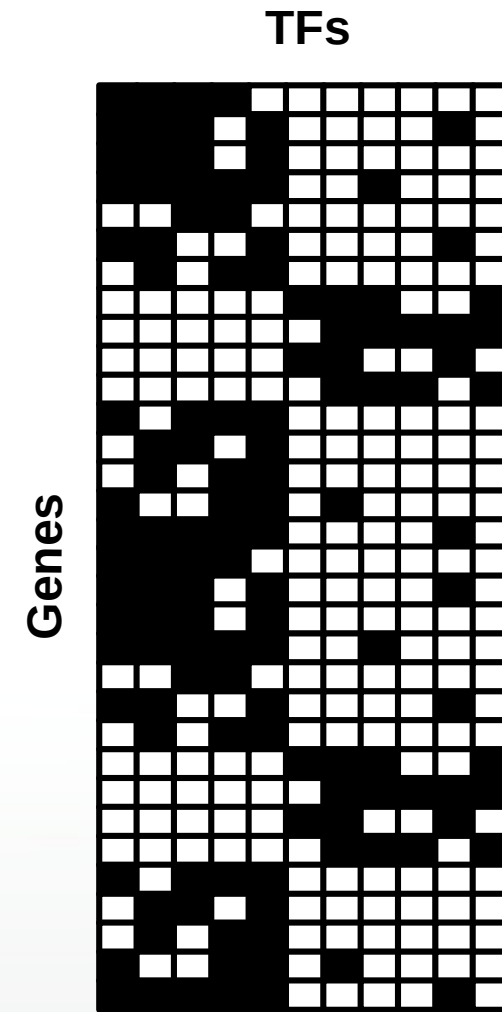
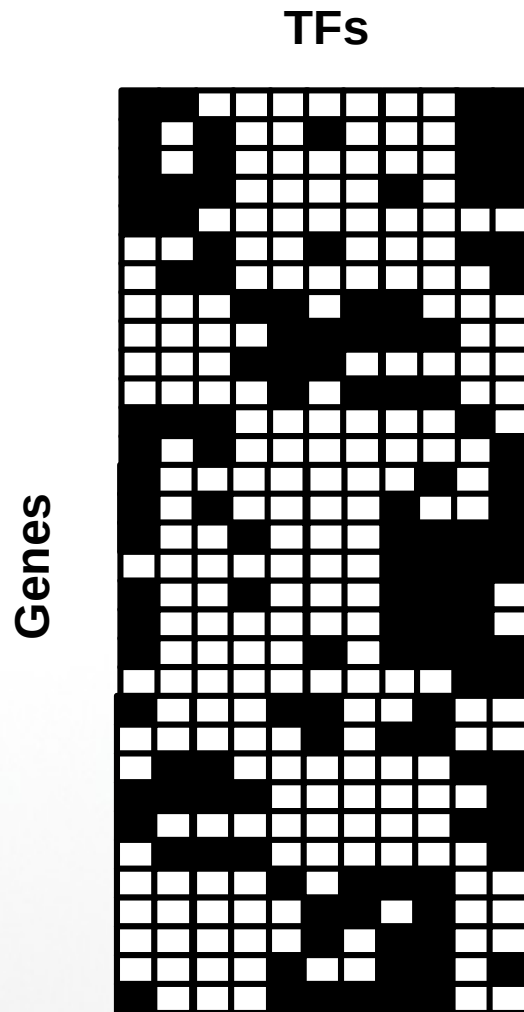
We can view the problem as a dimension reduction problem.





# Transition Matrix Approach

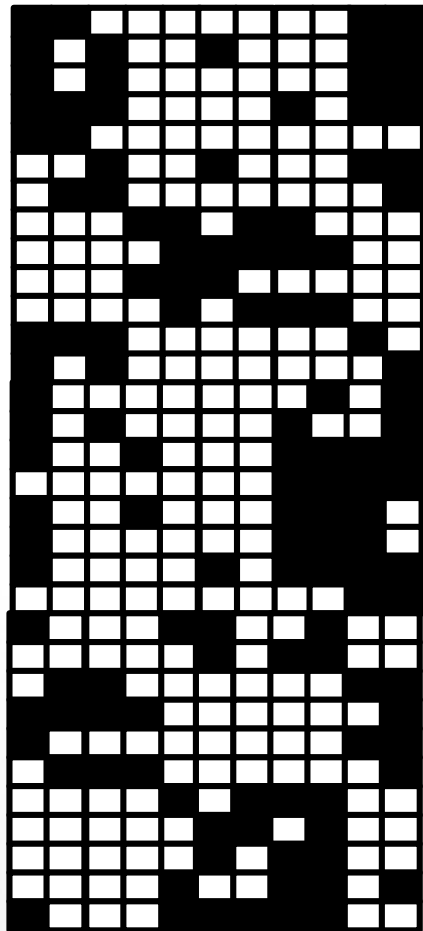
Consider two adjacency matrices...



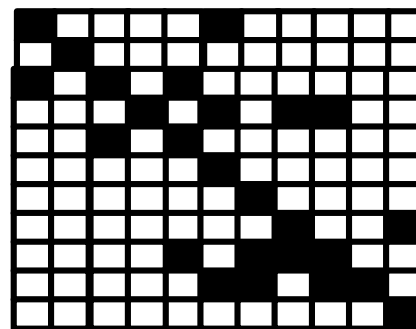
# Transition Matrix Approach

Consider two adjacency matrices...

Smoker Control



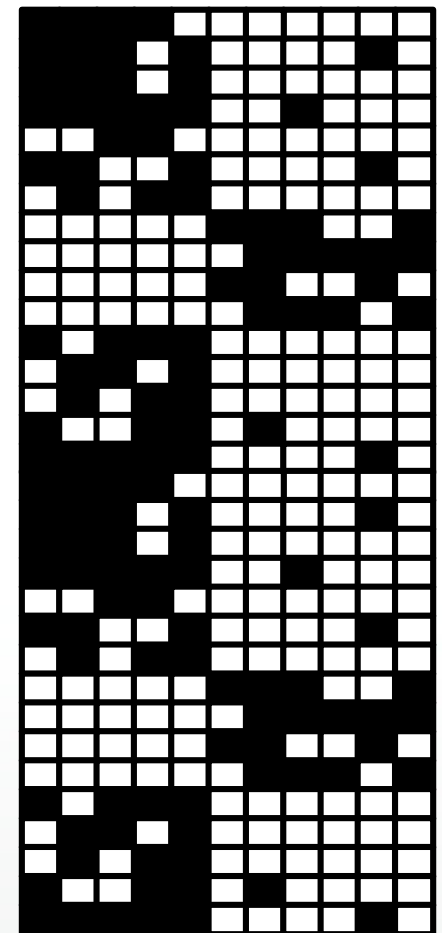
Tau



X

~

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  $(p \times m)$  representing the set of  $p$  genes targeted by  $m$  TFs. We seek a matrix, **T**, such that

$$\mathbf{B} = \mathbf{AT} + \mathbf{E}$$

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



# 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.

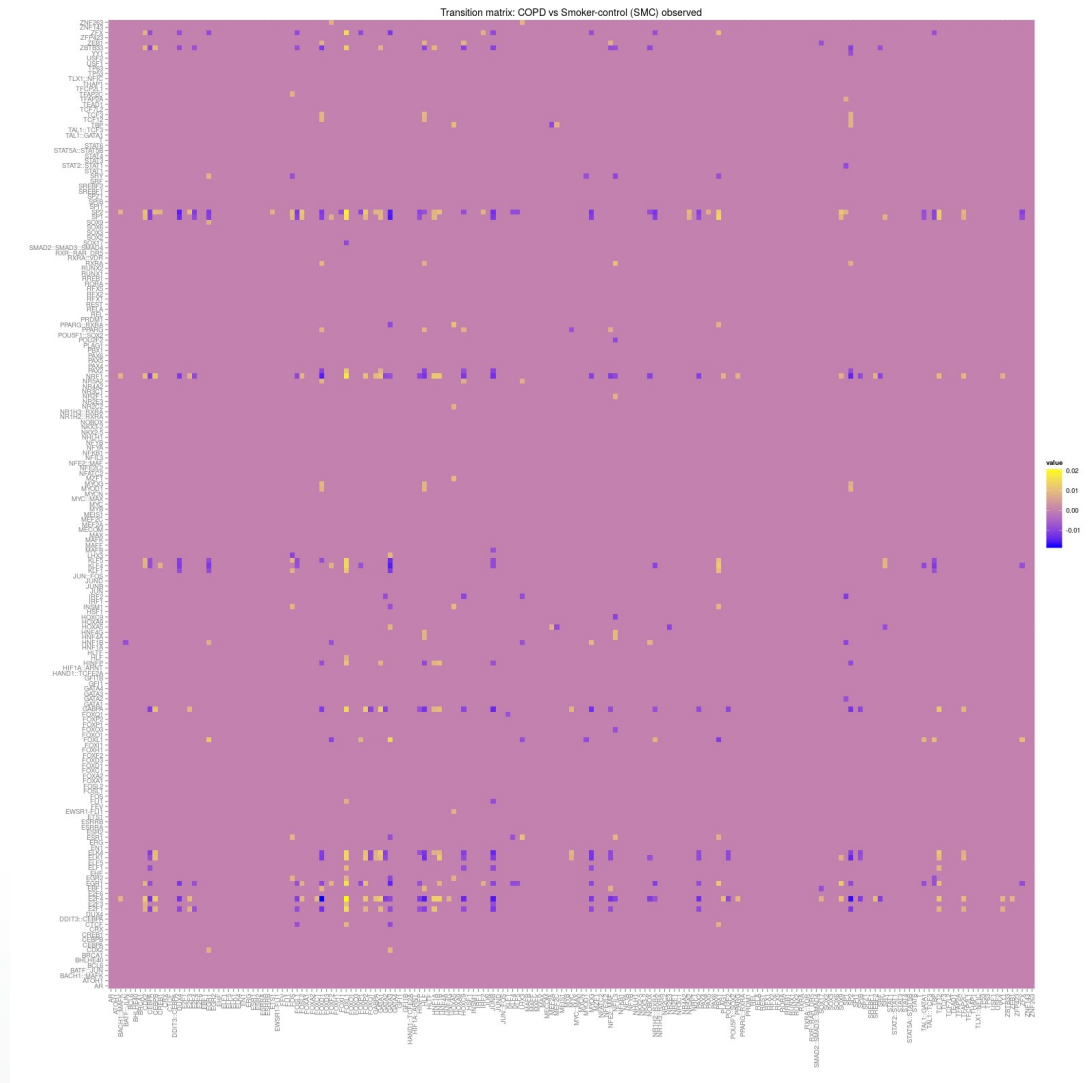
$$\sum_{i=1}^p \left( \mathbf{B}_{i,k} - \sum_{j=1}^m A_{i,j} \mathbf{T}_{j,k} \right)^2 + \lambda \beta' \mathbf{Q} \beta$$

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

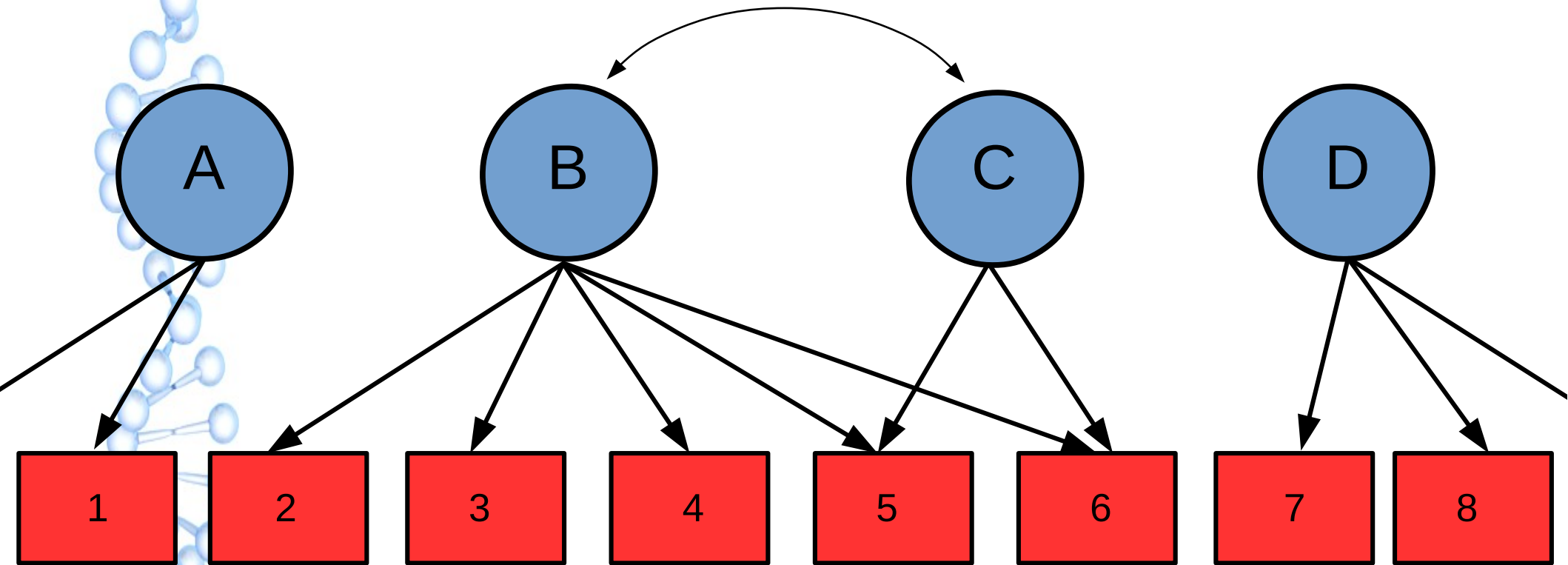
- Penalty model matrix is a diagonal matrix with value 0 for it's own TF and 1 for all others.







# An Example



Case Network

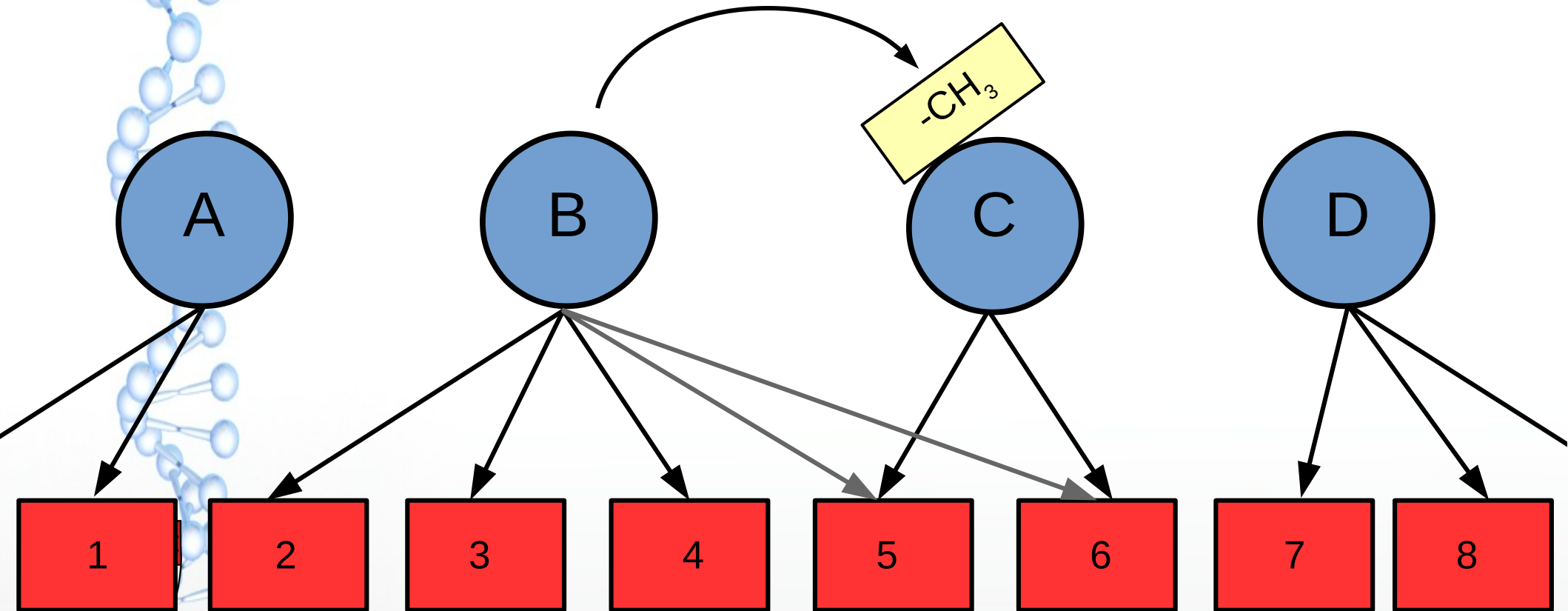
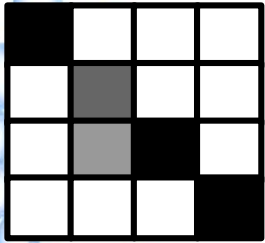
TM =

	A	B	C	D
A				
B				
C				
D				



# Biological Mechanism

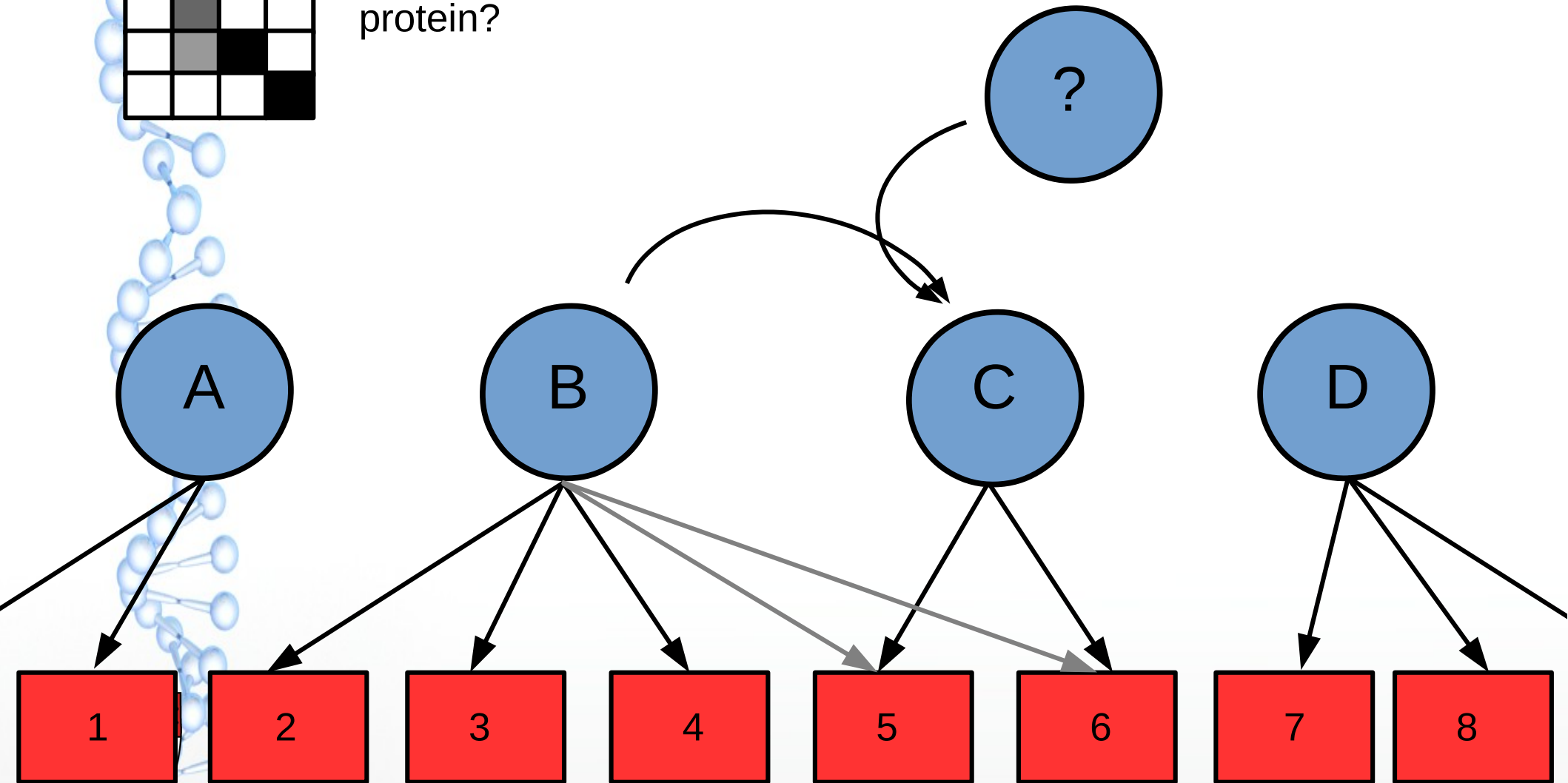
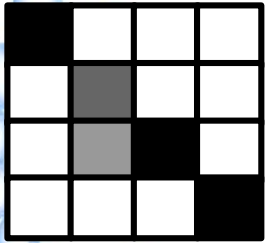
Suggested mechanism #1:  
Differential methylation of the gene for TF C?



# Biological Mechanism

Suggested mechanism #2:

Protein complex of B-C-? with unknown 3<sup>rd</sup> protein?





# 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

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

$s_j$  (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



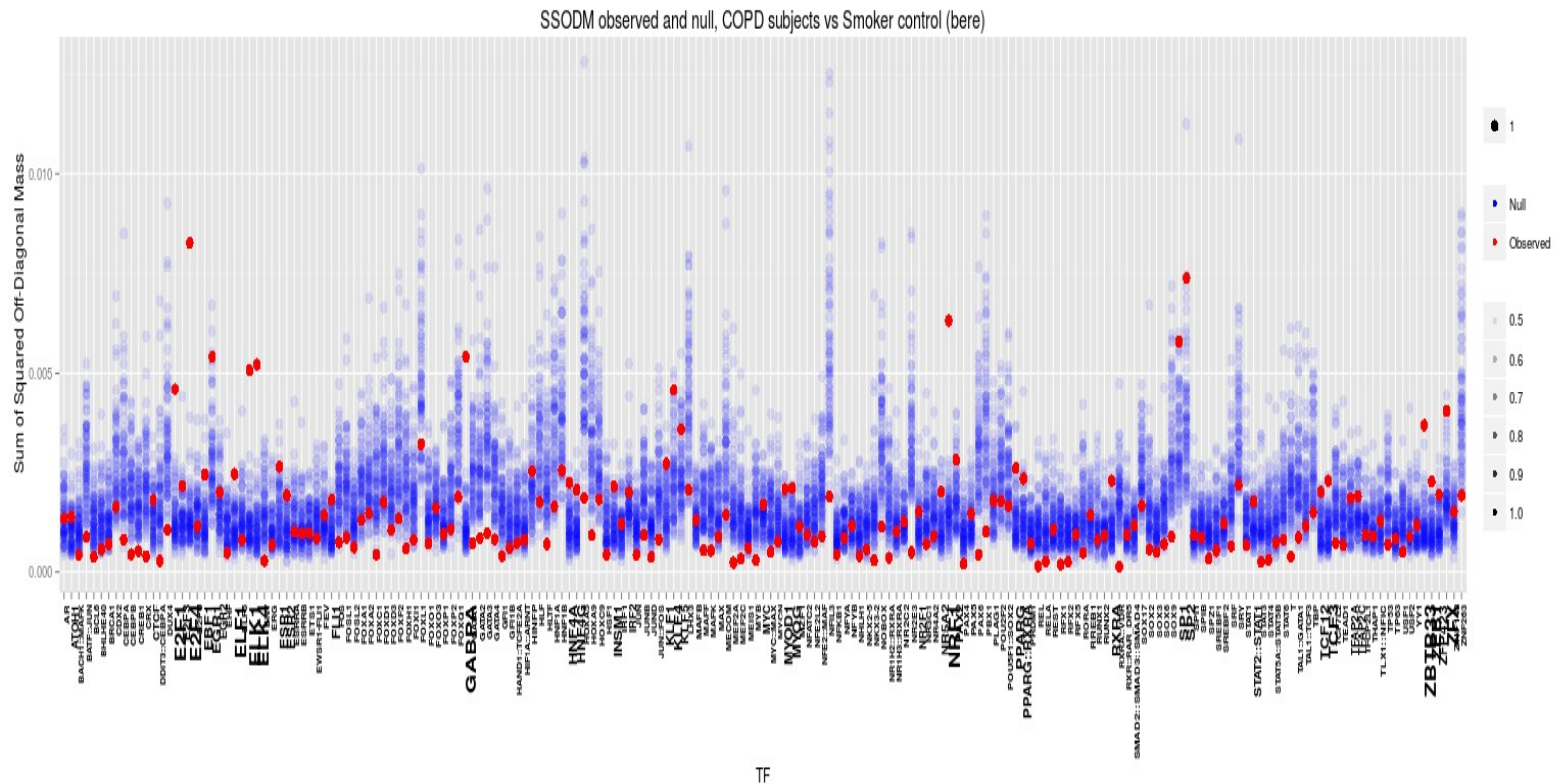


# Permutation inference on differential TFI statistic

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



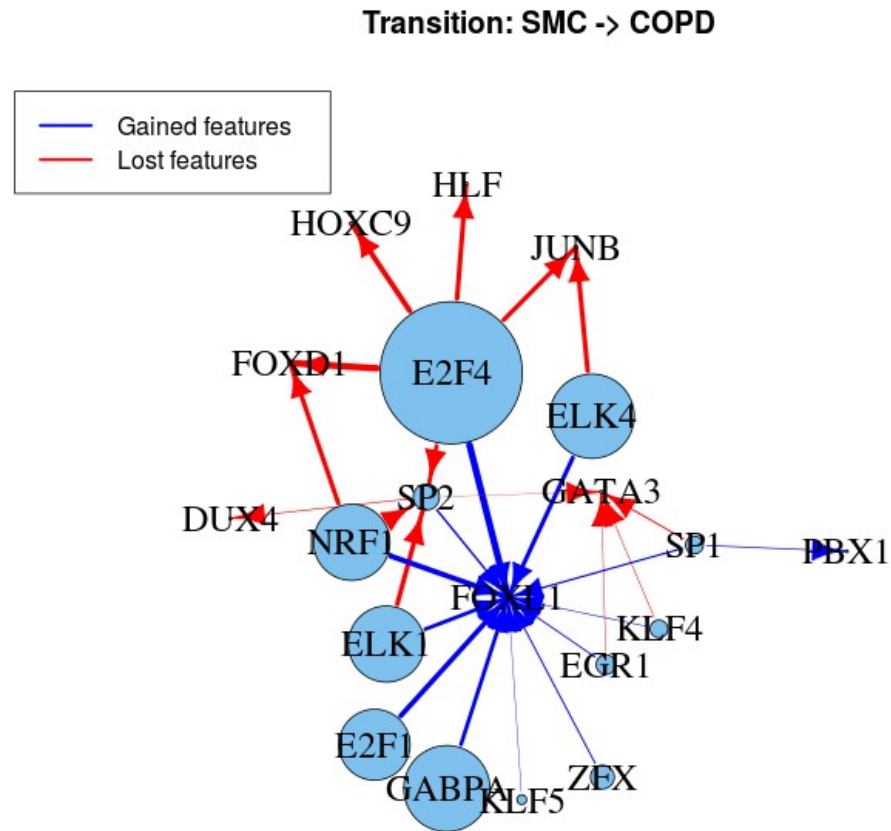
# Application to a case-control COPD study

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	FOX D1	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

