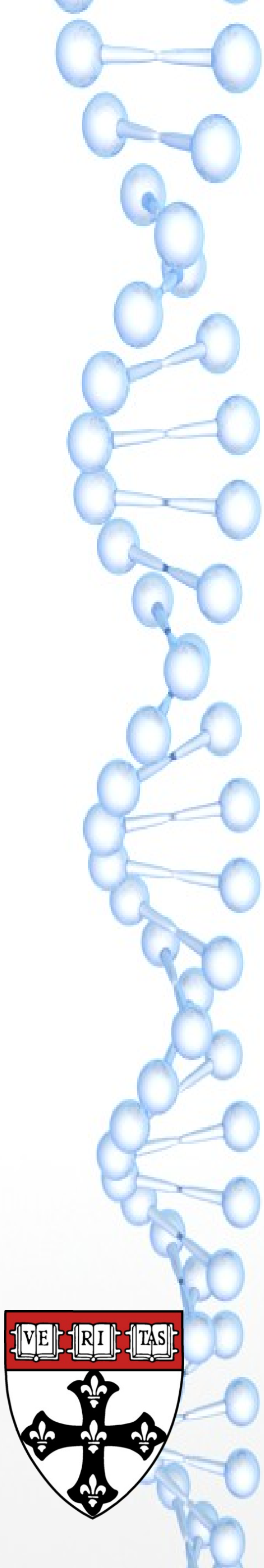
Methods in Case-Control Gene Regulatory Networks



RIP Meeting

July 29, 2015

Dan Schlauch

Outline

1) Why gene regulatory network inference?

2) The challenges of GRN inference.

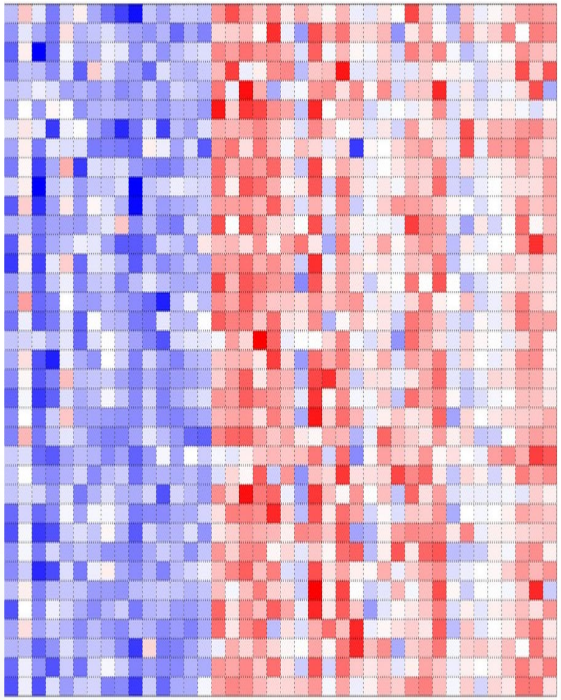
3) The challenges of GRN differentiation.

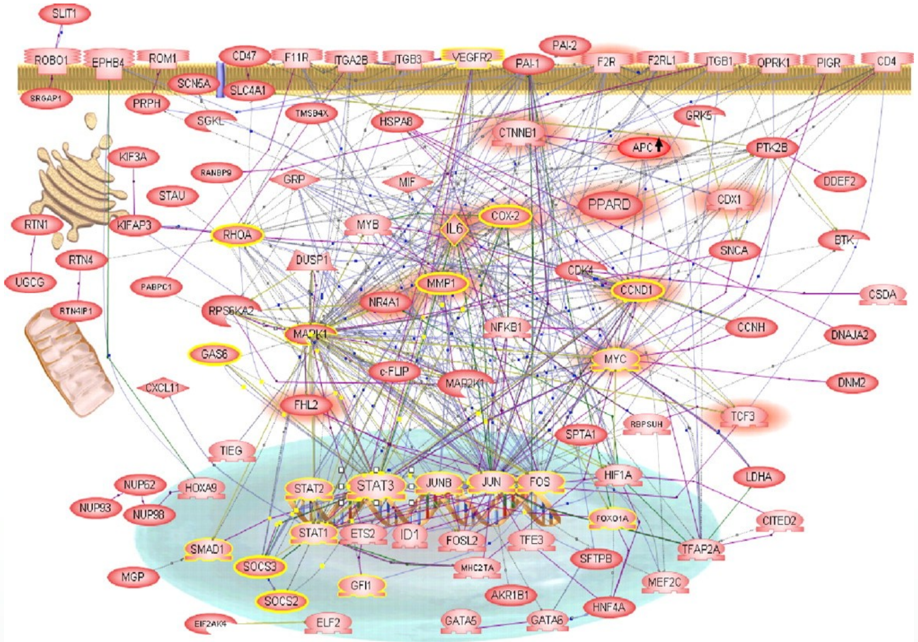
4) BERE, a novel GRN algorithm.

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

Why Gene Regulatory Network Inference?

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





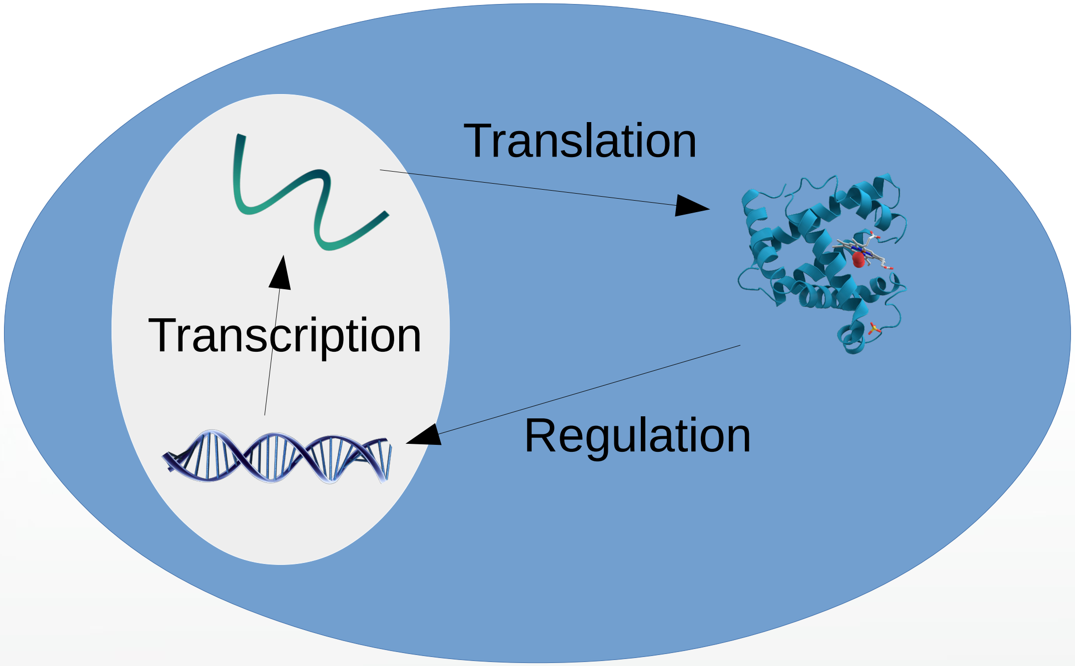
Amir Abdollahi et al. PNAS 2007;104:12890-12895

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

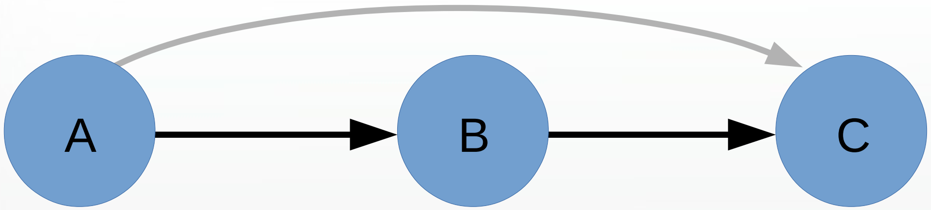
- Gene expression measurements are noisy, biased.
- Model complexity may require the estimate of too many model parameters.
    - May be computationally intractable.
    - May be statistically undetermined.

How to address dimensionality?

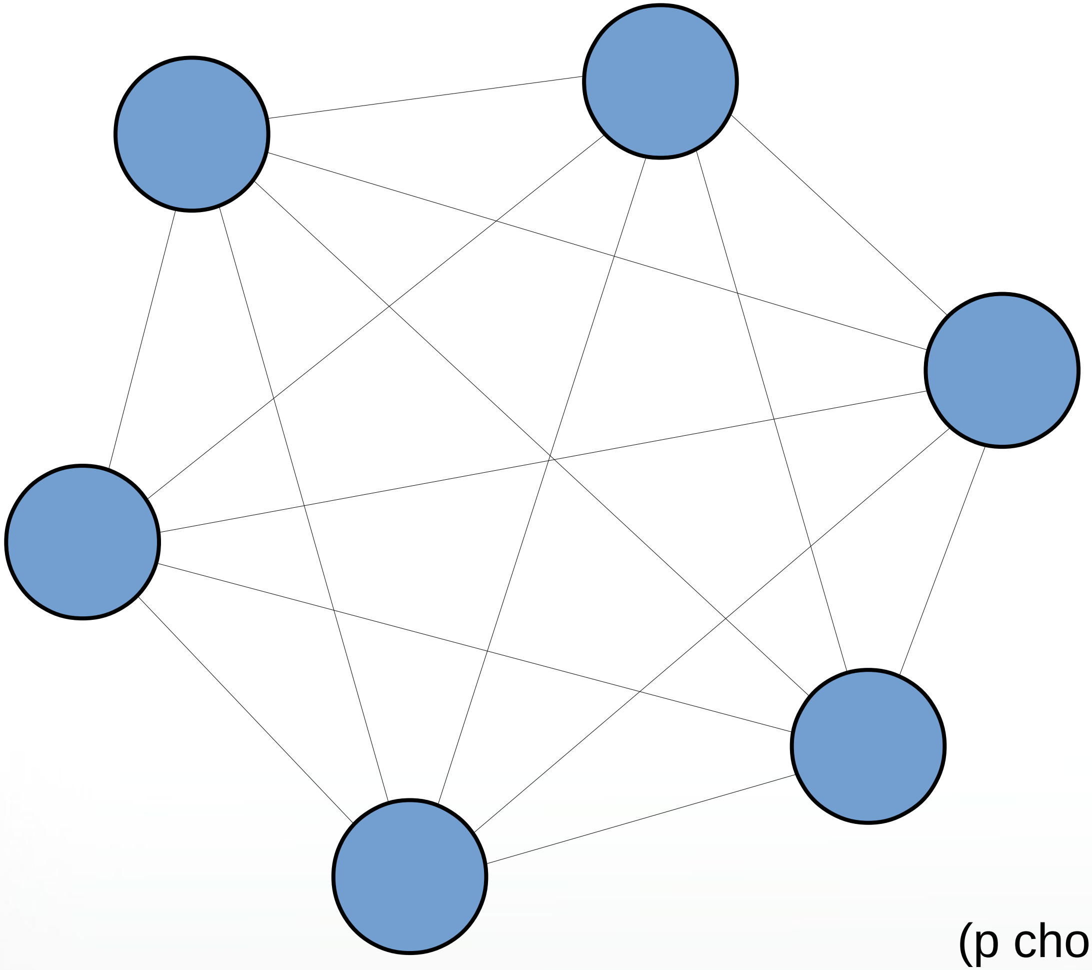
Assume sparsity.

- 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”.

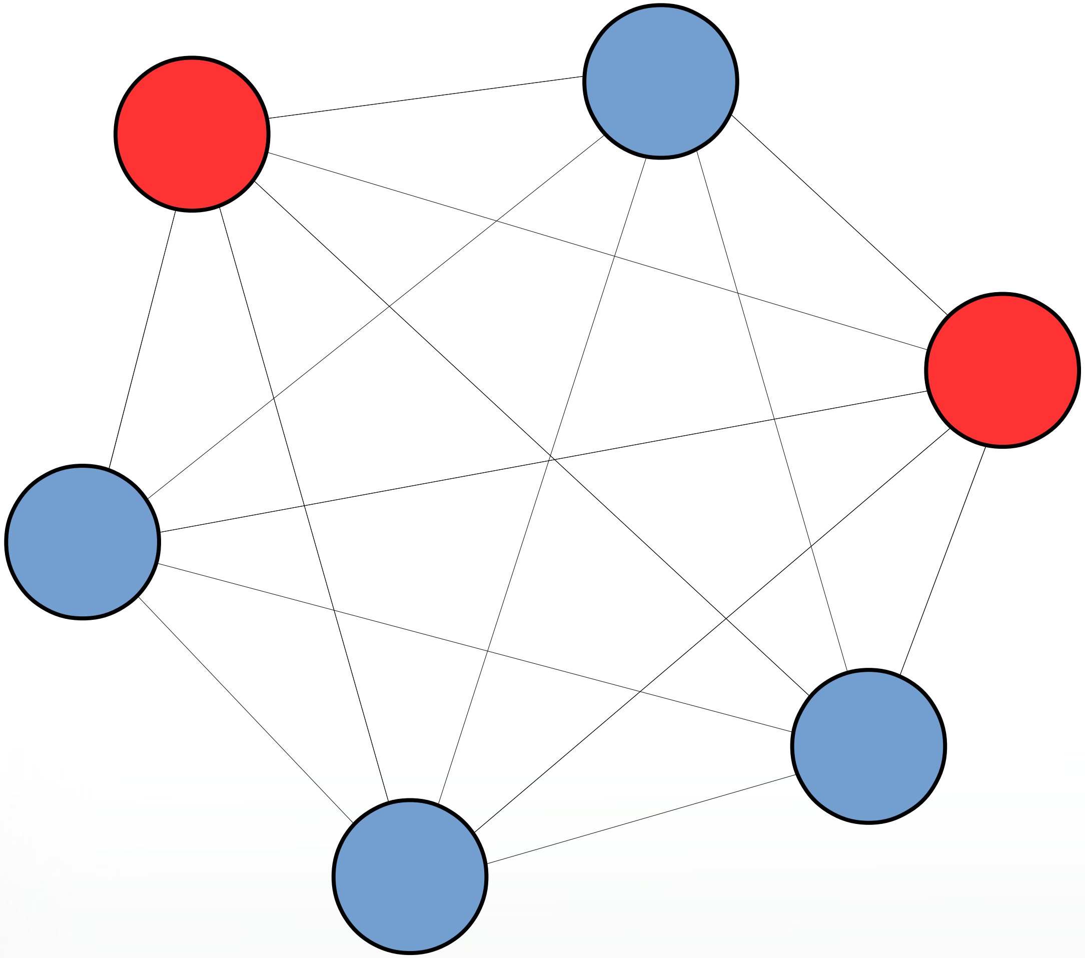


Challenges in GRN inference

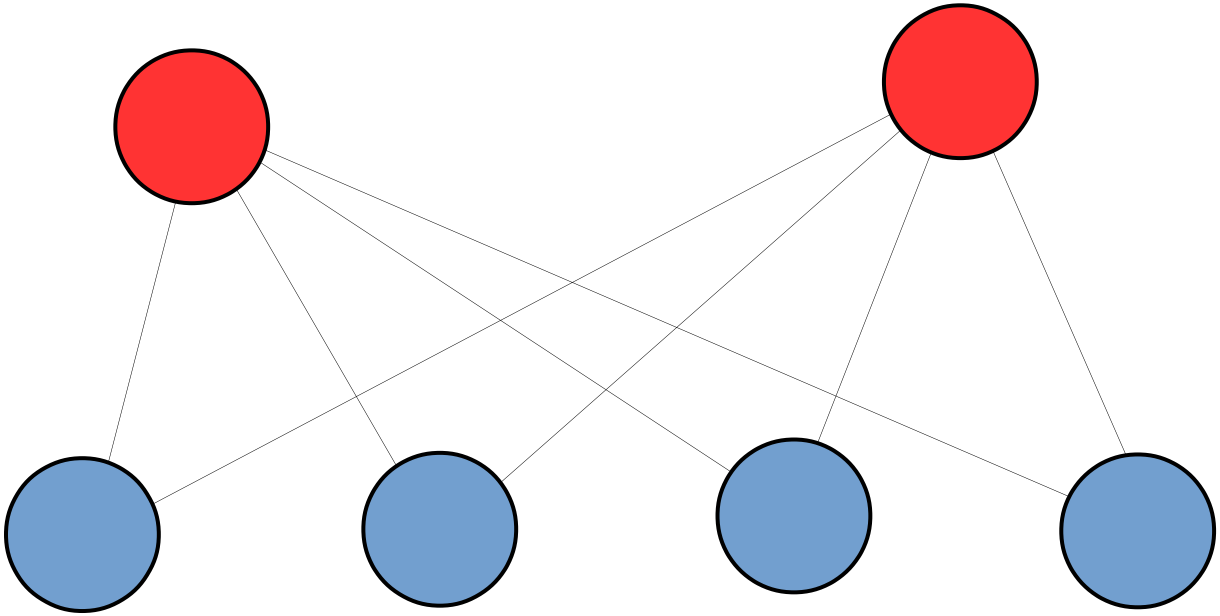


(p choose 2) edges

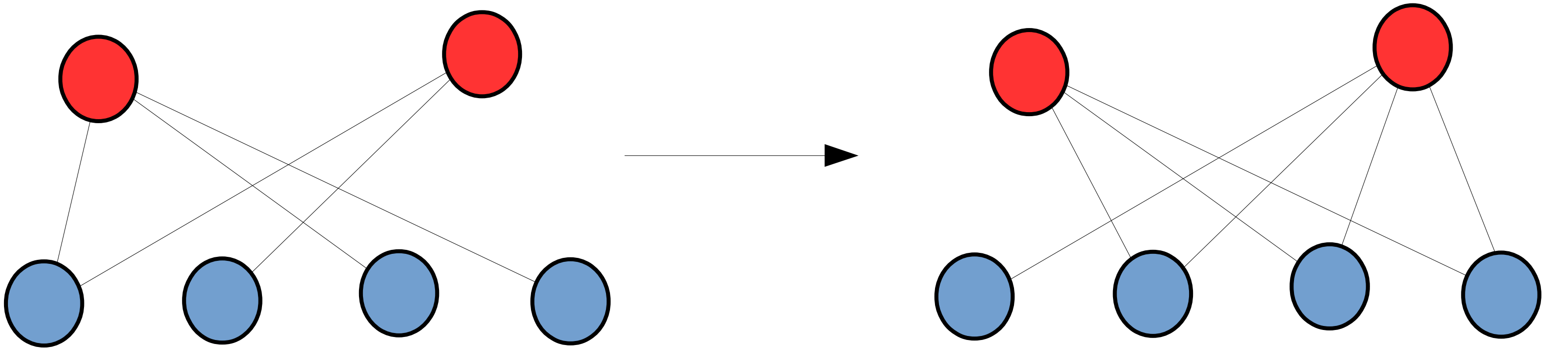
Challenges in GRN inference



Challenges in GRN inference



The network differentiation problem



Background:

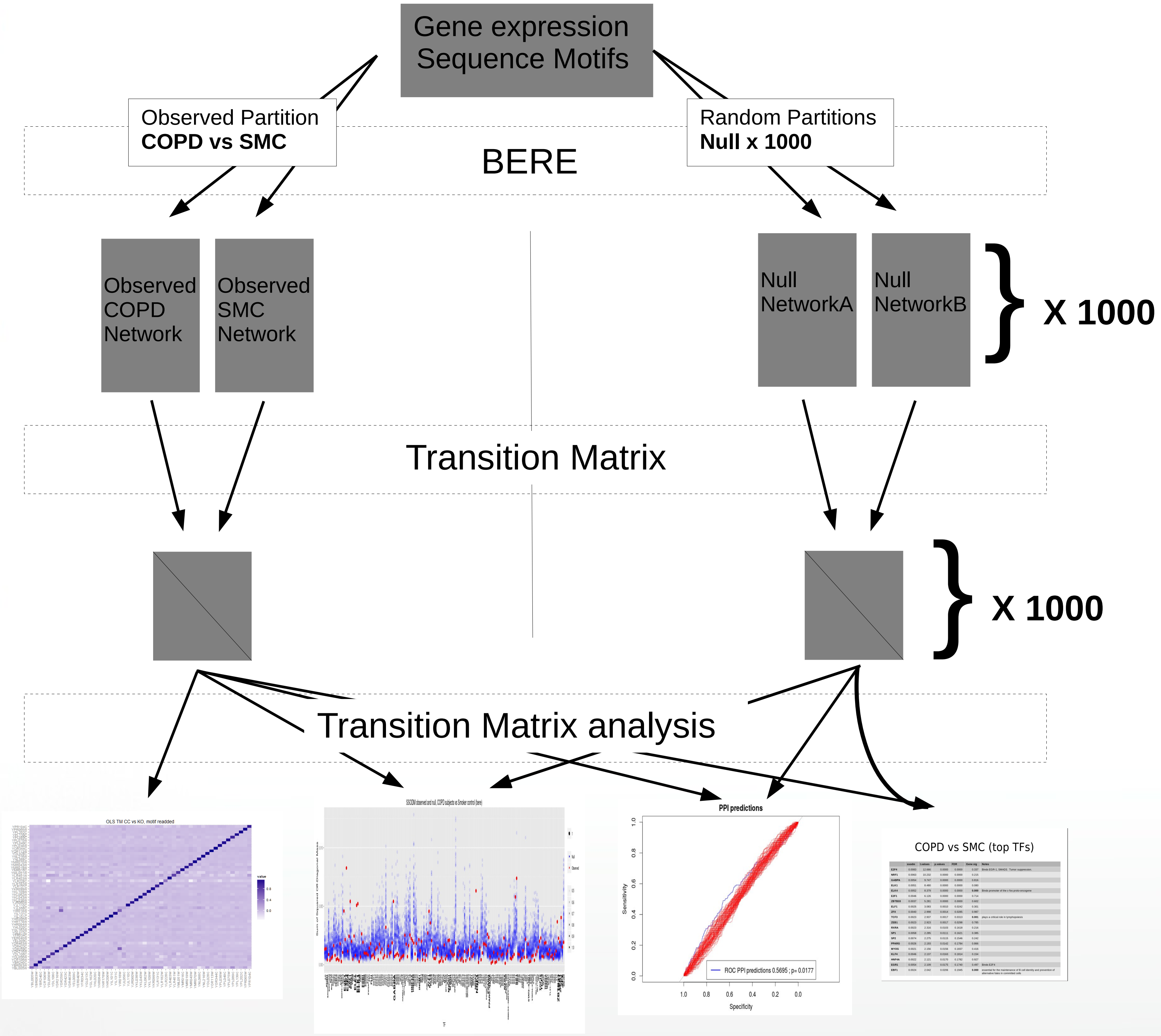
- Transcription factors may behave in different ways in different contexts.
- The targeted set of genes are defined by post-translational factors not measured by gene expression.
- These changes in “involvement” may not be readily observed using standard differential gene expression analyses.

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 1000

BERE

Observed COPD Network

Observed SMC Network

Null NetworkA

Null NetworkB

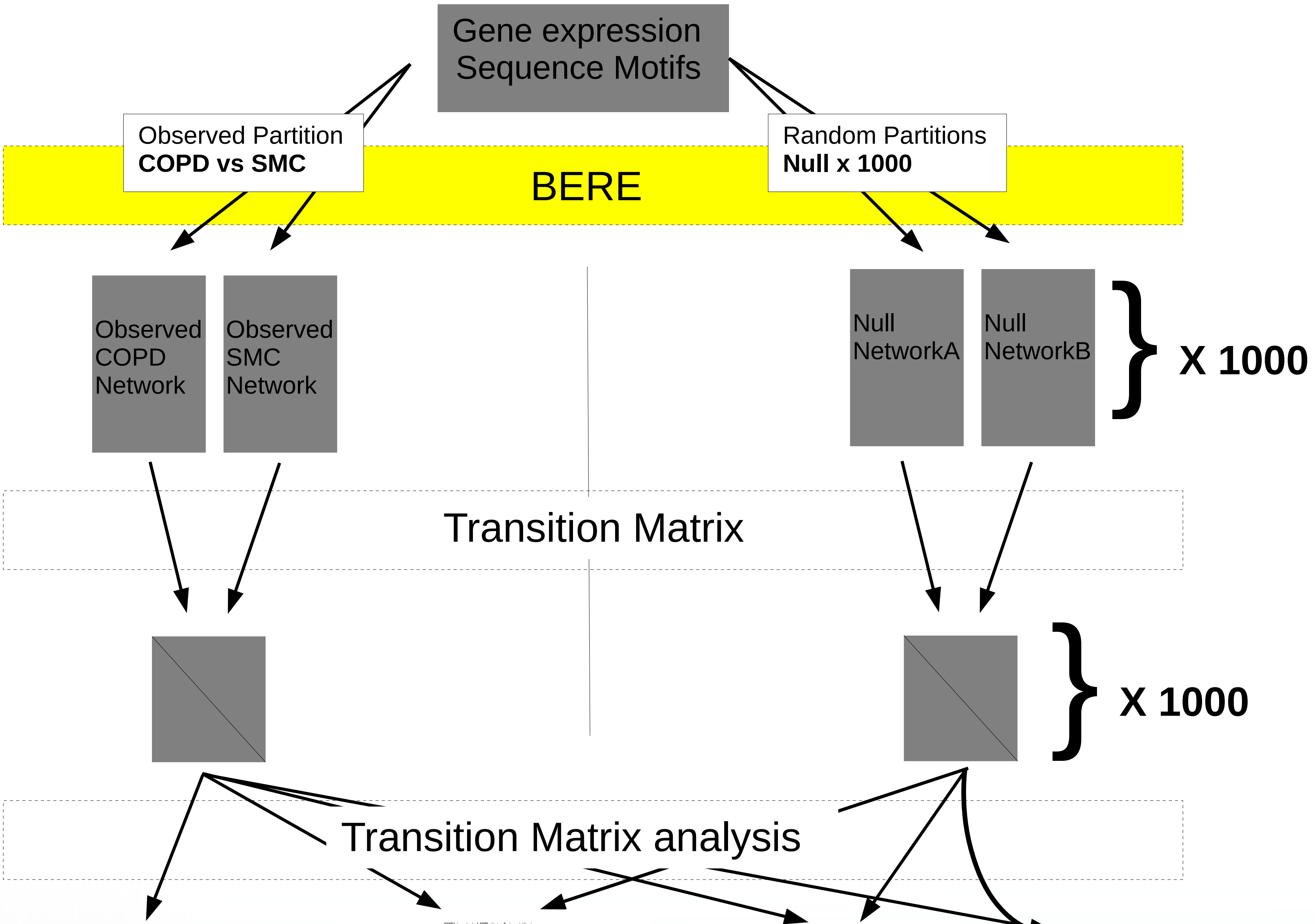
X 1000

Transition Matrix

X 1000

Transition Matrix analysis

OLS TM CC vs KO, motif readded
value
SSODM observed and null, COPD subjects vs Smoker control (bere)
Sum of Squared Off-Diagonal Mass
TF
Null
Observed
PPI predictions
Sensitivity
Specificity
ROC PPI predictions 0.5695 ; p= 0.0177
COPD vs SMC (top TFs)



Gene expression
Sequence Motifs

Observed Partition
COPD vs SMC

Random Partitions
Null x 1000

BERE

Observed COPD Network

Observed SMC Network

Null NetworkA

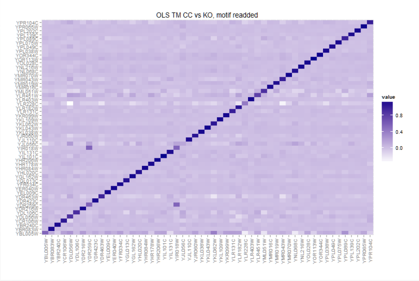
Null NetworkB

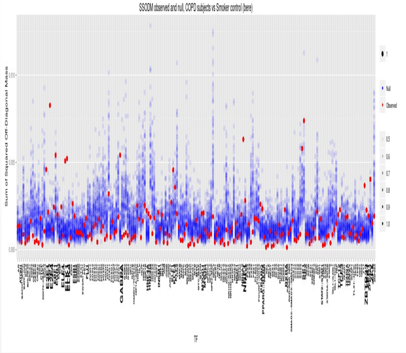
X 1000

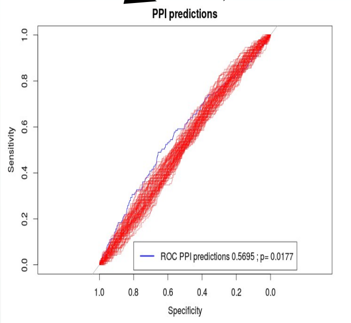
Transition Matrix

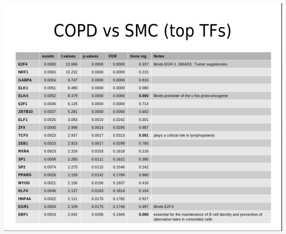
X 1000

Transition Matrix analysis





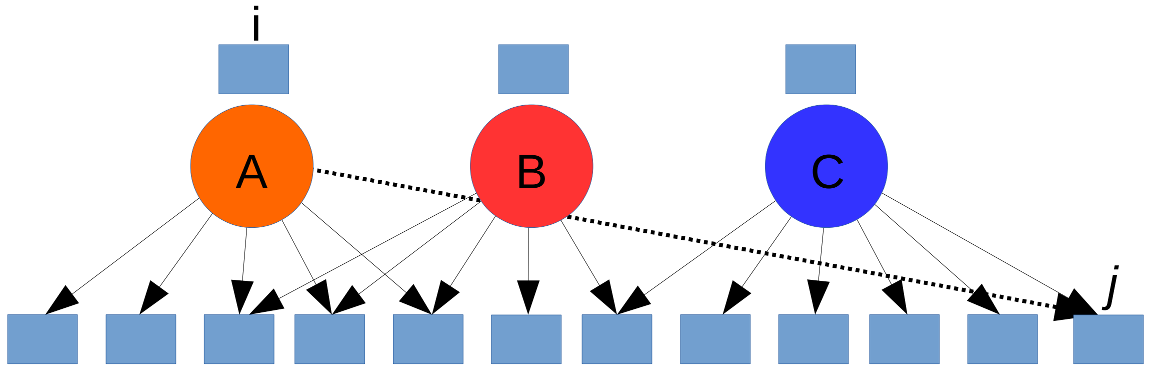


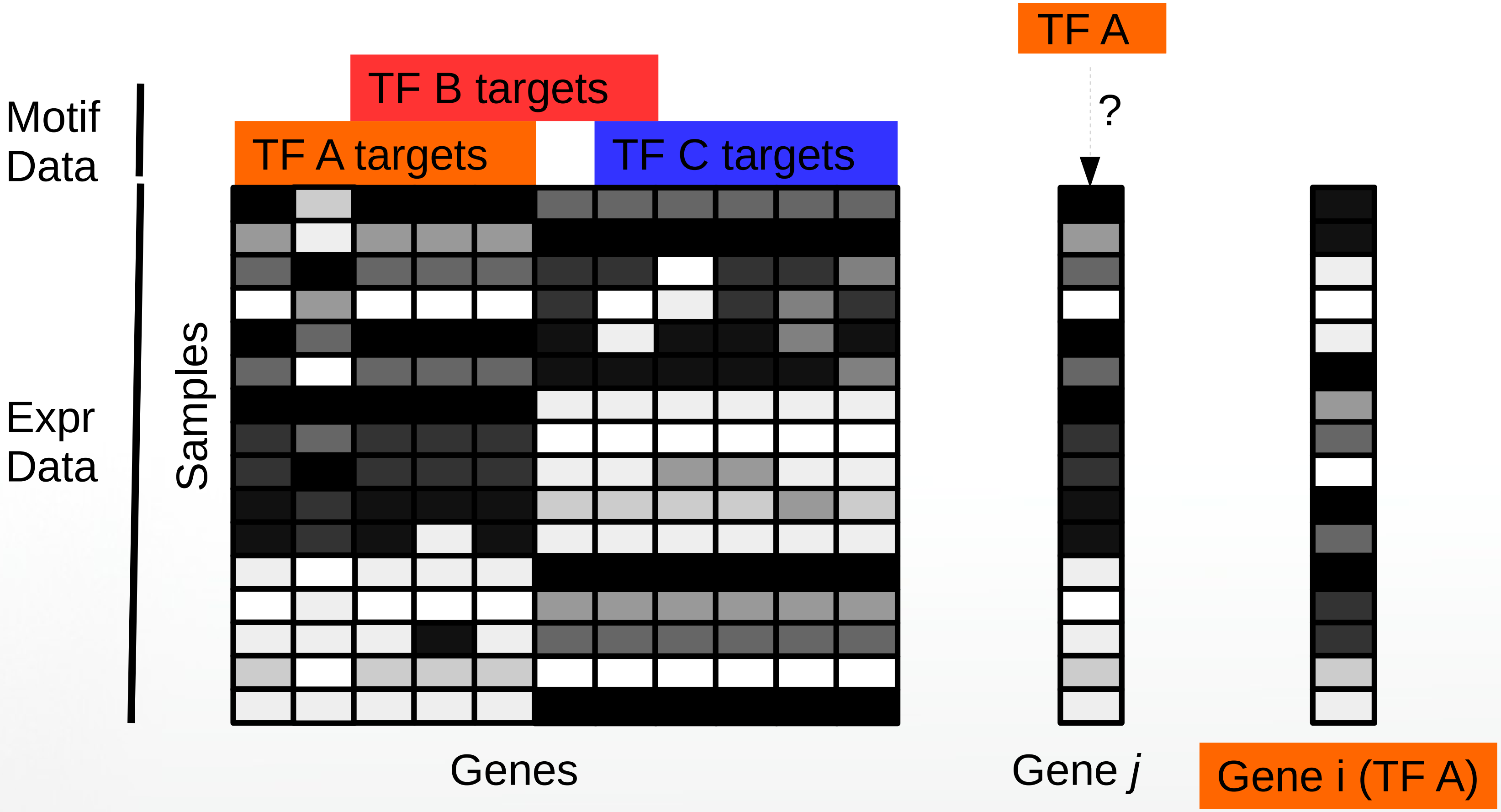


COPD vs SMC (top TFs)

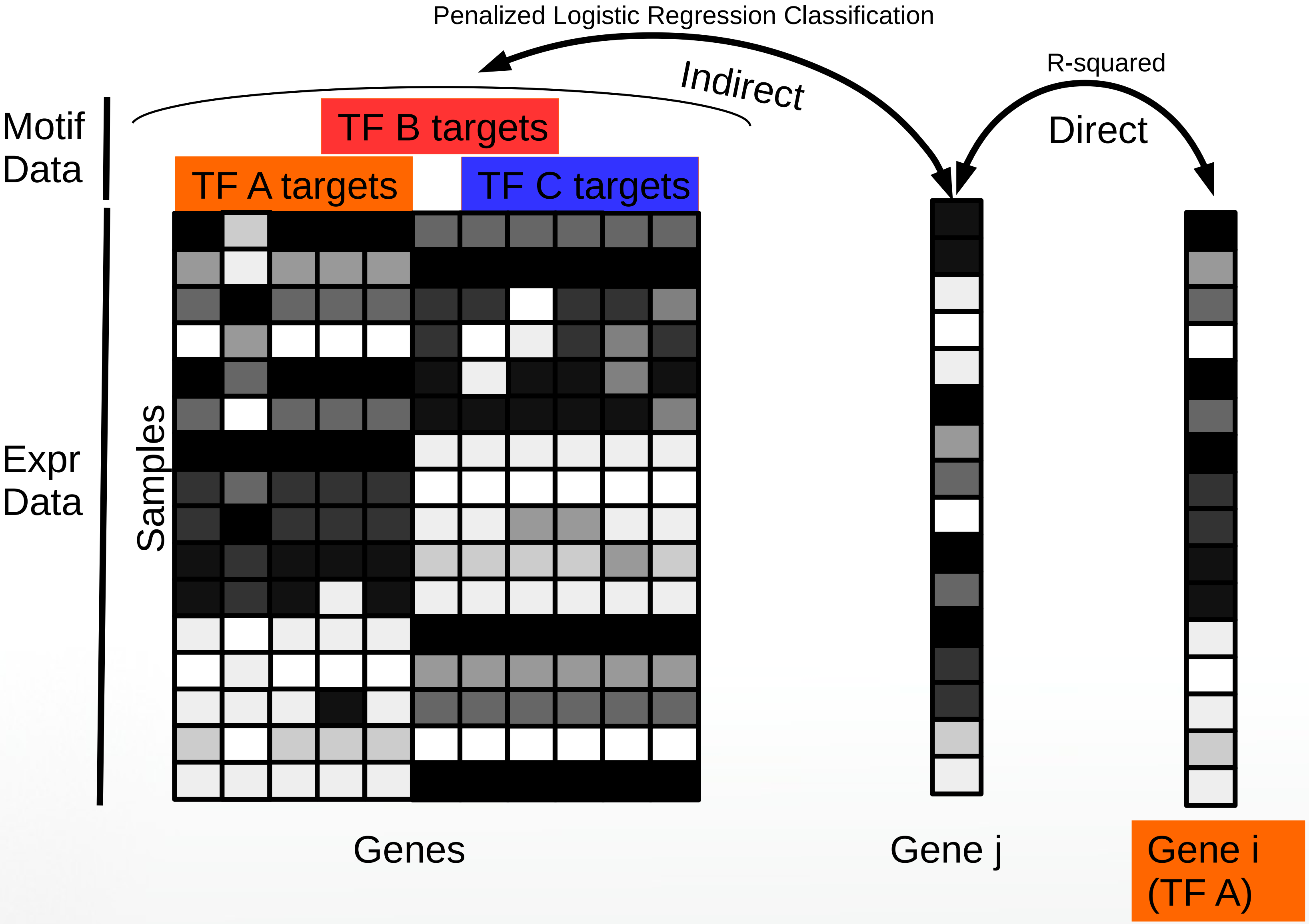
	ssodm	t.values	p.values	FDR	Gene sig	Notes
E2F4	0.0083	12.666	0.0000	0.0000	0.337	Binds EGR-1, SMAD3. Tumor suppression.
NRF1	0.0063	10.232	0.0000	0.0000	0.215	
GABPA	0.0054	9.747	0.0000	0.0000	0.816	
ELK1	0.0051	8.480	0.0000	0.0000	0.080	
ELK4	0.0052	8.379	0.0000	0.0000	**0.000**	Binds promoter of the c-fos proto-oncogene
E2F1	0.0046	6.126	0.0000	0.0000	0.714	
ZBTB33	0.0037	5.281	0.0000	0.0000	0.602	
ELF1	0.0025	3.083	0.0010	0.0242	0.301	
ZFX	0.0040	2.998	0.0014	0.0285	0.987	
TCF3	0.0023	2.937	0.0017	0.0313	**0.001**	plays a critical role in lymphopoiesis
ZEB1	0.0023	2.923	0.0017	0.0298	0.785	
RXRA	0.0023	2.316	0.0103	0.1618	0.216	
SP1	0.0058	2.285	0.0111	0.1621	0.385	
SP2	0.0074	2.275	0.0115	0.1546	0.242	
PPARG	0.0026	2.193	0.0142	0.1784	0.866	
MYOG	0.0021	2.156	0.0156	0.1837	0.416	
KLF4	0.0046	2.137	0.0163	0.1814	0.154	
HNF4A	0.0022	2.121	0.0170	0.1782	0.927	
EGR1	0.0054	2.109	0.0175	0.1740	0.497	Binds E2F4
EBF1	0.0024	2.042	0.0206	0.1945	**0.000**	essential for the maintenance of B cell identity and prevention of alternative fates in committed cells

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





BERE


Penalized Logistic Regression Classification
Indirect
R-squared
Direct
Motif
Data
TF B targets
TF A targets
TF C targets
Expr
Data
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.

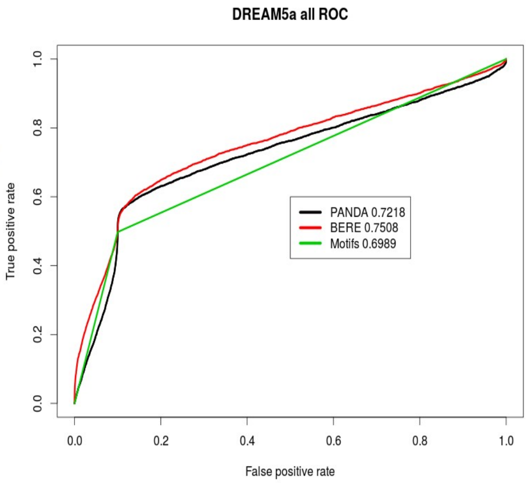
2.) Consider the direct evidence of regulation.
The squared conditional coexpression of gene i and gene j, where gene i is a transcription factor.

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.

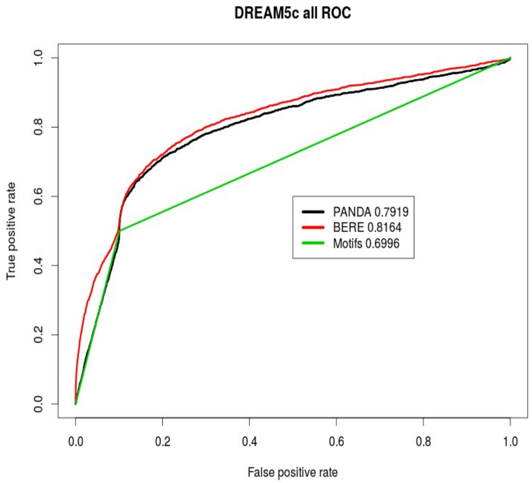
4.) Combine indirect and direct evidence into a score for network edgeweights.

BERE

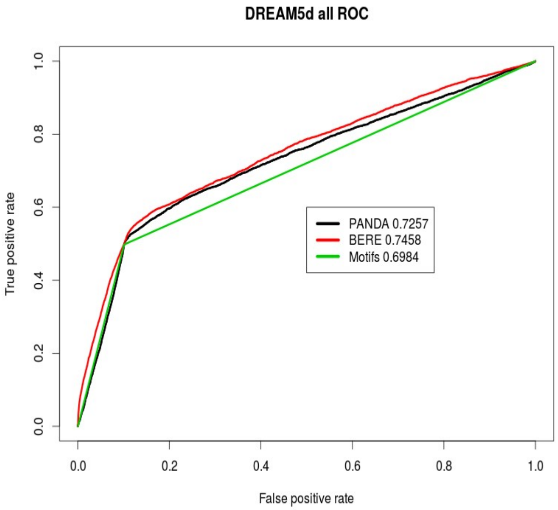
In Silico



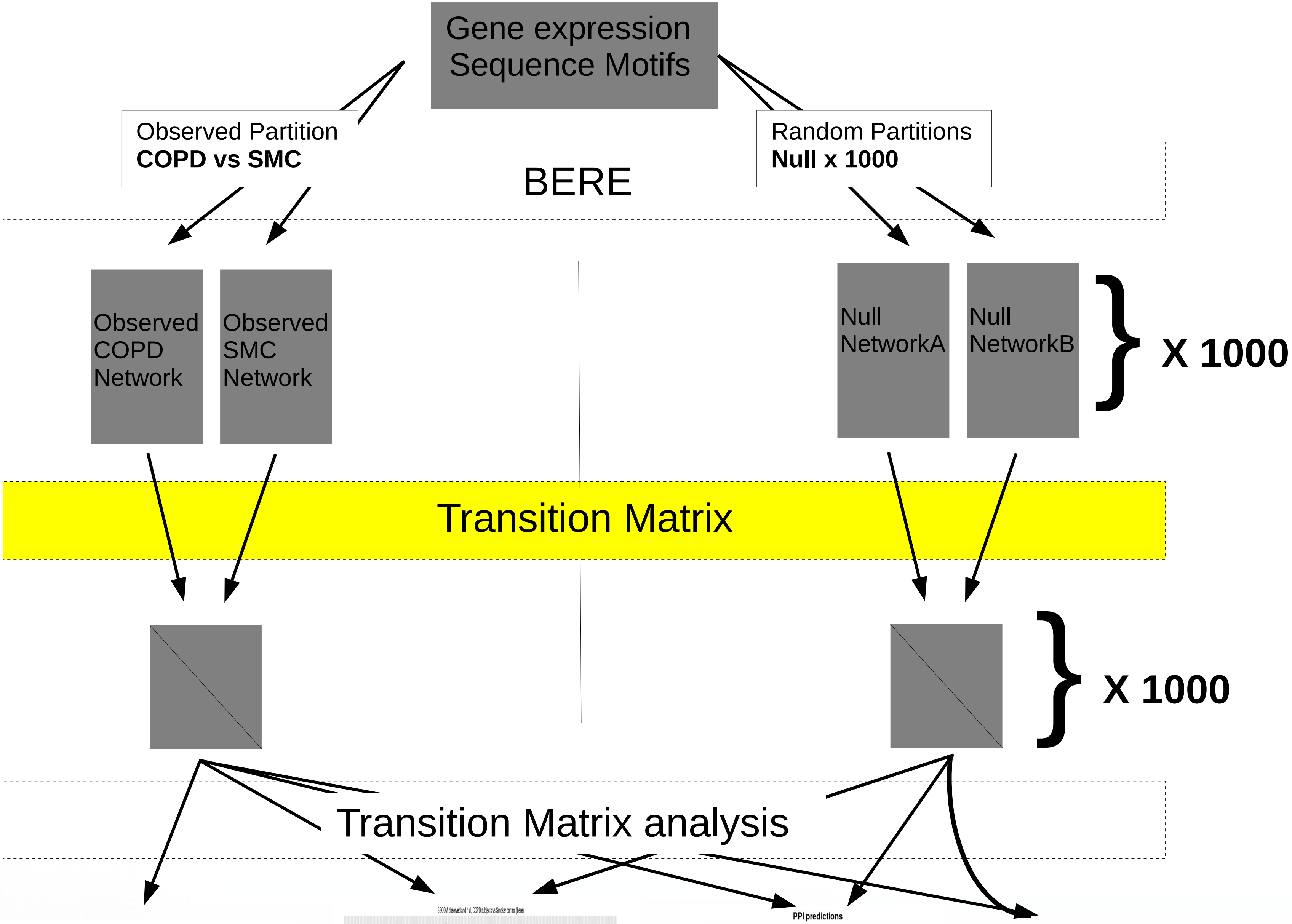
E. coli

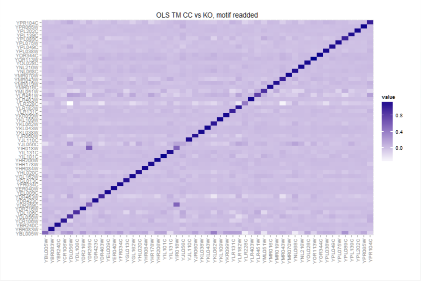


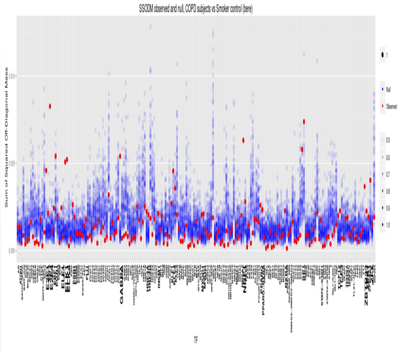
Yeast

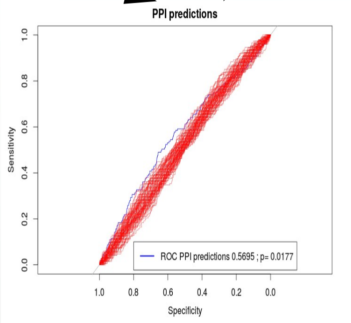


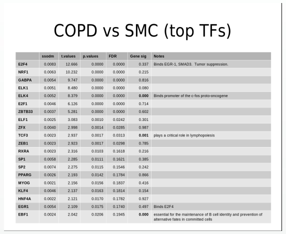
Running R package: 8GB RAM, 2.40Ghz	Time
2555 genes, 53 TF, 106 samples	11s
17342 genes, 189 TF, 226 samples	12m, 20s







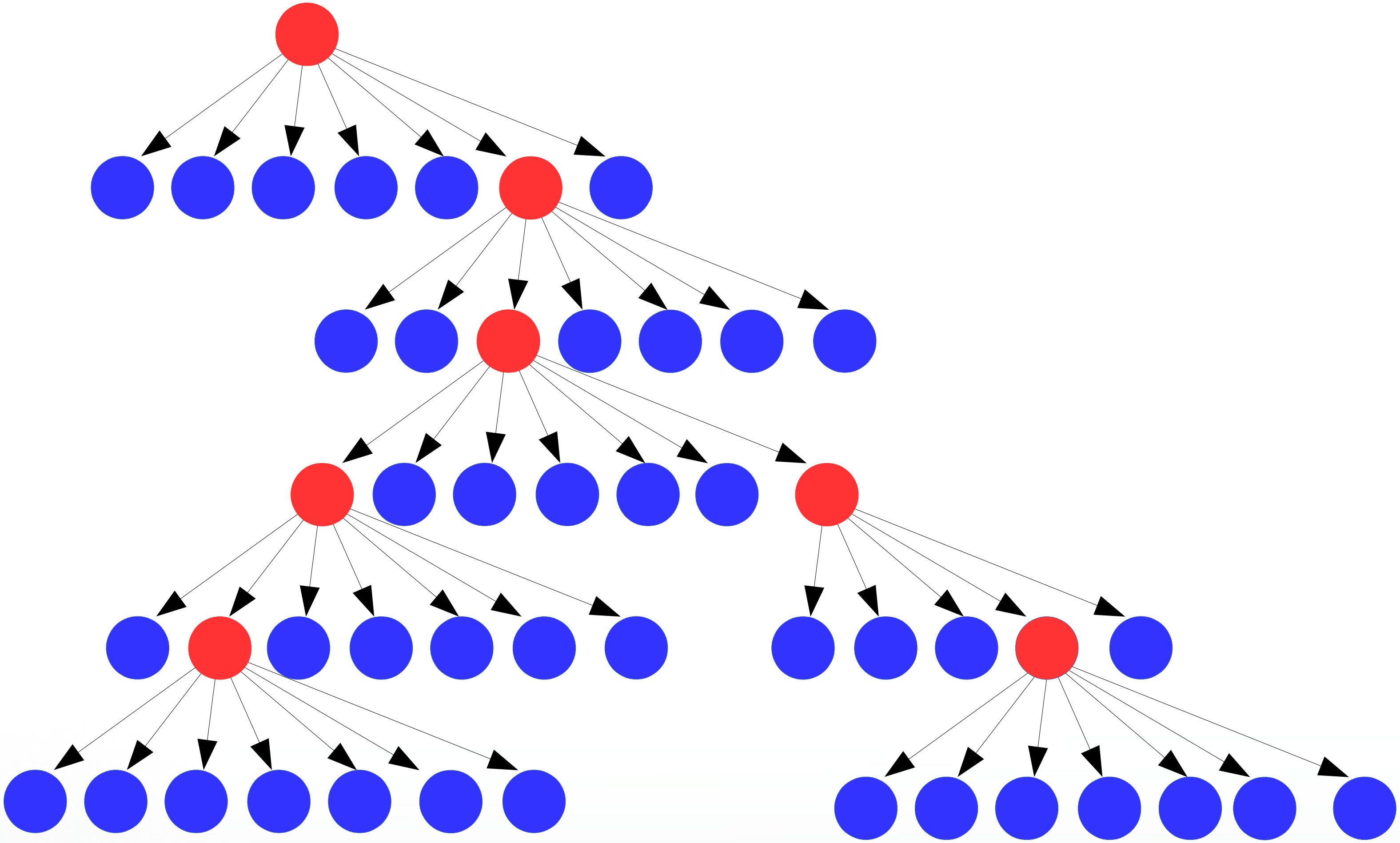




COPD vs SMC (top TFs)

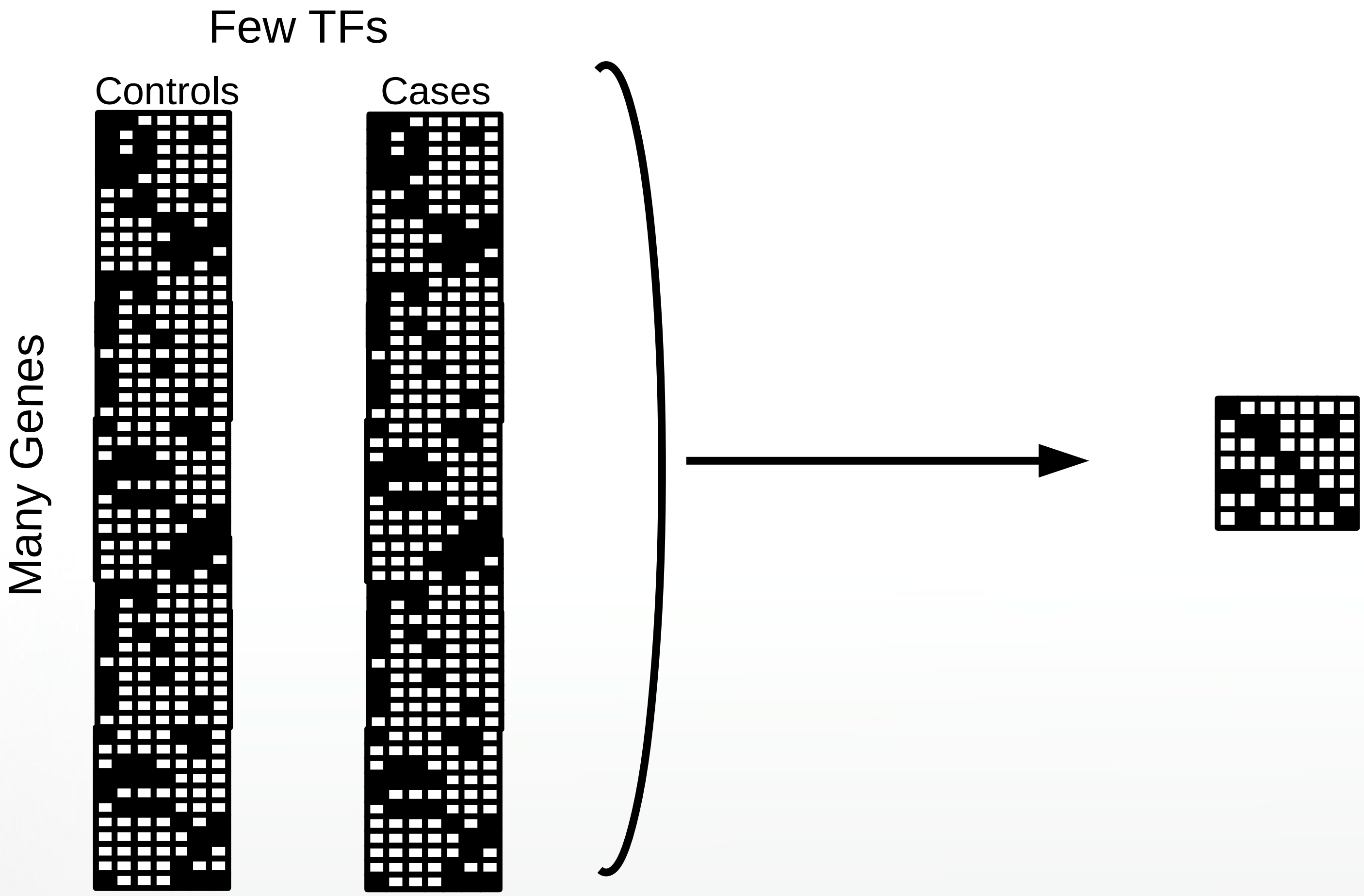
	ssodm	t.values	p.values	FDR	Gene sig	Notes
E2F4	0.0083	12.666	0.0000	0.0000	0.337	Binds EGR-1, SMAD3. Tumor suppression.
NRF1	0.0063	10.232	0.0000	0.0000	0.215	
GABPA	0.0054	9.747	0.0000	0.0000	0.816	
ELK1	0.0051	8.480	0.0000	0.0000	0.080	
ELK4	0.0052	8.379	0.0000	0.0000	**0.000**	Binds promoter of the c-fos proto-oncogene
E2F1	0.0046	6.126	0.0000	0.0000	0.714	
ZBTB33	0.0037	5.281	0.0000	0.0000	0.602	
ELF1	0.0025	3.083	0.0010	0.0242	0.301	
ZFX	0.0040	2.998	0.0014	0.0285	0.987	
TCF3	0.0023	2.937	0.0017	0.0313	**0.001**	plays a critical role in lymphopoiesis
ZEB1	0.0023	2.923	0.0017	0.0298	0.785	
RXRA	0.0023	2.316	0.0103	0.1618	0.216	
SP1	0.0058	2.285	0.0111	0.1621	0.385	
SP2	0.0074	2.275	0.0115	0.1546	0.242	
PPARG	0.0026	2.193	0.0142	0.1784	0.866	
MYOG	0.0021	2.156	0.0156	0.1837	0.416	
KLF4	0.0046	2.137	0.0163	0.1814	0.154	
HNF4A	0.0022	2.121	0.0170	0.1782	0.927	
EGR1	0.0054	2.109	0.0175	0.1740	0.497	Binds E2F4
EBF1	0.0024	2.042	0.0206	0.1945	**0.000**	essential for the maintenance of B cell identity and prevention of alternative fates in committed cells

A Pathway model



Transition Matrix Approach

We can view the problem as a dimension reduction problem.



Few TFs

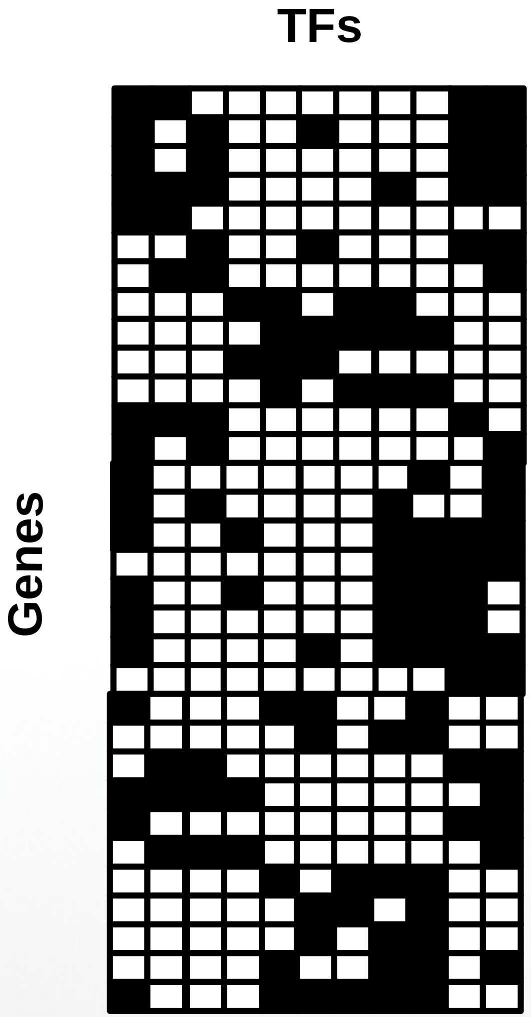
Controls

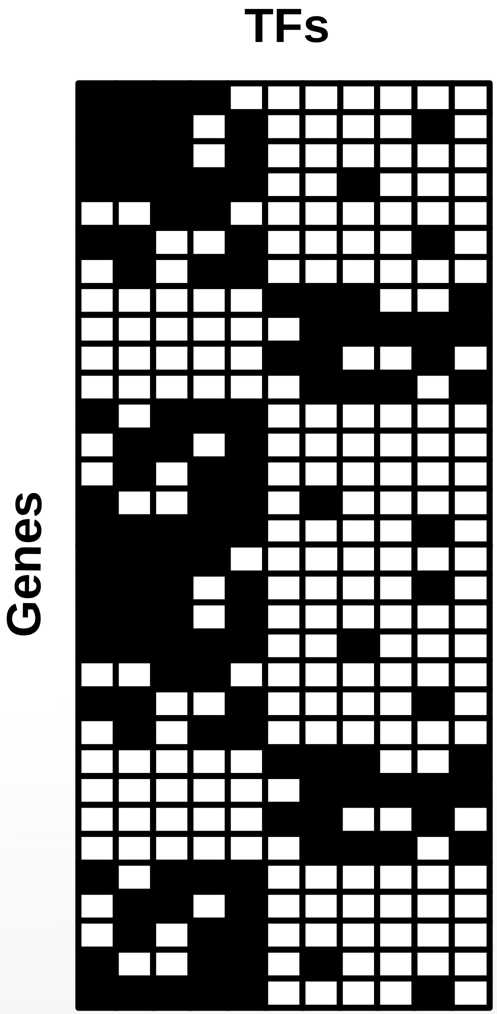
Cases

Many Genes

Transition Matrix Approach

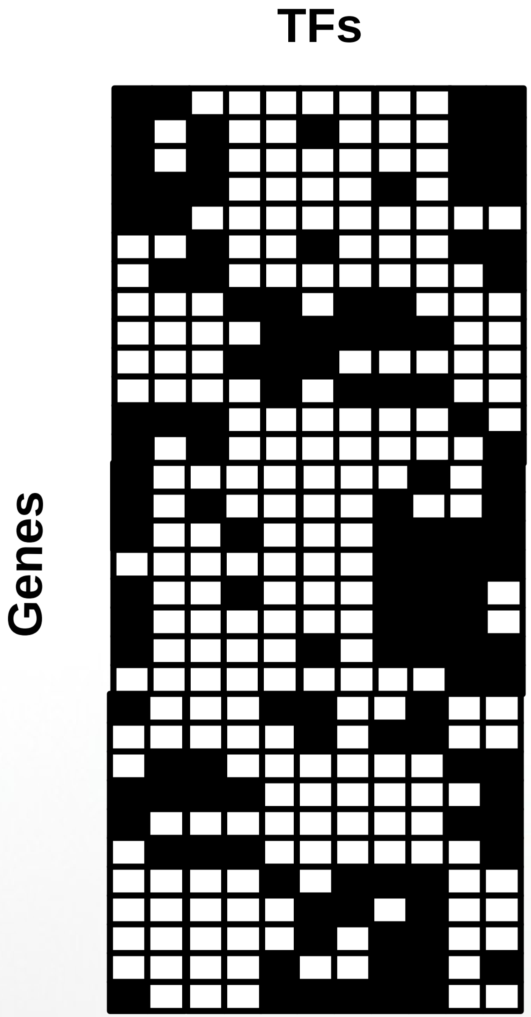
Consider two adjacency matrices...

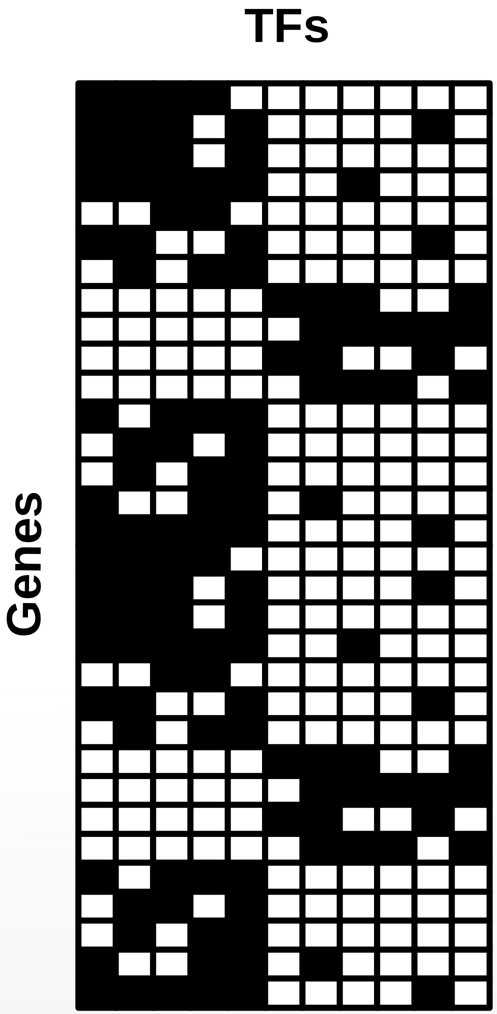




Transition Matrix Approach

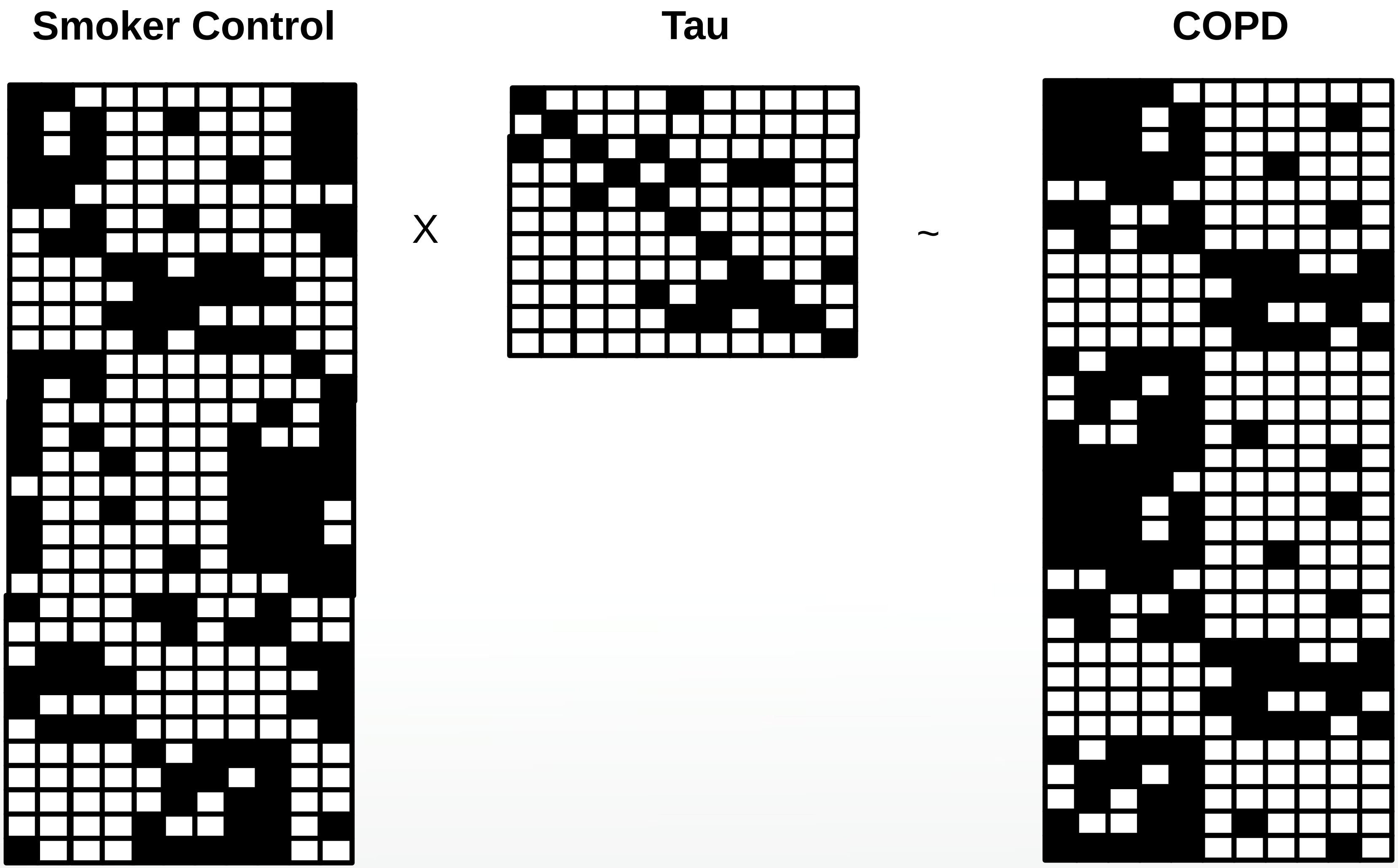
Consider two adjacency matrices...





Transition Matrix Approach

Consider two adjacency matrices...



The Transition Matrix (Tau)

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

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

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

The Transition Matrix (Tau)

Interpretation: 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.

Reasoning: We want to focus on changes in targeting behavior of a TF in terms of biologically recognized alternative targets.

Constraints:

- 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)

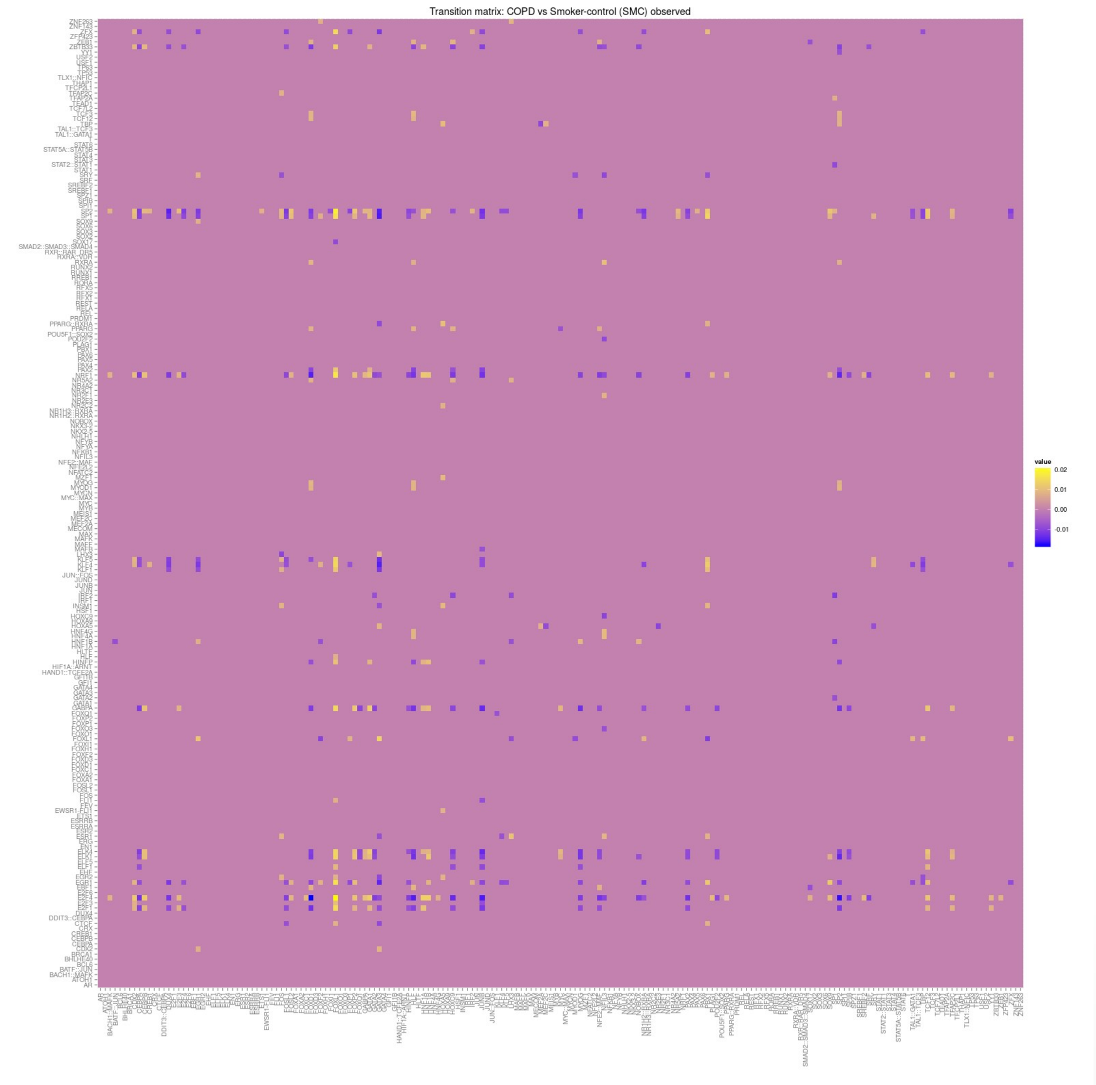
- We can satisfy these properties with an L_1 regularization.
  - For a column, *k*, we perform the following error minimization.

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

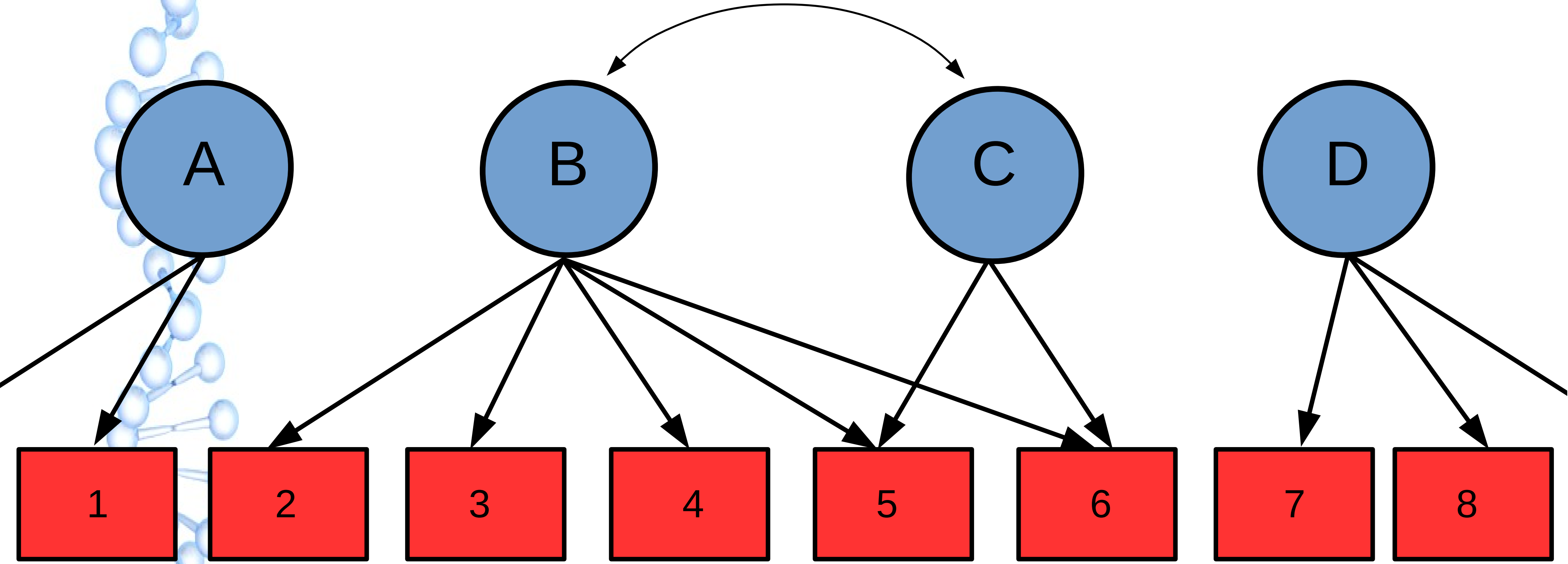
$$\mathbf{Q}_{i,j}=\begin{cases}1 & for\, i=j\neq k\\ 0 & elsewhere\end{cases}$$

- 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

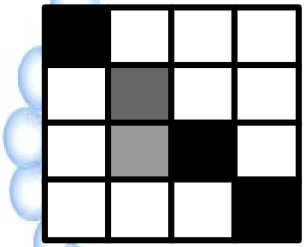


Control Network

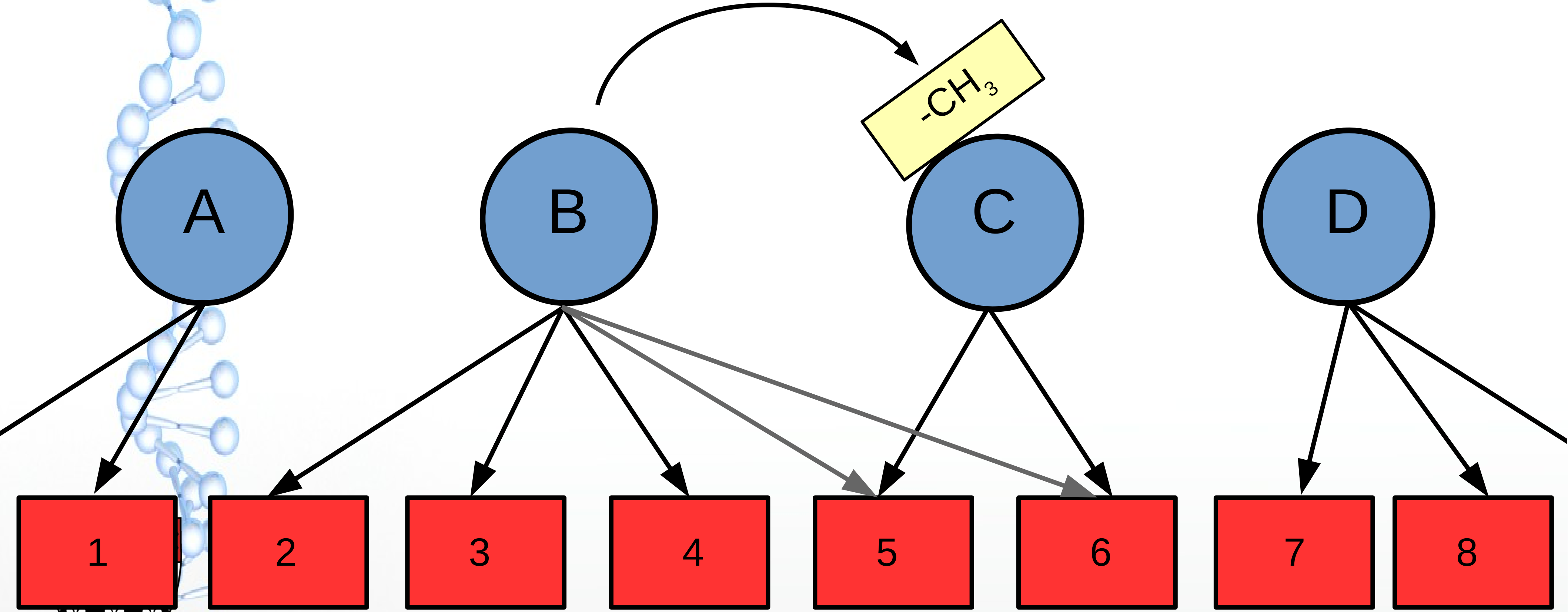
Cases Network

$$TM = \begin{array}{c|cccc} & A & B & C & D \\ \hline A & \text{■} & & & \\ B & & \text{▓} & & \\ C & & \text{▒} & \text{■} & \\ D & & & & \text{■} \end{array}$$

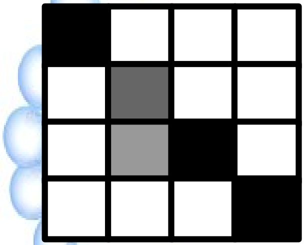
Biological Mechanism



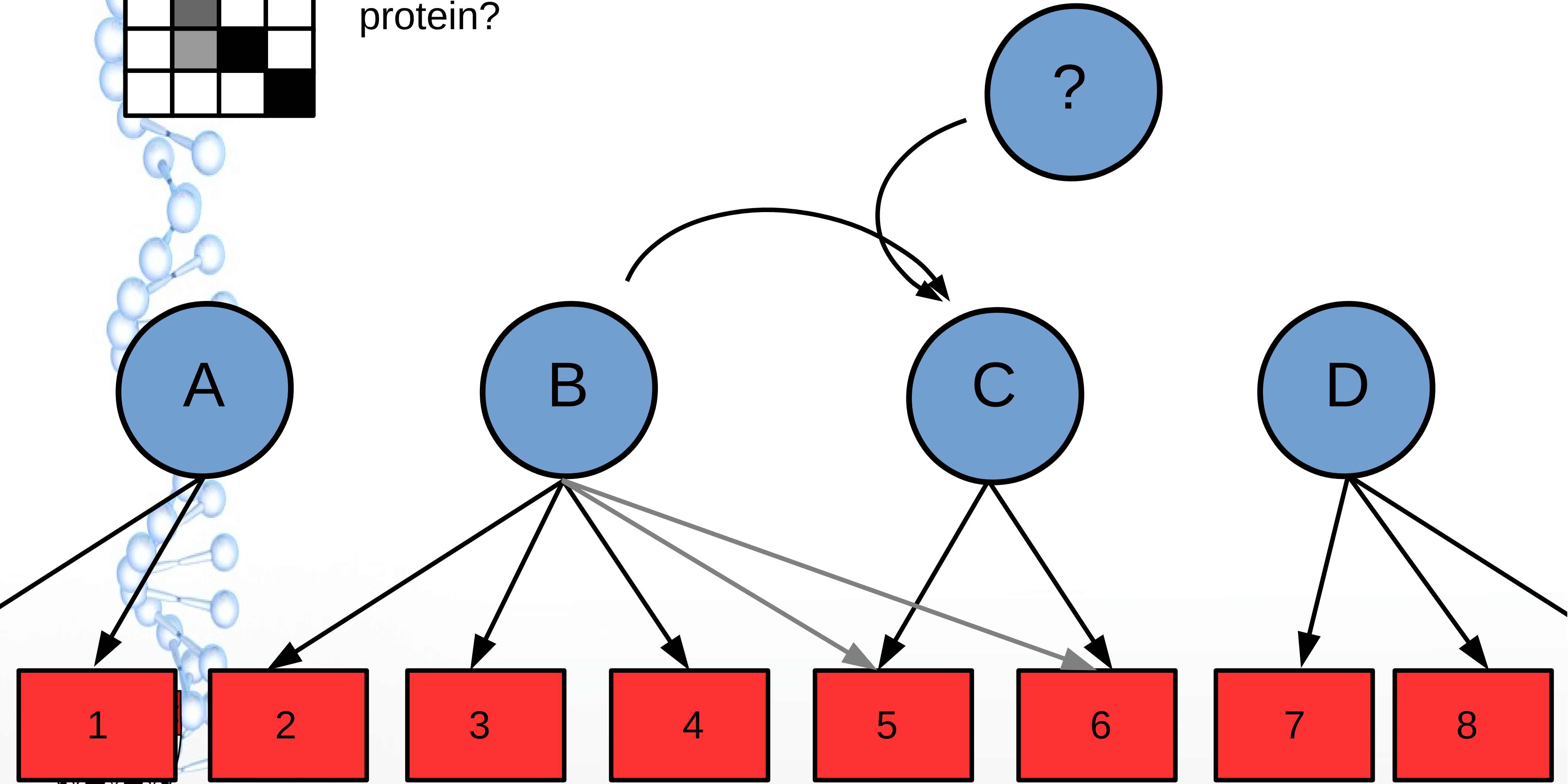
Suggested mechanism #1:
Differential methylation of the gene for TF C?

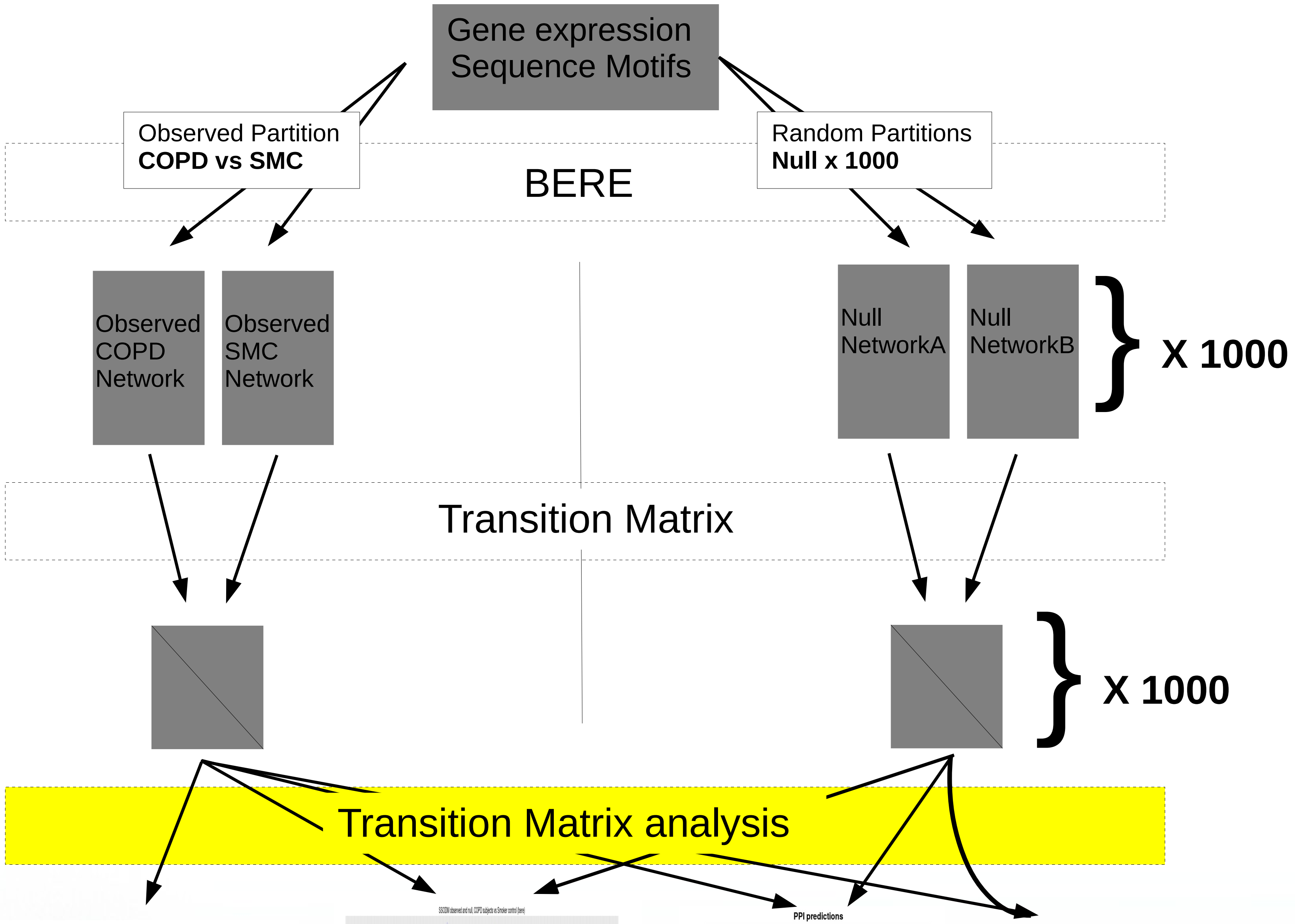


Biological Mechanism



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 1000

BERE

Observed COPD Network

Observed SMC Network

Null NetworkA

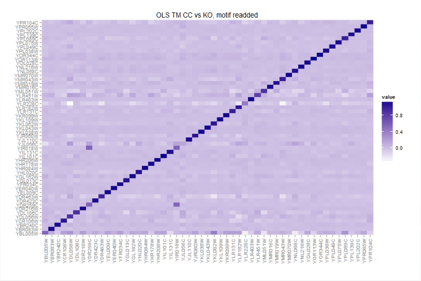
Null NetworkB

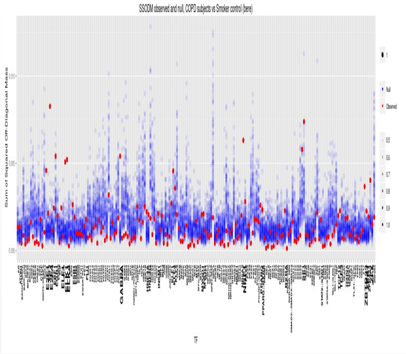
X 1000

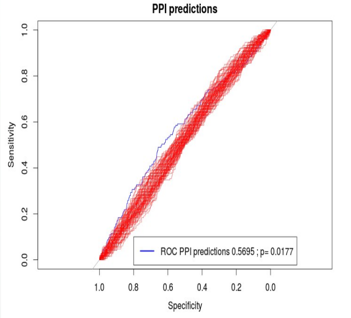
Transition Matrix

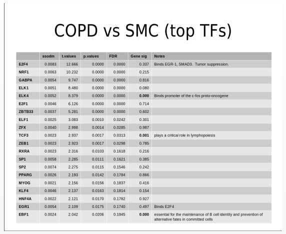
X 1000

Transition Matrix analysis








COPD vs SMC (top TFs)

	ssodm	t.values	p.values	FDR	Gene sig	Notes
E2F4	0.0083	12.666	0.0000	0.0000	0.337	Binds EGR-1, SMAD3. Tumor suppression.
NRF1	0.0063	10.232	0.0000	0.0000	0.215	
GABPA	0.0054	9.747	0.0000	0.0000	0.816	
ELK1	0.0051	8.480	0.0000	0.0000	0.080	
ELK4	0.0052	8.379	0.0000	0.0000	**0.000**	Binds promoter of the c-fos proto-oncogene
E2F1	0.0046	6.126	0.0000	0.0000	0.714	
ZBTB33	0.0037	5.281	0.0000	0.0000	0.602	
ELF1	0.0025	3.083	0.0010	0.0242	0.301	
ZFX	0.0040	2.998	0.0014	0.0285	0.987	
TCF3	0.0023	2.937	0.0017	0.0313	**0.001**	plays a critical role in lymphopoiesis
ZEB1	0.0023	2.923	0.0017	0.0298	0.785	
RXRA	0.0023	2.316	0.0103	0.1618	0.216	
SP1	0.0058	2.285	0.0111	0.1621	0.385	
SP2	0.0074	2.275	0.0115	0.1546	0.242	
PPARG	0.0026	2.193	0.0142	0.1784	0.866	
MYOG	0.0021	2.156	0.0156	0.1837	0.416	
KLF4	0.0046	2.137	0.0163	0.1814	0.154	
HNF4A	0.0022	2.121	0.0170	0.1782	0.927	
EGR1	0.0054	2.109	0.0175	0.1740	0.497	Binds E2F4
EBF1	0.0024	2.042	0.0206	0.1945	**0.000**	essential for the maintenance of B cell identity and prevention of alternative fates in committed cells

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.

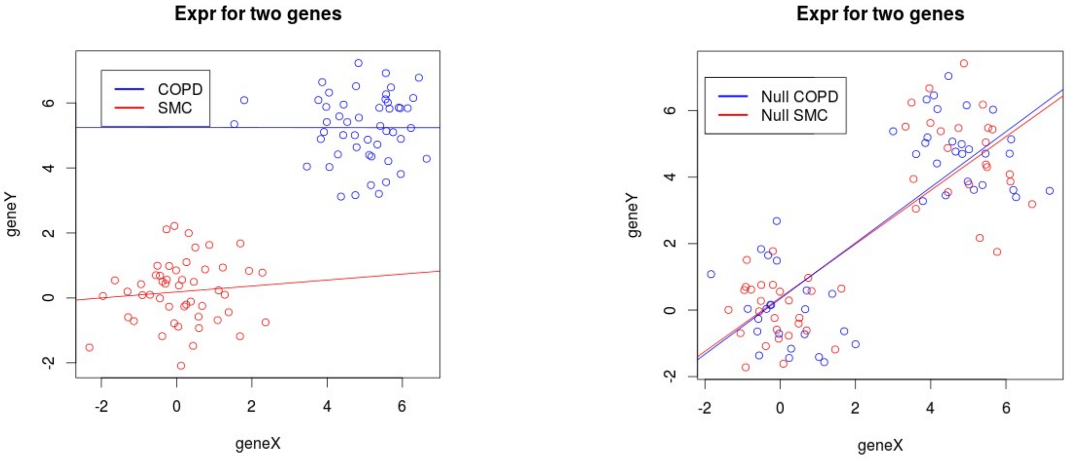
5. Repeat 1-4 1000 times.

Is this a valid estimation of the null distribution of dTFI?

Concern:

Variance of test statistic may be inflated.

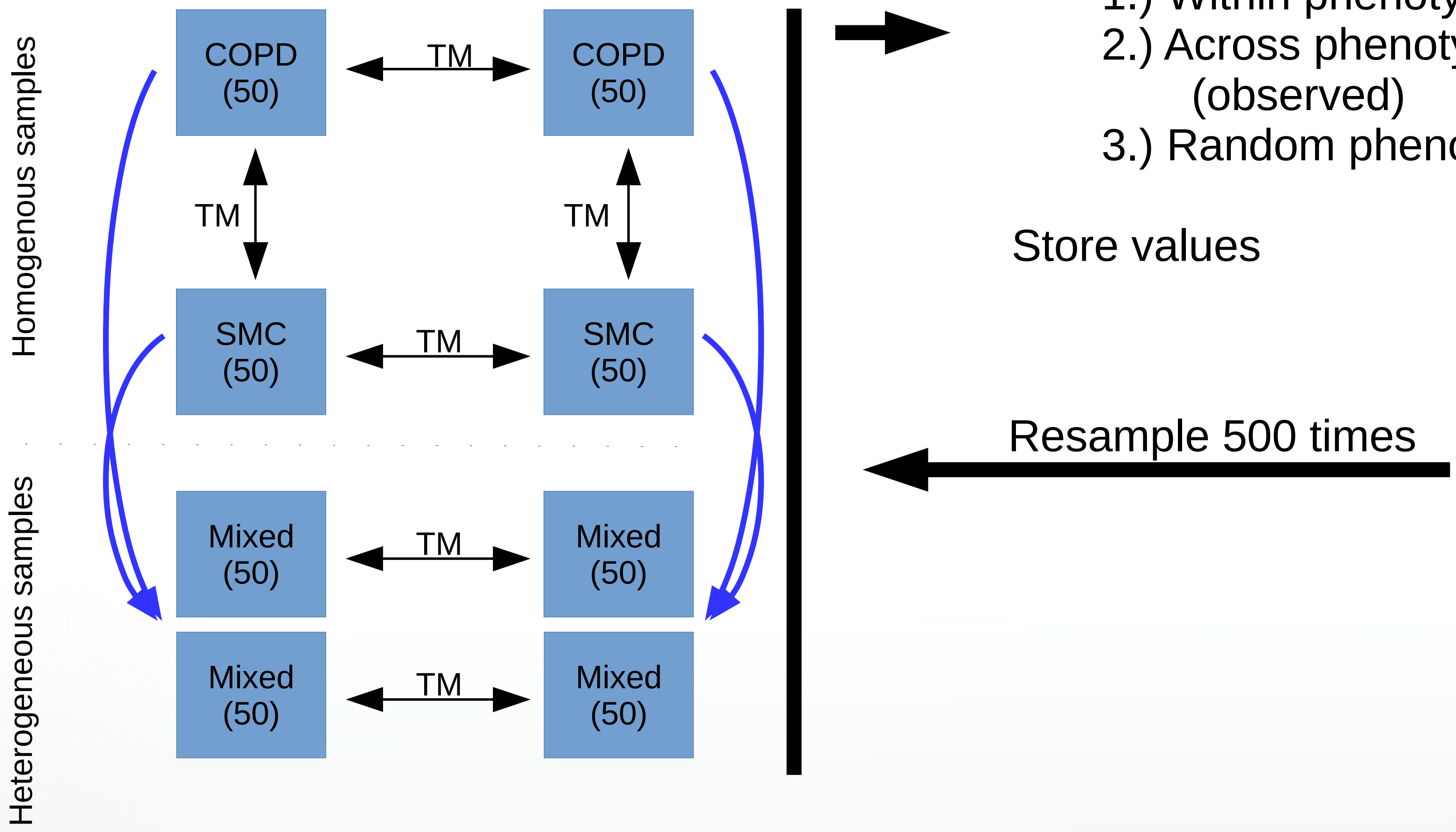
Example: Two highly correlated genes



- Power to detect interaction will be **greater** under the null
- Edges under the null may be **more stable**.
- Transition may be **less variable** under the null
- p-values for observed transition may become **smaller**,
  i.e. variance inflation

Is this a valid estimation of the null distribution of dTFI?

Randomly select 200 samples



Calculate dTFI:

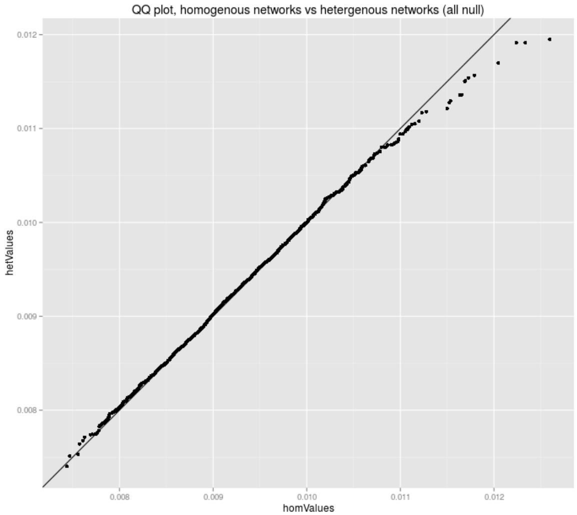
1.) Within phenotype
2.) Across phenotype
(observed)
3.) Random phenotype

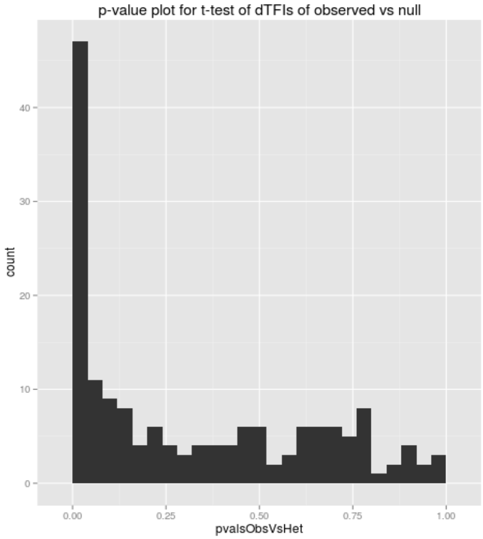
Store values

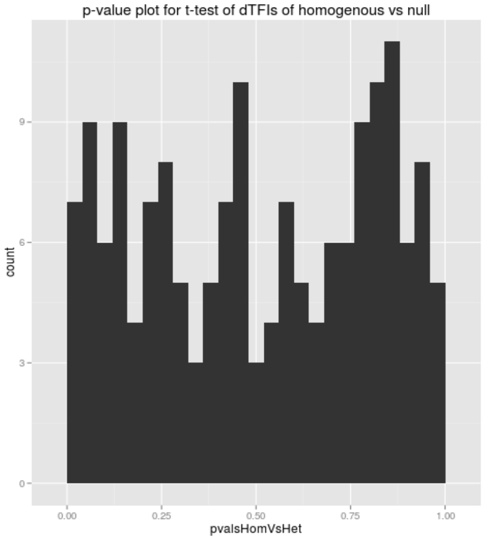
Resample 500 times

Is this a valid estimation of the null distribution of dTFI?

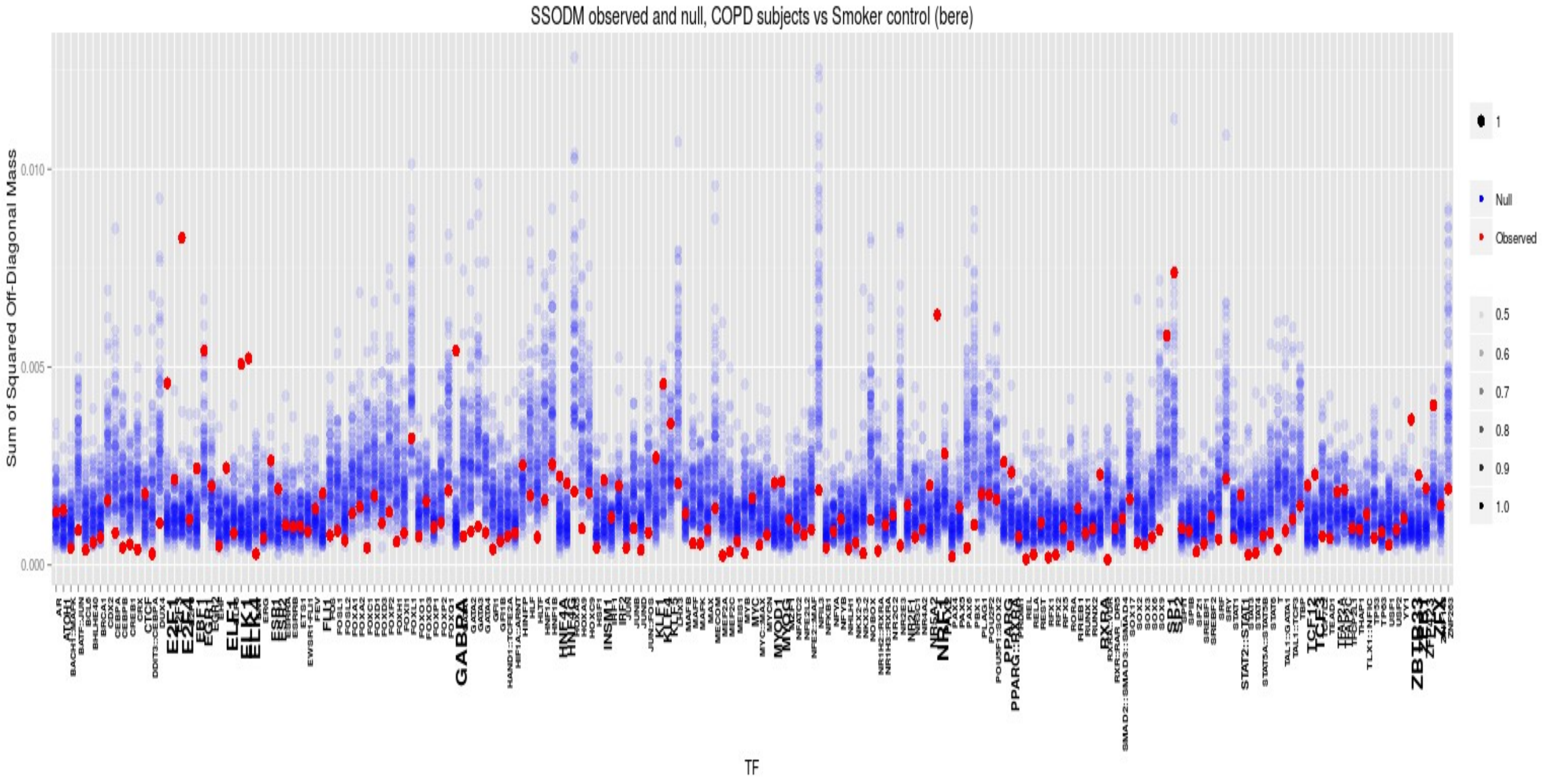
Yes! ... we think


QQ plot, homogenous networks vs hetergenous networks (all null)


p-value plot for t-test of dTFIs of observed vs null

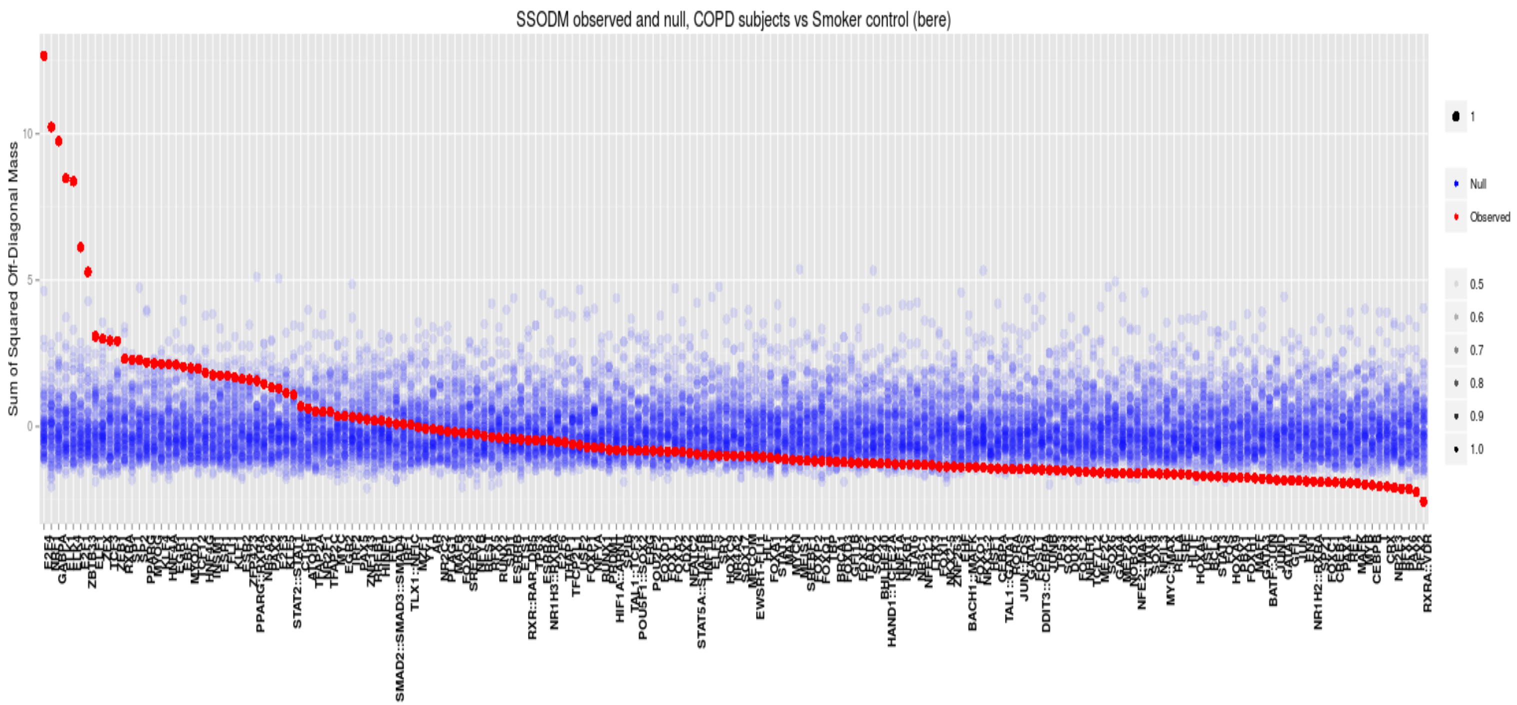

p-value plot for t-test of dTFIs of homogenous vs null

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.

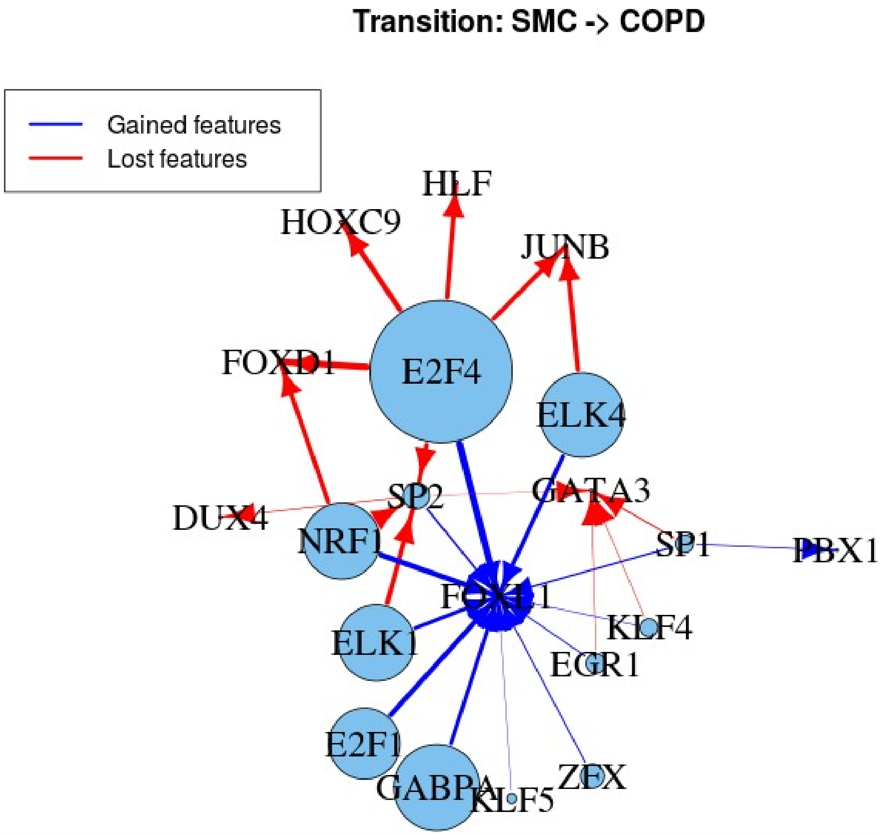
Application to a case-control COPD study

	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

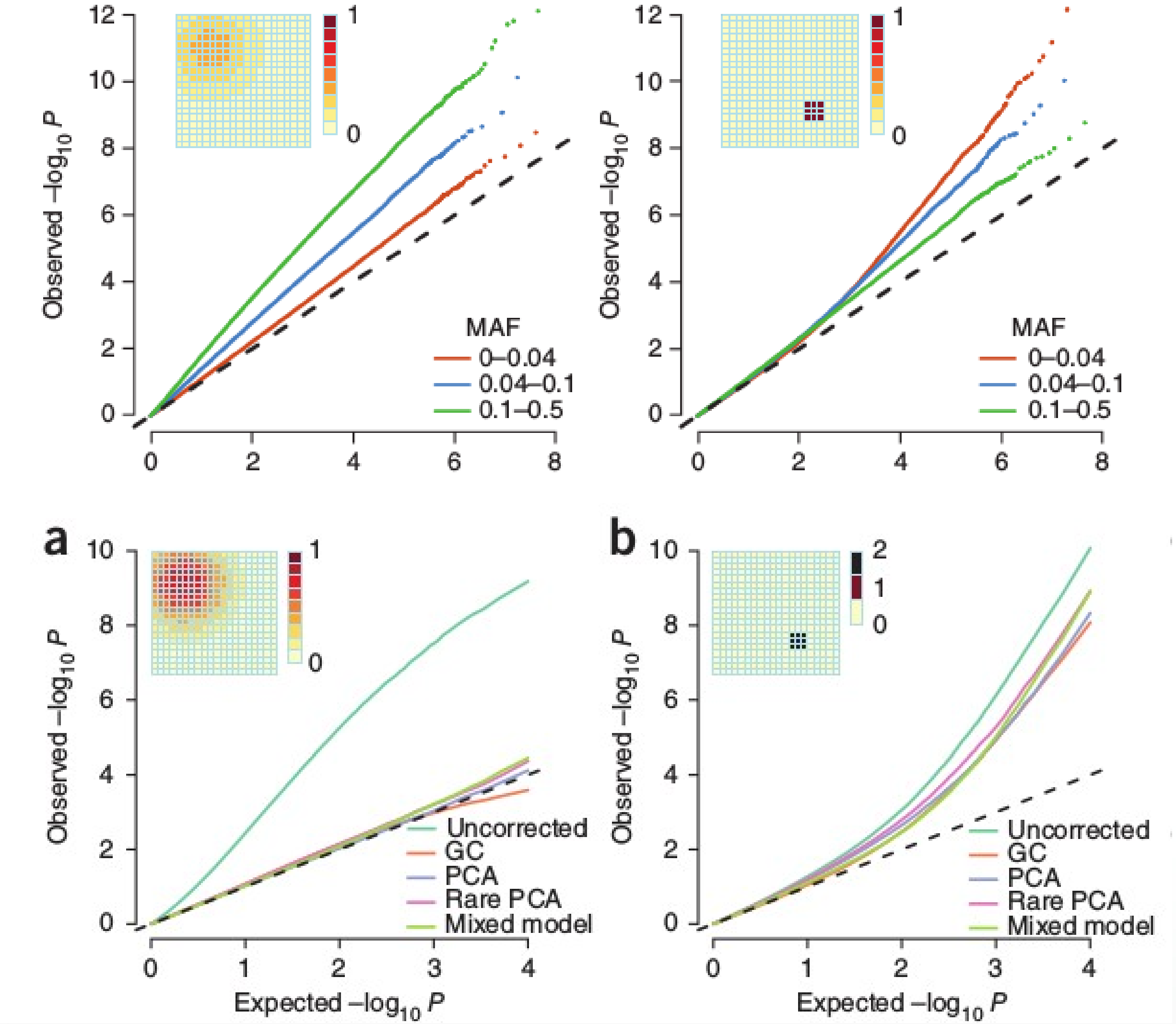
Application to a case-control COPD study



Part 2:

Variance inflation for non-genetic associations of sharply defined, spatially separated phenotypes with rare, spatially separated alleles in GWAS

The Problem



Mathieson 2013

The issue is not just stratification!

Distribution of **non-genetic**, **non-stratified** armitage trend test is not chi-sq(1) for rare alleles.

E(ATT)=1

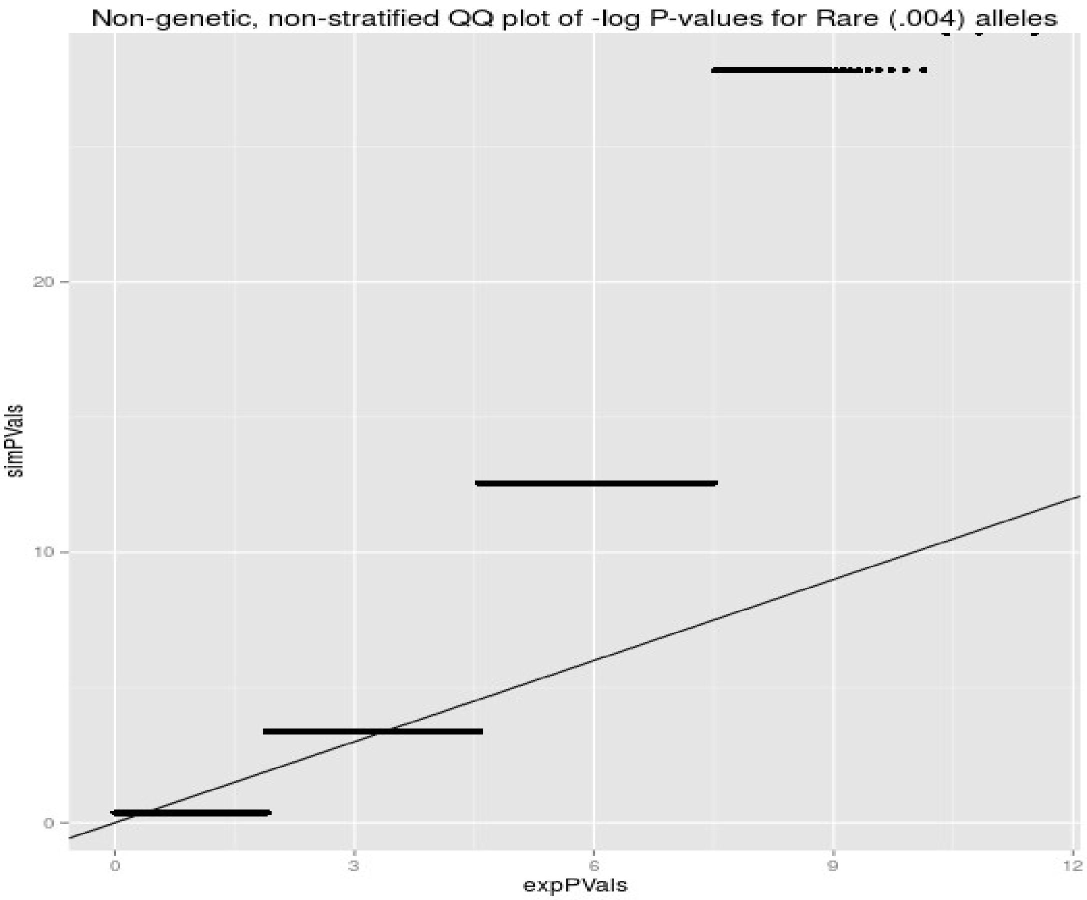
Var(ATT)!=2

For example, with no population stratification:
MAF=.004, PhenoFreq=.01

mean(ATT)=.997

Var(ATT)=3.1

This is a finite sample size issue.



Non-genetic, non-stratified QQ plot of -log P-values for Rare (.004) alleles

But it's also stratification

However, correcting for stratification (even perfect correction) is insufficient to stop inflation.

Example:

After a perfect correction for stratification, for a genotype and phenotype that appear in only one subpopulation, for a non-genetic risk,
ATT~n/n1*chi-sq(1)

There are at least 3 issues

1.) Common correction methods use linear functions to define risk and **do not distinguish subpopulations well**.

2.) **Finite sample sizes** yield inflated variance of ATT statistic.

3.) **Differential genotype/phenotype variances** lead to scaling of null test statistic distribution

The Approach

Find a superior method for more precise subpopulation identification

- Use rare alleles *only*
- Use Jaccard similarity instead of Variance-Covariance matrix
  - (Choose top eigenvectors based on eigendecomp)

Scale by a loci-specific variance inflation factor

Apply to most recent 1000GP data (much improved quality)

Inflation issues

Distribution of **non-genetic**, **non-stratified** armitage trend test is not chi-sq(1) for rare alleles.

E(ATT)=1

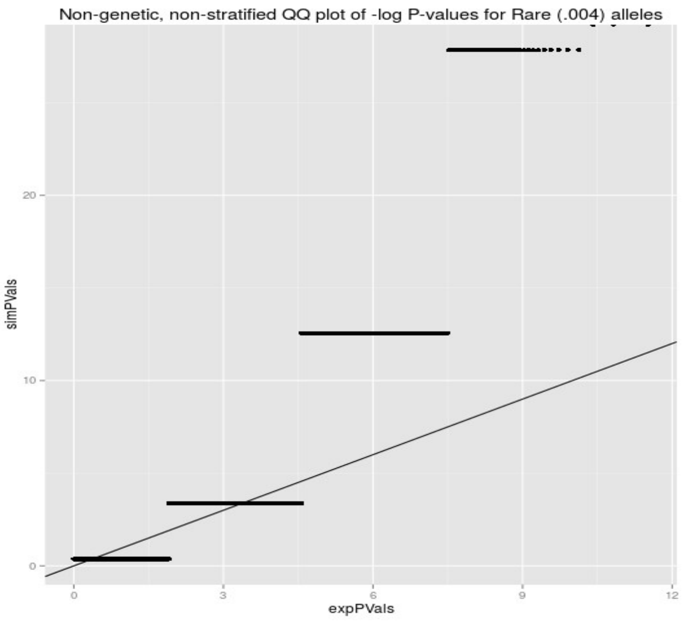
Var(ATT)!=2

For example, with no population stratification:

MAF=.004, PhenoFreq=.01

mean(ATT)=.997

Var(ATT)=3.1



Individual SNP variance inflation

So we have a variance inflation factor of 1/n1 for that particular SNP.

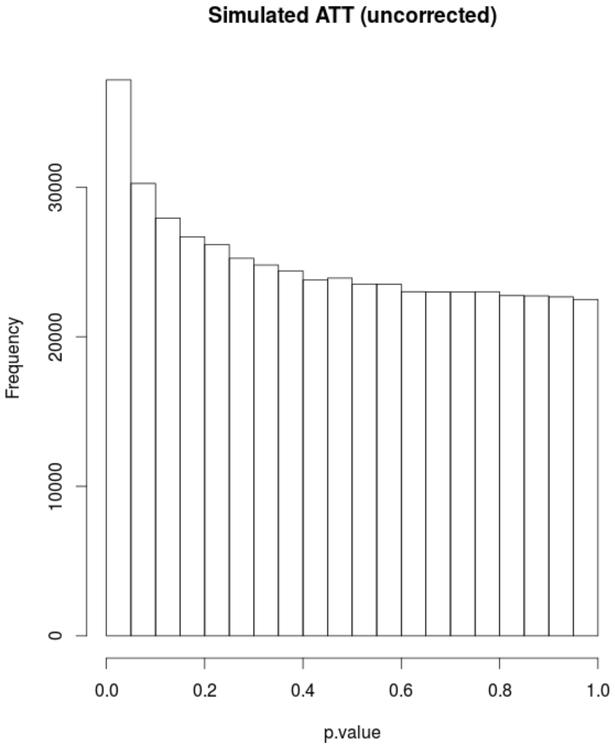
Generalizing this we have

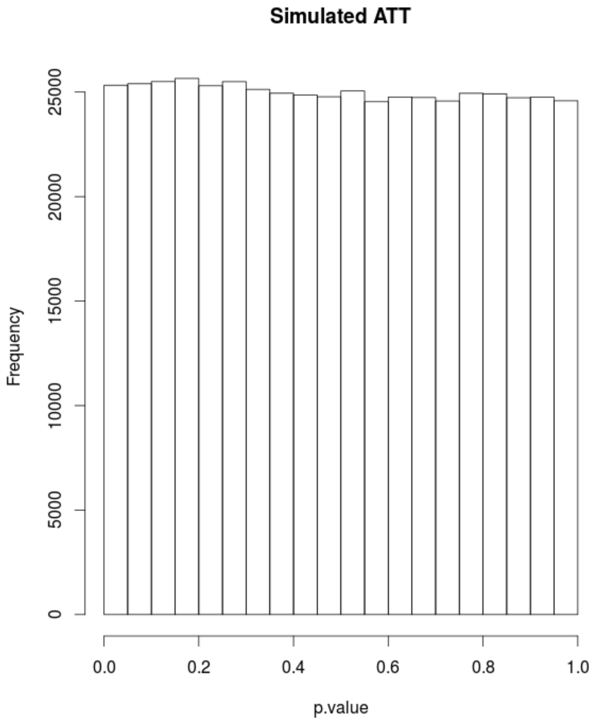
$$
\begin{aligned}
VIF_i &= \frac{\left[\sum_{k=1}^{N} Leverage_{i,k}\right]^2}{\sum_{j=1}^{N}\left[Leverage_{i,j}^2\right]} \\
Leverage_{i,j} &= Var\left(geno_{i,j}\right) \times Var\left(pheno_{i,j}\right)
\end{aligned}
$$

Where the leverage is the product of variance of genotype(i,j) and the variance of phenotype(i,j).
The variances are found by p(1-p) where p is the fitted value from the population correction.

Correcting variance scaling

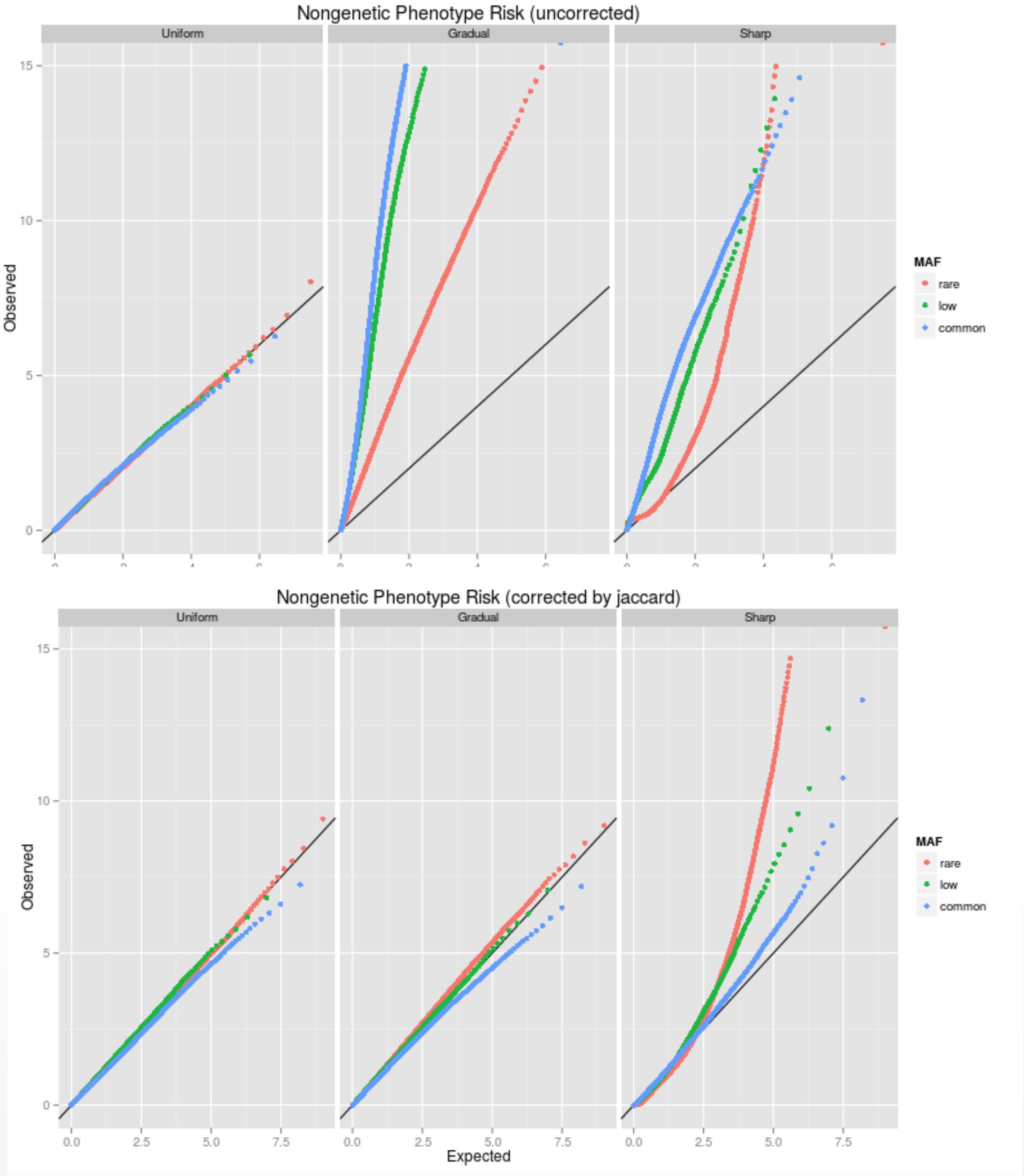
```
z <- (1:100)/200
zvar <- z*(1-z)
sumviSq <- sum(zvar)^2
sumSqvi <- sum(zvar^2)
vFactor <- 1/(sumSqvi/sumviSq)
hist(replicate(500000, {
    x <- rbinom(100,1,prob=z)-z
    y <- rbinom(100,1,prob=z)-z
    1-pchisq(vFactor*cor(x,y)^2, 1)
    vFactor*cor(x,y)^2
}))
```





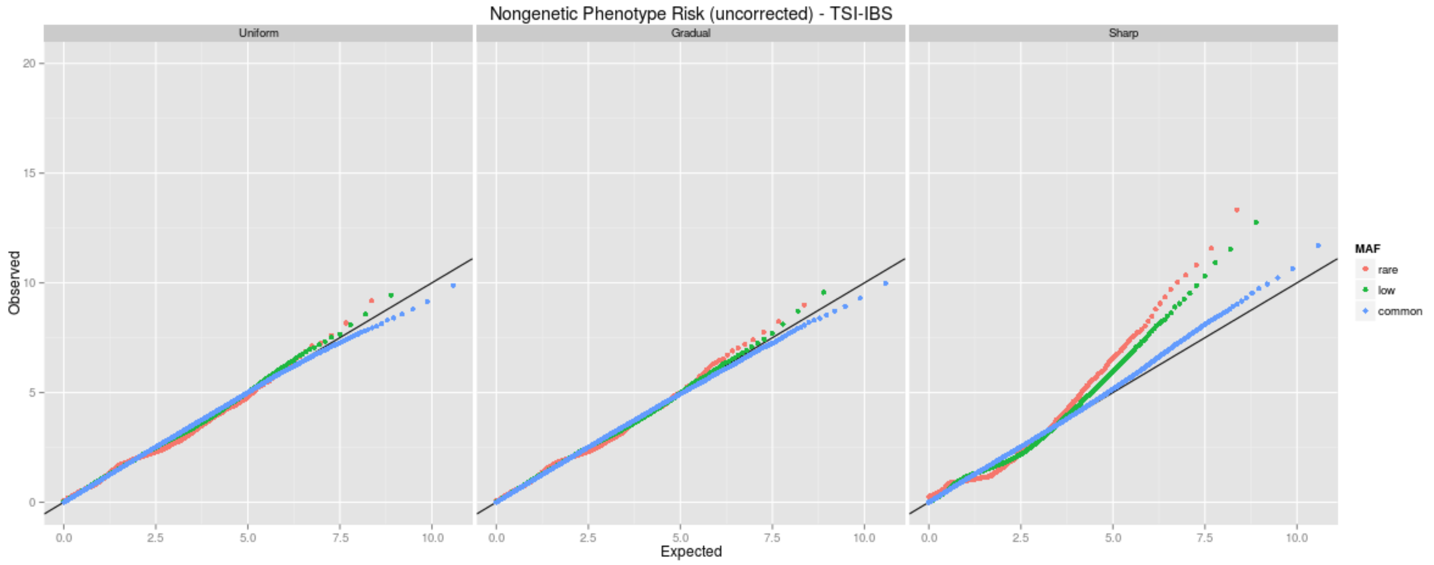
Variance inflation for associations of sharply defined, spatially separated phenotypes with rare, spatially separated alleles

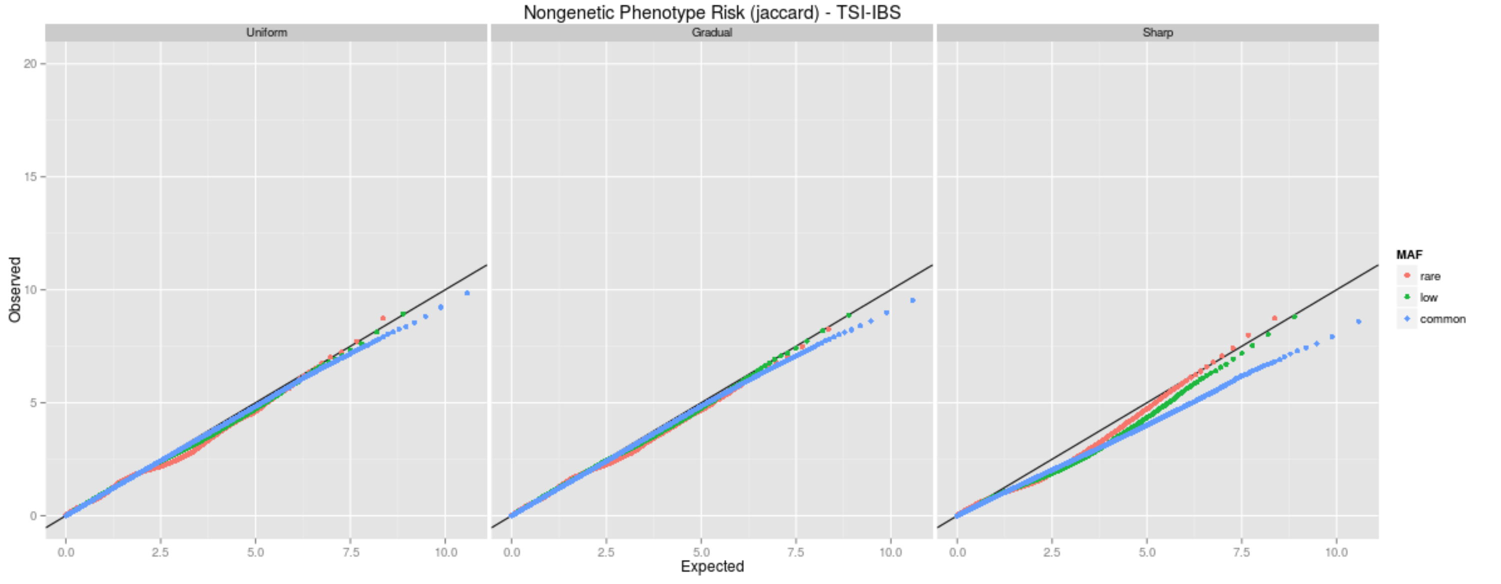
1000GP



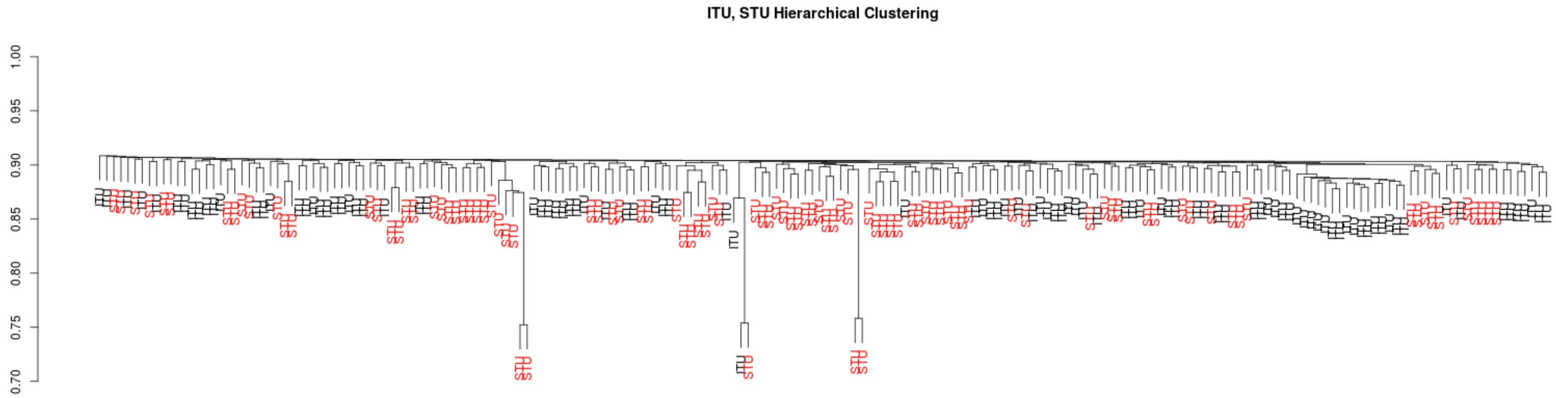
Variance inflation for associations of sharply defined, spatially separated phenotypes with rare, spatially separated alleles

1000GP

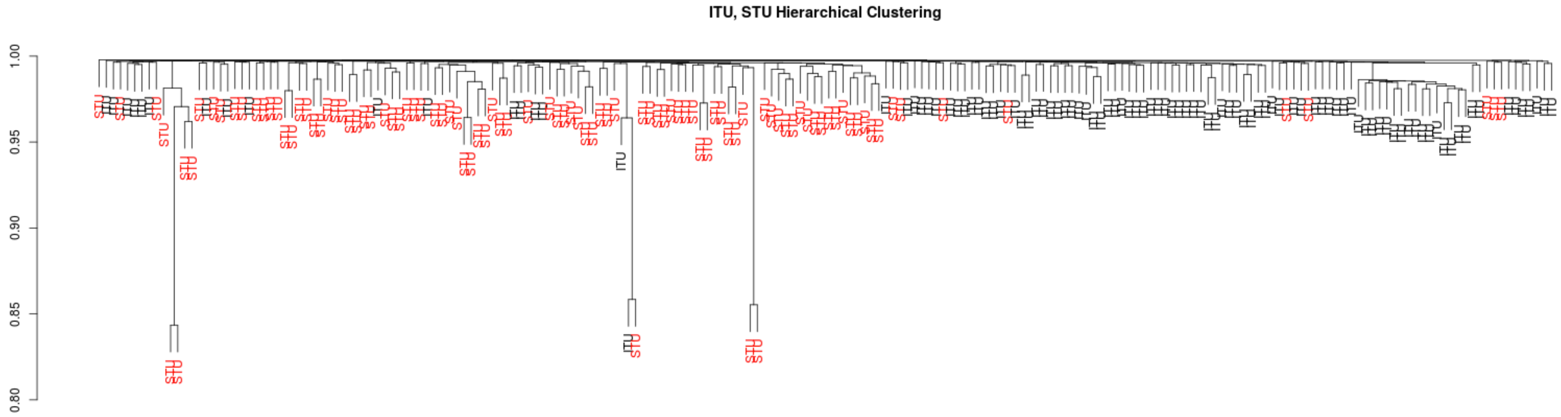




Variance inflation for associations of sharply defined, spatially separated phenotypes with rare, spatially separated alleles



Common (~10% MAF) alleles Hierachical Clusting via Jaccard similarity



Rare (<1% MAF) alleles Hierachical Clusting via Jaccard similarity