Methods for Estimating Hidden Structure and Network Transitions in Genomics

Daniel Schlauch, PhD Candidate

Department of Biostatistics
Harvard School of Public Health

April 24, 2017

Table of Contents

Batch Effect on Covariance Structure

Batch effect on covariance structure confounds gene coexpression

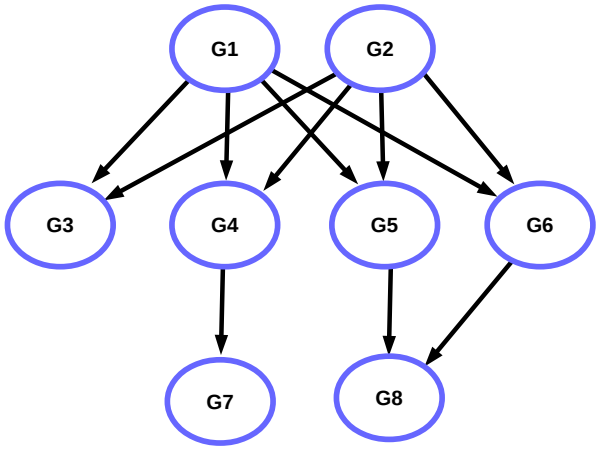
Daniel Schlauch [1,2], Joseph N. Paulson [2], Kimberly Glass [2,3], and John Quackenbush [1,3]

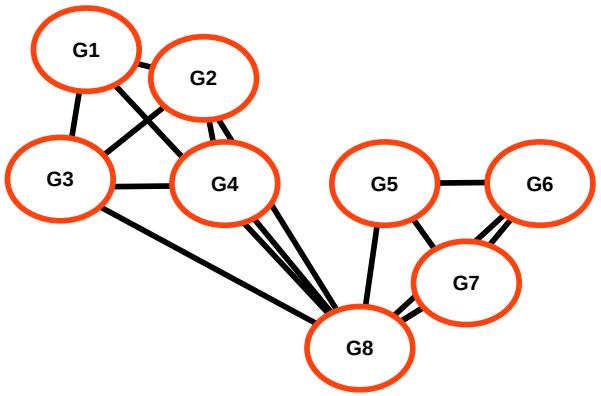
[1] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA
[2] Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA
[3] Department of Medicine, Harvard Medical School, Boston, MA

Differential Network Inference

How do we model functional interactions?

- Gene Regulatory/Coexpression Networks (GRN/GCN)
- Directed/undirected graph
- May imply a sort of physical interaction
- Guilt by association

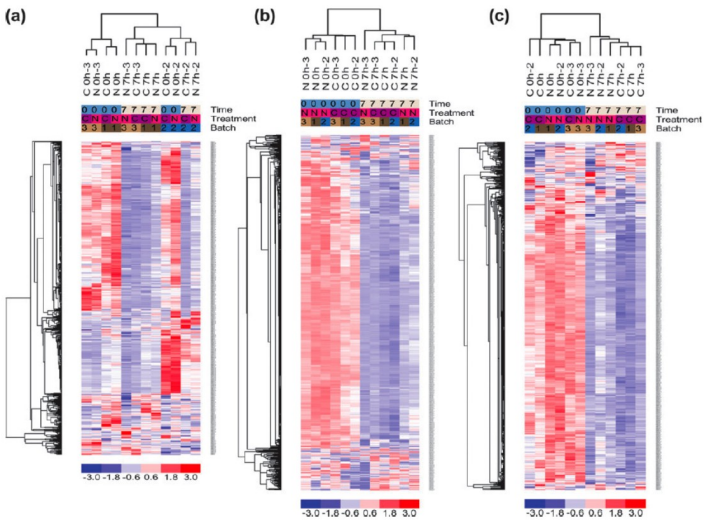




Batch Effect

Batches

Experimental Conditions



Johnson et al.(Biostatistics 2007)

Unwanted variation from:

- Laboratory Conditions
- Circadian Rhythm
- Cell cycle
- Reagents
- Atmospheric Ozone
- Differential biological variation
- Etc.

Methods for Controlling Batch Effect

Location scale model:

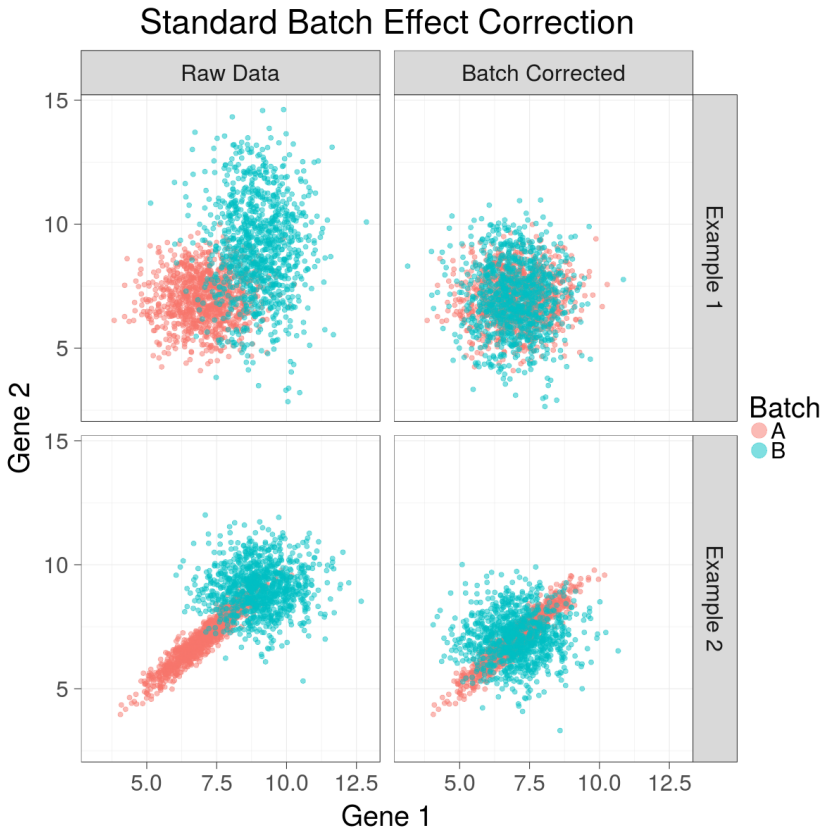
$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg}$$

Limitations:

- Gene-specific location/scale assumptions
- Independent effects
- *Differential coexpression*

Batch effect removal methods typically return a corrected gene expression matrix (e.g. ComBat) or a correction vector (e.g. SVA).

Limitations to common batch effect correction methods



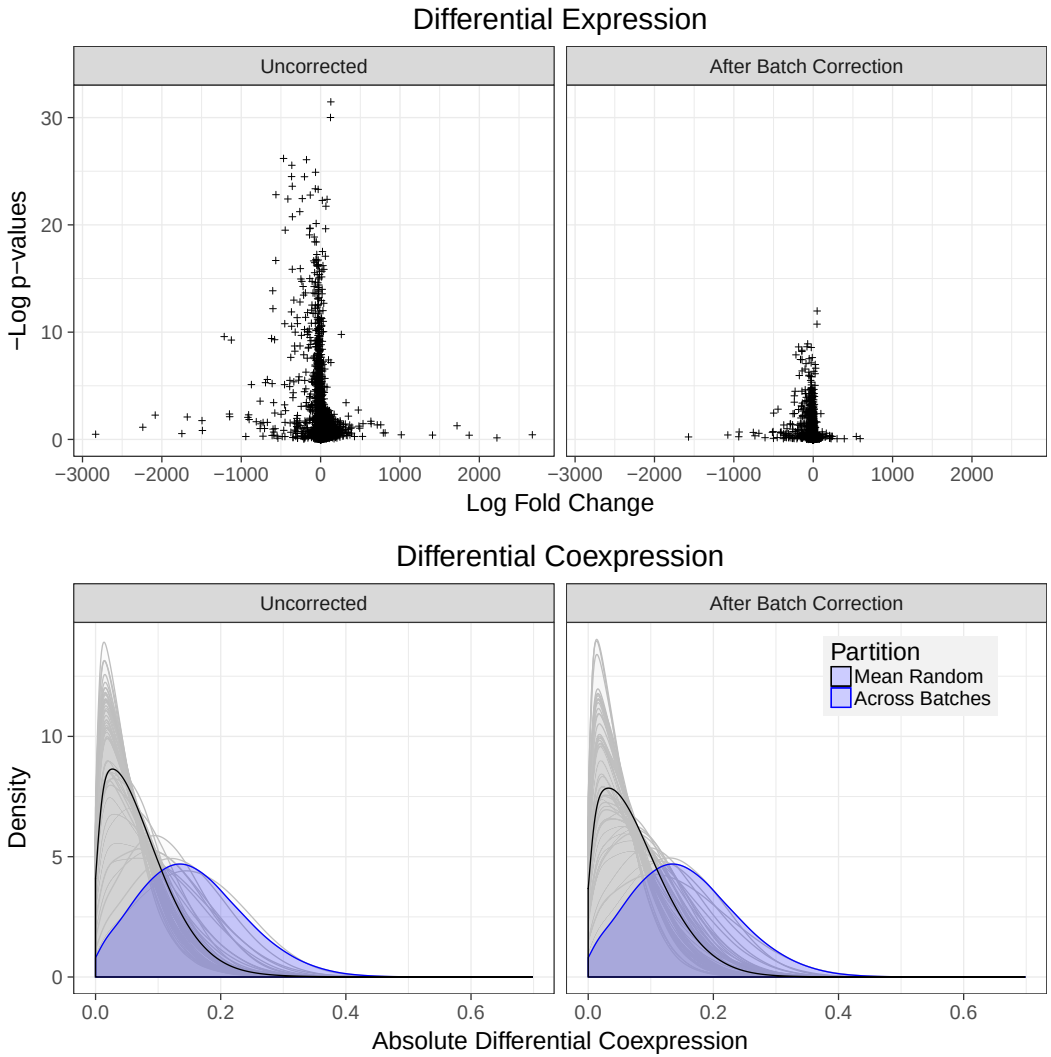
Standard corrections:

$$f\left[Gene1|BatchA\right] = f\left[Gene1|BatchB\right]$$

$$f\left[Gene2|BatchA\right] = f\left[Gene2|BatchB\right]$$

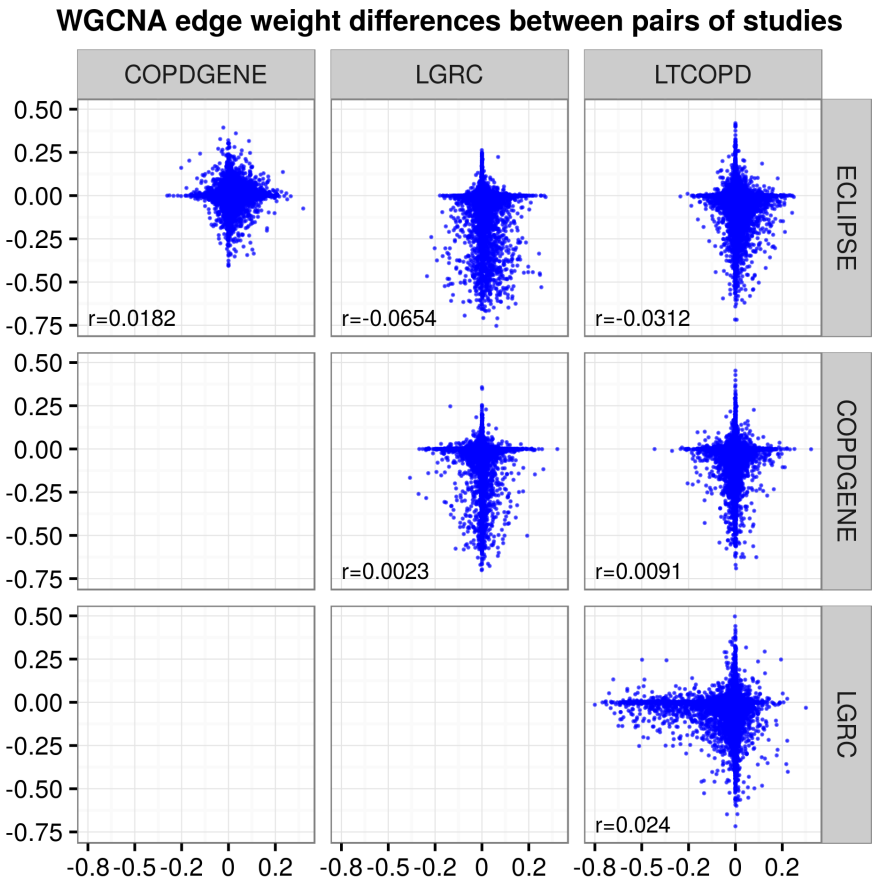
$$f\left[Gene1, Gene2|BatchA\right] \neq f\left[Gene1, Gene2|BatchB\right]$$

Limitations to existing batch effect correction methods



ENCODE Project:
50k genes
126 samples (63 patients RNA-seq'ed at 2 centers)

Challenges with batch effect on differential coexpression



- Ultra-high dimensionality
- Differential coexpression
- Modularity

Estimating the conditional coexpression matrix

Motivating concepts:

- Provide a regression framework for for the coexpression matrix
- Estimate a reduced number of parameters
- Exploit modular nature of gene expression patterns

Our proposal:

- Define our parameters as functions of components of variation.
- Estimate the eigenvalue contribution of each eigenvector.

Model

Consider a set of N samples with q covariates measuring gene expression across p genes. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})$ denote the covariates for sample i and let $\mathbf{g}_i = (g_{i1}, \ldots, g_{ip})^T$ denote the gene expression values for sample i for the p genes.
We can express a model for the gene expression as

$$\mathbf{g}_i = \beta^T \mathbf{x}_i + \epsilon_i \text{ for } i = 1, \ldots, N$$

where $\epsilon_i \sim MVN_p(\mathbf{0}, \Sigma_i)$. Notably, the covariance of ϵ_i differ according to i.

$$\Sigma_i = \mathbf{Q}\mathbf{D}_i\mathbf{Q}^T$$

where $\mathbf{D}_i$ is a diagonal matrix with diagonal defined as $\mathbf{X}_i\Psi_{q\times p}$

Likelihood Function

$$\mathcal{L}(\mu, \Sigma) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{G}_i - \mu)^T \Sigma_i^{-1} (\mathbf{G}_i - \mu)}$$

Where we define Σ_i

$$\Sigma_i = \mathbf{Q}\mathbf{D}_i\mathbf{Q}^T$$

Where $\mathbf{Q}$ is a matrix with columns defined as the eigenvectors of the estimated coexpression matrix, $\mathbf{G}^*\mathbf{G}^{*T}/N$.

Least Squares Estimator

To calculate least squares solution, $\hat{\Psi}$, we solve separately for each column of $\mathbf{Q}$.
and note that the residual matrices should be orthogonal to the hyperplane spanned by X^T.
Recall $0 = \mathbf{X^T}\left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)$

$$\mathbf{0}_q = \sum_{i=1}^{N} \mathbf{X}_i^T \left[\mathbf{Q}_h^T \left[\mathbf{G}_i^* \mathbf{G}_i^{*T} - \mathbf{Q}_h \mathbf{X}_i \hat{\Psi}_h \mathbf{Q}_h^T\right] \mathbf{Q}_h\right] \quad (1)$$

$$\hat{\Psi}_h = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \sum_{i=1}^{N} \left[\mathbf{X}_i^T \mathbf{Q}_h^T \mathbf{G}_i^* \mathbf{G}_i^{*T} \mathbf{Q}_h\right] \quad (2)$$

The Corrected Coexpression Matrix

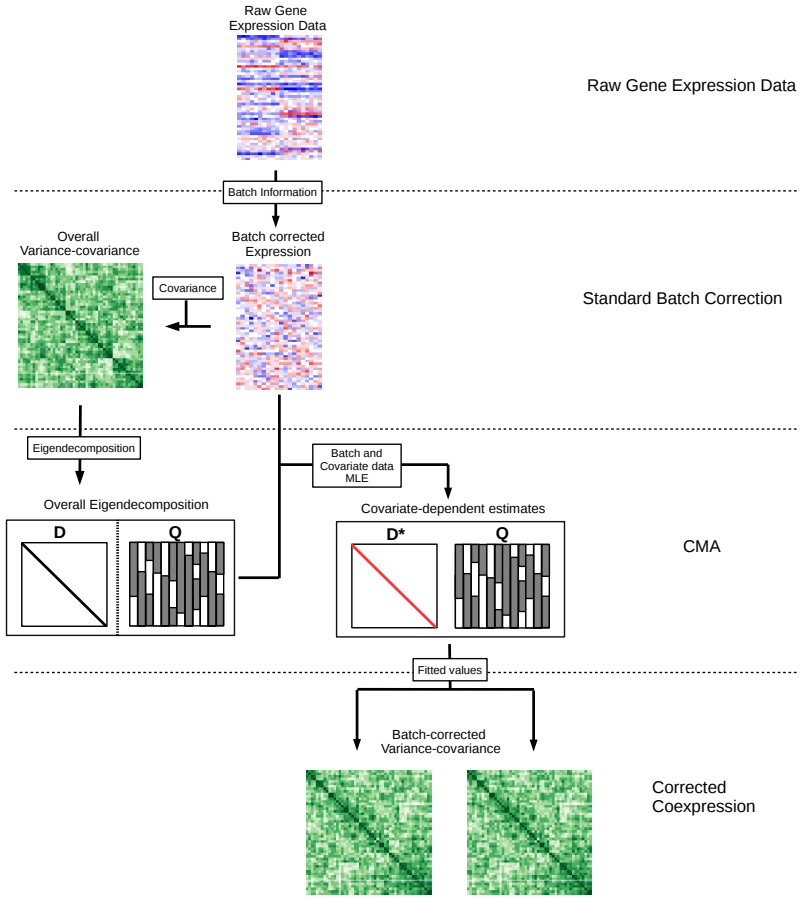
With the estimates obtained with our method, it is straightforward to see how fitted values for the coexpression matrix for each sample or experimental condition can be obtained. Given an estimate for Ψ, $\hat{\Psi}$, we can now estimate the batch-independent coexpression structure as

$$\hat{\mathbf{S}} = \mathbf{Q} diag\left(\bar{\mathbf{X}}\hat{\Psi}\right)\mathbf{Q}^T \text{ or } \hat{\mathbf{S}} = \sum_{i=1}^{p} \bar{\mathbf{X}}\hat{\Psi}_i \mathbf{Q}_i \mathbf{Q}_i^T$$

The differential coexpression matrix between two conditions, defined in binary as column 2 of $\mathbf{X}$, is computed

$$\hat{\mathbf{W}} = \mathbf{Q} diag\left(\hat{\Psi}_{2,\cdot}\right)\mathbf{Q}^T$$

Example Workflow

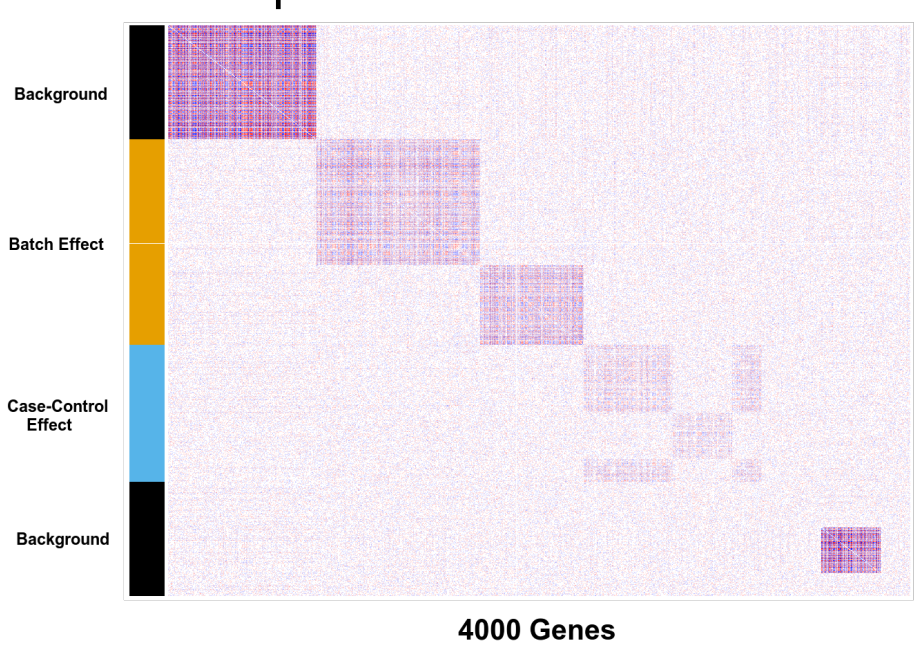


Simulations

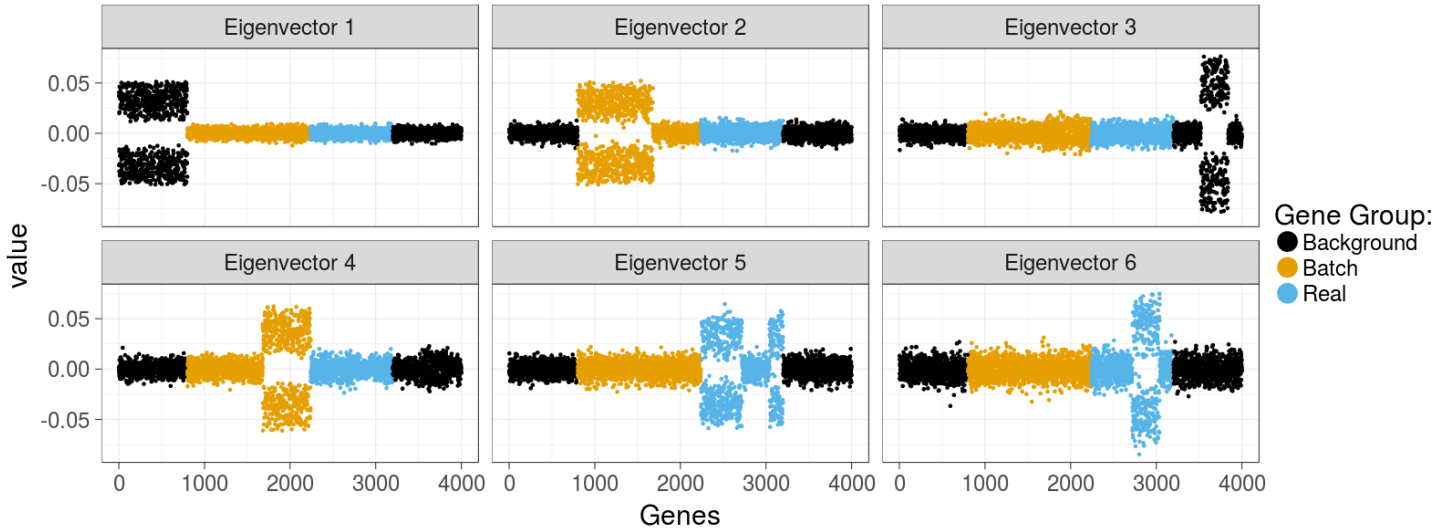
$$g_i \sim MVN_{4000}(\mathbf{0}, \Sigma_i)$$

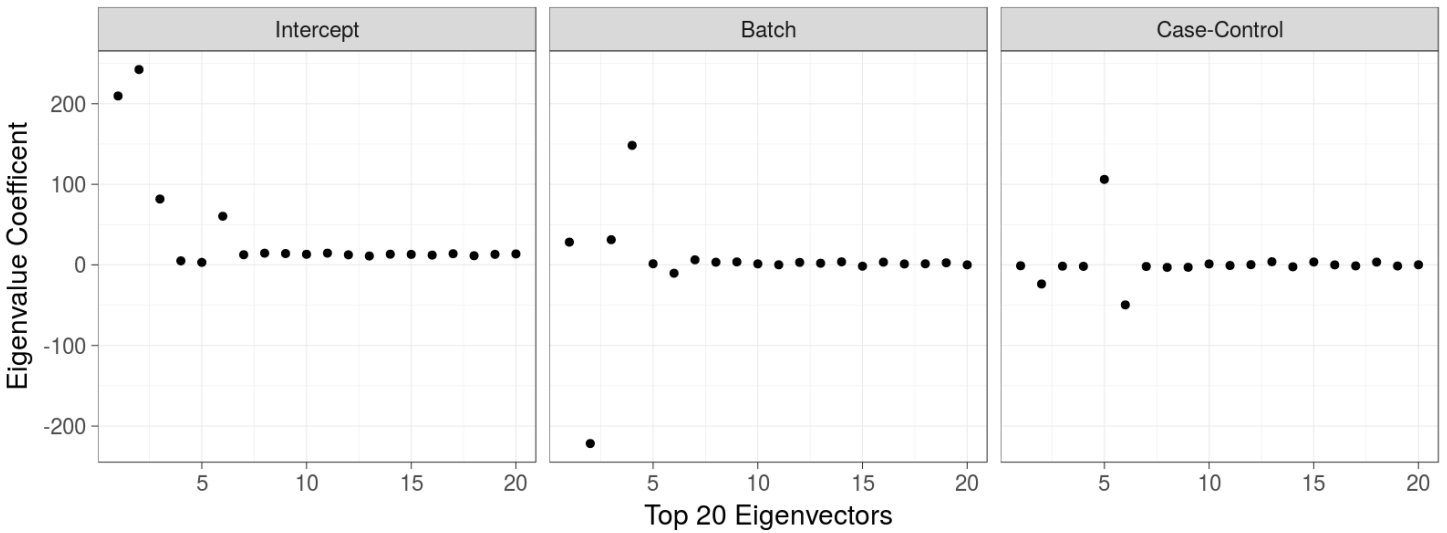
Σ_i describes the covariance of 4000 genes and contains modules which are classified as

- Background
- Batch
- Case-Control

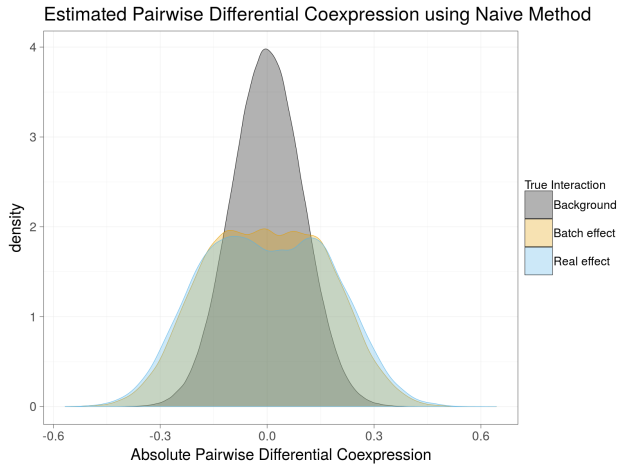


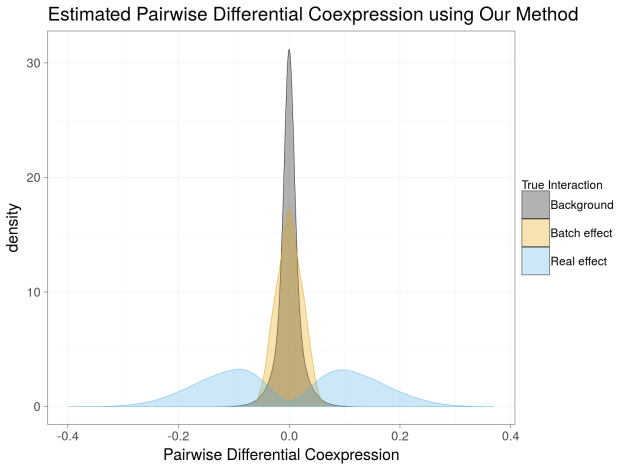
Simulations

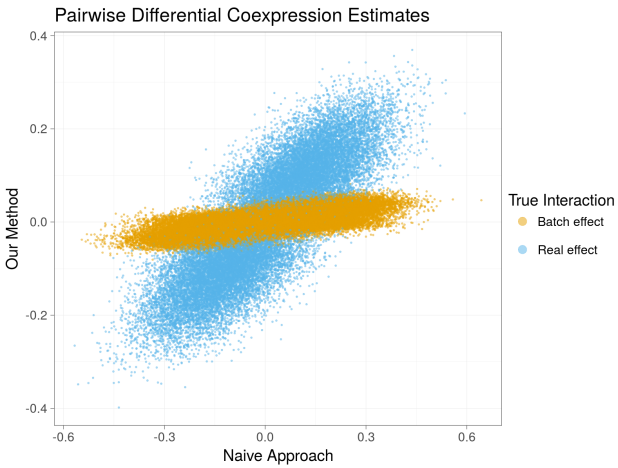


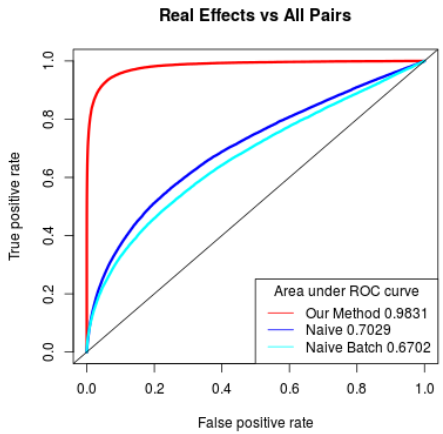


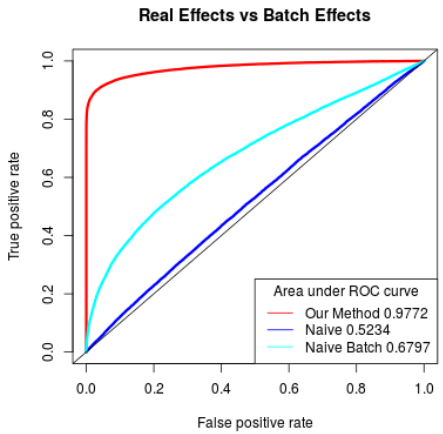
Simulations











Application to data from COPDGene Study

GO Term	Count	%	Enrichment	FDR
anatomical structure development	309	0.26	1.29	2.58E-05
single-organism developmental process	309	0.26	1.29	2.73E-05
anatomical structure morphogenesis	168	0.14	1.46	1.60E-04
single-multicellular organism process	324	0.27	1.25	4.01E-04
system process	132	0.11	1.50	1.46E-03
regulation of cellular process	514	0.43	1.12	5.86E-03
single organism signaling	328	0.28	1.21	7.89E-03
regulation of localization	151	0.13	1.40	1.15E-02
regulation of multicellular organismal process	156	0.13	1.35	6.56E-02

Table : GO categories for differential coexpression in COPDGene identified with CPBA found with FDR$<$0.1.

Identifying Genetic Outliers

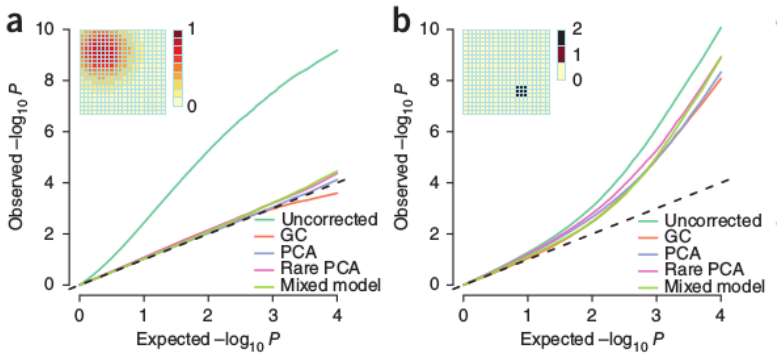
Identification of genetic outliers due to sub-structure and cryptic relationships

Daniel Schlauch [1,2,4], Heide Fier [1,3] and Christoph Lange [1,4]

[1] Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA 02115
[2] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02115
[3] Institute of Genomic Mathematics, University of Bonn, Bonn, Germany
[4] Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA 02115

Background

- Individuals may be too similar (due to cryptic relatedness) or too different (due to population structure).
- Both features may lead to spurious results, inflation of type I error.
- Many methods exist for addressing some of these concerns (e.g. PCA, LMM).
- Limitations exist, such as with rare alleles and sharp localized effects, or with the assumption of linear or discrete population structure.



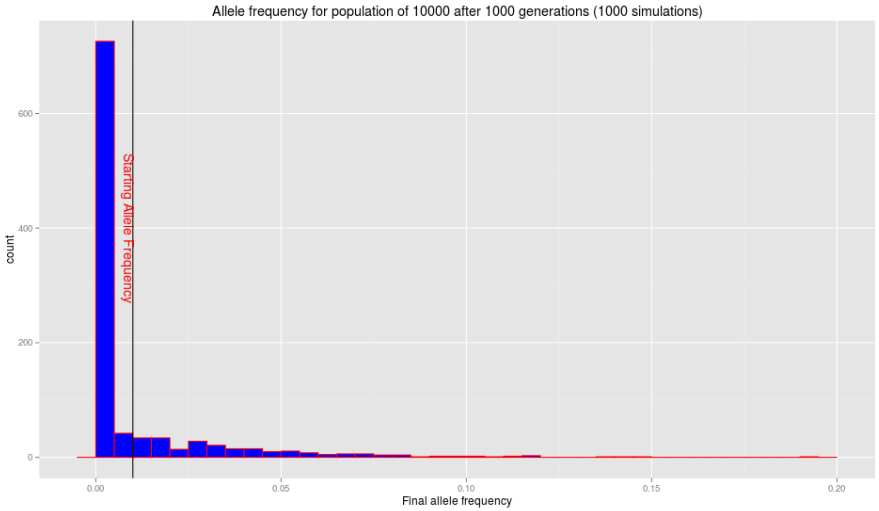
Mathieson, Nature Genetics 2012

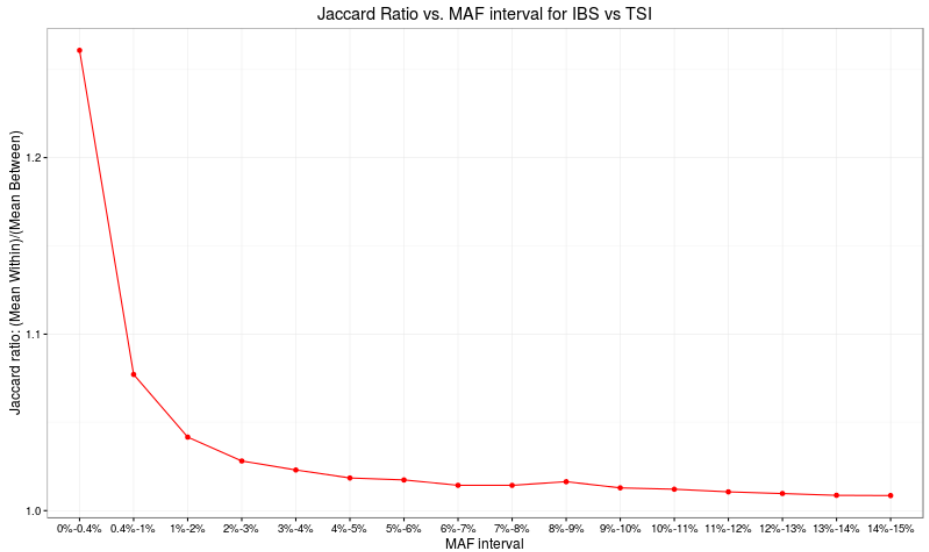
We want to create a similarity measure that...

- is more sensitive to fine scale population stratification
- can be used as a formal test for cryptic relatedness
- can be used as a formal test for population structure

Basis for measure

- Rare variants are recent variants.
- In the absense of selection, rare variants become fixed at 0% with high probability over a relatively short timeframe.





P[Fixation|n=10000,g=1000,maf=.01]=.678

- Key Idea: **Less** frequent variants are **more** informative of ancestry.

Test Statistic

$$s_{i,j} = \frac{\sum_{k=1}^{N} w_k \mathbf{G}_{i,k} \mathbf{G}_{j,k}}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]}$$

where

$$w_k = \begin{cases} \frac{\binom{2n}{2}}{\binom{\sum_{l=1}^{2n} \mathbf{G}_{l,k}}{2}} & \sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1 \\ 0 & \sum_{l=1}^{2n} \mathbf{G}_{l,k} \leq 1 \end{cases}$$

$$E\left[s_{i,j}\right] = 1$$

Test Statistic

In the absence of population structure, cryptic relatedness and dependence between loci the distribution of the similarity index, $s_{i,j}$

$$s_{i,j} \sim N\left(1, \sigma^2_{i,j}\right)$$

Where the variance of s_{ij} can be estimated by

$$\hat{\sigma^2_{i,j}} = \hat{Var}\left(s_{i,j}\right) = \frac{\sum_{k=1}^{N}\left(w_k - 1\right)}{\left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^2}$$

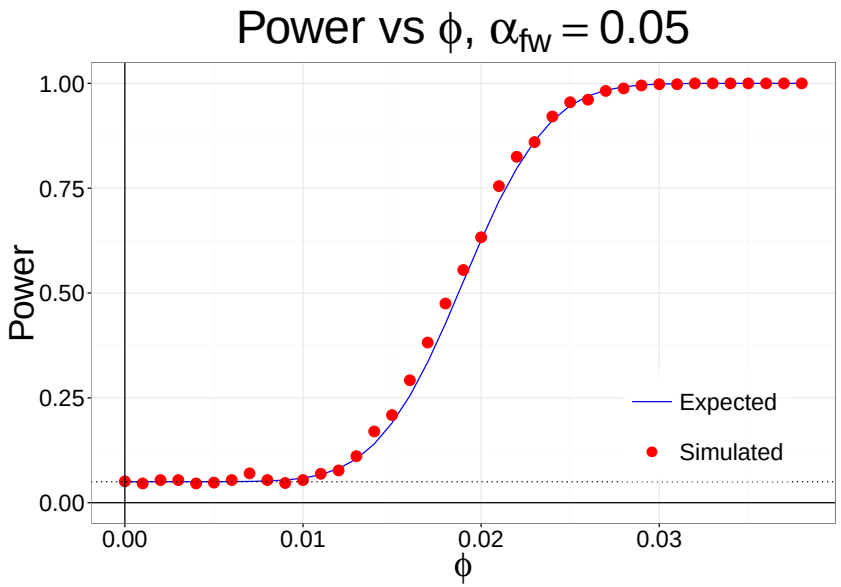
$$s_{i,j}^{(diploid)} = \frac{\sum_{k=1}^{N}\left[w_k \mathbf{H}_{i,k}\mathbf{H}_{j,k}\right]/4}{\sum_{k=1}^{N} I\left[\left(\sum_{l=1}^{n} \mathbf{H}_{l,k}\right) > 1\right]}$$

Tests of Heterogeneity

$$\hat{\phi}_{i,j} = \frac{s_{i,j} - 1}{\left[\frac{\sum_{k=1}^{N} \hat{p_k} w_k}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]} - 1 \right]}$$

$$R : max\left(s_{i,j}\right) > 1 - probit\left(\frac{\alpha}{\binom{n}{2}}\right)$$

$$P\left(Reject\ H_0 | \phi_{i,j} = \gamma\right) = \alpha + (1 - \alpha)\left(1 - \Phi\left(\frac{\mu_{i,j} - 1}{\sqrt{\hat{\sigma^2}_{i,j}}}\right)\right)$$



Tests of Heterogeneity

$$H_0 : \mu_{i,j} = 1 \forall i, j \in 1 \ldots n$$

$$H_A : \exists i, j \in 1 \ldots n | \mu_{i,j} \neq 1$$

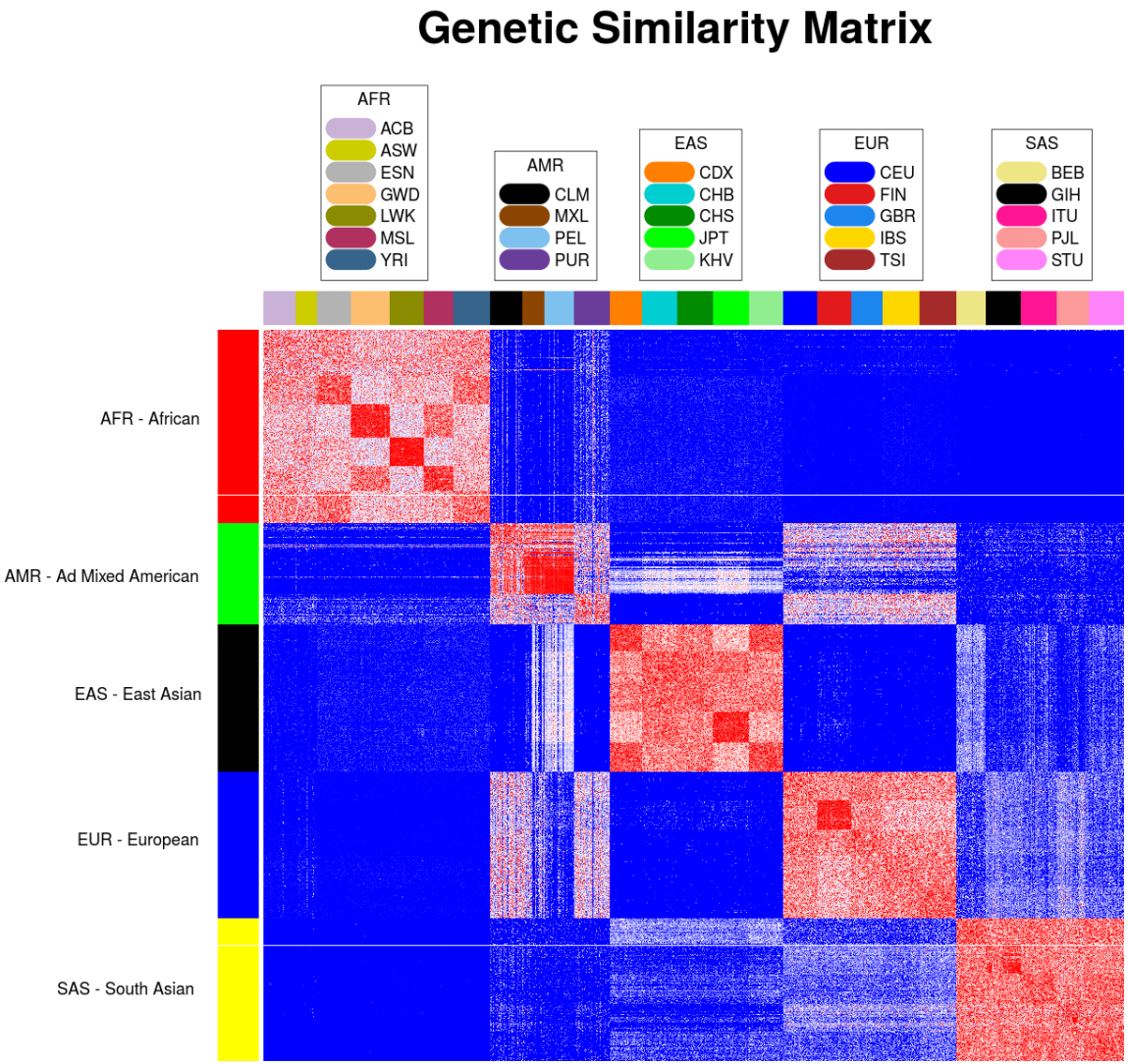
Test for population structure:

$$K = \sup_x |F_s(x) - \Phi(x)|$$

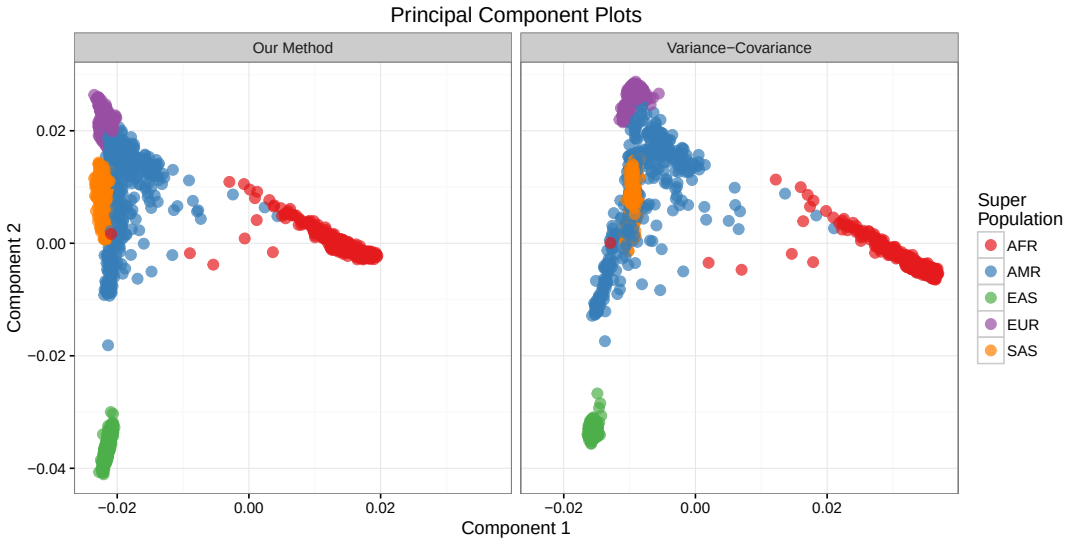
Test for cryptic relatedness:

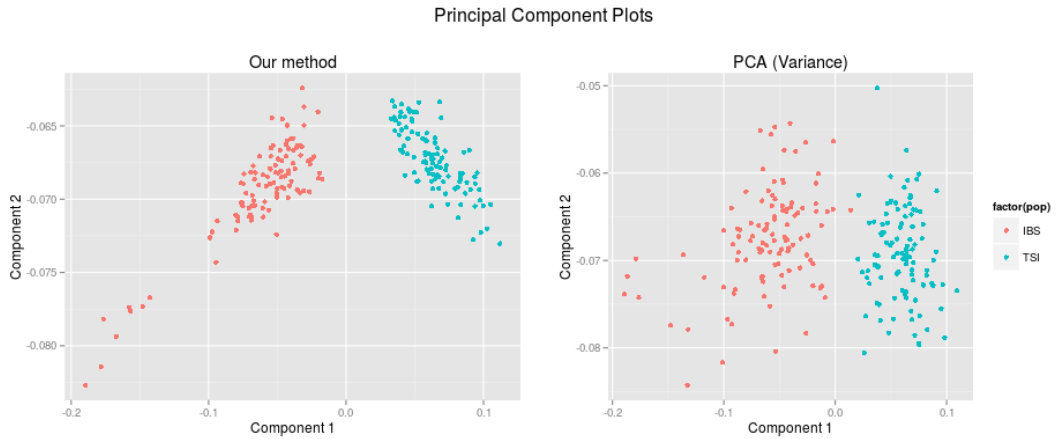
$$R : max(s_{i,j}) > 1 - probit\left(\frac{\alpha}{\binom{n}{2}}\right)$$

Application to 1000 Genomes Project



Application to 1000 Genomes Project



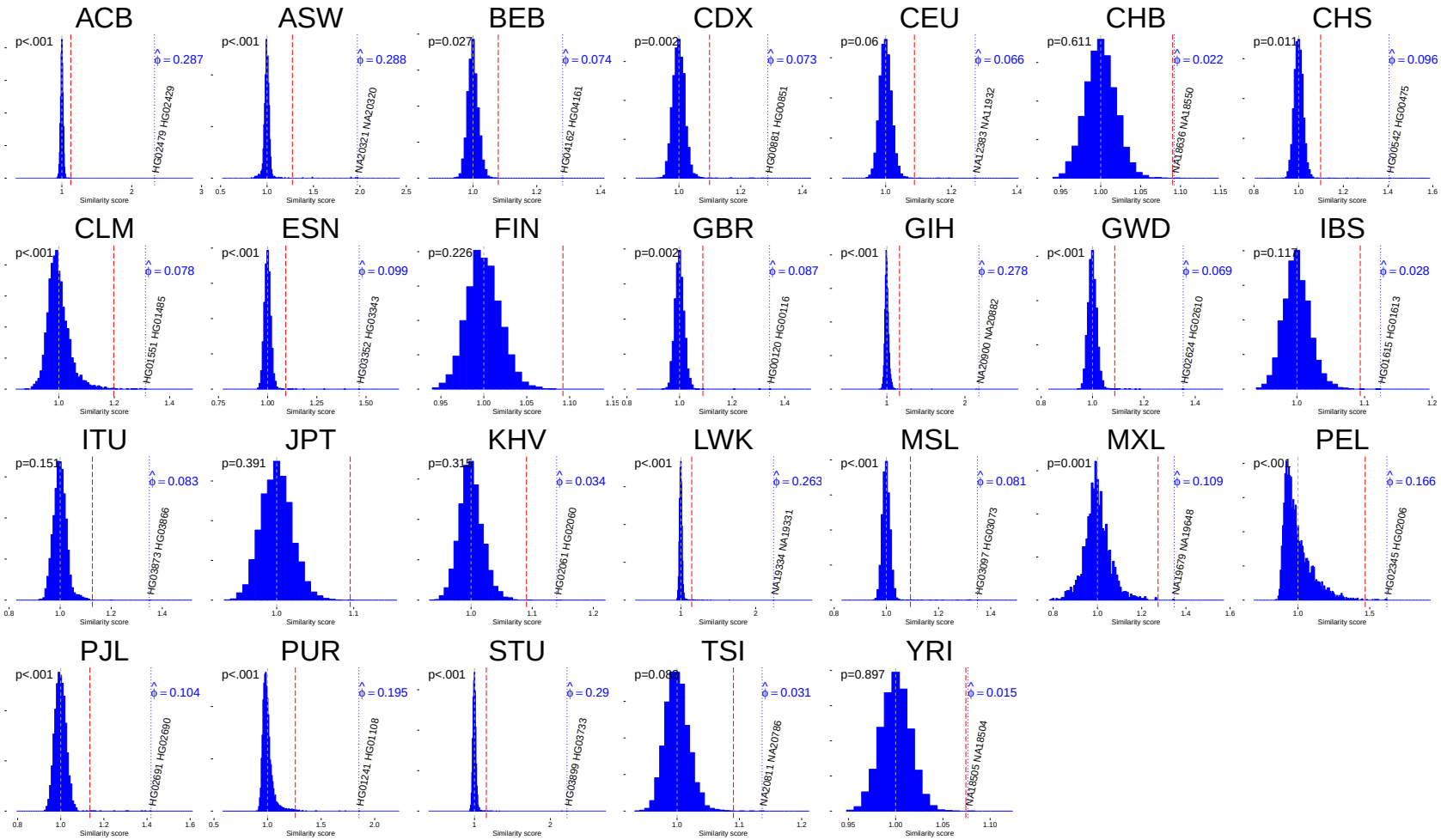


STEGO is comparable to PCA when applied on a global scale.

But produces superior separation for recently related populations.

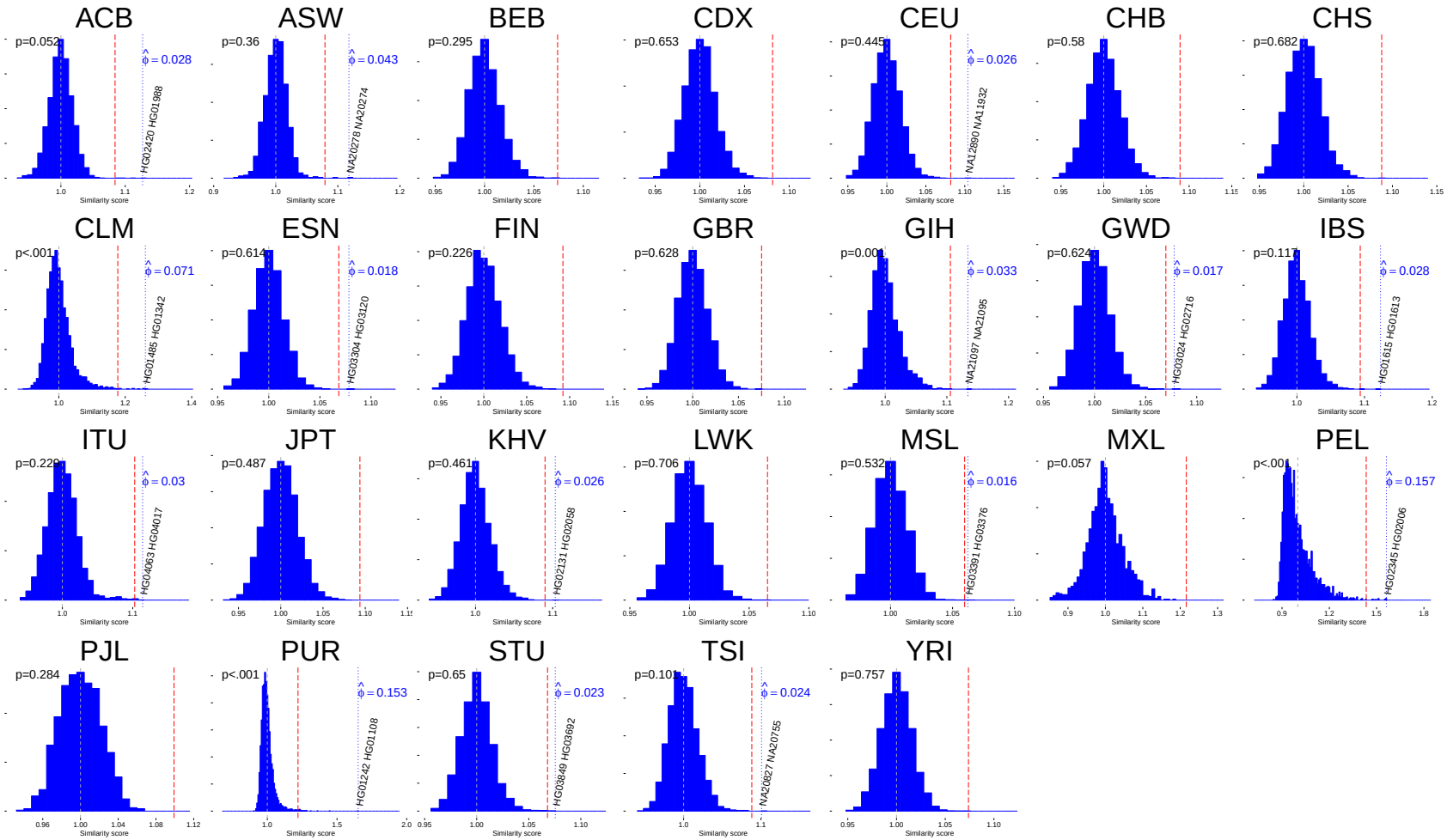
Application to 1000 Genomes Project

Distribution of s statistics by population



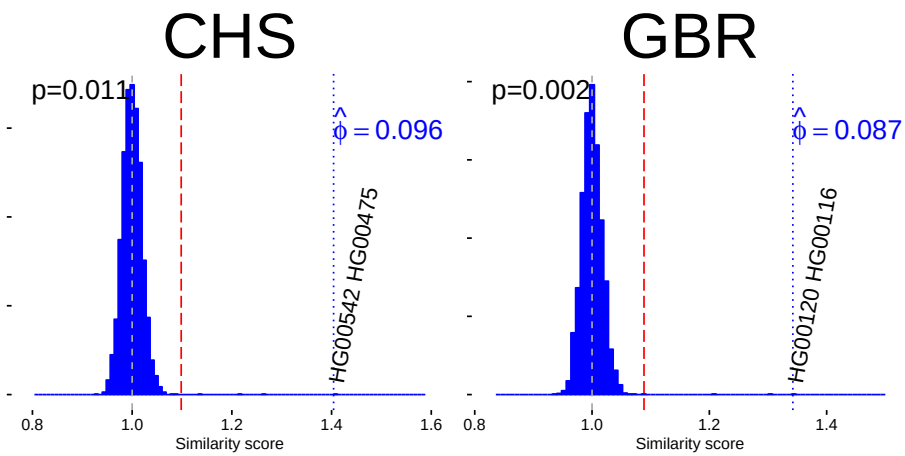
Application to 1000 Genomes Project

s statistics by population after removal of related pairs

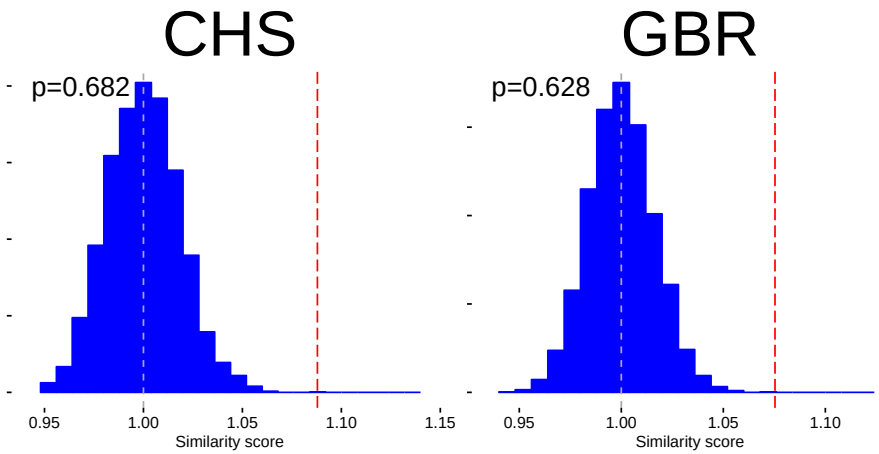


Application to 1000 Genomes Project

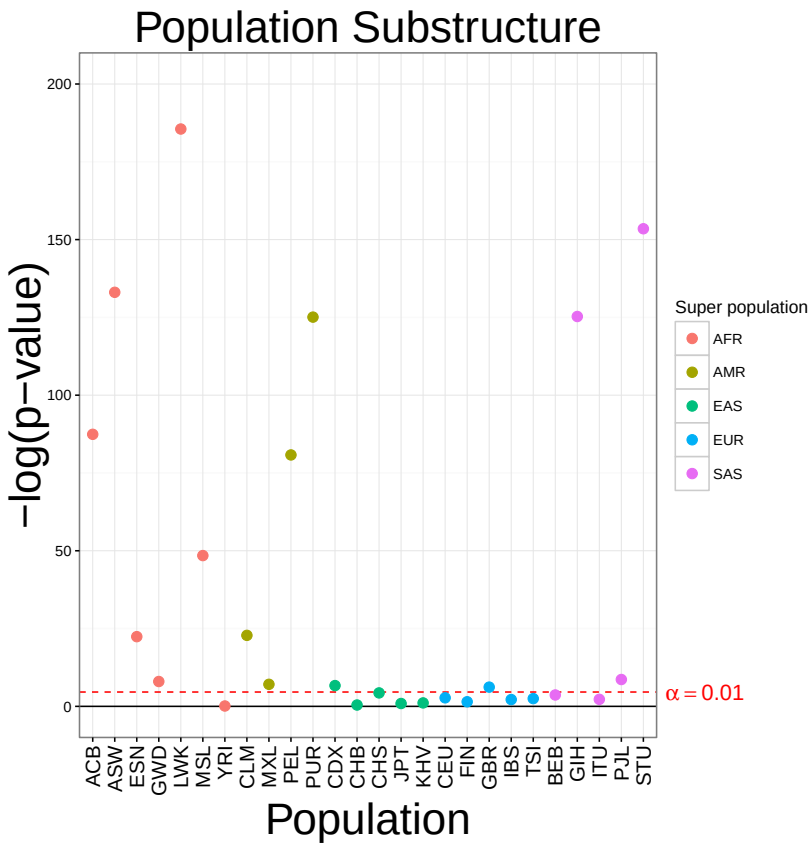
Original Data

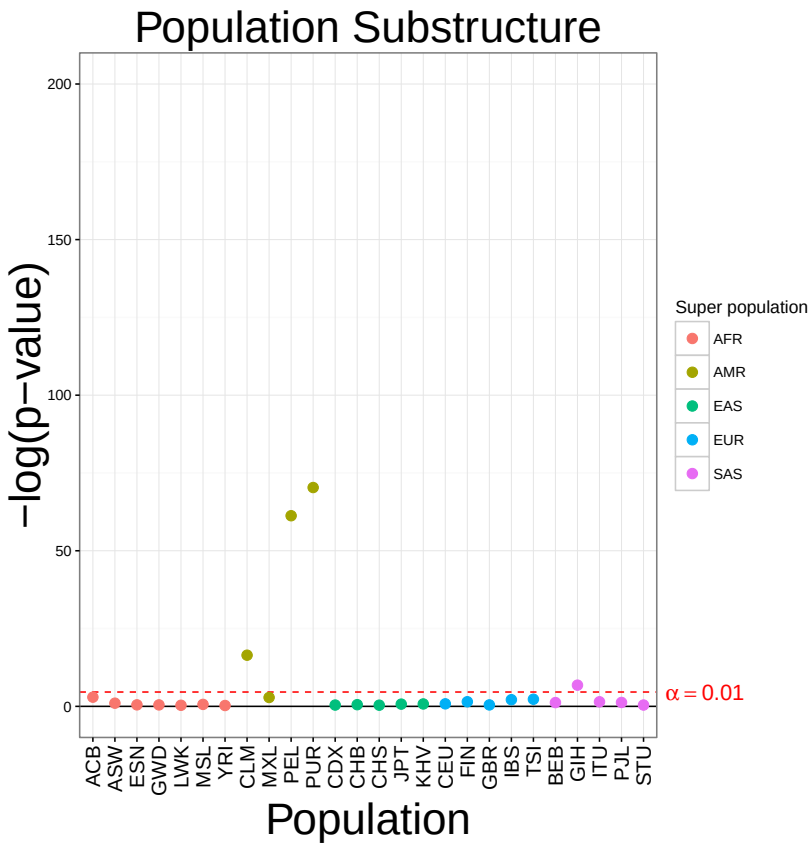


After removal of related pairs



Application to 1000 Genomes Project





State Transitions Using Gene Regulatory Network Models

Estimating Drivers of Cell State Transitions Using Gene Regulatory Network Models

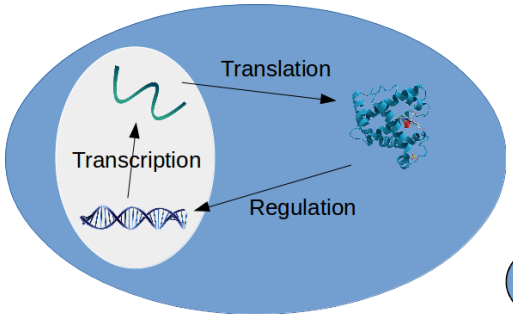
Daniel Schlauch [1,2], Kimberly Glass [2,3], Craig P. Hersh [2,3,4], Edwin K. Silverman [2,3,4] and John Quackenbush [1,2,3]

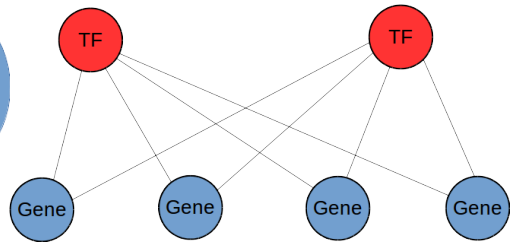
[1] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA
[2] Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA
[3] Department of Medicine, Harvard Medical School, Boston, MA
[4] Pulmonary and Critical Care Division, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

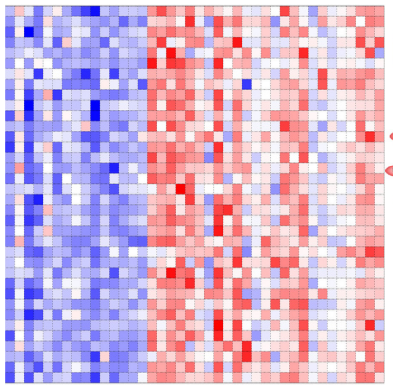
Background

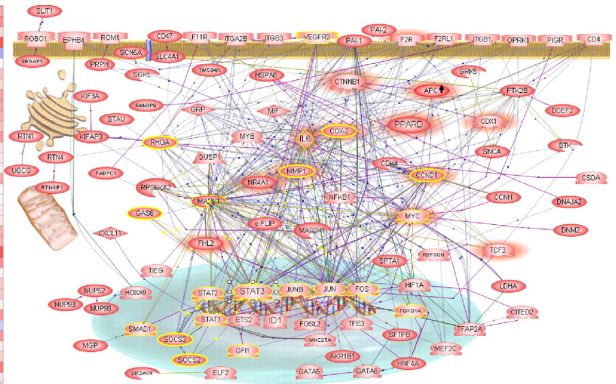
Why Study Gene Regulatory Networks?

- Genes are not independent objects.
- Regulation of higher level pathways and processes.









Abdollahi et al. PNAS 2007

Background

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"

Background

Biological Challenges

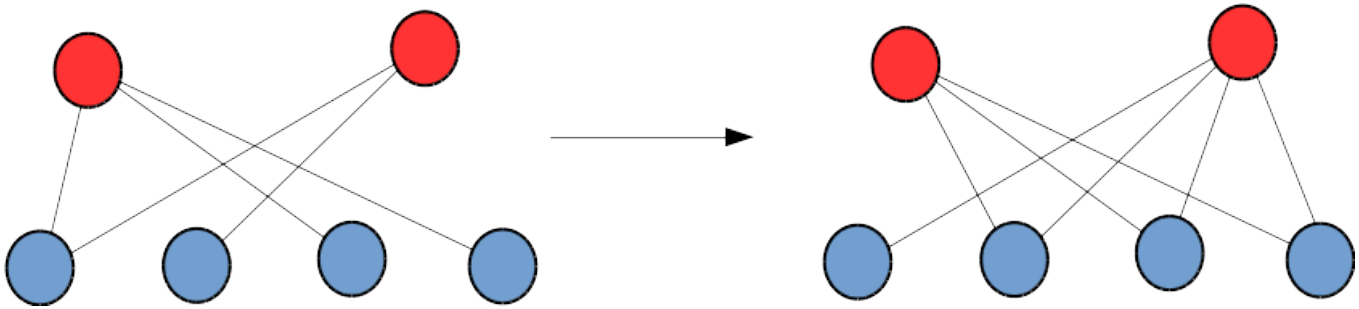
- Measurements of gene expression are at the mRNA level.
- Measurements only consist of mRNA abundance.
- Experimental data is collected as static snapshots.
- Biological variability can be difficult to induce

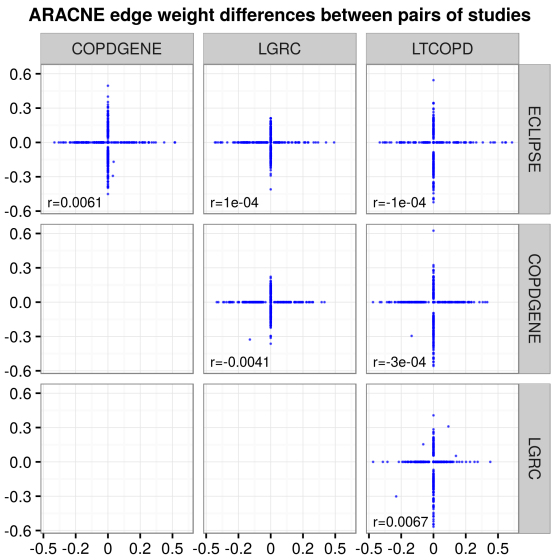
Statistical Challenges

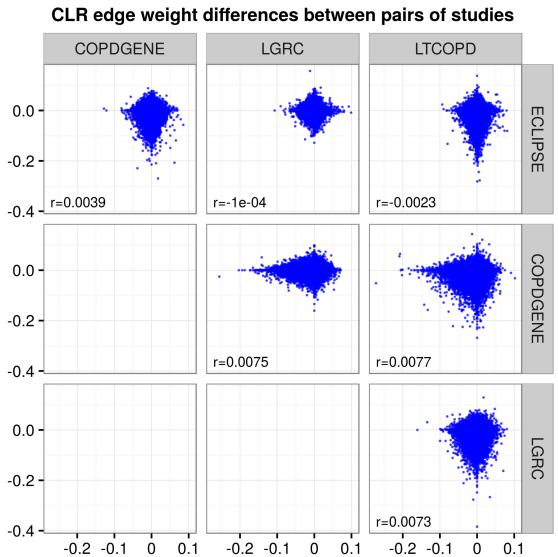
- Gene expression measurements are noisy.
- Model complexity may require the estimate of too many model parameters.
- May be computationally intractable.
- May be statistically undetermined. "The curse of dimensionality"

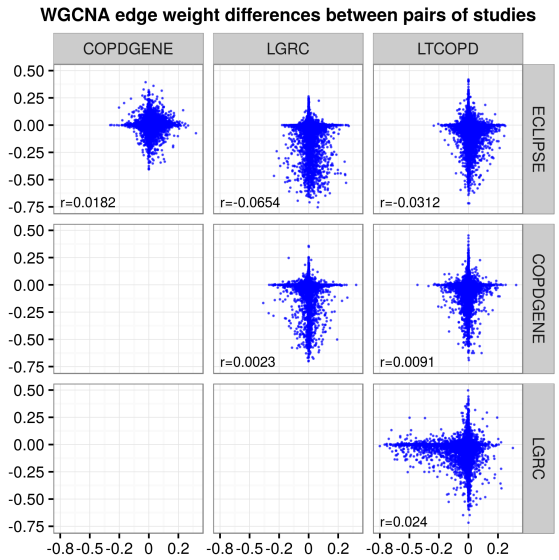
Replication of inferred networks in COPD

Differential edgeweights do not correlate between studies of COPD

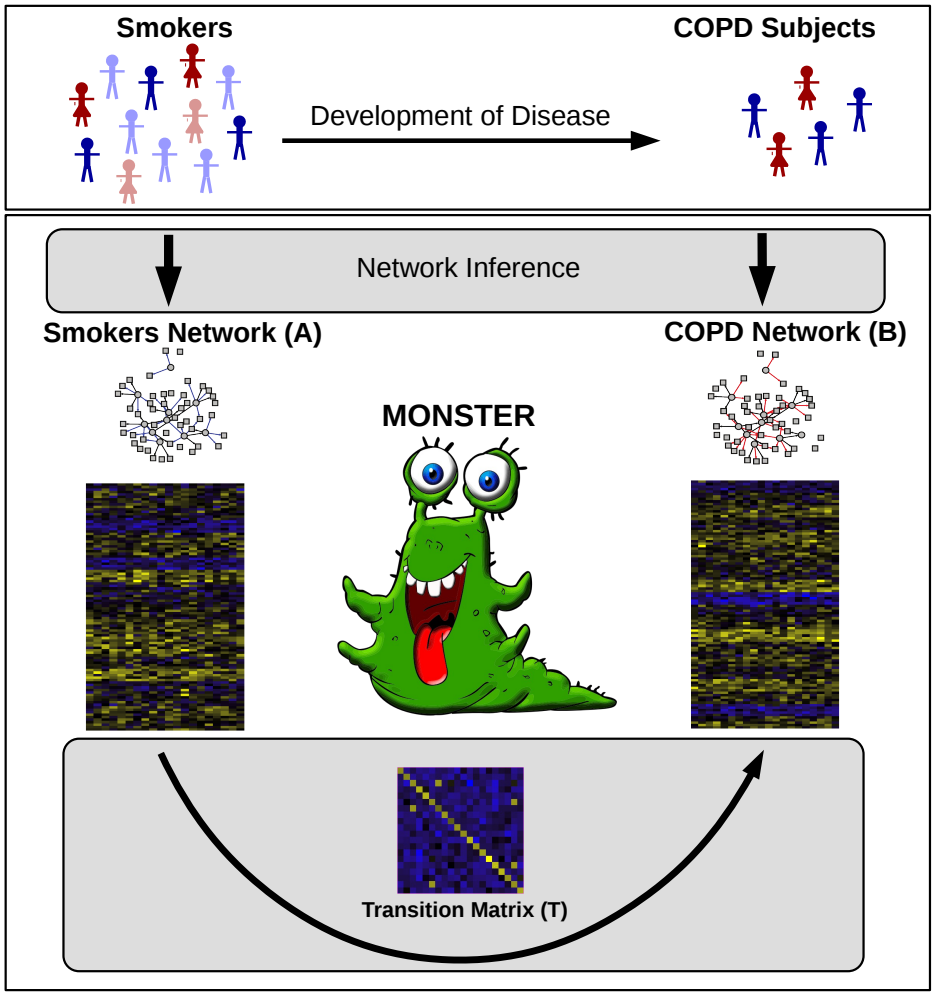




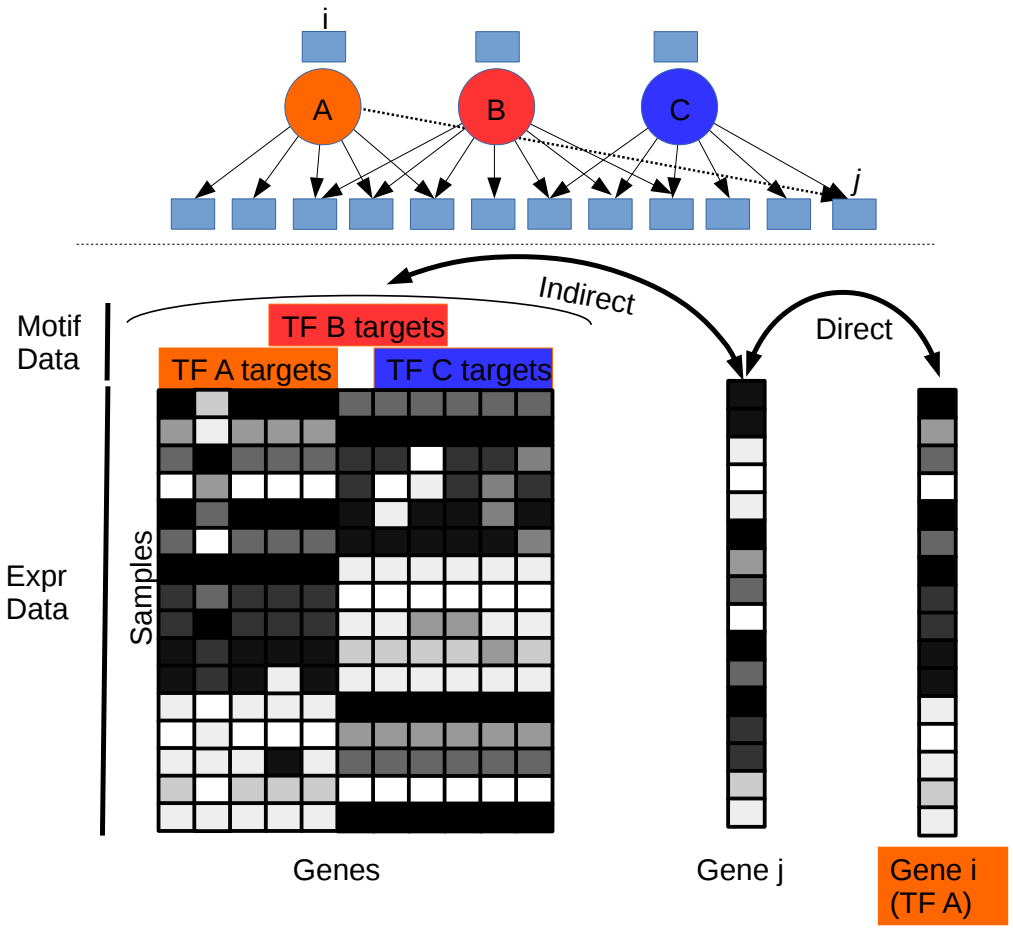




Algorithm Overview



Network Inference



Network Inference

Direct Evidence:

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

Indirect Evidence:

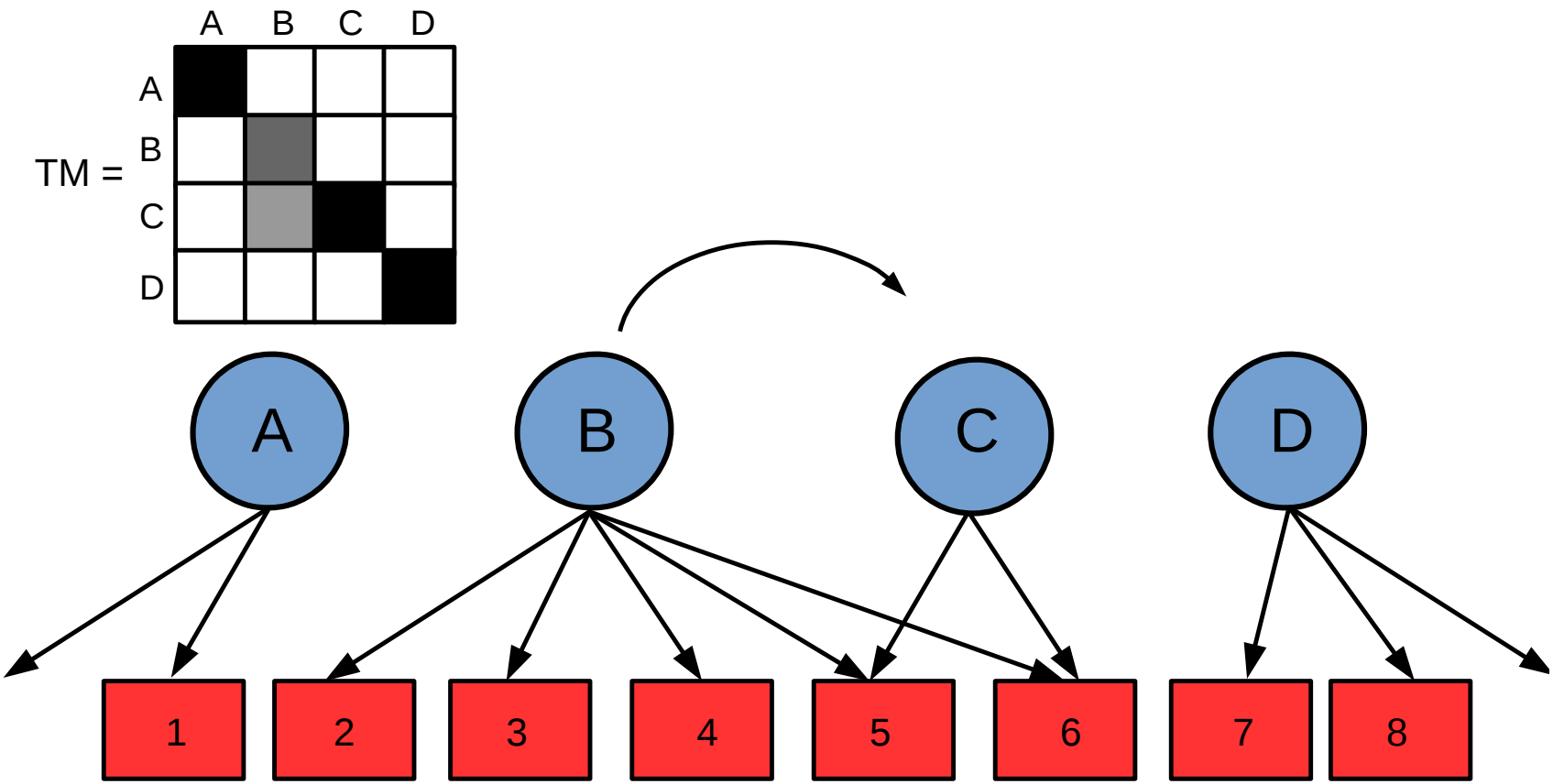
$$logit\left(E\left[M_i\right]\right) = \beta_0 + \beta_1 g_{(1)} + \cdots + \beta_N g_{(n)}$$

$$e_{i,j} = \frac{1}{1 + e^{\beta_0 + \beta_1 g_{j,(1)} + \cdots + \beta_k g_{j,(k)}}}$$

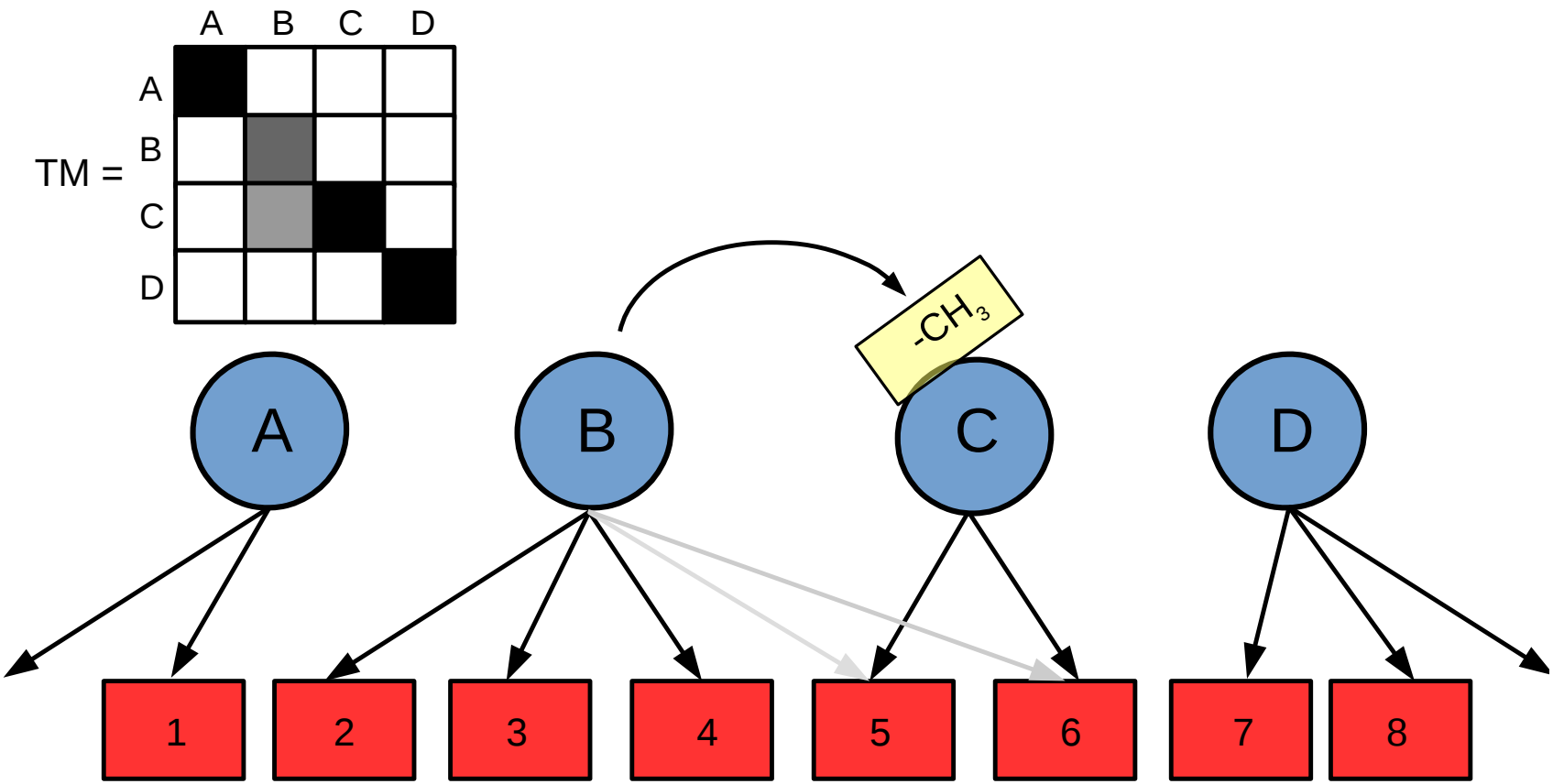
Edgeweight:

$$w_{i,j} = (1 - \alpha)\left[d_{i,j}\right] + \alpha\left[e_{i,j}\right]$$

Network Transition



Network Transition



Network Transition

$$E[b_i - a_i] = \tau_{1,i} a_1 + \cdots + \tau_{m,i} a_m$$

where b_i and a_i are column-vectors in $\mathbf{B}$ and $\mathbf{A}$ that describe the regulatory targeting of transcription factor i in the final and initial networks, respectively.

In the simplest case, this can be solved with normal equations,

$$\hat{\tau}_i = \left(A^T A\right)^{-1} A^T (b_i - a_i)$$

to generate each of the columns of the transition matrix $\mathbf{T}$ such that

$$\hat{\mathbf{T}} = [\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_m]$$

Network Transition

Regularization:

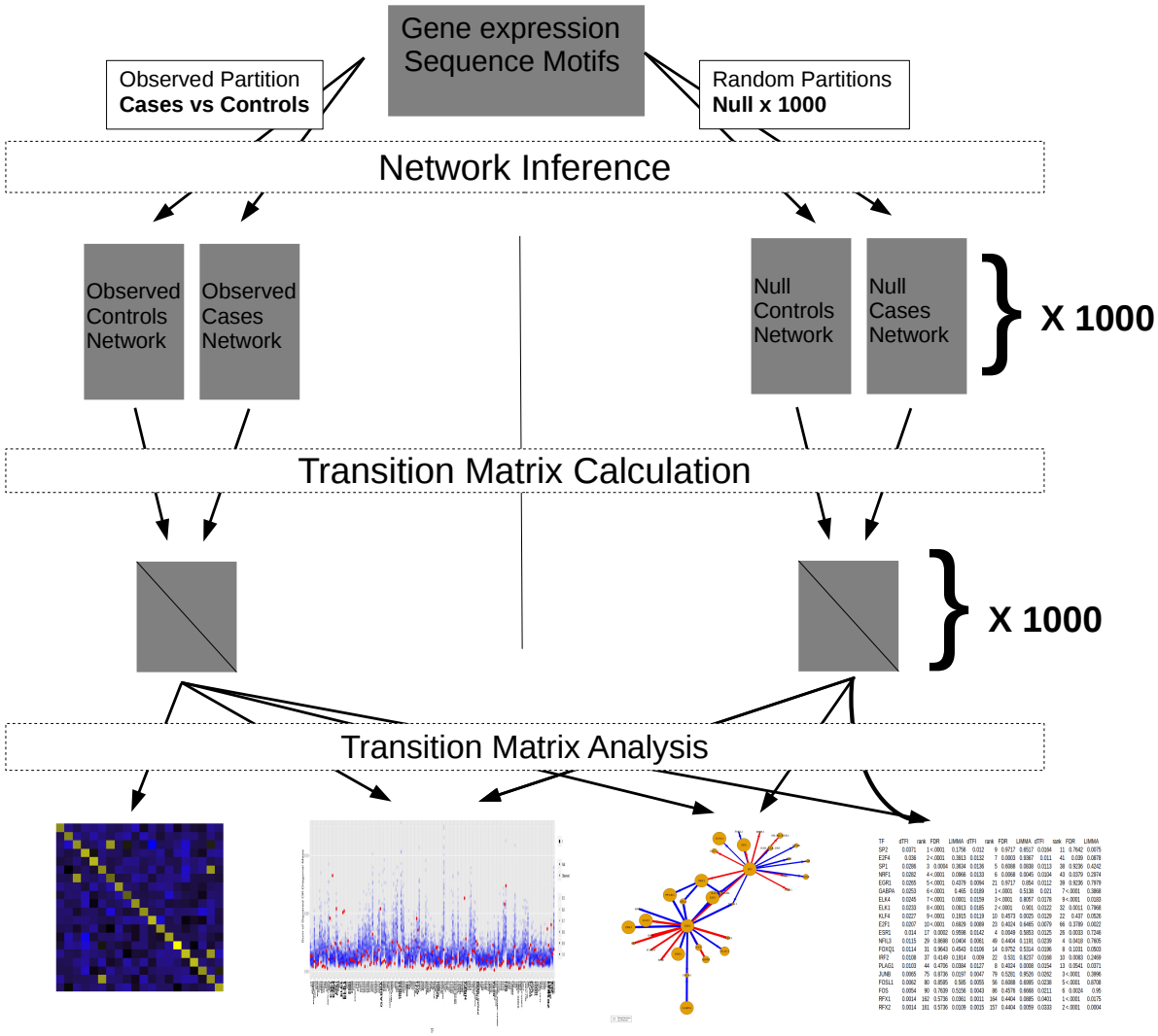
$$\mathbf{Q}_{i,j} = \begin{cases} 1 & \textit{for } i = j \neq k \\ 0 & \textit{elsewhere} \end{cases},$$

which results in the minimization of the penalized residual sum of squares

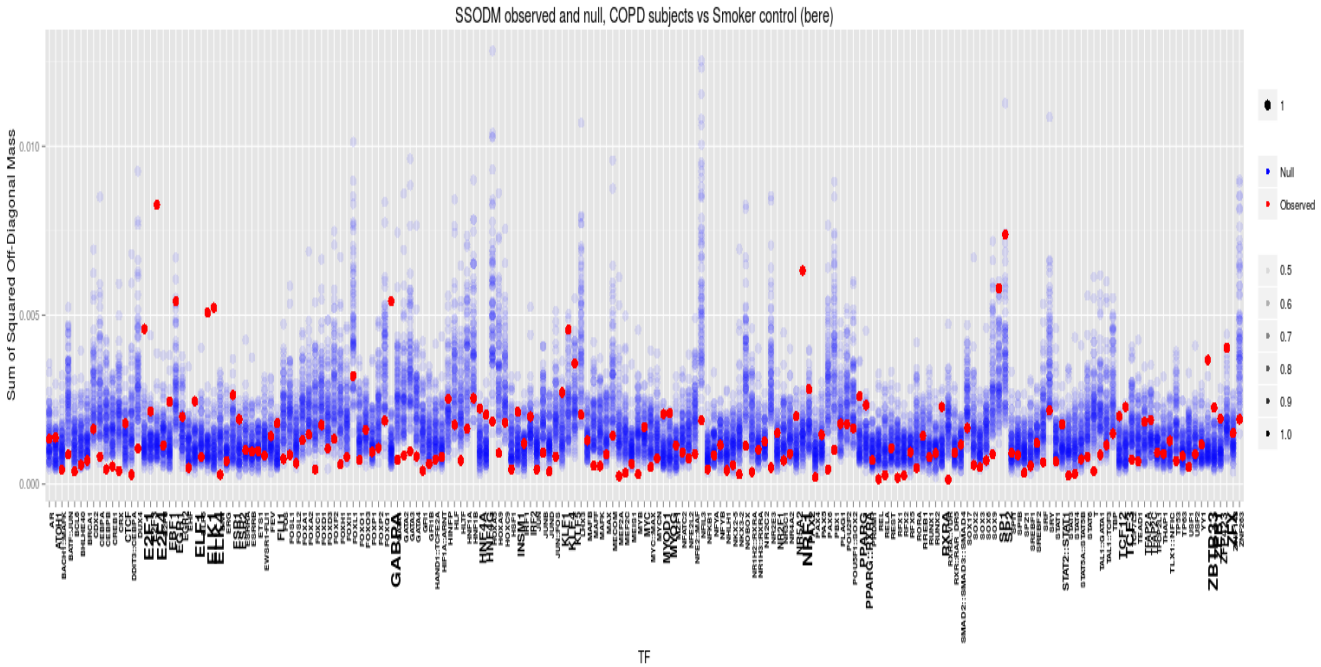
$$PRSS\left(\mathbf{T}_{\cdot,\mathbf{k}}\right) = \sum_{i=1}^{p} \left(\mathbf{B}_{i,k} - \sum_{j=1}^{m} A_{i,j} \mathbf{T}_{j,k} \right)^2 + \lambda \sqrt{\mathbf{T}'_{\cdot,k} \mathbf{Q} \mathbf{T}_{\cdot,k}}$$

An implementation of this extension is available in the R package MONSTER.

Network Transition

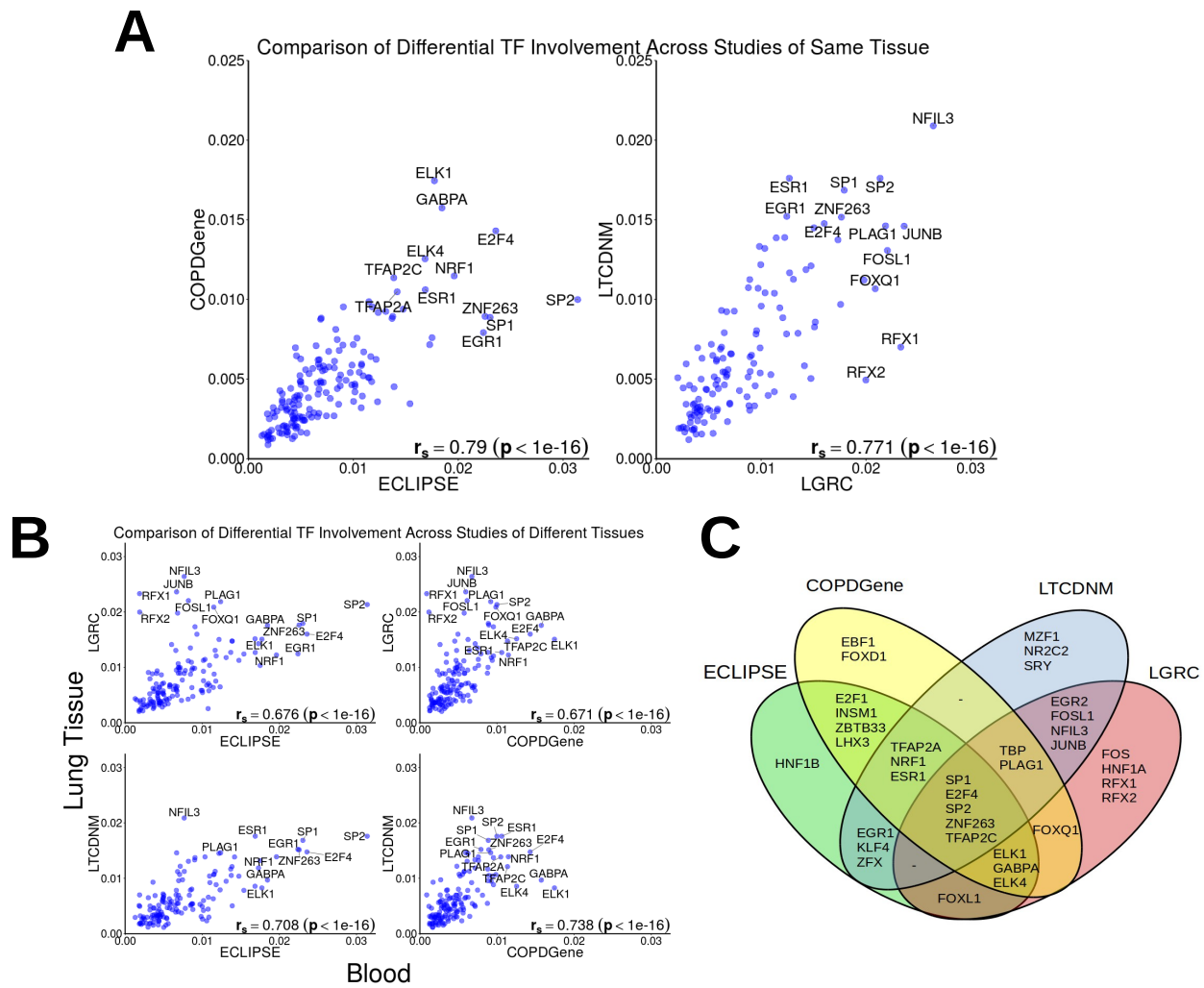


Evaluating Transition Matrix

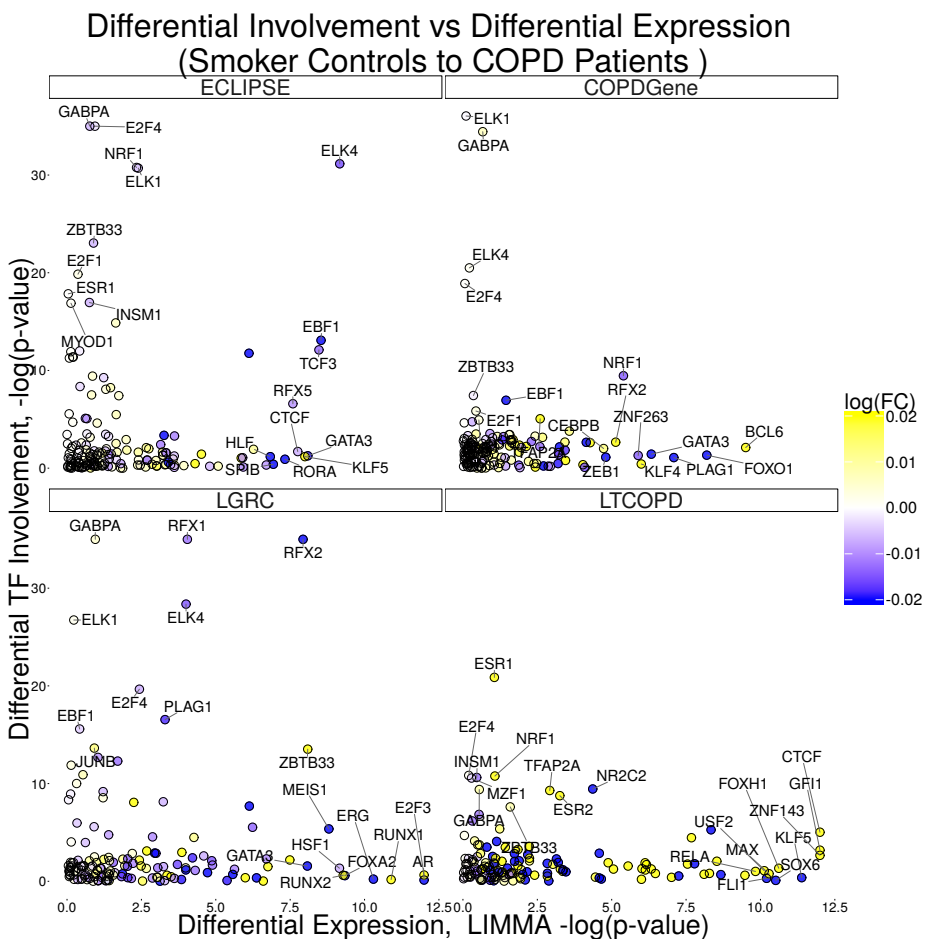


$$d\hat{T}FI_j = \frac{\sum_{i=1}^{m} I\left(i \neq j\right) \hat{\tau}_{i,j}^2}{\sum_{i=1}^{m} \hat{\tau}_{i,j}^2}$$

Reproducibility and novel results



Reproducibility and novel results



Differential TF Involvement

	ECLIPSE			COPDGene			LGRC			LTCDNM		
transcription factor	dTFI	rank	FDR	dTFI	rank	FDR	dTFI	rank	FDR	dTFI	rank	FDR
SP2	.0314	1	.0357	.0100	9	.6812	.0213	6	.3752	.0176	2	.7438
E2F4	.0236	2	<.0001	.0143	3	<.0001	.0160	14	.037	.0148	7	<.0001
SP1	.0230	3	.1551	.0089	18	.7721	.0179	10	.3594	.0169	4	.5516
ZNF263	.0226	4	.311	.0089	16	.3372	.0177	11	.7716	.0152	6	.927
EGR1	.0224	5	.1242	.0079	23	.7597	.0124	28	.6892	.0152	5	.5305
NRF1	.0196	6	<.0001	.0115	5	.0304	.0122	30	<.0001	.0139	11	.0558
GABPA	.0185	7	<.0001	.0157	2	<.0001	.0176	12	<.0001	.0097	32	.0853
ELK1	.0177	8	<.0001	.0174	1	<.0001	.0151	17	<.0001	.0083	40	.2099
ZFX	.0175	9	<.0001	.0076	24	.8366	.0103	40	.4348	.0132	16	.2739
KLF4	.0173	10	.1025	.0072	28	.8142	.0143	21	.2312	.0119	20	.5516
ESR1	.0169	11	.0357	.0106	7	.0941	.0127	27	.0888	.0176	3	<.0001
ELK4	.0168	12	<.0001	.0125	4	<.0001	.0152	16	<.0001	.0086	39	.1318
TFAP2C	.0139	17	.0656	.0114	6	.0941	.0148	19	.037	.0121	19	.2099
PLAG1	.0124	21	.263	.0092	15	.4136	.0219	5	<.0001	.0146	8	.1554
FOXQ1	.0115	28	.9318	.0099	10	.7905	.0209	7	.2846	.0107	27	.927
FOSL1	.0082	57	.9175	.0061	41	.6166	.0220	4	.037	.0131	17	.3496
NFIL3	.0077	62	.2365	.0067	33	.0304	.0264	1	.4669	.0209	1	.7121
FOS	.0068	73	.9175	.0057	48	.5212	.0198	9	.037	.0112	24	.5139
JUNB	.0067	77	.9318	.0059	43	.6392	.0236	2	<.0001	.0146	9	.2299
RFX1	.0019	159	.3532	.0009	164	<.0001	.0233	3	<.0001	.0070	48	.3496
RFX2	.0019	158	.4041	.0012	163	.0482	.0200	8	<.0001	.0049	81	.6245

Differential Expression

	ECLIPSE		COPDGene		LGRC		LTCDNM	
transcription factor	dTFI rank	LIMMA p	dTFI rank	LIMMA p	dTFI rank	LIMMA p	dTFI rank	LIMMA p
SP2	1	.1756	9	.6517	6	.0075	2	.0009
E2F4	2	.3913	3	.9367	14	.0878	7	.8232
SP1	3	.3634	18	.0838	10	.4242	4	.9759
ZNF263	4	.9834	16	.0028	11	.0271	6	.1859
EGR1	5	.4379	23	.8540	28	.7979	5	.0378
NRF1	6	.0966	5	.0045	30	.2974	11	.3418
GABPA	7	.4650	2	.5138	12	.3868	32	.5771
ELK1	8	.0913	1	.9010	17	.7968	40	.0005
ZFX	9	.8253	24	.5795	40	.0474	16	.1572
KLF4	10	.1915	28	.0025	21	.0526	20	.1159
ESR1	11	.9598	7	.5853	27	.7246	3	.3477
ELK4	12	.0001	4	.8057	16	.0183	39	.7314
TFAP2C	17	.2318	6	.9574	19	.5853	19	.6754
PLAG1	21	.0384	15	.0008	5	.0371	8	.9523
FOXQ1	28	.4543	10	.5314	7	.0503	27	.5340
FOSL1	57	.5850	41	.6995	4	.8708	17	.3686
NFIL3	62	.0404	33	.1191	1	.7605	1	.8650
FOS	73	.5156	48	.6668	9	.9500	24	.7891
JUNB	77	.0197	43	.9526	2	.3996	9	.6077
RFX1	159	.0361	164	.0885	3	.0175	48	.8285
RFX2	158	.0109	163	.0059	8	.0004	81	.1345

Acknowledgements

Dissertation Committee

- John Quackenbush
- Christoph Lange
- Kimberly Glass

Channing Division of Network Medicine

- Ed Silverman
- Craig Hersh

JQ Lab

- John Quackenbush
- Joe Barry
- Joey Chen
- Maude Fagny
- Marieke Kuijjer
- Camila Lopes-Ramos
- Megha Padi
- Joe Paulson
- John Platig
- Heather Selby
- Nicole Trotman