A batch correction method for differential gene network analyses

Dan Schlauch, PhD Candidate

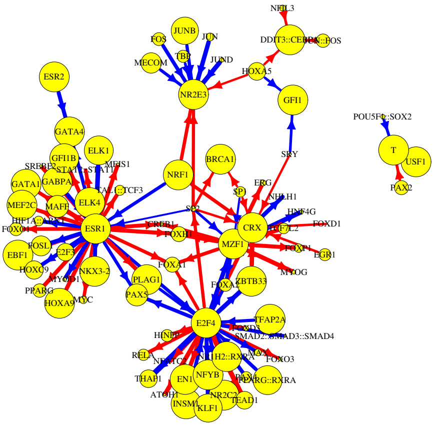
Department of Biostatistics
Harvard School of Public Health

March 28, 2017

Outline

1. Gene Networks
   - Gene Network Analysis
   - Measuring Association
2. Controlling for Batch Effect
   - What is Batch Effect?
   - Batch Effect Methods
   - Limitations
3. Correcting the coexpression matrix
   - General Idea
   - Coexpression Correction Method
4. Results
   - Simulation Results
   - Real Data Results

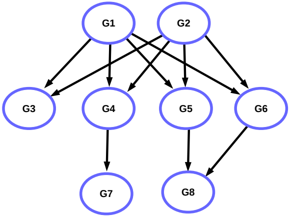
Inferring gene expression networks

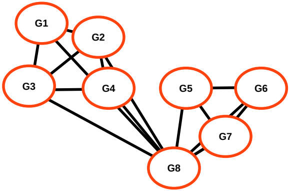


Types of Network Inference Approaches

Typically, network inference methods fall into two categories:

1. Gene Regulatory Networks (GRNs)
    - Directed graph
    - Imply a sort of physical interaction
2. Gene Coexpression Networks (GCNs)
    - Undirected graph
    - Imply a more general common pathway or process

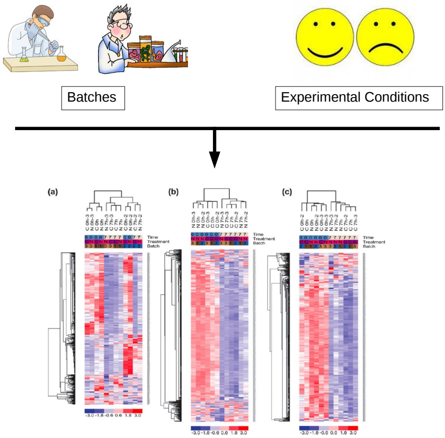




Measuring association

- Pearson Correlation
    - Linearity, outliers, etc.
- Spearman Correlation
    - More robust, less sensitive to outliers
- Euclidean Distance
- Mutual Information
    - Non-linear
- Partial Correlation
    - Direct effects

Batch Effect



Batches

Experimental Conditions

Johnson et al.(Biostatistics 2007)

- Laboratory Conditions
- Circadian Rhythm / cell cycle
- Reagents
- Atmospheric Ozone
- Etc. etc.

Methods for Controlling Batch Effect

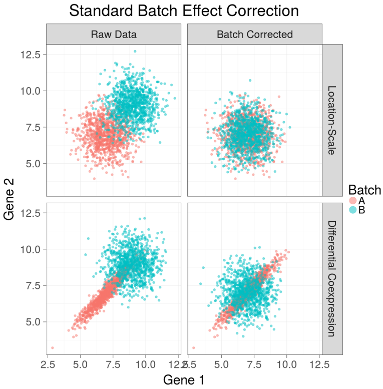
- COMBAT - Empirical Bayes approach for location/scale adjustment
- Surrogate Variable Analysis (SVA) - SVD approach for estimating
- Reference based (RATIO-G) methods -Scales sample measurements by the geometric mean of a group of reference measurements
- Distance Weighted Discrimination (DWD) -Based on SVM
- Many more...

Limitations of existing batch effect correction methods

- Perfect confounding
- Location/scale assumptions
- Independent effects
- Batches must be known or
- Batches must be estimated (SVA)
- *Differential coexpression*

Batch effect removal methods typically return a corrected gene expression matrix.

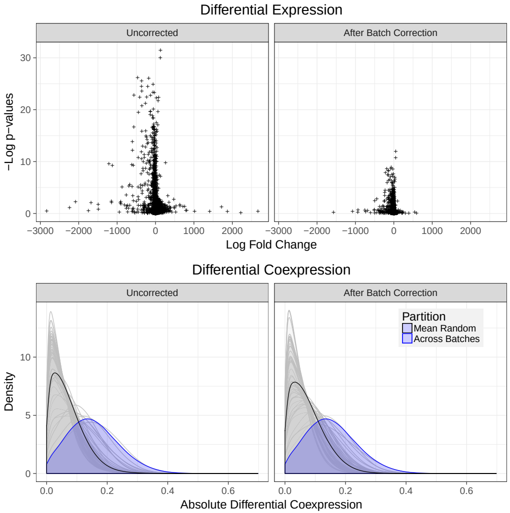
Limitations to existing batch effect correction methods



- Protocol induced coexpression?
- Differential biological variation

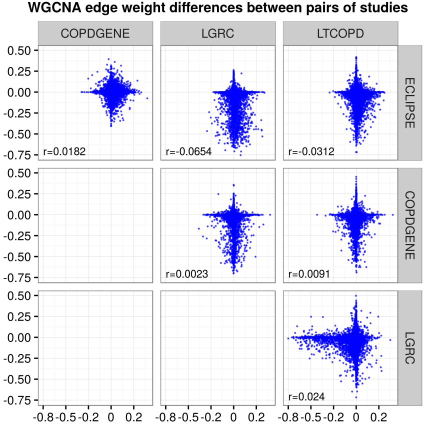
$$f\left[Gene1 \middle| BatchA\right] = f\left[Gene1 \middle| BatchB\right]$$

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:

1.) Provide a regression framework for for the coexpression matrix.
2.) Estimate a reduced number of parameters.
3.) 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

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

Where we define Σ_i

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

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

Least Squares Estimator

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

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

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

$$\sum_{i=1}^{N} \left[\mathbf{X}_i^T \mathbf{X}_i\right] \hat{\Psi}_h = \sum_{i=1}^{N} \left[\mathbf{X}_i^T \mathbf{Q}_h^T \mathbf{G}_i \mathbf{G}_i^T \mathbf{Q}_h\right]$$

$$\mathbf{X}^T \mathbf{X} \hat{\Psi}_h = \sum_{i=1}^{N} \left[\mathbf{X}_i^T \mathbf{Q}_h^T \mathbf{G}_i \mathbf{G}_i^T \mathbf{Q}_h\right]$$

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

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

The Corrected Coexpression Matrix

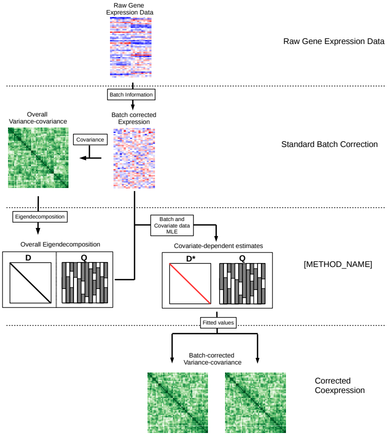
With the estimates obtained with our method, it is straightforward to see how fitted values for the coexpression matrix for each sample or experimental condition can be obtained. Given an estimate for Ψ, $\hat{\Psi}$, we can now estimate the batch-independent coexpression structure as

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

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



Raw Gene Expression Data

Batch Information

Overall Variance-covariance

Covariance

Batch corrected Expression

Standard Batch Correction

Eigendecomposition

Batch and Covariate data MLE

Overall Eigendecomposition

D

Q

Covariate-dependent estimates

D*

Q

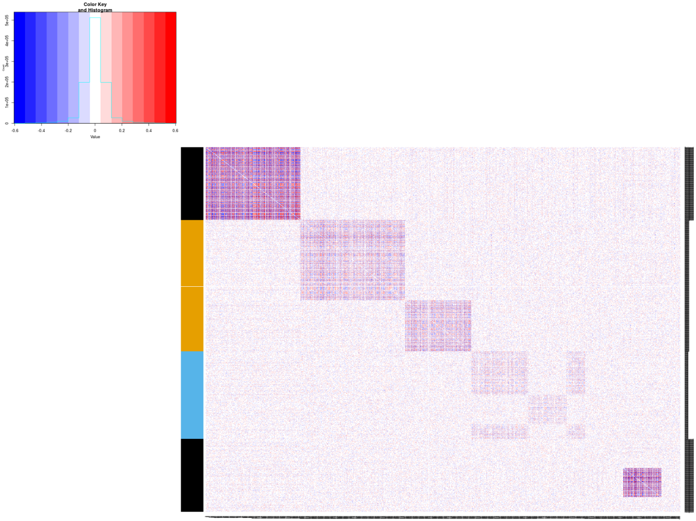
[METHOD_NAME]

Fitted values

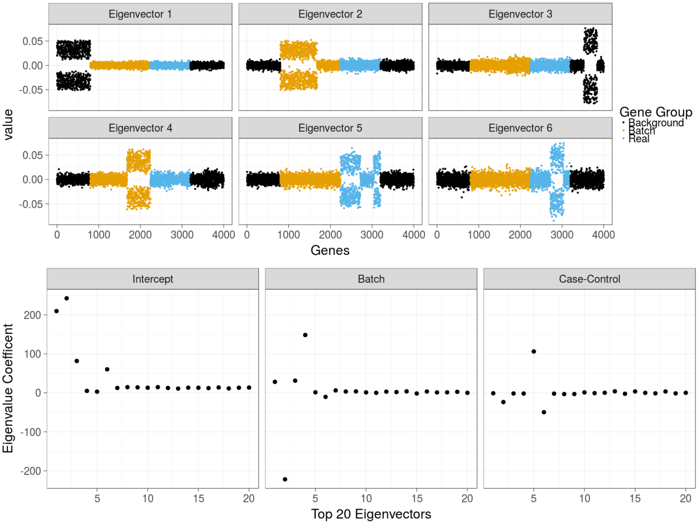
Batch-corrected Variance-covariance

Corrected Coexpression

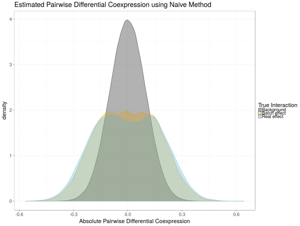
Simulations

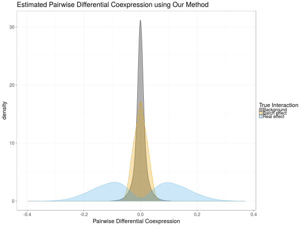


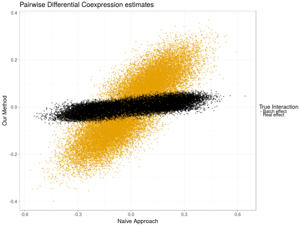
Simulations

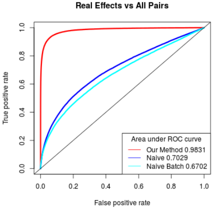


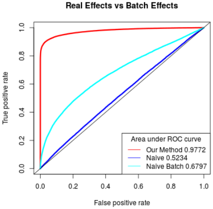
Simulations











Immuno-navigator

Soon...

Acknowledgements

Thanks to:

- John Quackenbush
- Kimbie Glass
- Joe Paulson